

Correspondence

Understanding a Semiconductor Process Using a Full-Scale Model

Jerald Hunter, Deana Delp, Donald Collins, and Jennie Si

Abstract—A full-scale semiconductor manufacturing plant model was developed from a SEMATECH dataset using the computer software package EXTEND. The model was generated to study the complex interactions and characteristics of a semiconductor fabrication process. Equipment downtimes, process flow routes, and machine processing times were used to validate the model. Pilot runs of the model were used to determine simulation run times and data collection rates for the initial inventory and product cycle time measurements. The product cycle time results from the model at 95% capacity were within 63 hours (or 7%) of the SEMATECH cycle time measurements. These results demonstrate the accuracy of the simulation model built from the SEMATECH dataset. The full-scale model was set up to run special scenarios showing the effects of eliminating maintenance and changing product types. The full-scale model was compared to a small-scale model based on the same dataset to demonstrate the inadequacy of the validated small-scale model. A full-scale model is also useful for analyzing scheduling routines, detecting bottlenecks, and understanding machine relations in the semiconductor industry.

Index Terms—Computer simulations, factory management, modeling, scheduling, semiconductor manufacturing.

I. INTRODUCTION

Semiconductor manufacturing is one of the most complex manufacturing processes in the world. Random yields and rework, complex product flows, time-critical operations, batching, simultaneous resource possession, and rapidly changing products and technologies make semiconductor manufacturing difficult to schedule and model. Typical wafers undergo hundreds of processing steps, reentering the same processing machines multiple times as each layer is successively added. Often, some processes are skipped, repeated, or completed in a different order. This results in a complex, highly reentrant process flow that is difficult to manually schedule in an optimum manner.

The main focus of manufacturing strategies in the semiconductor industry is minimizing production cost and increasing productivity while improving both quality and due date delivery performance. Semiconductor plants have tended to operate in a make-to-stock mentality, with production lots rarely being associated with a specific customer order or due date. Together with the high capital costs of equipment, this has resulted in a major emphasis on maintaining high throughput and equipment utilization, while reducing both the mean and variance of cycle times. The more work-in-process (WIP) inventory there is on the factory floor, the longer the mean cycle time of a lot. Less cycle time variability allows for greater accuracy of proper due date delivery. Input release policies attempt to achieve shorter, more reliable flow times by releasing work to the shop in a controlled manner. Testing proposed solutions has always plagued schedulers. One way to test a method is to

put it into practice. This in turn would require a large implementation effort. Another option is to build a realistic plant simulator. Evaluation of such scheduling policies is typically done with computer simulations to avoid the expense of real factory experiments.

This paper addresses the issues of understanding the relationships and trends in a semiconductor manufacturing plant through modeling and simulation of an entire plant. It entails the general setup, initialization, and verification of the model. Details of the SEMATECH dataset used to develop the full-scale model, as well as the implementation of the model into software are included. Three tests were used to validate the model to the dataset and several capacity and special scenario runs were completed. A small-scale model was developed using the same dataset. Several comparisons of the models are discussed followed by conclusions.

II. FULL-SCALE MODEL

A simulator of a full-scale semiconductor manufacturing plant based on Dataset 1 from SEMATECH was developed using the commercially available software package EXTEND [1]–[5]. Miller [6] developed a simulation model to emulate an IBM semiconductor production facility in the late 1980s. This experiment uses the SEMATECH dataset since it represents a readily available source of validated semiconductor manufacturing process parameters and information. The following aspects are included in the factory:

- process cycle time for each product type and process;
- setup time for each product type and process;
- travel time between process steps;
- process steps and sequence required for each product type;
- number of machines available for each process;
- MTBF and MTTR for each machine type;
- setup logic on the bottleneck machine group;
- batching on batch processing machines;
- product rework and scrap.

A. SEMATECH Dataset

For the purposes of this study, it was assumed that all the data in Dataset 1 adequately reflects a typical semiconductor factory. The SEMATECH factory includes the process plans and equipment data for a two-product, nonvolatile memory chip factory. The model contains 83 machine groups and two product flows, with product 1 having 210 steps and 14-mask layers, and product 2 having 245 steps and 16-mask layers. The initial lot size is constant for product 1 and product 2 with 48 wafers per lot. Batch sizes are all constant, using the minimum batch sizes listed in Dataset 1, except for machine 30. This was the only batch machine group that was over 75% utilized. In a real factory, a batch machine would only be a bottleneck if the batch size could not be increased. Thus, for this machine a larger batch size was used to reduce the average utilization to 72.7%. The model was set up with a deterministic product release rate of 2.19 h per lot release and first-in first-out (FIFO) at the queues. The bottleneck, an implant station, was scheduled using FIFO/SA (setup avoidance). The setup logic was only utilized on machine group 10 since it was the primary bottleneck. Dataset 1 lists both machine groups 10 and 11 as using the setup logic. The setup logic slowed the model dramatically, so it was not added to machine group 11. Not using the setup logic on machine group 11 had the effect of giving machine groups 10 and 11 similar

Manuscript received February 13, 2001; revised February 18, 2002. This work was supported by the National Science Foundation under Grant ECS-0002098.

J. Hunter and D. Collins are with the Department of Manufacturing and Aeronautical Engineering, Arizona State University East, Mesa, AZ 85212 USA.

D. Delp and J. Si are with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-7606 USA.

Publisher Item Identifier S 0894-6507(02)04454-8.

utilizations, 90.5% and 88.1%, respectively, for the initial setup. For the purposes of this study, the setup logic does not bias the results, as the factory will be the same for all treatment conditions.

In fact, shifting bottlenecks represent a key factory dynamic, so bottleneck machines close to the same utilization were actually desired. The target factory bottleneck utilization for the simulator was 90% including downtime. The plant is fully automated with no labor constraints. The goal was to see if policies could improve performance by tracking downtime; thus, the labor variability was an unnecessary variable. The plant operates 24 hours a day, 7 days a week. All processing, setup, and move times are constant. The mean time to repair (MTTR) and mean time between failure (MTBF) given in Dataset 1 have exponential distributions, and machine breakdowns do not result in damaged products. The likelihood of scrap and rework were based on the probabilities defined in Dataset 1 for each machine.

B. Implementation

The model measures machine group utilization, WIP, cycle time, and cycle time variability. Product lots are released by a generator and sent to a router. Each lot is assigned the following attributes: type (product 1 or product 2), step number, and lot size. The router sends the product to the proper machine group based on the product type and step number. Within each machine group lots wait in a queue until a machine is available. The setup, load, processing, and unload times are set within the machine group based on the product type and step number. Each machine has a random downtime generator, set according to the MTBF and MTTR from Dataset 1. Lots in process at a machine during a breakdown will remain at that machine until it is available again for processing. Once a lot is processed, it may be randomly selected for scrap or rework. Product scrap and rework were modeled using a normal distribution with the mean matching the probability of a lot being scrapped or reworked as stated in Dataset 1. Some steps may scrap and/or rework either entire lots or individual wafers. If the lot is acceptable, then it returns to the router and continues to the next step in the process flow. If the lot or wafers are scrapped, they exit the factory. If the entire lot is selected for rework, then it is routed for rework. In the special case where only select wafers are reworked, the remainder of the lot waits until the wafers are reworked before proceeding to the next step in the process flow. Some machine groups have transportation times, as defined in Dataset 1, to simulate transit time to the next processing step. When a lot has completed all steps required, it exits the factory. At this time the cycle time is calculated for that lot.

The batch processing subroutine required unique logic since the EXTEND batching icons could not accommodate the variable lot sizes properly. Scrap and rework could result in lots ranging in size from 1 to 48 wafers. Thus, it was not possible to predict the number of lots required to meet the minimum batch size. A batching subroutine was developed to collect lots until the minimum number of wafers was accumulated.

EXTEND did not have an icon to adequately reflect all the logic that goes into setup decisions. The wafer lots have two types of setup that are required at the implant machines (machine groups 10 and 11). The setup times for a chemical change are either 30 min or an hour, while the setup for a recipe change is 7 min. Due to this difference, the setup logic minimizes the number of chemical change setups, as this provides the most benefit. The setup logic attempts to reduce setups by only sending lots to machines that are already set up for the same chemical. Each machine within the group has its own FIFO queue where lots are stored for use on that machine only. To keep the simulation run times as short as possible, the setup reduction was only used on machine group 10 (medium current implant) and not on machine group 11 (high current implant), as listed in Dataset 1. Machine group 10 is the primary bottleneck in the factory, even with the setup logic. Machine group 11

without the setup logic remains less utilized than machine group 10 with the setup logic.

C. Validation

The model validation consisted of comparing model predictions to expected results for special factory conditions. Three validation tests were performed on the model to demonstrate the correct product process flows, individual machine downtimes, and process times via machine utilization. The model was validated against historical data given in the SEMATECH dataset.

The first test verified the correct process flow for both products. Each product has a detailed process flow path that requires visiting a large number of machine groups, with multiple visits to most groups. The model was run to verify that each machine group was visited the correct number of times by each product. The rework and scrap functions were turned off for this test to keep the number of machine visits for the individual lots regular. This test consisted of releasing 50 lots of product 1 and 50 lots of product 2 into the system. The step number and machine group visits were monitored for each product with 48 lots completing all 210 steps for product 1 and 46 lots completing all 245 steps for product 2 in the process flows. The lots remaining in the system (2 for product 1 and 4 for product 2) were stranded at batching operations [1].

The second verification test consisted of comparing the machine downtimes listed in Dataset 1 to the simulated machine downtimes in the model. Machine downtimes were modeled using random number generators included in the EXTEND modeling software. Validation consisted of comparing the percentage of machine downtimes from Dataset 1 to the percent downtimes from running the simulation. The run time for this validation test was 20 000 h, which is considerably longer than the longest MTTR (66 h, 49 min) or MTBF (798 h, 19 min) in the system. With this simulation length, every machine produced a breakdown. The MTTR and MTBF were modeled as exponential distributions based on the data given in the SEMATECH dataset. The simulation was run three times with a different random seed generator to measure repeatability for this test. Eleven machine groups had a difference of 4.3% to 1%, while 72 machine groups had less than 1% difference from the simulated downtime to the downtime listed in Dataset 1 [1].

The third validation test analyzed the processing time on each machine through machine utilization values by comparing Dataset 1 machine utilizations with the simulated machine utilizations. The process times for the machines were constant values in the simulator and Dataset 1. The amount of processing time on each machine dictates the throughput and utilization of each machine group. The utilization for the machine group was the average of the individual machine utilizations within the group. Machine breakdowns and downtimes were considered utilized. The tests for product 1 and product 2 were run separately with a constant release rate and with no downtime, scrap, or rework. Product 1 was run for 1000 h and product 2 was run for 1500 h to allow a sufficient number of lots to cycle through the process. For testing product 1 all machine groups had a utilization difference of 1.5% or less, 80 machine groups were less than 1% difference. For testing product 2 all machines had a utilization difference of 1.3% or less, 82 machines were less than 1% difference [1].

III. PILOT CAPACITY RUNS

Several pilot simulation runs were made with the full-scale model to establish appropriate factory ramp-up time and data collection intervals. The selection of ramp-up period and sampling plan was based on the work of Torsina [7], who performed a detailed statistical analysis. Several others validated this ramp-up time in other research

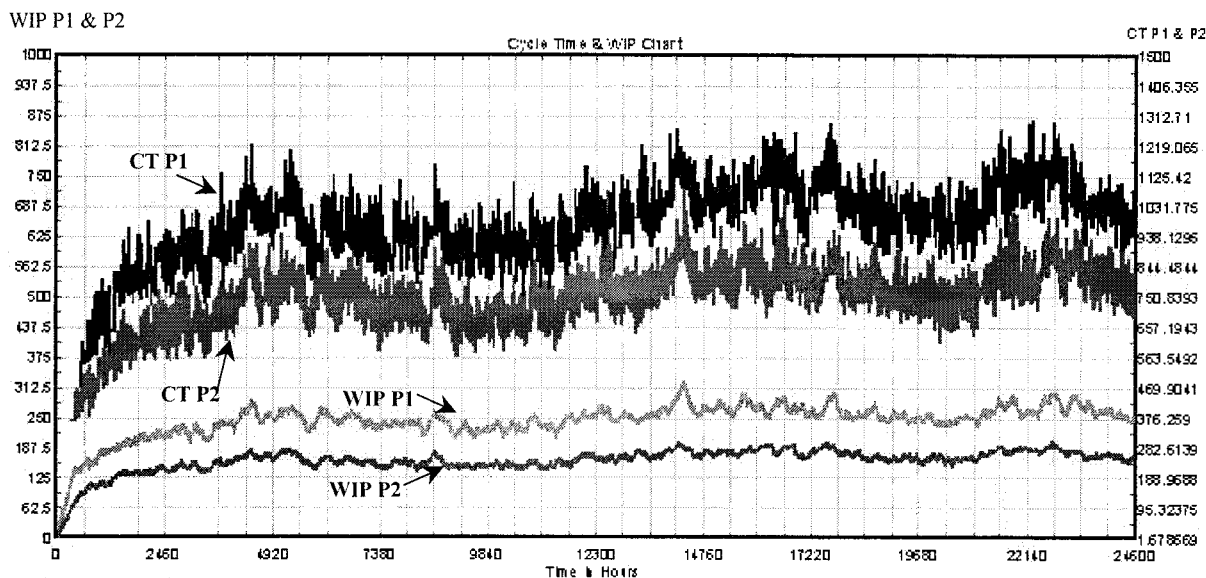


Fig. 1. WIP and cycle time measurements for products 1 and 2 of the pilot run at 98% capacity.

TABLE I
TEST RUNS—30 MONTH AVERAGE

Run Capacity	Average WIP (total lots)	Average Cycle Time (hours)	
		Product 1	Product 2
90%	310	642	825
95%	381	744	970
98%	431	785	1042

[3], [4]. Torsina’s analysis stemmed from recommendations of Pegden *et al.* [8] that stated statistically independent results can be obtained using a batch means approach. Fig. 1 shows the factory WIP and cycle time results from a pilot run. Based on this graph, a ramp-up period of 125 days (3000 h) was selected. A simulation time of 30 months (30 days/month, 720 h/month), with monthly averages collected for the WIP, cycle time, and cycle time variability measurements, was used. Thus, the total run time for each simulation was 125 days + (30 months × 30 days/months) = 1025 days or 24 600 h.

All the pilot runs used the same product mix (2/3 of product 1 and 1/3 product 2). All statistics were reset at the start of each month. The averages of the cycle time and WIP measurements are shown in Table I for the model at 90%, 95%, and 98% capacity. A comparison of the average cycle times between the model simulations at 90% and 95% capacity and the SEMATECH Dataset 1 is shown in Table II.

The results for Dataset 1 were collected on a system where the bottleneck machine was 95% utilized. The lot arrival rates to the model were constant and produced only 90% utilization at the bottleneck machine in the first simulation on the model. This reduction in bottleneck utilization leads to a reduction of system WIP and cycle time. The cycle time measurements for this simulation were within 82 h (or 9%) of the SEMATECH dataset measurements. The second simulation run also had constant lot arrival rates, but the product arrival rates were increased to produce 95% utilization at the bottleneck machine to match the capacity of Dataset 1. The increased release rate produced an increase in the system WIP and product cycle times. The cycle time measurements on the 95% capacity model were within 63 h (or 7%) of the SEMATECH dataset results. The third simulation run at 98% utilization further increased the system WIP and cycle time for both products. Dataset 1 supplies only the minimum and maximum number

of wafers allowed for batching at each batch-processing machine; thus, it is an open parameter in the model. The number of wafers is not exactly known or constant for each batching machine within the process. This is a cause for discrepancy in the cycle times between the model simulations and the dataset.

IV. SPECIAL SCENARIO RUNS

An aspect of flexible modeling is to accommodate for special scenarios and forecasting. The special scenarios are easy to implement and useful for determining the effects caused by product shifts and machine breakdowns or replacements. Running various scenarios can be used as a planning aid for adding new customer products and machines to the process, and determining the resulting bottleneck shifts and capacity in the semiconductor manufacturing plant. The full-scale model was run to demonstrate a change in the product mix and a theoretical factory with no equipment failures or emergency maintenance. A change in product mix is used to determine the effect of losing a product, due to changing technology or loss of a customer. Products 1 and 2 were run separately in the full-scale model to create the change in product mix. Table III shows the results of the single product runs. The WIP and cycle times were lower for the single product scenarios compared to the two-product mix. Most notably, the bottleneck changed in the product 2 run. This created a shift in the bottleneck from machine group 11, an implant station, to machine group 30, an oxide station. The dual product mix and single product mixes were run for a theoretical factory having no equipment failures or emergency maintenance. Table IV shows the results for the theoretical runs. Without failures and maintenance the system WIP and product cycle times were lower, as well as the WIP and cycle times for the single product mixes. Again, the most prominent change was the shifting bottleneck. Machine group 78 is an inspection station and was the bottleneck for the dual product and product 1 runs. Without machines being shutdown for maintenance the inspection station becomes heavily utilized at 91% for the dual product mix. Machine group 30, the oxide station, remained the bottleneck for the product 2 run.

V. SMALL-SCALE MODEL

An alternate model was developed using Dataset 1 from SEMATECH. This model was built to catch the essence of the

TABLE II
AVERAGE PRODUCT CYCLE TIMES, MODEL AT 90% AND 95% CAPACITY

	Product 1	% Difference (%)	Product 2	% Difference (%)
Dataset 1 Avg. CT (hours)	702	-	907	-
Model at 90% Avg. CT (hours)	642	8.5	825	9.0
Model at 95% Avg. CT (hours)	744	6.0	970	6.9

TABLE III
FULL-SCALE MODEL WITH SINGLE PRODUCT MIXES

Run	Bottleneck Machine Group	Average WIP (total lots)	Average Cycle Time (hours)	
			Product 1	Product 2
Both Products	MG #11	431	785	1042
Product 1	MG #11	186	453	-
Product 2	MG #30	89	-	539

TABLE IV
THEORETICAL FULL-SCALE MODEL WITH DUAL AND SINGLE PRODUCT MIXES

Run	Bottleneck Machine Group	Average WIP (total lots)	Average Cycle Time (hours)	
			Product 1	Product 2
Both Products	MG #78	251	379	474
Product 1	MG #78	152	344	-
Product 2	MG #30	71	-	406

same semiconductor manufacturing plant without the detail of the full-scale model. The small-scale model consisted of six machine groups, which include a stepper, stripper, two batching ovens, implant station, and an inspection station. The model takes into account reentrant product flows, rework and scrapped wafers, setups, batching, machine downtime, and setup logic at the implant station. The model contains two product flows with product 1 having 75 steps and 6-mask layers, and product 2 having 85 steps and 8-mask layers. The initial lot size is constant for product 1 and product 2 with 48 wafers per lot. The model was set up with a deterministic product release rate and FIFO at the queues. The implant station was scheduled using FIFO/SA. The plant operates 24 hours a day, 7 days a week. All processing, setup, and move times are constant. Scrap and rework were also modeled. The MTTR and MTBF have exponential distributions, and machine breakdowns do not result in damaged products. The small-scale model ran at the same input release rate as the full-scale model for 24 600 h of production time with a 3000-h ramp-up time, and the same product mix (2/3 of product 1 and 1/3 of product 2).

The model validation consisted of comparing the process flows of the small-scale model to the process flows of Dataset 1. The process flows for the small-scale model were correlated to match the process flows of the full-scale model reentering the six machine groups the same number of times. A validation test was performed on the small-scale model to show the correct product process flows. The machine downtimes and process times via machine utilization remain the same for the small-scale model as for the full-scale model. Validation for the machine downtimes and process times were previously demonstrated. The small-scale model was run to verify that each machine group was visited the correct number of times by each product. The process flow from the small-scale model matched the process flows given in the SEMATECH dataset for both products. Only machine group 78 had a difference in the number of visits as compared to the visits in Dataset 1 for each product. Machine group 78, an inspection station, had fewer

TABLE V
FULL-SCALE AND SMALL-SCALE MODEL COMPARISON

Run	Bottleneck Machine Group	Average WIP (total lots)	Average Cycle Time (hours)	
			Product 1	Product 2
Full-Scale 98%	MG #11	431	785	1042
Small-Scale 87%	MG #8	84	162	205

TABLE VI
SMALL-SCALE MODEL WITH SINGLE PRODUCT MIXES

Run	Bottleneck Machine Group	Average WIP (total lots)	Average Cycle Time (hours)	
			Product 1	Product 2
Both Products	MG #8	84	162	205
Product 1	MG #10	32	98	-
Product 2	MG #10	20	-	122

visits since in the full-scale model inspection is carried out from processes other than the five machine groups in the small-scale model.

VI. FULL-SCALE AND SMALL-SCALE MODEL COMPARISON

The small-scale model gave an overall look at a semiconductor manufacturing plant with the advantage of quick implementation. There were several notable discrepancies between the small-scale model and full-scale model results as shown in Table V. The most significant difference is the predicted bottleneck. The small-scale model shows machine group 8, the stripper associated with rework, as the bottleneck as opposed to machine group 10, the implant station. The bottleneck in the full-scale model, machine group 11, is an implant station. By reducing the number of processing steps while developing the small-scale model, the process flow, product cycle times, and WIP became dependent on the amount of lot rework in the system. Using the same release rate, the bottleneck utilization was lower for the small-scale model than for the full-scale model. The cycle times for both products and WIP in the small-scale model were lower than in the full-scale model, as would be expected due to the shorter process flow, thus producing lower utilization.

The small-scale model ran the same special case scenarios as the full-scale model. Table VI shows the results of the single product runs. The WIP and cycle times were lower for the single product scenarios compared to the two-product mix. The most significant change was the bottleneck shift from machine group 8, the stripper, to machine group 10, an implant station, for both single product runs. The dual product mix and single product mixes were also run in a theoretical factory setting having no equipment failures or emergency maintenance. Table VII shows the results for the theoretical runs. Without maintenance, the system WIP and product cycle times were lower. The theoretical factory kept a constant bottleneck at machine group 8, the stripper, for all product mixes. The stripper machine group was heavily utilized at 92% for the dual product mix.

TABLE VII
THEORETICAL SMALL-SCALE MODEL WITH DUAL AND SINGLE
PRODUCT MIXES

Run	Bottleneck Machine Group	Average WIP (total lots)	Average Cycle Time (hours)	
			Product 1	Product 2
Both Products	MG #8	43	82	100
Product 1	MG #8	22	68	-
Product 2	MG #8	14	-	88

The models were simulated on a 1 GHz Pentium III computer. The 24 600-h simulation runs required 1 h and 43 min to run for the full-scale model and 41 min to run for the small-scale model. The accuracy of the results and the relatively quick simulation time for the full-scale model outweigh the fast implementation advantage of the small-scale model. The trends and characteristics of a plant can be predicted in a couple hours. This is beneficial when determining the long-term effects of a 30-month schedule on a full-scale semiconductor manufacturing plant. With today's computing power the full-scale model takes an hour longer to run, with more accurate and beneficial results.

VII. CONCLUSION

Dynamic event models can be designed to match the characteristics of a real-life system to evaluate numerous strategies for operation. EXTEND software was selected for the development of a semiconductor manufacturing plant simulator due to the ease of its object oriented model creation and wide array of pre-made modeling subroutines. A fully automated semiconductor manufacturing plant was developed, and flow routes, machine downtimes, and processing times were debugged and verified with a SEMATECH dataset. Several initial simulations were used to determine ramp-up times, simulation times, and data collection intervals, thus giving initial WIP and cycle time measurements for the full-scale semiconductor manufacturing plant. Although several minor discrepancies appeared between the model and

the dataset, the initial measurements produced quality results within $\pm 10\%$ of the SEMATECH dataset given the specific constraints from the dataset. The model also identified shifting bottlenecks and trends in the system WIP and product cycle times when special case scenarios were run. A quickly developed small-scale model demonstrated the inaccuracies in factory modeling when assumptions to reduce the number of machine groups are applied to simplify the modeling process. The purpose of the full-scale simulator was to develop individual machines and machine groups, and to integrate them as a whole to produce a functional model. The overall result was a valid model for predicting trends and analyzing results and schedules for a semiconductor manufacturing plant.

REFERENCES

- [1] J. Hunter, "Release and Queuing Policy impact on semiconductor factory performance," Masters thesis, Arizona State Univ. East, Mesa, AZ, 2000.
- [2] V. Palmiri and D. W. Collins, "Analysis of the K -step ahead[®] minimum inventory variability policy using SEMATECH semiconductor manufacturing data in a discrete-event simulation model," in *6th Annu. IEEE Int. Conf. Emerging Technologies and Factory Automation*, 1997, pp. 520–527.
- [3] D. Delp, J. Si, D. W. Collins, J. Hunter, and J. Fowler, "Development of a full scale semiconductor manufacturing model for advanced input control and bottleneck queuing," in *Proc. Int. Conf. Modeling and Analysis of Semiconductor Manufacturing (MASM 2000)*, 2000, pp. 220–225.
- [4] D. W. Collins, T. Torsina, and R. Balgemann, "A simulation study to compare minimum inventory variability policies (MIVP[®]) and first-in-first-out (FIFO) algorithm," in *Proc. IFAC'99, 14th World Congr. IFAC Int. Federation of Automatic Control*, 1999, pp. 461–466.
- [5] J. Fowler. (1997) Modeling and Analysis for Semiconductor Manufacturing Laboratory: SEMATECH Dataset 1. Arizona State Univ., Tempe, AZ. [Online]. Available: www.eas.asu.edu/~masmlab/
- [6] D. J. Miller, "Simulation of a semiconductor manufacturing line," *Commun. ACM*, vol. 33, pp. 98–108, Oct. 1990.
- [7] T. Torsina, "Simulation study to compare minimum inventory variability policies and first-in-first-out algorithm," Masters thesis, Arizona State Univ., Tempe, AZ, 1997.
- [8] D. Pegden, R. Shannon, and R. Sadowski, *Introduction to Simulation Using SIMAN*, 2nd ed. New York: McGraw-Hill, 1995.