Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

DETERMINING AN APPROPRIATE NUMBER OF FOUPS IN SEMICONDUCTOR WAFER FABRICATION FACILITIES

Jens Zimmermann
Lars Mönch

Chair of Enterprise-wide Software Systems
Dept. of Mathematics and Computer Science
University of Hagen
58097 Hagen, GERMANY

Scott J. Mason

Department of Industrial Engineering
University of Arkansas
Fayetteville, AR, 72701, USA

John W. Fowler

Department of Industrial Engineering
Arizona State University
Tempe, AZ, 85287, USA

ABSTRACT

In this paper, multiple orders per job type formation and release strategies are described for semiconductor wafer fabrication facilities (wafer fabs). Different orders are grouped into one job because orders of an individual customer very often fill only a portion of a Front-Opening Unified Pod (FOUP). A FOUP is assigned to each job and is used to move the job throughout the wafer fab after the job formation. We determine an appropriate number of FOUPs for a given order release rate that will yield acceptable values for on-time delivery performance, cycle time, and throughput via discrete event simulation.

1 INTRODUCTION

In semiconductor manufacturing, integrated circuits (IC) are produced on silicon wafers. This type of manufacturing is very capital intensive. The process conditions are very complex (cf. Gupta et al. 2006). We have to deal with parallel machines (also referred to as tools), different types of processes (batch processes and single wafer processes), sequence-dependent setup times, prescribed customer due dates for the jobs, and re-entrant process flows.

It is common for a fab to produce a large number of different products and the product mix may change often. This is especially in wafer fabs that specialize in application specific circuits (ASIC foundries).

In most semiconductor companies wafers travel through 300-mm wafer fabs in FOUPs containing a maximum of 25 wafers. Due to the increase in wafer size, a FOUP full of 300-mm wafers can weigh in excess of 40 pounds. Therefore, full factory automation given by Automated Material Handling Systems (AMHS) is required as operators will be unable to physically carry FOUPs safely.

In addition to economic concerns, the potential revenue to be earned from the larger 300-mm wafers forces companies to restrict manual transfer of production wafers, whether by hand or by cart. These technological challenges make the already challenging planning and scheduling problems even more complex.

The combination of decreased line widths and more area per wafer result in fewer wafers being needed to fill IC orders of some customers. Each wafer fab will have only a limited number of FOUPs as they are expensive. A large number of FOUPs have the potential to cause overload in the AMHS. In addition, some tools have the same processing times regardless of the number of wafers in the batch. Therefore, it is generally not reasonable to assign an individual FOUP to each order.

Therefore, 300-mm manufacturers often have the need and the incentive to group orders from different customers into one or more FOUPS to form production jobs. These jobs have to be scheduled on the various types of tool groups in the wafer fab and processed together. This class of integrated job formation and scheduling problems are called “multiple orders per job” scheduling problems.

To date, multiple orders per job” scheduling problems have only been considered for single, parallel, and flow-shop machine environments in the literature (cf., for example, Qu and Mason 2005).

The interdependency between multiple orders per job formation problems and the number of FOUPs has apparently not been considered in the literature so far. In order to investigate this dependency we use discrete event simulation as an appropriate technique to deal with the stochastic and dynamic nature of a full wafer fab problem.

The paper is organized as follows. In the next section, we describe the researched problem in some detail. Then we continue with a literature review in Section 3. We present the suggested solution methodology in Section 4. The used framework for FOUP simulations is presented in Sec-

tion 5. The results of the performed simulation study are analyzed and discussed in Section 6.

2 PROBLEM DESCRIPTION

We describe the problem in Section 2.1. Then we discuss related literature. Finally, we analyze the problem.

2.1 Problem Statement

Customers of a semiconductor manufacturer place their orders for specific product types at different points in time by phone, fax, e-mail, or the Internet. We denote the set of all existing orders by $O := \{1, \ldots, n\}$. Each order has the following attributes associated with it:

- s_o: the size of order *o* measured in number of wafers that are necessary to fulfill the required number of IC's accounting for production yield,
- d_o: the due date of order *o*,
- r_o: the release time of order *o*,
- w_o: the weight of order *o* that is a measure for the customer importance.

The capacity of a FOUP is denoted by *K* and is measured in wafers. The attribute s_o can vary widely across different types of wafer fabs. In a high volume, low mix commodity – type wafer fab, s_o will be usually greater than *K*. In high mix ASIC or foundry type wafer fabs , usually $s_o < K$ holds. Therefore, orders with the same z_o will be aggregated to better use the capacity of the FOUP. This leads to better FOUP utilization and reduces the number of FOUPs needed. Throughout the rest of the paper, we denote the number of FOUPs by n_f.

The following assumptions are made within this research. Only orders with the same process flow can be used to form a job. This means that we know the route of the FOUP after a job is assigned to the FOUP.

When a job is formed then this decision cannot be changed, i.e., a split or merge of jobs is not allowed.

Different types of processes are used within the wafer fabs of interest. The processing times of most non-batching tools depend on the number of wafers within the job. Batching tools are run in a FOUP based manner, i.e., the maximum batch size is measured in number of FOUPs. Only jobs with the same process flow can be batched together.

FOUPs only pick up orders that are waiting for processing. Therefore, information on future order arrivals will not be taken into account. This has the consequence that in some situations the number of wafers that will be transferred via one FOUP is small.

A job will be formed only in two different situations. The job associated with a FOUP is completed and the FOUP is newly available. When orders are in the order pool, then a new job will be formed. When the order pool is empty and a FOUP is available, then jobs will be formed whenever a new order is released into the order pool. Note that the first situation will be more important because usually the number of FOUPs is limited.

The space for the storage of FOUPs waiting in stockers and mini-stockers is large enough. Therefore, blocking of the tools because of missing stocker space is not possible.

A large number of FOUPs leads to congestion of the AMHS. We simply introduce additional transportation time to model this congestion instead of modeling the AMHS in detail.

The problem studied in this paper is to determine an appropriate number of FOUPs given a certain order release rate such that we can maintain a certain throughput (TP), have a small cycle time (CT), and a small total weighted tardiness (TWT). The measure TWT is given by the summation of the tardiness of all completed orders *o* multiplied with their weight w_o.

Besides determining the number of FOUPs we have to look for strategies to form the jobs. We expect that job formation will have an impact on CT and TWT, especially in situations, when the number of FOUPs is small.

2.2 Problem Analysis

For a fixed number of FOUPs, the wafer fab can be considered approximately as a CONWIP system. Therefore, the problem to find an appropriate number of FOUPs is similar to determine an appropriate work in process (WIP) level in a CONWIP system.

For simple CONWIP manufacturing systems, CT and TP can be determined simultaneously in a recursive manner by mean value analysis using Little's law (Hopp and Spearman 2000). The throughput of the CONWIP system can be considered approximately as its release rate. Based on this release rate we are able to determine the waiting time of the orders in the order pool by using queuing theory.

When the order release rate in the order pool is larger than the release rate into the CONWIP system, then the number of orders in the order pool will increase over time.

We expect that the cycle time of the orders in the order pool will be large when the number of FOUPs is small. At the same time, we expect that the number of wafers within one FOUP will increase when the number of FOUPs is small. The impact of job formation rules will be important in the case of a full order pool.

We expect that a large number of FOUPs leads to a smaller number of wafers within a FOUP. At the same time, the cycle time will increase with an increasing num-

ber of FOUPs because of CONWIP system properties and because of AMHS congestion.

2.3 Related Literature

Because of our findings from the problem analysis section, we have to discuss literature with respect to job formation strategies in multiple orders per job scheduling problems.

Multiple orders per job scheduling problems are considered by Qu and Mason (2005), Laub et al. (2007), and Erramilli, and Mason (2006). Several decomposition approaches are suggested for single, parallel, and flowshop machine environments that address the job formation, job assignment, and job sequencing problem independently. However, job formation and release strategies are not considered on the entire wafer fab level so far.

A survey related to order release strategies in semiconductor manufacturing is presented by Fowler et al. (2002). This survey contains especially a discussion of CONWIP approaches in semiconductor manufacturing. Framinan et al. (1999) also discuss various approaches to analyze CONWIP systems in different industries. It turns out that very often simulation based approaches are appropriate. This approach is also used by Gillard (2002) for analyzing the CONWIP strategy in one of Intel's wafer fabs. Therefore, we will use discrete-event simulation to determine an appropriate number of FOUPs.

3 SIMULATION BASED METHODOLOGY

In this section, we present a simulation framework that can be used to determine an appropriate number of FOUPs and assess the performance of different job formation heuristics. Furthermore, we also present the used job formation heuristics.

3.1 Simulation Framework

We use the simulation engine AutoSched AP because most of process specifics of the semiconductor industry can be modeled in an appropriate way. Because there is no clear separation made between orders and lots in AutoSched AP, we extend the simulation framework suggested by Mönch et al. (2003) to the case of orders and FOUPS.

The main idea of the framework is a blackboard-type data model that is between the simulation engine and production control approaches. The blackboard acts as a mirror of the business objects found in the simulation model like machines, products, and lots (as moving entities).

In order to avoid a time consuming customization of AutoSched AP, we introduce orders only within the blackboard. Because we have a fixed number of FOUPs, the treatment of FOUPs is similar to lot handling in CONWIP approaches for wafer fabs. Therefore, our implementation is based on the AutoSched AP customization suggested by Mönch (2005).

FOUPS will be represented by lots in AutoSched AP. Each FOUP object contains pointers to its order objects. The overall procedure can be described as follows:

1. Generate orders within the blackboard.
2. Create FOUPS whenever orders are created and the maximum number of FOUPS is not reached or when a FOUP gets available. Choose the content of a FOUP based on heuristics, set the pointers to the order objects.
3. Launch and process a lot in AutoSched AP that represents the FOUP from Step 2.
4. Destroy FOUPS at the end of the simulation and collect FOUP and order related statistics.

The described framework is shown in Figure 1.

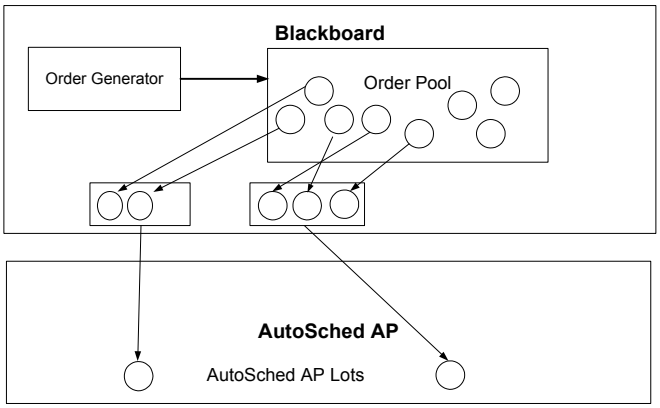

Figure 1: Simulation Framework

3.2 Solution Approach

Given a certain order release rate λ we consider a wafer fab with a fixed number of FOUPs n_f and a given heuristic h for forming the jobs. We measure the values of the performance measures throughput (TP), cycle time (CT), and total weighted tardiness (TWT).

A First in First Out (FIFO) type heuristic first sorts the orders with respect to increasing ready times r_o of the orders. Then the maximum number of orders is taken from the beginning of this list to form job J such that $\sum_{o \in J} s_o \le K$ is valid. We call this heuristic FIFO-H in the remainder of this paper.

The next heuristic considers simultaneously the weight, the due date, and the release date of each order. The priority index of order o is given by:

$$I_{WDD,o}(t) := \frac{w_o}{d_o}\left(2 - \frac{r_o}{t}\right). \quad (1)$$

The orders are sorted with respect to a decreasing index $I_{WDD,o}$. The motivation for the second term in the product is as follows. In case of orders that stay for a long time in the order pool, r_o is small. Hence, the index $I_{WDD,o}$ is large. A maximum number of orders are taken from the beginning of this list to form job *J* such that $\sum_{o \in J} s_o \le K$ is valid. This heuristic is called the Weighted Due Date heuristic (WDD-H).

The following heuristic is similar to WDD-H. We use the following index to assess order *o*:

$$I_{WSS,o}(t) := \frac{w_o}{s_o}\left(2 - \frac{r_o}{t}\right). \quad (2)$$

Note that we simply replace d_o from expression (1) by s_o in expression (2). This heuristic is called the Weighted Shortest Size heuristic and we use the abbreviation WSS-H.

4 COMPUTATIONAL EXPERIMENTS

In this section, we present the computational results of our simulation experiments. We describe our design of experiments in Section 4.1. Then, the computational results are presented and discussed.

4.1 Design of Experiments

We use a modification of the MiniFab simulation model developed by El Adl et al. (1996) with 24 tools organized in nine tool groups for the experiments. The model contains three products. We simulate 15 months with the first three months used for warm up. Exponentially distributed tool breakdowns are considered. We take five independent replications for each simulation run to obtain stochastically significant results.

An order is created on average every 38 minutes. The quantity s_o is distributed with 50 percent probability according to U[2,8] and with 50 percent probability according to U[3,11].

We use a fixed weight scheme for the orders. The discrete distribution D_1 describes a situation where many lots have small or medium weight. D_1 is given by the expression

$$D_1 := \begin{cases} w_j = 1 & p_1 = 0.5, \\ w_j = 5 \quad \text{with} & p_2 = 0.35, \\ w_j = 10 & p_3 = 0.15. \end{cases} \quad (3)$$

The due dates of the orders are calculated using the flow factor concept. The flow factor *FF* is defined as the ratio of the cycle time and the raw process time. We calculate due dates of order *o* by the expression:

$$d_o := RN \cdot FF \sum_{k=1}^{u_o} p_{ok} + r_o, \quad (4)$$

where we denote the processing time of processing step *k* by p_{ok}. The quantity u_o denotes the number of processing steps of order *o*. *RN* is distributed according to *U[0.5, 1.1]*. We use a flow factor $FF = 1.5$.

The transportation time of the FOUPs is modeled as a specific additional load and unload time. This offset depends on the number of FOUPs within the system and the original load and unload times.

We carry out simulations with a fixed maximum number of FOUPs $n_f = 10,20,\ldots,110$ for each job formation rule. The tools within the wafer fab are controlled with the FIFO dispatching rule.

4.2 Computational Results

In Figure 2, we present the results of our experiments for the average weighted tardiness (AWT). The AWT values are obtained by simply dividing the corresponding TWT value by the number of completed orders.

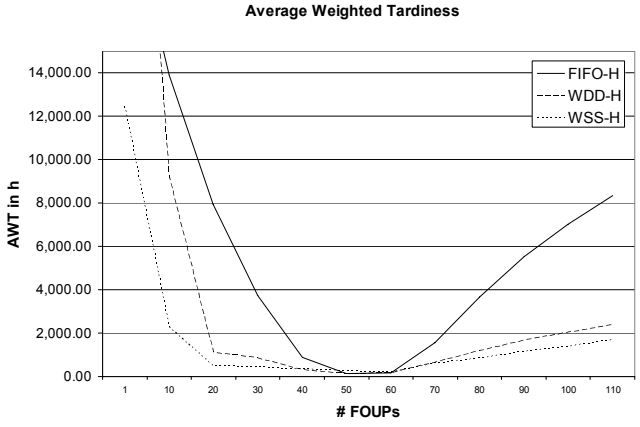

Figure 2: AWT Values

We use the performance measure AWT instead of TWT because AWT also takes the number of completed orders into account. Therefore, AWT is a more fair measure for comparison in situations where the throughput is different for the job formation heuristics.

It turns out that we obtain the smallest values for AWT in case of 50 and 60 FOUPs for all job formation rules. We observe an increase in the AWT value when more than 70 FOUPs are used because of the additional transportation time. A large number of FOUPs leads to a congested transportation system. The transportation times increase and the FOUPs wait a long time for transportation. This causes tools to not be able to process wafers because no orders are available at the tools. This leads also to higher cycle times as shown in Figure 3.

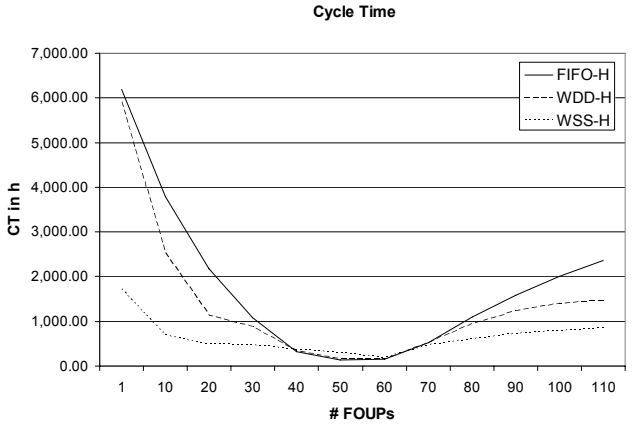

Figure 3: CT Values

Figure 3 shows that the cycle time increases when more than 70 FOUPs are used. A higher cycle time leads to a lower throughput. The results for the throughput are presented in Figure 4.

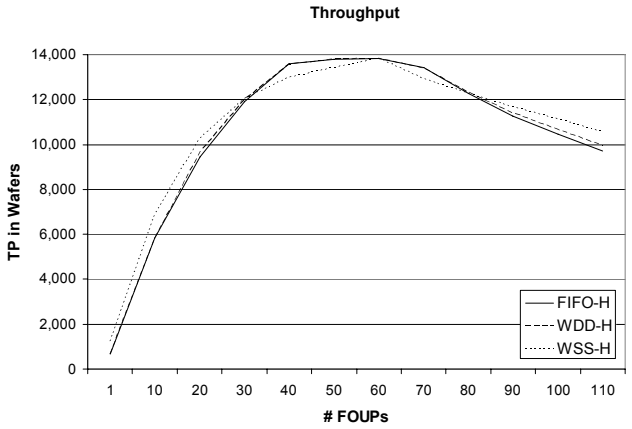

Figure 4: TP Values

We see that the throughput decreases when more than 70 FOUPs are used. The throughput increases up to approximately 50 FOUPs. The number of FOUPs is large enough to produce all orders in an adequate time.

The WSS-H rule leads to small AWT values in the cases of few and many FOUPs, a small cycle time and a high throughput in comparison with the other rules. In case of 40 and 50 FOUPs, the WSS-H rule leads to slightly worse results for AWT. FIFO-H and WDD-H provide better results. The results of the WDD-H rule are better or equal to the FIFO-H rule. All tested rules provide essentially the same results for AWT, cycle time, and throughput in the case of 60 FOUPs.

In Table 1, we analyze the FOUP utilization measured in orders, the average waiting time of the orders in the order pool, and the cycle time of the FOUPs. The cycle time of the FOUPs is similar for all assessed rules.

Table 1: FOUP Utilization, Waiting Time, and Cycle Time

FOUP utilization in orders						
FOUPs	1	10	20	30	40	50
FIFO	4.12	4.09	4.09	4.13	4.12	3.92
WDD	4.17	4.07	4.19	4.15	4.11	3.92
WSS	7.61	4.82	4.46	4.17	3.93	3.81
FOUPs	60	70	80	90	100	110
FIFO	3.61	4.13	4.13	4.11	4.10	4.09
WDD	3.61	4.12	4.14	4.16	4.18	4.20
WSS	3.61	3.97	4.12	4.26	4.37	4.45
Average waiting time in order pool in hour						
FOUPs	1	10	20	30	40	50
FIFO	5234	3389	1989	961	221	14
WDD	5232	3412	1947	934	241	23
WSS	5059	3057	1703	886	450	248
FOUPs	60	70	80	90	100	110
FIFO	7	341	849	1291	1643	1941
WDD	9	352	833	1260	1596	1896
WSS	56	517	812	1116	1386	1642
FOUP cycle time in hours						
FOUPs	1	10	20	30	40	50
	53.9	61.4	76.1	90.8	106.3	124.3
FOUPs	60	70	80	90	100	110
	137.2	188.7	235.4	287.8	344.1	406.1

The utilization of the FOUPs decreases up to a number of 60 FOUPs. Then, the utilization increases. This is understandable because the cycle time increases and more orders are in the manufacturing system. The average waiting time follows the same trend.

We see from Table 1 that the FOUP cycle time increases almost linearly with the number of FOUPs. The utilization and the waiting time in the order pool are nearly the same for the FIFO-H and WDD-H rule. The utilization for the WSS-H rule is higher and the waiting time is lower in the cases of few and many FOUPs. This rule selects orders with a small number of wafers. Hence, more orders are completed.

We identify an optimal number of FOUPs for the applied wafer fab model and a fixed order release rate. This number is 60 FOUPs because we get a small average

weighted tardiness, a small cycle time, and a high throughput.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we described a simulation study to find an appropriate number of FOUPs in wafer fabs where multiple orders are used to form a job. We presented a simulation environment that can be used to model the multiple order per job decisions. The simulation study clearly shows that there is a appropriate number of FOUPS given a certain order release rate such that the values of performance measures like TWT, CT, or throughput are within a certain range.

There are some directions for future research. It seems possible to relax some of the assumptions described in Section 2.1. For example, it seems a good strategy to wait for future order arrivals to increase the number of wafers within a FOUP.

Furthermore, simulation studies with more realistic simulation models, especially with respect to number of tools and number of products, need to be performed. It seems likely that including the AMHS in the simulation study will be beneficial.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the simulation efforts of Marco Rosario Scopelliti.

REFERENCES

Angrawal, G. K., and S. S. Heragu. 2006. A Survey on automated material handling systems in 300-mm semiconductor fabs. *IEEE Transactions on Semiconductor Manufacturing,* 19(1):112-119

El Adl, M.K., A. A. Rodriguez, and K. S. Tsakalis. 1996. Hierarchical modeling and control of re-entrant semiconductor manufacturing facilities. In *Proceedings of the 35th Conference on Decision and Contro*l, Kobe, Japan.

Erramilli, V., and S. J. Mason. 2006. Multiple orders per job compatible batch scheduling. *IEEE Transactions on Electronics Packaging Manufacturing* 29(4):285-296.

Fowler, J.W., G. L. Hogg, and S. J. Mason. 2002. Workload control in the semiconductor industry. *Production Planning & Control* 13(7):568-578.

Framinan, J. M., P. L. Gonzalez, and R. Ruiz-Usano. 1999. The CONWIP production control system: review and research issues. *Production Planning & Control* 14(3): 255-265.

Gillard, W. G. 2002. A simulation study comparing performance of CONWIP and bottleneck-based release rules. *Production Planning & Control*, 13(2):211-219.

Gupta, J.N.D., R. Ruiz, J. W. Fowler, and S. J. Mason. 2006. Operational planning and control of semiconductor wafer production. *Production Planning & Control*, 17(7):639-647.

Hopp, W. J., and M. L. Spearman. 2000. *Factory physics*, 2nd ed. Boston: McGraw-Hill.

Laub, J. D., J. W. Fowler, and A. B. Keha 2007. Minimizing makespan with multiple-orders-per-job in a two-machine flowshop. *European Journal of Operational Research 182*(1):63-79.

Mönch, L., O. Rose, and R. Sturm. 2003. Simulation framework for performance assessment of shop-floor control systems. *SIMULATION: Transactions of the Society of Modeling and Computer Simulation International* 79(3):163-170.

Mönch, L. 2005. Simulation-based assessment of order release strategies for a distributed shifting bottleneck heuristic. In *Proceedings of the 2005 Winter Simulation Conference,* 2186-2093, Orlando, USA.

Qu, P., and S. J. Mason. 2005. Metaheuristic Scheduling of 300mm Jobs Containing Multiple Orders. *IEEE Transactions on Semiconductor Manufacturing* 18(4):633-643.

AUTHOR BIOGRAPHIES

JENS ZIMMERMANN is a Ph.D. student in the Department of Mathematics and Computer Science at the FernUniversität in Hagen, Germany. He received a master's degree in information systems from the Technical University of Ilmenau. He is interested in semiconductor manufacturing, simulation, multi-agent-systems, and machine learning. He is a member of GI (German Chapter of the ACM). His email address is <Jens.Zimmermann@-fernuni-hagen.de>.

LARS MÖNCH is a Professor in the Department of Mathematics and Computer Science at the University in Hagen, Germany and heads the chair of enterprise-wide software systems. He received a master's degree in applied mathematics and a Ph.D. in the same subject from the University of Göttingen, Germany. His current research interests are in simulation-based production control of semiconductor wafer fabrication facilities, applied optimization and artificial intelligence applications in manufacturing and logistics. He is a member of GI (German Chapter of the ACM), GOR (German Operations Research Society), SCS, INFORMS, IIE, and serves as an Associate Editor of IEEE Transactions on Automation Science and Engineering. His email address is <Lars.Moench@fernuni-hagen.de>.

SCOTT MASON is an Associate Professor in the Department of Industrial Engineering <http://www.ineg.-uark.edu/> at the University of Arkansas. He received his

Ph. D. in Industrial Engineering from Arizona State University after earning Bachelor's and Master's degrees from The University of Texas at Austin. Dr. Mason's areas of focus include scheduling and large-scale systems modeling, optimization, and algorithms, with domain expertise in semiconductor manufacturing and transportation logistics. He is an Associate Editor for IEEE Transactions on Electronics Packaging Manufacturing <http://www.cpmt.org/-trans-/trans-epm.html> and currently serves as Associate Department Head of Industrial Engineering and Chair of Graduate Studies at Arkansas.

JOHN FOWLER is a Professor of Industrial Engineering at Arizona State University (ASU) and was the Center Director for the Factory Operations Research Center that was 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 IIE, INFORMS, 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 Semiconductor Manufacturing. He is an IIE Fellow, the Vice President of Chapters/Fora for INFORMS, the Trasurer of Omega Rho, and is on the Winter Simulation Conference Board of Directors.