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

A QUEUEING NETWORK APPROXIMATION OF SEMICONDUCTOR AUTOMATED MATERIAL HANDLING SYSTEMS: HOW MUCH INFORMATION DO WE REALLY NEED?

Theresa M. Roeder

Graduate School of Management
University of California, Davis
Davis, CA 95616, U.S.A.

Nirmal Govind

Intel Corporation
Chandler, AZ 85226, U.S.A.

Lee W. Schruben

Dept. of Industrial Engineering and Operations Research
University of California, Berkeley
Berkeley, CA 94720, U.S.A.

ABSTRACT

Queueing networks are sometimes used to model material handling in flexible manufacturing systems. We explore the use of a closed queueing network model to approximate an intrabay automated material handling system (AMHS) in semiconductor manufacturing. Rather than solving the model analytically, we propose simulating it. Current industry models are very complex and require long development and run times. The simulated approximation can be used as an easy and fast alternative. To compare the approximation with the detailed models in use, we employ an information taxonomy to classify AMHS models based on the amount and types of information needed to model the system, and to obtain desired output. This classification aids modelers in determining the level of detail to incorporate in a model based on the objectives of the simulation study.

1 INTRODUCTION

The need for and dependence on Automated Material Handling Systems (AMHSs) in the semiconductor industry has been increasing in recent years. With the move to 300mm wafers and fully-automated wafer fabrication facilities ("fabs"), the ability to efficiently and accurately design and control material handling is becoming more critical (Nadoli and Pillai 1994; Jefferson and Pillai 1999).

Analytic approaches that address the design and control of semiconductor AMHSs are limited. Moreover, the complexity of these systems coupled with the uncertainties in a fab make analytical modeling intractable. This makes simulation an attractive option for studying such systems, and it is no surprise that simulation modeling has become the primary method for analyzing AMHSs (and fab operations in general). Unfortunately, the level of detail and complexity usually associated with AMHS simulation models can cause excessive runtimes. Mackulak and Savory (2001) describe a study in which the desired AMHS experiments took over 250 hours of simulation time. Many different approaches have been taken to combat this problem, as well as the problem of extensive simulation development times. For example, Mackulak and Savory (2001) and Jefferson and Pillai (1999) combine simulation models with statistical design of experiments (DOE) to reduce the number of required runs. Paprotny et al. (2000) distribute simulations across machines to reduce runtimes. Mackulak et al. (1998) describe generic AMHS simulations that can be built upon to significantly reduce the development time.

Though analytic work on semiconductor AMHS is limited, there exists considerable work on analytic models for material handling systems in flexible manufacturing systems (FMSs) (see Johnson and Brandeau (1996) for a survey of work in this area). Many models use $M/G/c$ approximations for the system. Other approaches include integer and nonlinear programming models. A disadvantage of math programming is that it does not capture the variability inherent in AMHSs easily.

To incorporate variability, Tanchoco and Egbelu (1987) use a closed queueing network approximation software to determine the number of automatic guided vehicles (AGVs) required in a material transportation system. They compare the analytic results of this queueing approximation to simulation results and find that the approximation produces lower bounds on the number of vehicles. Curry et al. (2003) develop a closed queueing network model of a fixed-route material handling system. They show that the analytic

results are statistically identical to simulation results, but that the time to solve the analytic model is exponential in the number of transporters in the system. In this paper, we propose a *simulation-based* closed queueing network approximation of an intrabay AMHS that can help alleviate some of the problems with the current approach for simulating semiconductor AMHS.

Long simulation run times are one of the primary problems with current semiconductor AMHS simulations. Rather than trying to reduce run times by distributing runs across machines or reducing the number of required runs, some simulation methodologies take an alternate approach to the simulation modeling itself. For example, simplified simulation models can be used to approximate system behavior. Because they underestimate the variability, safety factors are introduced to compensate. In Wu et al. (1999), the authors use data on existing systems to approximate the size of the safety factor. A similar approach is taken in Rasmidatta et al. (2002), where the authors use a simplified simulation known as a resource-driven (RD) simulation as a control variate for the full simulation, known as a job-driven (JD) simulation (Schruben and Roeder 2003). RD simulation runtimes tend to be shorter than JD runtimes, and the authors develop calibration metrics to adjust the RD output to be more in-line with the JD output.

While the common modeling approach for semiconductor AMHS has been to simulate in great detail, there are several reasons for using simplified or approximate simulation models:

1. *Use of appropriate level of detail:* Often, AMHS simulation models are used to analyze systems that do not yet exist. In other cases, the simulation models systems that are changing, making the model or original data used obsolete. In both cases, a minimal amount of accurate input data is available (Jefferson and Pillai 1999). Building a detailed model of a system where the details may not be known, or are known with limited accuracy, is a poor use of modeling resources. It may be wiser to use an approximate model that captures the important elements of the system, rather than implementing details and gaining a false sense of confidence in the results of the simulation.
2. *Shorter development time:* Because less detail is included in the model, development time can be significantly reduced.
3. *Reduced number of errors:* Including fewer details in the model reduces the opportunity to introduce errors into the model; this includes both programming and modeling errors. (The approximation itself will not capture all system aspects. In contrast to programming bugs, these "errors" are introduced explicitly and intentionally in the decision to use the approximation.)

In this paper, we investigate using a closed queueing network simulation model as an approximation of the intrabay AMHS. The objective of this study is to determine if the approximation can provide either estimates of performance measures that are within reasonable degrees of error, or reliable and useful bounds on the performance measures. A more detailed simulation model is developed to aid in the estimation of error from the approximation based simulation model. To compare the two models, we introduce an information taxonomy that explicitly shows the difference between the two models in terms of the information used. The approximation model uses less information than the detailed simulation model; hence one of the important aspects that we investigate is the amount of information (level of detail) that is necessary to achieve reasonably accurate estimates of desired performance measures.

In Section 2, we describe semiconductor AMHS in general, and intrabay AMHS in particular. Section 3 introduces and justifies the proposed closed queueing network approximation. Section 4 outlines the information taxonomy, which is then used in Section 5 to compare the two simulation models. We give concluding remarks in Section 6.

2 AMHS IN SEMICONDUCTOR MANUFACTURING

Automated material handling in semiconductor manufacturing facilities has come a long way since its inception. Floor-based systems such as the Rail Guided Vehicles (RGVs) and Automated Guided Vehicles (AGVs) have been replaced with overhead systems. Jefferson and Pillai (1999) describe several reasons why overhead systems are superior to floor-based systems, e.g., greater personnel safety, smaller footprint, and better scalability. The advent of 300mm wafers have made the AMHS a critical component of the fab; it is no longer practical to hand-carry a lot, which typically consists of 20-25 wafers in an enclosed container known as a Front Opening Unified Pod (FOUP) and can weigh up to 16 lbs (compared to around 8 lbs in 200mm).

Processing equipment (tools) in fabs are located in bays; the number and types of tools in a bay can vary. The AMHS consists of interbay and intrabay systems. The interbay system is responsible for transporting lots between bays for processing or storage. The systems interface at stockers, which also serve as storage for lots that cannot be processed immediately. Lots are dropped off at stockers by the interbay system. The Overhead Hoist Vehicles (OHVs) of the intrabay system pick up the lots and move them to a tool for processing. Once processing is complete, OHVs return the lots to the stocker, where they are picked up by the interbay system.

The intrabay system, the focus of this paper, consists of an overhead track on which OHVs travel. An OHV carries one lot in a FOUP and is able to interface with the stocker and the processing equipment to pick up or unload a FOUP. We will collectively refer to the overhead track, the OHVs, the stockers, and the processing equipment within a bay as the intrabay system. A generic intrabay system is depicted in Figure 1.

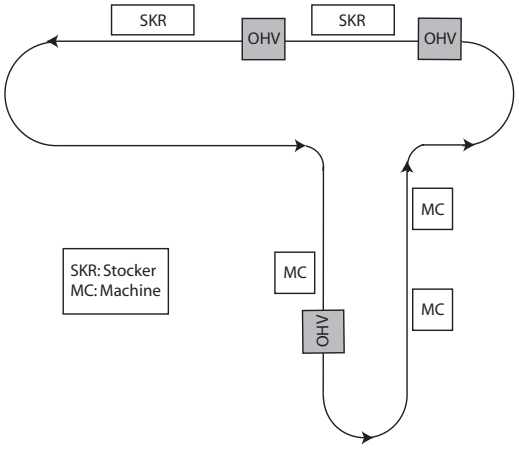

Figure 1: Intrabay System

We now present a closed queueing network approximation of the intrabay system.

3 CLOSED QUEUEING NETWORK APPROXIMATION

Closed queueing networks are often used to model production systems (Duenyas and Hopp 1990; Govind 2004). In closed queueing network models, the numbers of jobs (N) and machines (M) are fixed, with the jobs cycling around and visiting the machines in sequence as shown in Figure 2. In a semiconductor AMHS, the cycling of jobs can be abstracted by considering the "jobs" to be the OHVs in an intrabay system (rather than the lots). An OHV may pick up a load from a stocker and drop it off at a machine, traversing machines on the way; or it may pick up a load from a machine and drop it off at a stocker, again traversing machines on its way. We see the same behavior in closed queueing networks.

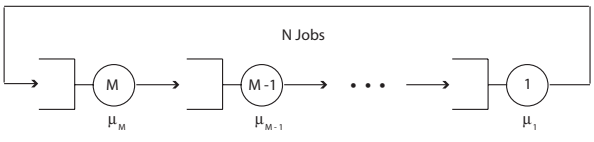

Figure 2: Closed Queueing Network

We model the intrabay AMHS using this closed queueing network approach as follows. The stocker in an intrabay system is modeled as one of the M machines in Figure 2, say machine M. The arrival of lots to the stocker from the interbay system can be approximated as arrivals to this machine. Note that the system is still a closed network since we are modeling the cycles of resources, the OHVs, and not the flow of jobs or lots. This is in the spirit of RD simulations discussed earlier. Say an OHV in the real system needs to transport a lot from the stocker (machine M) to a tool, say machine 2. In the closed queueing network, the OHV will proceed from machine M to machine $M-1$, get "processed" at machine $M-1$ with a zero processing time, move on to machine $M-2$, again get processed with a zero processing time, and so on until it reaches machine 2. At machine 2, the OHV will unload the lot (get "serviced" at the machine), and is then free to pick up a waiting lot (if it exists) from machine 2 or machine 1 to transport the lot to the stocker. If no lots are waiting, the OHV continues cycling through the system with zero processing time at the machines until it finds a lot at the stocker or a machine.

This is clearly a simplification of the real system. The fact that the closed queueing network is only a simplified approximation of the real system brings up the question of the usefulness of such a model, especially in terms of the accuracy of the results that can be obtained. We will provide a comparison of the approximation and a more detailed simulation model in Section 5, after introducing an information taxonomy in Section 4 to aid in the comparison.

4 INFORMATION TAXONOMY

While clearly there will be differences between a detailed model and the approximation in Section 3, an explicit, quantifiable list of the differences is helpful for comparing the models. To develop this list, we use an information taxonomy proposed in (Roeder 2004). We do not use the term "taxonomy" in its strictest sense. Rather, we are using it to describe a means of systematically organizing the information contained in a system or a model of a system. The classification is independent of the implementation chosen (e.g., a process interaction versus an activity scanning-based simulation), but does depend on the purpose of the model.

There are several characteristics by which we can categorize information. Each group of characteristics is exclusive and exhaustive, but the groups are not mutually exclusive. The groups are described below, and the taxonomy is shown in Figure 3.

General system or entity-specific: Information is either general information about the system (e.g., number of resources), or related specifically to entities (e.g., job j's waiting time in queue). We refer to general information as "non-subscripted" and to entity-specific information as "subscripted." Subscripted information is typically relevant

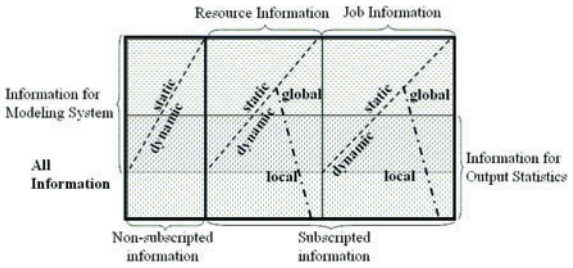

Figure 3: Information Taxonomy

if we are interested in the differences between system entities. If the differences are less important in the context of the current study than the similarities, it may not be necessary to maintain this information. Aggregating information on similar entities into general system information can reduce the memory and computational effort required for the simulation.

Subscripted: Resource or job: Subscripted information can be further subdivided into information pertaining to resources (e.g., last failure time) and jobs (e.g., sojourn time). Usually, the number of resources is fixed during a simulation run, while the number of jobs may vary. The number of jobs can be significantly larger than the number of resources, especially for large or congested systems. In these cases, the memory and computational cost of maintaining job information can be significant.

Subscripted: Local or global: Subscripted information is either local or global. (All non-subscripted information can be considered global.) An example of local job information is the waiting time at the current queue, while global job information is the job's arrival time to the system. Local information need only be maintained while the job is at its current location. In contrast, global information must be tracked during the job's entire existence in the system. In some cases, we may be able to reduce the complexity of the model by only tracking jobs explicitly in certain parts of the model, e.g., at the bottleneck. Most resource information is global.

Modeling, statistics, or both: The information included may be needed to model the system (e.g., job routing), to collect the desired output statistics (e.g., the cumulative time spent in repair), or for both (e.g., resource utilization). This categorization is of interest because it can highlight that certain pieces of information are not necessary. For example, if we wish to estimate the average queue size for a $G/G/s$ queueing system, the only variable required to model the system is the number of jobs in the system. (Collecting statistics requires an additional variable, the cumulative area under the queue versus time curve.) Explicit job information need not be stored.

Static or dynamic: Information may be either static (usually input parameters) or dynamic. For example, the $G/G/s$ model mentioned above has one piece of static information, s, and two pieces of dynamic information, the number of jobs in queue and the queue accumulator.

Although we will not do so here, the information taxonomy can be used for complexity analyses of models (Roeder 2004). In the next section, we describe the differences between the approximation and a detailed simulation model using the taxonomy's formalism.

5 COMPARISON

To answer the question regarding the usefulness of the approximation, we developed two simulation models: a simulation model that captures the detailed workings of the intrabay system, and an approximate model based on the closed queueing network.

The detailed model is an attempt at an AMHS intrabay simulation at the level of detail commonly used in the semiconductor industry. This model explicitly tracks each OHV's location, models intelligent assignment of lots to OHVs, etc. There is a significant difference in the development time for the two models. While the approximation was developed in only a few days, the "full" model used as a comparison required several weeks. The development times highlight the value of considering alternate approaches. Both models were developed in SIGMA, an event scheduling simulation package (Schruben and Schruben 2001).

The biggest difference between the two approaches is in how OHVs are modeled. In the full model, they are modeled explicitly: Both local and global subscripted information is used. This explicit modeling allows us to capture OHV blocking (when an OHV is prevented from moving forward by another OHV), which uses dynamic global job information; in order to model blocking, we must know the locations and statuses of all OHVs in the bay. Local OHV information is used to model OHV behavior at machine loadports.

This same local OHV information is the only OHV information used in the approximation. Because we do not store global OHV information, we are not able to explicitly model blocking. However, the amount of time OHVs are blocked is captured in the amount of time vehicles spend in queue at machines. The difficulty of modeling OHV blocking is circumvented, a significant savings in both development time and data manipulations required is achieved; only a fraction of the OHV information tracked in the full model is used (local information for a subset of OHVs versus local and global information for all OHVs).

The second large set of information not used in the approximation is lot information. In the approximation, lots are represented implicitly at machines as busy/idle loops, and using integer counts of the number of lots waiting at the different positions in the system, global system information. In contrast, the full model maintains a record for each job;

unlike in the approximation, job waiting times and order in queue are known. The memory requirements for the full model increase as a function of the number of lots in the system. The memory requirements and execution speed for the RD approximation are insensitive to system congestion.

Both models include the same information on machines and loadports.

Vehicle blocking is not the only thing that cannot be explicitly modeled because of lack of information in the approximation. For example, because we are not explicitly modeling lots or OHVs in the approximation, lots are not assigned to OHVs when the lots arrive at the stocker, or complete service at machines. Rather, passing OHVs check whether there is a queue of lots waiting. This tends to underestimate the average time that a lot waits for an OHV pickup. The lack of information also reduces the available output statistics. For example, because the approximation does not track lots, we do not know the distribution of lot waiting times; however, mean waiting times by lot type are available (see Roeder (2004) for a method for estimating delay distributions in RD models).

The approximation can capture the major elements of the intrabay system such as the movement of the OHVs, dropoff and pickup at machines and stockers, interaction of the OHVs with the machine loadports, and processing of lots at machines. As mentioned above, blocking is captured via the machine queues. In addition, a simulation of a closed queueing network model can provide us with estimates of performance measures such as the average delivery time and the average waiting time for an OHV.

Although we do not provide detailed empirical results here, the preliminary results from the comparison of the two models are promising. For example, the number of moves for the systems are comparable, and the approximation appears to provide a good lower bound on average queue sizes. A full comparison of the approaches will be presented in a future publication.

6 CONCLUSIONS

In this paper, we propose an alternate closed queueing network-based approach to simulating intrabay AMHS in semiconductor manufacturing. The advantages of the approach introduced here are that it is simple to implement, and has far fewer data requirements than an explicit model of the system. Using an information taxonomy, we are able to quantify the differences between the explicit AMHS simulation and the queueing network approximation both in terms of the information needed to develop the models, and in terms of the output statistics that are available from the two models. The approximation is able to provide estimates of most of the major performance measures that are tracked in the complex AMHS models in use today.

An approximation that can provide estimates that are comparable with those of the complete model could be a valuable addition to the AMHS design and analysis toolbox because of its ease of development and use. The statistics that can be obtained from it may be sufficient to eliminate certain system configurations without using valuable resources.

ACKNOWLEDGMENTS

We wish to thank Tim Quinn, Chuck Golla, and Paul Cherry at Intel Corporation for helping us with their expertise in semiconductor AMHS.

REFERENCES

Curry, G., B. A. Peters, and M. Lee. 2003. Queueing network model for a class of material-handling systems. *International Journal of Production Research* 41 (16): 3901–20.

Duenyas, I., and W. Hopp. 1990. Estimating variance of output from cyclic exponential queueing systems. *Queueing Systems* 7 (3-7): 337–54.

Govind, N. 2004. Robust parameter design with imperfect experimental control of noise. Ph.D. Dissertation, The Pennsylvania State University.

Jefferson, T., and D. Pillai. 1999. Throughput analysis and modeling of 300mm intrabay transport vehicles. In *International Conference on Semiconductor Manufacturing Operational Modeling and Simulation*. San Francisco.

Johnson, M. E., and M. L. Brandeau. 1996. Stochastic modeling for automated material handling system design and control. *Transportation Science* 30 (4): 330–50.

Mackulak, G. T., F. P. Lawrence, and T. Colvin. 1998. Effective simulation model reuse: A case study for amhs modeling. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. Watson, J. S. Carson, and M. S. Manivannan, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, 979–84.

Mackulak, G. T., and P. Savory. 2001. A simulation-based experiment for comparing amhs performance in a semiconductor fabrication facility. *IEEE Transactions on Semiconductor Manufacturing* 14 (3): 273–80.

Nadoli, G., and D. Pillai. 1994. Simulation in automated material handling system design for semiconductor manufacturing. In *Proceedings of the 1994 Winter Simulation Conference*, ed. J. D. Tew, M. S. Manivannan, D. A. Sadowski, and A. F. Seila, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, 892–99.

Paprotny, I., J. Yngve, G. T. Mackulak, and R. J. Gaskins. 2000. Applying conservative distributed simulation to a large scale automated material handling design. In *Proceedings of Western Multiconference 2000 on Communication Networks and Distributed Systems Modeling and Simulation*, 107-112.

Rasmidatta, C., S. Murray, J. W. Fowler, and G. T. Mackulak. 2002. New approaches for simulation of wafer fabrication: The use of control variates and calibration metrics. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. Charnes, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, 1414–22.

Roeder, T. M. 2004. An information taxonomy for discrete event simulations. Ph.D. Dissertation, University of California, Berkeley.

Schruben, D., and L. W. Schruben. 2001. *Graphical simulation modeling using SIGMA*. 4th Edition ed. Custom Simulations.

Schruben, L. W., and T. M. Roeder. 2003. Fast simulations of large-scale highly congested systems. *Simulation: Transactions of the Society for Modeling and Simulation International* 79 (3): 1–11.

Tanchoco, J., and P. Egbelu. 1987. Determination of the total number of vehicles in an agv-based material transport system. *Material Flow* 4 (1987): 33–51.

Wu, S., J. Rayter, I. Paprotny, G. T. Mackulak, and J. Yngve. 1999. Increasing first pass accuracy of AMHS simulation output using legacy data. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. Nembhard, D. T. Sturrock, and G. W. Evans, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, 784–89.

AUTHOR BIOGRAPHIES

THERESA M. ROEDER is a Visiting Professor at the Graduate School of Management at the University of California, Davis. She holds a Ph.D. in Industrial Engineering and Operations Research from the University of California, Berkeley. Her email address is <tmroeder@ucdavis.edu>.

NIRMAL GOVIND is a Senior Systems Engineer at Intel Corporation, and is involved in strategic factory operations. He holds a Ph.D. in Industrial Engineering and Operations Research from the Pennsylvania State University, and an M.S. in IE & OR from the University of California at Berkeley. His email address is <nirmal.govind@intel.com>.

LEE W. SCHRUBEN is Chancellors Professor and Department Chairman of the Department of Industrial Engineering and Operations Research at the University of California at Berkeley. Before joining the Berkeley faculty, he was at Cornell University in the School of Operations Research and Industrial Engineering. His email address is <schruben@ieor.berkeley.edu>.