CAPACITY PLANNING FOR SEMICONDUCTOR WAFER FABRICATION WITH TIME CONSTRAINTS BETWEEN OPERATIONS

Jennifer K. Robinson

Chance & Robinson, Inc.
7171 Buffalo Speedway #639
Houston, TX 77025, U.S.A.

Richard Giglio

The University of Massachusetts at Amherst
Department of Mechanical and Industrial Engineering
Amherst, MA 01002, U.S.A.

ABSTRACT

Planning capacity for wafer fabrication is complicated by time constraints between process steps. For example, if certain baking operations are not started within two hours of a prior cleaning then the lot in question must be sent back to be cleaned again. For two-element systems an approximation based on M/M/c queuing formulas is developed and compared with results from discrete event simulations. The approximation performs well in predicting the probability of reprocessing and provides a bound that can easily be included in the spreadsheet capacity models often employed by manufacturers. For multi-element systems, the results of a fluid model used to understand general system characteristics are summarized. Discrete event simulation was used to validate the results of the analytic models and provide guidelines for operating time-constrained systems.

1 INTRODUCTION

The efficient production of semiconductors is the driving force in electronic technology. As competition increases, semiconductor manufacturers must pay close attention to production costs. New facility construction can cost upwards of a billion dollars, with equipment alone accounting for up to 80% of the total cost (Padillo and Meyersdorf 1998). With some types of equipment costing several million dollars each, capacity planning decisions have an immediate impact on the bottom line. Operating capacity is critical to maintaining profitability. If demand exceeds capacity then revenue is lost when facilities are not run at maximum capacity. On the other hand, overloading a factory is costly because of long cycle times, missed delivery dates, excessive inventory, and possibly lower yields (Srinivasan *et. al.* 1995).

Operating a wafer fabrication facility (fab) is highly complex, with technologies and market conditions constantly changing. Planners are continually juggling cost, capacity and cycle time trade-offs, but data are often

difficult to gather, and their sheer volume makes validation arduous. Issues such as setups, batch tools, reentrant flow, and shared tools across tool groups make planning fab capacity difficult. Many of these issues have been addressed (with varying success) by available capacity planning tools, which include spreadsheets, analytic models, and simulation models.

One issue that is not generally addressed by current capacity models is the presence of time constraints between process steps, also called *time bound sequences*. In a time bound sequence (TBS), there exists a step that must be completed within some fixed time of an earlier step. There may or may not be intervening operations between the two steps. In semiconductor manufacturing, an example is a baking operation that must be started within two hours of a prior clean operation. If more than two hours elapse, the lot must be sent back to be cleaned again.

The capacity of a system is the maximum feasible arrival rate of work to the system, or, equivalently, the maximum achievable throughput rate of the system. The behavior of a time bound sequence with more than two operations is difficult to predict except at very low equipment utilizations. In this case, lots flow through with few delays, and are rarely sent back for reprocessing. At higher arrival rates, or for highly variable systems, time bound sequences can rapidly become unstable. Once a few lots are delayed enough to be sent back for reprocessing, these lots increase the arrival rate to the earlier equipment. This in turn increases queuing delays, and makes it more likely that other lots will be sent back. A "vicious cycle" ensues, making predicting system capacity difficult.

Determining the capacity of a time bound sequence, even one with only two operations, requires understanding the distribution of lot cycle times. Such knowledge can not easily be derived from spreadsheet models, which usually include only static data such as mean cycle times. Even analytic models, such as queuing models, customarily rely on the first and second moments of arrival and service times, not on the entire distribution. Therefore, to understand the behavior of a time-constrained system,

capacity planners must turn to simulation. Most commercially available factory simulators, however, do not include time constraints between process steps, so capacity planners must often ignore this effect, and hope for the best. The goal of this research is to provide capacity planners with an alternative to "hoping for the best."

For time bound sequences that involve only two operations a simple approximation based on M/M/c queuing formulas is developed and compared with results from a discrete event simulation for various system parameters. The approximation is shown to perform well in predicting the probability of reprocessing for highly variable systems. It provides a bound that can easily be included in spreadsheet capacity models. For time bound sequences with intermediate operations, a fluid model was used to understand system behavior. The results were validated using discrete event simulation, and are summarized here in the form of operational recommendations for time bound sequences.

2   BACKGROUND

The manufacture of integrated circuits consists of four basic steps: wafer fabrication, wafer probe, assembly (packaging), and final testing. The most expensive phase is wafer fabrication, in which circuits are layered through successive operations onto a smooth, typically silicon, wafer. This involves a sequence of as many as 300-600 highly complex processing steps. Many of the intermediate steps are repeated for each layer of circuitry, different circuits require different sequences of steps, and each operation can include multiple sub-operations on different machines.

Some of the processing steps are performed on individual wafers, others on lots (groups) of wafers, and still others on batches (collections) of lots. A lot generally consists of 24 or 48 wafers, while a typical batch contains up to six lots. The collection of lots into batches results in a non-smooth product flow. The situation is further complicated by the existence of re-entrant flow, a characteristic that makes wafer fabrication different from traditional manufacturing. As different layers are added to the surface of a typical semiconductor device, lots at different stages of production return to the same processing equipment many times. Capacity planning of a fabrication facility, therefore, may involve analyzing production sequences and processing time recipes for several products, each with non-smooth, re-entrant flow.

2.1   Problem Definition

Figure 1 shows a time-constrained system in which lots flow through two operations in series, with each operation performed on a single-machine group. Lots must begin processing on Machine 2 within a pre-defined time after completing processing on Machine 1. Otherwise, they must go back and repeat processing on Machine 1. The elapsed time between completing processing on Machine 1 and starting on Machine 2 is denoted as *TE*. The capacity of this system is the number of lots that can be processed during a given time window (e.g. lots per week). Suppose that processing on Machine 1 requires six minutes per lot, while processing on Machine 2 requires five minutes per lot, and that all lots must go through the two machines in sequence. Assume also that the machines are both available for the same number of hours per week. In this case, Machine 1 is the bottleneck, and the maximum capacity of the system is 10 lots per hour.
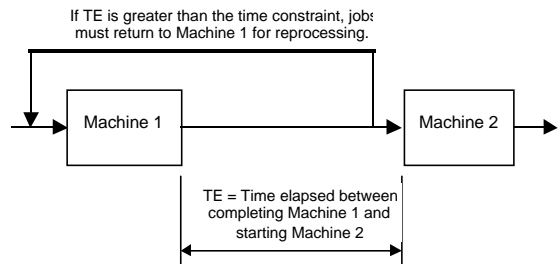

Figure 1: Sample Diagram of a Time-Constrained System

The presence of the time constraint, combined with variation, may reduce actual capacity to less than 10 lots per hour. If the time constraint is 3 minutes and Machine 2 has highly variable processing times, then waiting lots may easily have a *TE* greater than 3 minutes. In this case, lots will have to be sent back to Machine 1, increasing the load on that machine, and thereby decreasing the throughput of the system.

The situation becomes even worse if there are *n* machines, *n>2*, and lots must begin processing on Machine *n* within a pre-defined time after completing processing on Machine 1. If the delay exceeds TE while in queue for any machine, say machine *k*, $1<k<n$, then the lot will be sent back to Machine 1, increasing the load on machines 1 through *k*-1. This increased load will increase queuing at Machines 2 through *k*-1 (all of which are subject to the time constraint), and make it more likely that still more lots will be sent back for reprocessing. This will lead to a vicious cycle in which system performance degrades rapidly.

Unless a TBS has very low variability, there generally will be some positive probability that lots will exceed the time constraint, and be sent back for reprocessing. Once this happens, system behavior is likely to degrade further. To predict stability and determine the capacity of a general TBS requires knowledge of the entire distribution of lot cycle times. However, this research has found that the behavior of two-operation TBS, which have no intervening operations, can be approximated reasonably well.

3   LITERATURE REVIEW

Although no references have been found describing the specific problem of time bound sequences for semiconductor manufacturing, numerous studies depict capacity planning for wafer fabs. Neacy *et. al.* (1994) describe a survey of over 200 participants from companies across the United States and Europe, highlighting difficulties with current capacity planning methods, as well as factors that contribute to capacity loss in wafer fabs. Uzsoy *et. al.* (1992) provide an excellent review of the relevant issues in production planning for semiconductor fabs.

Several authors describe case studies in which multiple methodologies are used for capacity planning: Brown *et. al.* (1997), Domaschke *et. al.* (1998), Burman *et. al.* (1986), Johal (1996), and Grewal *et. al.* (1998). Many studies describe the application of simulation to capacity planning decisions in wafer fabs: Spence and Welter (1987), Tullis *et. al.* (1990), and Potti and Mason (1997) are a few examples. Other researchers have applied analytic models to questions related to capacity planning for wafer fabs, including Srinivasan (1995), Chen *et. al.* (1988), and Connors *et. al.* (1996). For a more detailed review of these papers, see Robinson (1998).

4   TWO-ELEMENT TIME BOUND SEQUENCES

This section describes a series of simulation experiments constructed to understand the behavior of two-operation time bound sequences. A two-operation system is of interest because time bound sequences in actual wafer fabs sometimes do include only two operations, and because the basic representation can serve as a building block for more complex layouts. For example, suppose that the middle machine in a three-operation system has a mean significantly less than the first machine or a variance small compared to *TE*. In this case, there is little probability that a lot will be recycled due to Machine 2, and the means and variances of Machines 1 and 2 can be aggregated into one virtual machine.

Simulation results are compared with M/M/c approximations for the probability of lots being reworked. The simulation models were developed using SIGMA for Windows, an event graph simulator developed by Lee Schruben (1995). Models were converted to standard C code. Models were converted to standard C code. This conversion increased run speed by a factor of more than 300, and allowed large experiments to be run through the use of batch files.

4.1 Computation of Expected Results

The system modeled is a clean and bake sequence. Lots arrive every t_a minutes (where t_a is an independent, identically distributed, or I.I.D., exponential random variable) and are cleaned at a sink for t_c minutes (t_c is also an I.I.D. exponential random variable) before being baked in an oven for t_b minutes (and t_b is an I.I.D. exponential random variable). The cleaning and baking workstations each consist of a number of identical machines. After a part is cleaned, the baking operation must start within t_r minutes (where t_r is a constant), or else it must be returned to be cleaned again. Note that exponential processing times are not a realistic approximation for wafer fabrication, where processing times are fairly deterministic. This assumption will be relaxed later in the paper.

If there were no reprocessing due to the time constraint, the model described above would be an open Jackson network, and the decomposition method (Whitt 1983a, 1983b) could be used with the bake workstation treated as a simple M/M/c queue. Reich (1957) proved that the output process of an M/M/c queue is a Poisson process. For a FIFO M/M/c queue, the distribution of customer waiting times is well-known, and it is possible to calculate the probability of an individual customer waiting in the bake operation queue for a time less than or equal to t_r (Gross and Harris 1985).

With reprocessing, customers with bake operation waiting times greater than t_r are pulled out of the bake operation queue, and sent back to the clean operation. As long as the clean queue is stable, customers who leave the bake queue return later, and, therefore, the total number of lots serviced at the bake workstation, in steady state, is not affected by reprocessing. This is the crux of why the two-operation model is more tractable than the three-operation model. Of course the distribution of arrivals to the bake workstation is no longer Markovian. The intent of this experiment was to determine the magnitude of inaccuracy introduced by this non-Markovian behavior.

Let μ_b be equal to the service rate of an individual server in the bake workstation, and let λ_b be equal to the external arrival rate into system ($\lambda_b = 1/E[t_a]$). Note that λ_b is not equal to the arrival rate into the queue for the clean operation, because of reprocessing, but is equal to the arrival rate to the bake workstation (because lots are never processed more than once at the bake workstation). Let c_b be equal to the number of servers in the bake workstation.

Employing Equation 2.43 in Gross and Harris (1985), it can be shown (see Robinson 1998) that the probability that an arriving customer will be reprocessed at the clean step before going through the bake workstation is

$$\Pr(REDO) = 1 - w_q(t_r)$$

where

$$w_q(t) = \frac{\left(\dfrac{\lambda_b}{\mu_b}\right)^{c_b} \mu_b e^{-(\mu_b c_b - \lambda_b)t}}{(c_b - 1)!} p_0 \qquad (1)$$

In Equation (1) $w_q(t)$ is the probability that an arriving lot will have time in queue less than or equal to t, and hence will not have to be reprocessed. p_0 is the probability that the bake workstation is idle. The arrival rate to the cleaning station now consists of the external arrival rate to the system, λ_b, plus the arrival rate of reprocessed lots. Denoting this arrival rate as λ_c yields

$$\lambda_c = \lambda_b \left(1 + \Pr(REDO)\right) \qquad (2)$$

Here the arrival rate of reprocessed lots is equal to the arrival rate to the system, multiplied by the probability of reprocessing. Letting c_c be the number of servers in the cleaning workstation, and u_c be the service rate of an individual cleaning operation, the condition for stability of the cleaning operation is

$$\frac{\lambda_c}{c_c \mu_c} < 1 \qquad (3)$$

Equation (3) can be used to determine the stability of a two-process time-constrained system, where the maximum stable input rate is the system's capacity. To compute the system capacity, first define λ_c^* as the value obtained from Equation (3) at equality. A search algorithm can then be applied to Equation (2) to find the value of λ_b that results in equality when λ_c^* is substituted for λ_c. (Note that Pr(REDO) is a function of λ_b). Since the right-hand side of Equation (2) can be shown to be monotonically increasing in λ_b, deriving this algorithm is fairly straightforward. However, the result is only an approximation, because the arrival process to the bake workstation is not Markovian. In the next section, the predicted probability of reprocessing will be compared with the simulated probability of reprocessing in order to determine those circumstances where the approximation performs acceptably.

4.2 Experimental Design

Although a two-operation TBS is the most basic system of interest, that simple model contains several variables: the time constraint, t_r, the mean interarrival time of lots to the system, $E[t_a]$, the number of identical tools in the clean workstation, c_c, the number of identical tools in the bake workstation, c_c, and the service rates of each workstation. Other possible parameters of interest include the distributions of interarrival and service times, and the dispatch rule followed at each workstation. A series of experiments was conducted to evaluate the impact of these different variables.

4.2.1 Experiment 1

In the first experiment, the total service rates at both the cleaning and baking workstations were always held equal

to 1/0.90 lots per hour. From this overall service rate, the mean clean process time, *CPT*, and the mean bake process time, *BPT*, were calculated according to the number of tools in each workstation. Service at both queues followed a FIFO dispatch rule. Interarrival times and service times were exponentially distributed. Three factors were examined in this experiment: time constraint (t_r), mean interarrival time ($E[t_a]$), and number of servers in each workstation. The latter were always changed together, so that there were, for example, two cleaning stations and two cleaning operations, or three of each. The service rates were held constant at each workstation, so that varying the mean interarrival time to the system was like varying the traffic intensity at each server. Because of the reprocessed lots the actual traffic intensity at the clean workstation could not be known ahead of time. Each simulation replication was run until 100,000 lots had entered the system. Each design point was replicated three times, and the results averaged.

The estimated REDO probability from the simulation, *Pr(REDO)*, was defined as the average number of REDO events observed, *REDOS*, for each replication, divided by the number of arrivals to the system, *LIMIT*. Six levels of t_r, five levels of $E[t_a]$ and four levels of number of servers were simulated for this experiment, using a full factorial design. The levels of each factor are shown in Table 1. The experiment had 120 design points.

Table 1: Factor Settings for First Two-Operation Experiment

Time Constraint	Mean Interarrival Time	Number of Servers
1	2	1
0.9	1.8	2
0.8	1.6	3
0.7	1.4	4
0.6	1.2	
0.5		

4.2.2 Experiment 2

All parameters in Experiment 2 were identical to those in Experiment 1, except that the dispatch rule at the oven was LIFO instead of FIFO.

4.2.3 Experiment 3

The third experiment examined how variability in the service time distributions affected system performance. This experiment tested the effect of radically violating the Jacksonian network assumption of Markovian service times. Three different models of service time were proposed: constant, uniform, and exponential service at one

or both workstations, yielding nine combinations of service time distribution, one of which was identical to Experiment 1. The uniform distribution employed the same mean values used in Experiment 1, translated to a range of mean +/- 100%. This resulted in a coefficient of variation of 0.33 for all values. The interarrival time distribution to the system was always exponential because a distribution with low variability does not meaningfully represent the highly variable environment of a wafer fab. These runs used a FIFO dispatch rule.

4.3 Results

For each design point, expected system characteristics were computed using the approximation described in Section 4.1, and compared with observed values from the simulation. An Analysis of Variance (ANOVA) was also performed on the simulation results from Experiment 1 to determine the significance of the input variables.

4.3.1 Experiment 1

The input variables for this experiment were time constraint, mean interarrival time, and number of servers at the cleaning and bake workstations. The results for each level of each input variable were averaged across all values of other input variables, to investigate the overall effect of each variable on predicted and simulated reprocessing probability. The results are shown in Figures 2 to 4. Note in particular Figure 3. Here the predicted and simulated results for $Pr(REDO)$ are very close at all but the highest traffic point $E[t_a]= 1.2$. Investigation reveals that the clean operation is unstable for experiments at this interarrival time. To determine whether or not the approximation fit better when only stable points were considered, the highest interarrival time was eliminated from the results for certain calculations.

To confirm the graphical results regarding the appropriateness of the approximation, the time series $\{d_i\}$ was computed, where for each design point d represented the analytically approximated probability of reprocessing minus the simulated probability of reprocessing. A 95% confidence interval for d was then computed for all 120 design points. The confidence interval did not contain zero, indicating evidence of a statistically significant difference between the analytically approximated probability of reprocessing and the true probability of reprocessing (estimated via simulation). This is not surprising, since Figure 3 clearly shows the M/M/c approximation exceeding the simulated probability of reprocessing for the highest traffic case.
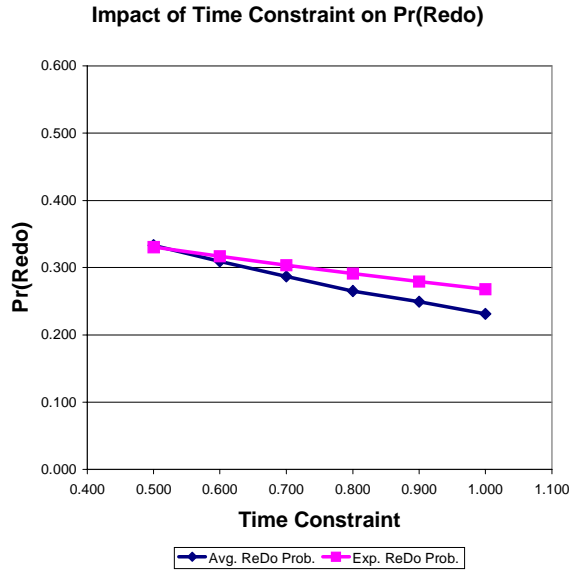
Impact of Time Constraint on Pr(Redo)

Figure 2: Impact of Time Constraint on Pr(Redo)

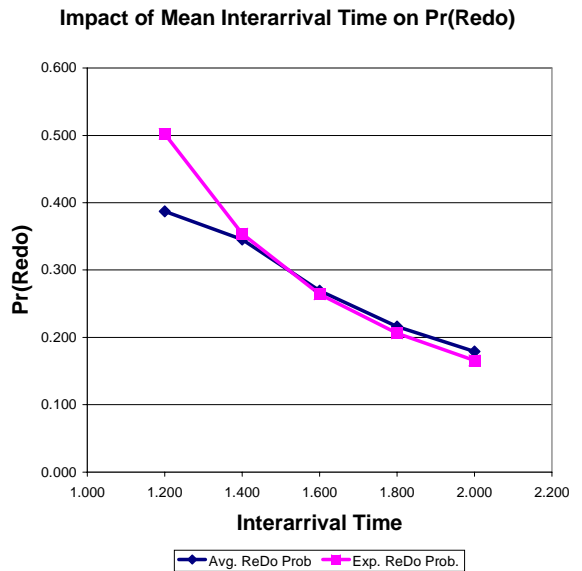**Impact of Mean Interarrival Time on Pr(Redo)**

Figure 3: Impact of Mean Interarrival Time on Pr(Redo)

A second confidence interval was computed using only the 96 stable design points. For those points the 95% confidence interval contained zero, indicating no evidence of a statistically significant difference between the analytically approximated probability of reprocessing and the true probability of reprocessing.
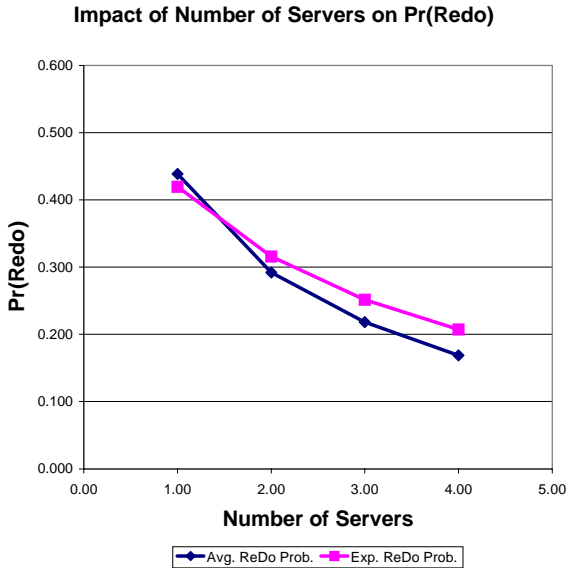
Impact of Number of Servers on Pr(Redo)

Figure 4: Impact of Number of Servers on Pr(Redo)

In addition to showing the accuracy of the M/M/c approximation, Figures 2 to 4 illustrate the influence of the three input variables on the probability of reprocessing. Tighter time constraints, smaller interarrival times (more heavily loaded systems) and smaller numbers of tools per workstation all appear to be associated with increased probability of reprocessing. To confirm the significance of this influence, a 3-factor ANOVA was conducted on the data and the results are displayed in Table 2.

Table 2: ANOVA Table for Experiment 1

ANOVA Table - Experiment 1	F Statistic	.999 Percentile F Stat.
Time Constraint Effect	1319.95	4.42
Interarrival Time Effect	8253.82	4.95
Number of Servers Effect	19019.80	5.78
TC/InterArr Interaction	12.37	2.53
InterArr/NumServers Interact.	264.58	3.02
NumServers/TC Interaction	91.05	2.78
3 Way Interaction	5.97	1.95

The ANOVA found that all three main effects were highly significant. The two and three way interaction effects were also significant, to a lesser extent. This indicates that drawing general conclusions regarding probability of reprocessing is a complex endeavor. *Pr(REDO)* for a workstation depends upon the magnitude of the time constraint, the utilization of the workstation, and the number of servers at the workstation. Moreover, it depends upon how these characteristics interact with one another. The single-server system, for example, is much more sensitive to changes in the other parameters than the multi-server systems.

Overall, this experiment shows that for systems where the interarrival and processing times are exponential, the dispatch rule is FIFO, the service rates of the two workstation are equal, and there are no random failures, an M/M/c approximation provides a reasonable guide for estimating whether or not a time-constrained system will be stable. This approximation is particularly valuable given that the probability of reprocessing for an actual workstation varies considerably depending on the parameters of the system. This result could be strengthened by looking at a wider range of time constraint values, and possibly by looking at lower traffic systems.

4.3.2 Experiment 2

The experiment showed that the LIFO results for probability of reprocessing were slightly higher than the FIFO results. A one-sided t-test (with $\alpha = 0.05$) supported the conclusion that the difference was significant. That difference, however, is quite small (LIFO mean = 0.28866, FIFO mean = 0.277942) relative to the magnitude of the factor effects. The factors all remain significant under LIFO and cause *Pr(REDO)* to move in the same direction as before. Details can be found in Robinson (1998).

4.3.3 Experiment 3

When the service times at the two workstations were both constant, the observed number of reprocessed lots was always zero. For systems with processing variability, the M/M/c approximation tends to overestimate the probability of reprocessing in non-exponential cases. Table 3 shows the overall average reprocessing probability observed for the nine scenarios, sorted in descending order of reprocessing. In general, less variability in the system corresponds to a lower probability of reprocessing. Variability at the sink, the first workstation, appears to increase the probability of reprocessing more than variability at the oven.

Table 3: Overall Average Probability of Reprocessing for Various Combinations of Service Time Distributions

Sink Distribution	Oven Distribution	Average *PR(REDO)*
Exponential	Exponential	0.279
Exponential	Uniform	0.249
Uniform	Exponential	0.215
Exponential	Constant	0.214
Uniform	Uniform	0.168
Constant	Exponential	0.166
Uniform	Constant	0.124
Constant	Uniform	0.105
Constant	Constant	0.000

5 CONCLUSIONS

The experiments conducted indicate that for time-constrained systems with no intervening operations, a simple M/M/c approximation provides a conservative upper bound on the maximum allowable loading on the machines. This approximation could easily be coded into existing capacity planning spreadsheets or other analytic models, and would provide a considerable improvement over the current method of "hoping for the best." For systems with a high degree of variability in arrival and service times, the approximation is quite close to observed results, despite the non-Markovian behavior introduced by the reprocessed lots. For systems with less variability, the M/M/c approximation tends to overestimate the probability of reprocessing, providing a conservative bound for determining whether the first workstation in a system will be overloaded. Simulation is recommended for more detailed analysis.

The experiments also show that the impact of time constraints is worse for single-server systems, systems with high traffic intensities, and systems with a high degree of variability. This is not surprising, nor is it likely to disagree with the intuition of manufacturing personnel in wafer fabs. One-of-a-kind tools, for example (workstations that contain a single piece of equipment), are commonly known to increase cycle time. Similarly, as a fab increases production volumes for the same tool set, management expects the cycle time to increase. The impact of variability in service times is less commonly understood, but may be illustrated through graphs similar to those in Section 4. In general, trade-off curves between maximum possible loading and time constraint, number of tools per workstation, and coefficient of service time variability might be useful for making operational decisions regarding time-constrained process sequences.

Analytic and simulation models can also be used to better understand the behavior of time bound sequences with intervening operations, and to set guidelines for system design. Because of the highly correlated nature of customer queue times in a tandem system, once a time-constrained system is allowed to build up significant queues, behavior is likely to deteriorate rapidly. Therefore, conservative bounds for setting system loading should be used. Within such parameters, queueing approximations for waiting time distribution can be used to estimate the probability of reprocessing. These can perform quite well, as described in Robinson (1998).

This paper illustrates an application in which simulation is particularly appropriate. Simulation is used to estimate the "true" capacity for an analytically intractable system, and to validate an analytic model. Future analyses can then use the analytic approximation without requiring additional simulation.

REFERENCES

Brown, S., F. Chance, J. Fowler, and J. K. Robinson. 1997. A centralized approach to factory simulation. *Future Fab International* 3:83-86.

Burman, D. Y., F. J. Gurrola-Gal, A. Nozari, S. Sathaye, and J. P. Sitarik. 1986. Performance analysis techniques for IC manufacturing lines. *AT&T Technical Journal* 65(4):46-57.

Chen, H., J. Harrison, A. Mandelbaum, A. Van Ackere, and L. Wein. 1988. Empirical evaluation of a queuing network model for semiconductor wafer fabrication. *Operations Research* 36(2):202-215.

Connors, D., G. Feigin, and D. Yao. 1996. A queuing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing* 9(3):412-427.

Domaschke, J., S. Brown, J. Robinson, and F. Leibl. 1998. Effective implementation of cycle time reduction strategies for semiconductor back-end manufacturing. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 985-992. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Grewal, N. S., A. C. Bruska, T. M. Wulf, and J. K. Robinson. 1998. Integrating targeted cycle-time reduction into the capital planning process, In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 1005-1010. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Gross, D. and C. M. Harris. 1985. *Fundamentals of Queuing Theory: Second Edition*. New York: John Wiley & Sons.

Johal, S. S. 1996. Non-linearity and randomness in a semiconductor wafer fab, In *Proceedings of the 1996 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2-6.

Neacy, E., S. Brown, and R. McKiddie, 1994, Measurement and improvement of manufacturing capacity (MIMAC) survey and interview results. SEMATECH Technology Transfer #94052374A-XFR.

Padillo, J. M. and D. Meyersdorf. 1998. A strategic domain: IE in the semiconductor industry. *IIE Solutions*, March: 36-42.

Potti, K. and S. J. Mason. 1997. Using simulation to improve semiconductor manufacturing. *Semiconductor International*, July: 289-292.

Reich, E., 1957, Waiting times when queues are in tandem. *Annals of Mathematical Statistics* 28:768-773.

Robinson, J. K. 1998. Capacity planning in a semiconductor wafer fabrication facility with time constraints between process steps. Ph.D. dissertation, Department of Mechanical and Industrial Engineering, University of Massachusetts at Amherst, Amherst, Massachusetts.

Schruben, L. 1995. *Graphical simulation modeling and analysis using SIGMA for windows.* Danvers, MA: boyd & fraser.

Spence, A. M. and D. J. Welter. 1987. Capacity planning of a photolithography work cell in a wafer manufacturing line. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 702-708.

Srinivasan, K. 1995. Capacity expansion with discrete options for semiconductor manufacturing. SEMATECH Technology Transfer #95062883.

Srinivasan, K., R. Sandell, and S. Brown. 1995. Correlation between yield and waiting time: a quantitative study. In *Proceedings of the 17th IEEE/CPMT International Electronics Mfg. Technology Symposium*, 65-69.

Tullis, B., V. Mehrotra, and D. Zuanich. 1990. Successful modeling of a semiconductor R&D facility. In *Proceedings of the 1990 IEEE/SEMI International Semiconductor Manufacturing Science Symposium*, 26-32.

Uzsoy, R., C-Y Lee, and L. M. 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* 24(4):47-60.

Whitt, W. 1983a. Performance of the queuing network analyzer. *The Bell System Technical Journal* 63: 1911-1979.

Whitt, W. 1983b. The queuing network analyzer. *The Bell System Technical Journal* 62(9):2779-2814.

AUTHOR BIOGRAPHIES

JENNIFER ROBINSON is a principal of Chance & Robinson, Inc., a consulting firm specializing in capacity planning for the semiconductor industry. Her clients in the semiconductor industry have included SEMATECH, Digital Equipment Corporation, Seagate Technology, IBM, and Siemens AG. Jennifer holds a B.S. (1989) degree in civil engineering from Duke University, an M.S. (1992) degree in Operations Research from the University of Texas at Austin, and a Ph.D. degree in Industrial Engineering from the University of Massachusetts at Amherst (1998). Her research interests center on factory productivity measurement and improvement.

RICHARD GIGLIO is a Professor in the Department of Mechanical and Industrial Engineering at the University of Massachusetts at Amherst. He received his B.S. degree from MIT, and M.S. and Ph.D. degrees from Stanford University. He has conducted research to develop mathematical models to help plan large-scale systems. He has also conducted extensive research in service industries, including insurance and health care, with an emphasis on preventive medicine systems. His most recent research interests concern product costing in highly automated manufacturing facilities.