

**A SIMULATION-BASED AND AN ANALYTIC MODELING APPROACH TO CALCULATE
AN APPROPRIATE NUMBER OF FOUPS IN WAFER FABS**

Lars Mönch

John W. Fowler

Scott J. Mason

Jens Zimmermann

Chair of Enterprise-wide Software Systems
Dept. of Mathematics and Computer Science

Department of Industrial
Engineering

Department of Industrial
Engineering

University of Hagen
58097 Hagen, GERMANY

Arizona State University
Tempe, AZ, 85287, USA

University of Arkansas
Fayetteville, AR, 72701, USA

ABSTRACT

In this paper, we study a design problem found in 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 job formation. An appropriate number of FOUPs for a given order release rate is determined that will yield acceptable values for cycle time and throughput via discrete event simulation and analytic approaches. We show that the analytic modeling approach may help to reduce the simulation efforts. On the other hand, simulation modeling can be used to set certain parameters for the analytic approach.

1 INTRODUCTION

In wafer fabs, integrated circuits (IC) are produced on silicon wafers. This type of manufacturing is capital intensive. The process conditions are very complex. 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 (cf. Gupta et al. 2006).

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 ergonomic concerns, the potential revenue to be earned from the larger 300-mm wafers forces companies to restrict manual transfer of produc-

tion wafers, whether by hand or by cart. The combination of decreased line widths and more area per wafer result in fewer wafers being needed to fill IC orders of some customers. A 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 time regardless of the number of wafers in the batch and 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 (cf. Qu and Mason 2005, Erramilli and Mason 2006, and Laub et al. 2007 for related scheduling approaches).

The interdependency between multiple orders per job formation problems and the number of FOUPs has apparently been considered in the literature only by the present authors in Zimmermann et al. (2008). In order to investigate this dependency discrete event simulation is used as a technique to deal with the stochastic and dynamic nature of a full wafer fab problem.

In this paper, we suggest a complementary approach based on mean value analysis (MVA) to find an appropriate number of FOUPs. The second approach has the advantage that it reduces the computational burden of the discrete event simulation approach considerably. However, only a smaller level of detail is possible. We also discuss how simulation can be used to set certain parameters for the mean value approach.

The paper is organized as follows. In the next section, we describe the researched problem. We also discuss related literature. The suggested solution methodologies are

presented in Section 3. The results of the computational experiments are analyzed and discussed in Section 4.

2 PROBLEM DESCRIPTION

2.1 Problem Statement and Analysis

Semiconductor customers place orders for specific product types at different points in time by various means. Each order has the following attributes associated with it: s_o is the size of order o measured in number of wafers necessary to fulfill the required number of IC's, d_o is the due date of order o , and r_o is the release time of order o .

The capacity of a FOUP (in wafers) is denoted by K . It 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 s_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 a 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 a wafer fab. 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 are only assigned orders that are waiting for processing; 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 in a single 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.

We assume that 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 load 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) and have a small cycle time (CT) for the FOUps. Besides determining the number of FOUps we have to look for strategies to form the jobs. However, in this paper we assume a job formation strategy based on the FIFO rule.

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 (Hopp and Spearman 2000). The throughput of the CONWIP system can be considered approximately as its release rate in case of a stable system.

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.

2.2 Discussion of Related Work

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, discrete-event simulation seems to be appropriate to determine an appropriate number of FOUps.

A second stream of research is related to performance assessment issues for closed queuing networks (CQN). Reiser and Lavenberg (1980) suggested an exact MVA algorithm for multi-chain CQN. Several approximations are suggested that avoid determining the probability mass function for the queue length distribution exactly by a mean queue length estimate. Furthermore, approximate MVA (AMVA) for single class networks is based on the assumption that when a customer is removed from the CQN the proportion in which the customers are distributed across the network does not change (Suri and Hildebrandt 1984). For simple CONWIP manufacturing systems, CT and TP can be determined simultaneously in a recursive manner by MVA (Hopp and Spearman 2000).

This type of approximation is used for the design of flexible manufacturing systems where the number of pallets moving in the manufacturing system is interesting. For a review of convolution type algorithms and AMVA with applications to flexible manufacturing systems we refer to Tempelmeier and Kuhn (1993). An AMVA scheme for a

manufacturing system with reentrant flows, single machines, multiple products, and batch machines is suggested in Park et al. (2002). However, parallel machines are not covered. An extension of AMVA from single-server stations to networks with multiple-server stations is suggested in Suri et al. (2007). In this paper, parallel machines, reentrant flows and batching machines have to be addressed. However, this is not done before in the literature.

3 SOLUTION APPROACHES

3.1 Discrete Simulation based Approach

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 use the FOUP simulation framework suggested by Zimmermann et al. (2008). The center point of the framework is a blackboard-type data layer that is between the AutoSched AP 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. 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 has not been reached or when a FOUP becomes available. Choose the content of a FOUP, 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.

Given an order release rate λ , we consider a wafer fab with a fixed number of FOUPs n_f . A 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 $\sum_{o \in J} s_o \leq K$ is valid. Conducting simulations with a different values for n_f , we are able to determine the values of CT and TP. We are interested in a maximum value of TP and a minimum value of CT.

3.2 Analytic Approach

We assume that we consider only one product, i.e., all FOUPs take the same route. The following notation is used for the remainder of the paper:

M	: number of machine groups,
n_m	: number of machines of machine group m ,
B_m	: maximum batch size at machine group m ,
v_m	: number of visits at machine group m ,
p_m^*	: mean processing time for a single FOUP at machine group m for each single visit,
CT_m^*	: mean cycle time including processing and waiting time at machine group m at each single visit,
p_m	: mean processing time spent by a FOUP at machine group m over all visits,
CT_m	: mean cycle time including processing and waiting time at machine group m at each single visit,
CT	: mean cycle time of the FOUPs within the manufacturing system,
TP	: throughput (in FOUPs per hour),
Q_m	: mean queue length at machine group m ,
U_m	: utilization of machine group m ,
w_m	: mean time that a single FOUP has to be wait for FOUPs ahead to be completed at machine group m .

We obtain for the mean cycle time with respect to a single visit at machine group m :

$$CT_m^* = p_m^* + w_m. \quad (1)$$

A single FOUP arriving at machine group m will find on average $Q_m \frac{n_f - 1}{n_f}$ FOUPs ahead of it. Therefore, the

mean waiting time for a single visit is $Q_m \frac{n_f - 1}{n_f} p_m^*$. In

case of parallel machines, this value has to be corrected by a term Y_m . Based on the results in Suri et al. (2007), the term has the form $Y_m := U_m^{\alpha \frac{n_f}{n_m} - \gamma} / n_m$, i.e., it depends on the utilization and the number of machines of the machine group. The quantity Q_m has to be divided by B_m because of the fact that p_m^* is the processing time of entire batches. Note that the treatment of batches is more sophisticated in Park et al. (2002), i.e., they consider also the time needed to fill a full batch. However, our approach seems to be detailed enough when we treat only small batch sizes. Multiplying equation (1) with v_m leads to

$$CT_m = p_m \left(1 + \frac{n_f - 1}{n_f} Q_m \frac{Y_m}{B_m} \right). \quad (2)$$

The quantity TP can be calculated as:

$$TP = \frac{n_f}{CT} \quad (3)$$

based on Little’s Law (cf. Hopp and Spearman 2000)

where $CT := \sum_{m=1}^M CT_m$. The following expression is used to

update the mean queue length of machine group m :

$$Q_m = CT_m TP. \quad (4)$$

It is again a consequence of Little’s Law. Finally, the update of the utilization of machine group m is given by:

$$U_m = \frac{TP P_m}{B_m n_m}. \quad (5)$$

The suggested AMVA scheme consists in an iteration using equation (2) and (3) and the update equations (4) and (5) starting with $n_f = 1$, $U_m = 0.0$, and $Q_m = 1/n_m$.

3.3 Combining the Two Approaches

Performing all the simulation runs for the different number of FOUPs is very time-consuming especially for large scale wafer fabs. Therefore, a method is highly desirable that avoids these huge simulation efforts. Using AMVA, we are usually able to start the simulation with $n_f \gg 1$ and save a large amount of simulation time. In this research, we simply use the parameters α, β, γ from Suri et al. (2007) for determining the corrective factors Y_m . However, a small number of simulation runs is enough to fit these values from equation (2) by using least squares techniques. The analytic approach has the limitation that statistics related to individual orders or FOUPs like tardiness measures cannot be calculated. We see that a combined approach using different modeling paradigms simultaneously is of great value.

4 COMPUTATIONAL EXPERIMENTS

We use a modification of the MiniFab simulation model developed by Spier and Kempf (1995) with 24 tools organized in nine tool groups for the experiments. The model contains three products. The product flows of the products are similar. We simulate 15 months with the first three months used for warm up. Exponentially distributed tool breakdowns (both time to failure and time to repair) 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 according to $U[2,8]$ with 50 percent probability and according to $U[3,11]$ with 50 percent probability. The transportation time of the FOUPs is modeled as an additional load time. This offset depends on the number of FOUPs within the system and the original load times. We carry out simulations with a fixed maximum

number of FOUPs $n_f = 10, 20, \dots, 110$. The tools within the wafer fab are controlled with the FIFO dispatching rule for the sake of simplicity.

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 for the FOUPs as shown in Table 1.

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.

Table 1: Throughput and Cycle Time

Throughput (in orders per hour)						
FOUPs	1	10	20	30	40	50
	0.08	0.67	1.124	1.36	1.55	1.59
FOUPs	60	70	80	90	100	110
	1.59	1.53	1.40	1.29	1.19	1.11
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

We see from Table 1 that the FOUP cycle time increases almost linearly with the number of FOUPs. We identify an optimal number of FOUPs for the applied wafer fab model and a fixed order release rate. This number is between 50 and 60 FOUPs because we get a small average cycle time, and a high throughput. In Figure 1 and Figure 2, we show the corresponding results for CT and TP using the analytic approach of Section 3.2. We see that AMVA is able to find similar values as the much more detailed simulation. The throughput is given in orders per hour. We assume within the AMVA that approximately four orders are within a single FOUP. Based on the results presented in Zimmermann et al. (2008) this assumption is reasonable.

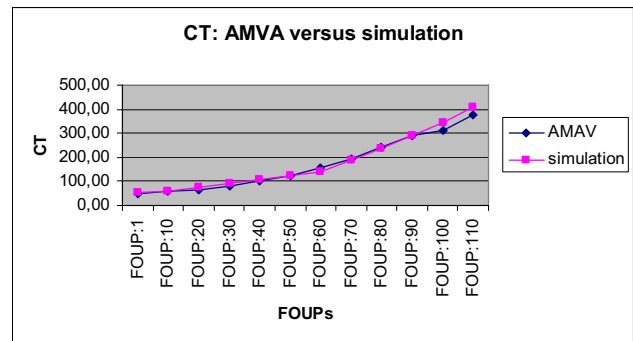


Figure 1: CT Values Obtained by the Two Approaches

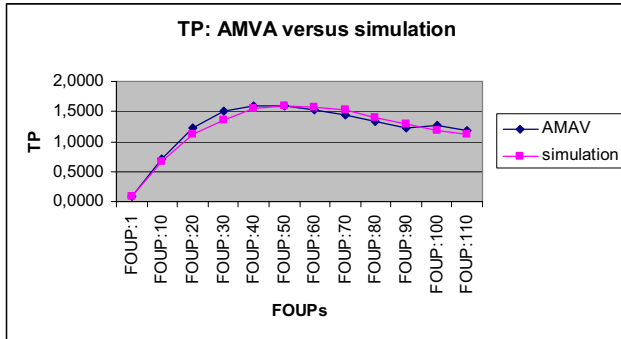


Figure 2: TP Values Obtained by the Two Approaches

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we described a simulation study to find an appropriate number of FOUPls 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 FOUPls given a certain order release rate such that the values of performance measures like CT, or TP for the FOUPls are within a certain range. We demonstrated that we are able to obtain similar results using AMVA techniques.

Simulation studies with more realistic simulation models, especially with respect to number of tools and number of products, need to be performed in future research. This is especially true for an assessment of the approximation qualities of AMVA.

REFERENCES

- 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., Fowler, J.W., and Keha, A.B. 2007. Minimizing makespan with multiple orders per job in a two machine flowshop. *European Journal of Operational Research* 182(1): 63-79, 2007.
- Park, Y., S. Kim, and C.-H. Jun. 2002. Mean value analysis of re-entrant lines with batch machines and multi-class jobs. *Computers & Operations Research* 29, 1009-1024.
- Qu, P., and S. J. Mason. 2005. Metaheuristic Scheduling of 300mm Jobs Containing Multiple Orders. *IEEE Transactions on Semiconductor Manufacturing* 18 (4), 633-643.
- Reiser, M., and S. S. Lavenberg, 1980. Mean value analysis of closed multi-chain queuing networks. *Journal of the Association of Computing Machinery* 27(2), 527-531.
- Spier, J., and K. Kempf. 1995. Simulation of emergent behavior in manufacturing systems. *Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 90-94.
- Suri, R., and R. R. Hildebrandt. 1984. Modeling flexible manufacturing systems using mean-value analysis. *Journal of Manufacturing Systems* 3(1), 27-38.
- Suri, R., S. Sahu, and M. Vernon. 2007. Approximate mean value analysis for closed queuing networks with multiple servers. In *Proceedings of the 2007 Industrial Engineering Research Conference*.
- Tempelmeier, H., and H. Kuhn. 1993. *Flexible manufacturing systems: decision support for design and operation*, New York: Wiley.
- Zimmermann, J., L. Mönch, S. Mason, and J. W. Fowler. 2008. Determining an appropriate number of FOUPls in semiconductor wafer fabrication facilities. In *Proceedings of the 2008 Modeling and Analysis of Semiconductor Manufacturing Conference (MASM 2008)*, 2164-2170.

AUTHOR BIOGRAPHIES

LARS MÖNCH is a Professor in the Department of Mathematics and Computer Science at the University in Hagen, Germany. His e-mail address is <Lars.Moench@fernuni-hagen.de>.

DR. JENS ZIMMERMANN is a simulation consultant at SimPlan AG, Maintal, Germany. His e-mail address is <Jens.Zimmermann@fernuni-hagen.de>.

JOHN W. FOWLER is a Professor of Industrial Engineering at Arizona State University (ASU). His e-mail address is <John.Fowler@asu.edu>.

SCOTT J. MASON is an Associate Professor in the Department of Industrial Engineering at the University of Arkansas. His e-mail address is <mason@uark.edu>.