Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

ALTERNATIVE LOADING AND DISPATCHING POLICIES FOR FURNACE OPERATIONS IN SEMICONDUCTOR MANUFACTURING: A COMPARISON BY SIMULATION

Elif Akçalı
Reha Uzsoy

School of Industrial Engineering
1287 Grissom Hall
Purdue University
West Lafayette, IN 47907, U.S.A.

David G. Hiscock
Anne L. Moser
Timothy J. Teyner

Intersil
1700 Fostoria Road
Findlay, OH 45840, U.S.A.

ABSTRACT

In semiconductor manufacturing, furnaces are used for diffusion and deposition operations. A furnace is a batch processing machine, which can simultaneously process a number of lots together as a batch. Whenever a furnace becomes available, scheduling the next batch involves decisions on both which operation to process next (dispatching policy) and how many lots to put into the batch (loading policy). A simulation model of a wafer fabrication facility is used to examine the effects of different loading and dispatching policies for diffusion operations. Results indicate that the loading policy has a significant effect on the average diffusion flow time as well as the overall cycle time of the products, whereas dispatching policy has a less significant effect. We show that the production volume of a product should be considered in setting the minimum number of lots needed to start a batch. We suggest that the diffusion flow time for a low volume product can be reduced by releasing the product in batches or by setting the minimum batch size such that the work-in-process of the product can be moved faster.

1 INTRODUCTION

In semiconductor wafer fabrication, furnaces are used for thermal processes, such as oxidation, diffusion, deposition, low pressure chemical vapor deposition, and annealing operations. A furnace is a batch processing machine that can simultaneously process a number of lots as a batch. Every furnace operation has a *recipe* associated with it that specifies the gas mixture and the temperature profile required for processing. The constant processing times of the furnace operations are usually long compared to other operations in wafer fabrication. Therefore, the flow time through furnaces may account for a significant portion of the total cycle time through the fab.

A furnace typically has several recipes associated with it and the set of recipes that can be assigned to a furnace determines the set of operations that the furnace can perform. The set of gases and the temperature required for processing dictate the set of recipes that can be assigned to a furnace. For example, since phosphorus oxychloride ($POCl_3$) is a highly contaminating substance, the recipes that use it and those that do not cannot be processed on the same furnace. On the other hand, as special boats are needed for the recipes that require extremely high temperatures, these recipes are assigned to the same furnace.

Coupled with these technological requirements, there is usually a limit on the maximum number of recipes that can be loaded on a furnace to eliminate the processing errors. The number of furnaces on which a recipe is loaded is determined by the total processing capacity required for the recipe. Given all these constraints, assignment of recipes to furnaces is a challenging strategic task beyond the scope of this study. In this research, we focus on the operational decisions for the furnaces.

During manufacturing, whenever a furnace becomes available, scheduling the next batch involves decisions on both which operation to process next (dispatching decision) and how many lots to put into the batch (loading decision). The dispatching decision refers to the prioritization of the operations or the lots that are put together in a batch. The loading decision, which considers the trade-off between starting the batch or waiting for more lots to arrive, further complicates the scheduling task. If the total number of lots in the buffer is less than the capacity of the furnace, starting a batch immediately underutilizes the furnace. However, delaying the initiation of processing until more lots arrive increases the queuing time for the lots that are currently waiting for processing. Once processing is

initiated, it cannot be interrupted to load other lots or change over the process. Therefore, given the long processing times for these operations, scheduling of furnaces should be given careful consideration.

The fab under study is planning to ramp the production volume of the plant to triple its current volume. Therefore, management is interested in implementing simple yet effective scheduling policies to reduce the flow time at diffusion operations so that this workcenter does not become the bottleneck of the system once the production volume is increased.

In our study, we consider the implementation of a set of loading policies for batch processing machines, called threshold policies. A *threshold policy* for a furnace specifies the minimum number of lots that should be present in the queue to initiate processing, which is less than or equal to the maximum number of lots that the furnace can process at a time.

Several researchers have studied the threshold policies for batch processing machines using stochastic modeling tools, such as stochastic dynamic programming or queuing models. Although these studies prove the optimality of threshold policies for simple systems, they do not provide much insight into actual implementation of these policies in a real world setting, where there may be more than one batch processing machine, an operation may not be performed by every machine, and operations may have different processing times.

Deb and Serfozo (1973) study threshold policies for a single batch processing machine, with exponential interarrival and identical and independently distributed batch service times. Glassey and Weng (1991) propose making predictions on the future arrivals and develop a heuristic that uses these predictions. They show that the heuristic performs better than the threshold policies for the single product problem. Fowler *et al*. (1992) present dispatch heuristics for single and multiple products that considers the next arrival only. Weng and Leachman (1993) propose a new methodology that handles both single and multiple products for a single batch processing machine again predicting future arrivals and show that it outperforms all other heuristics. Avramidis *et al*. (1998) develop computational procedures to minimize the expected work-in-process inventory level for a multiple product environment. All of these studies consider a single batch processor, which requires the methods be tested in multiple machine environments to see to what degree the results would be applicable in real life systems.

There are also a number of deterministic models for batch processing machines, which study more complex systems and propose more sophisticated control policies (Ikura and Gimple 1986, Uzsoy 1995). However, the complexity associated with these control policies and the uncertainty involved with the real world systems make the implementation of these deterministic policies a challenging task. Therefore, our goal is to provide some insight into the implementability of threshold policies in a wafer fabrication facility and test their robustness under several different dispatching policies.

The rest of the paper is organized as follows. In Section 2 we discuss the threshold and dispatching policies that we study and in Section 3 we present the simulation model used. Section 4 presents the experimentation and discusses the parameters of the simulation experiments as well as the performance measures we use. In Section 5 we discuss the experimental results and follow up by offering some conclusions and directions for future research in the last section.

2 FURNACE SCHEDULING

As we have mentioned earlier, a *threshold policy* specifies the number of lots that should be present in the queue to initiate processing on a furnace. This number is less than or equal to the maximum number of lots that can be processed at a time. In the facility under study, every time a furnace becomes available, the operator has to check the priorities of the lots. If there are at least two lots that require the recipe of the lot with the highest priority, the operator starts a batch. In practice, some furnace operators communicate with the operators at the upstream operations to see if more lots that require the recipe that they are about to start are expected to arrive within a certain time window. If so, the furnace operator may delay starting that particular batch. Suppose that for a recipe that would take six hours, there are two lots in the buffer. If two more lots will be arriving within the next two hours, then the operator delays processing until those lots arrive at the furnace. However, not all the furnace operators check the incoming lots from upstream operations and the management would like to define a rule that could be used by all the operators in all the shift consistently.

In this study we consider three different sets of threshold policies. The first is a fixed threshold policy, FTP(t), where the minimum batch size is set to t for all recipes. However, setting the threshold value too low for a high volume product may require more production runs than necessary. In addition, if the processing time for the recipe is considerably longer than the interarrival time of the lots for that recipe, the queuing time for some lots may increase. For a low volume product, on the other hand, setting the threshold value too high may increase the queuing time significantly. Therefore, we propose a variable threshold by product policy, VTPP(t_h, t_l), where the threshold values for the recipes of a low volume and a high volume product are set to t_l and t_h, respectively, where $t_l < t_h$. Finally, we propose another policy that considers the processing time of the recipe when setting the threshold values. We propose the following heuristic rule to set the threshold levels for the variable threshold by recipe policy, VTRP($\boldsymbol{t}$), where $\boldsymbol{t}$ is a one-dimensional vector of threshold

values for each recipe and t_r is the threshold value for a recipe r. For example, if the interarrival time is much greater than the processing time for a recipe, i.e., $\lambda_r >> p_r$, then the threshold value is set to 2.

$$t_r = \begin{cases} 2 & \lambda_r >> p_r \\ 3 & \lambda_r \geq p_r \\ 4 & \lambda_r < p_r \\ 5 & \lambda_r << p_r \end{cases}$$

In the facility motivating this study, the lots are prioritized according to their *critical ratio* (CR), which is given by

$$CR_l = \frac{d_l - t}{w_l}$$

where d_l is the due date of the lot, t the current time, and w_l the remaining theoretical processing time required to finish the lot. Theoretical processing time of a product is the sum of the constant processing times over all the operations that are required for the product. These processing times are specified by the engineers and do not vary.

Currently, the operators have to consult the lot prioritization routine supported by the factory information control system to decide which recipe to run next. The highest priority is given the recipe r that has the lot with the lowest critical ratio (LLCR). However, it is not uncommon for the operators to start a batch for the recipe with the longest queue to move as many lots as possible out of the work area. If there are more lots than the capacity of the furnace, the lots that are to be included in the batch are selected with first-come-first-serve rule (LNGQ-FCFS). We make this assumption to capture the operators' inclination not to consult the lot prioritization routines regularly. However, if lots are selected using their critical ratios, we refer to this dispatch policy as LNGQ-CR.

In addition to these rules, we propose two other dispatching rules for experimentation. Lowest Total Critical Ratio (LTCR) prioritizes the operation that has the total lowest critical ratio for all the lots in its queue. Another dispatching rule we test is Lowest Average Critical Ratio first (LACR) where the total critical ratio value for the queue is divided by the number of lots in the queue and the queue with the lowest average critical ratio per lot is processed next. Since furnace utilization is of concern to the manufacturers, we also devise a dispatching rule to quantify the utility of every minute spent by a furnace on an operation, lowest critical ratio per processing minute, (LPCR) where the total critical ratio value for the queue is divided by the processing time of the recipe. The calculation of all these measures are summarized in Table 1 where a lot l or a recipe (equivalently an operation), r, has the highest priority.

3 SIMULATION MODEL

We consider a wafer fabrication facility that produces a high variety of products with different processing complexities and varying production volumes. The simulation model for the plant has been built using the SIMAN simulation language (Pegden et al. 1995) supported by a Unix C language insert (Systems Modeling Corporation 1994). Transportation times among the stations are assumed to be negligible and are not included in the model. The operators are not included in the model either. The details of the model follow.

Table 1: Measures Used by Dispatching Policies

Dispatch Policy	Measure
LLCR	$\min_{r,l}\{CR_l\}$
LNGQ	$\max_{r}\{NQ_r\}$
LTCR	$\min_{r}\{\sum_{l} CR_l\}$
LACR	$\min_{r}\{\frac{\sum_{l} CR_l}{NQ_r}\}$
LPCR	$\min_{r}\{\frac{\sum_{l} CR_l}{p_r}\}$

3.1 Product Mix

The wide variety of product types that are in the product mix have been represented by an aggregated product mix, which has been verified by the process engineers to be representative of the complexity of the diffusion operations. We consider a basic product mix of 5 products. The number of mask layers (NM), wafer starts per quarter (WS) and the theoretical cycle time (TCT) in hours (for confidentiality reasons disguised) for each product are tabulated in Table 2.

Table 2: Product Mix Information

Product	NM	TCT	WS
Product 1	11	114	3,829
Product 2	13	129	4,743
Product 3	17	200	4,457
Product 4	19	288	3,371
Product 5	12	143	2,40

Lots are released into the fab at constant intervals based on the planned wafer starts. Each lot l is assigned a due date, d_l, at the time of its release, t, which is given by

$$d_l = t + (u)(TCT_l)$$

where $u \sim$ uniform(2, 4).

3.2 Furnace Operations

Low pressure chemical vapor deposition, oxidation, diffusion, deposition, and annealing operations are performed on diffusion furnaces. The diffusion station consists of seven substations, each of which corresponds to a family of recipes that require certain gas mixtures as explained in Section 1. Each substation has parallel servers, each of which can process a different set of recipes.

For the product mix we consider, there are approximately 100 unique recipes and every recipe has its individual buffer of infinite size. Actual theoretical processing times have been used for individual recipes and vary from 45 minutes to 24 hours per run. There are about 90 furnaces. The furnace capacity is five or eight lots for vertical and horizontal furnaces respectively. Empirical distributions have been derived from historical data to model downtimes for furnaces. Therefore, equipment availability losses due to furnace aborts, profiles, etc. have all been included in the model.

Recipes have different loading specifications that affect the loading time. The loading time component we model includes the time to download the recipe, load the wafers on the boats, and then load the boats on to the furnace paddles. Unloading time includes not only the wafer unloading time but also the time spent to inspect the test chips for oxide thickness and/or resistivity. However, not all recipes require inspection. If the recipe requires resistivity test, the test wafers have to be stripped before inspection. Since these wafers have to be sent to wet etch sinks for stripping, the lead time of stripping introduces a lot of variability to the process. In order to account for this variability, for these lots we generate the unload time from a uniform distribution. If the test chips fail inspection, the load may require a fix run or be discarded as scrap. However, the inspection failure rate is hard to determine since it cannot be attributed to recipes, but to furnaces that vary from month to month. Therefore, we do not model the reworked or scrapped runs.

3.3 Other Operations

We also include the photolithography, ion implantation, etching, and thin film operations in the model. For confidentiality reasons, we do not give the number of resources at each workcenter.

Photolithography consists of five substations for coating tracks, aligners, steppers, developing tracks, and inspection stations. No downtimes have been used for the photolithography equipment and processing times are generated from triangular distributions. Lots travel through the station visiting the coating tracks, steppers or aligners dictated by the technology of the part, developing tracks, and inspection stations.

Ion implantation consists of two substations for medium and high current operations with non-identical servers. Setup times for dose adjustments and gas changeovers have been modeled. Every piece of equipment has its primary gas species, and as the need arises with the fluctuations in WIP inventory levels, gas changeovers are performed. Constant theoretical processing times are used.

Etching has nine substations with infinite buffer capacity and parallel identical servers No downtimes have been modeled for the etching equipment and constant theoretical processing times have been used. Standard clean operations have not been included in the model since they are believed to have ample capacity.

Thin film has eight substations with infinite buffer capacity and parallel identical servers. Empirical distributions from historical data have been derived to model downtimes for thin film equipment. Some operations are performed on batch processing machines and others on unit processing machines. Theoretical processing times have been used, which is a function of the number of wafers in the run for the batch processing machines.

4 EXPERIMENTATION

Our simulation experiment is designed to determine the effects of different scheduling policies used at diffusion furnaces on average diffusion flow time, average tardiness, and product cycle time per mask layer. Three different experiments for the threshold policies, dispatch policies and lot release strategies have been made.

4.1 Simulation Experiments

The simulation experiments are conducted under steady-state conditions. The system is started empty and idle and a single run is made to collect 20 batches of data. A batch is defined as the duration of a fiscal year quarter (13 weeks). The first three quarters have been discarded to eliminate the initial transient period that has been determined by visual inspection of cycle time data series. Common random numbers are used for equipment downtimes to reduce variance between the experiments (Law and Kelton 1991). The mean and the precision of the estimate for the performance measures are calculated using the method of batch means (Law and Kelton 1991) and we discuss our results with the method of significant digits interpretation (Song and Schmeiser 1994).

4.2 Performance Measures

In our study, as a measure of local performance we use the average flow time through diffusion process. In order to assess the effect of furnace scheduling on the overall fab, we use the average cycle time per mask layer for the

products. Cycle time per mask layer, *CTML*, is one of the standard cycle time metrics that is widely used in wafer fabs and is given by

$$CTML_p = \frac{CT_p}{NM_p}$$

where CT_p is the average cycle time and NM_p is the number of mask layers for product p.

We also consider average tardiness, *AT*, to evaluate the due date performance of the facility and is computed by

$$AT = \frac{\sum_l \max(0, c_l - d_l)}{N}$$

where c_l is the completion time and d_l the due date for lot l, and N is the number of lots that are finished after their due date.

5 RESULTS

We have conducted three sets of experiments. The first experiment is aimed towards finding the optimal loading policy for the system. The second experiment is almost like a sensitivity analysis. We study different dispatching policies to see the impact on diffusion flow time and average tardiness measures of the dispatching policies we consider. With the final experiment we explore the effects of batch starts in an attempt to streamline the product flow with the loading policy of the batch processing machines.

5.1 Interpretation of Results

For the performance measures, we obtain the point estimators using the method of batch means. In order to give a measure of the variability of the point estimator, we also report the standard error associated with the point estimator. The standard error specifies the digits of the point estimator that are unlikely to change if the simulation experiment is to be repeated. Suppose the point estimate is 11.62 and the standard error associated with it is 0.02. Then using the significant digits approach, it can be claimed that the digits 11.6 are unlikely to change, if the simulation experiment is repeated. In our study, this is how we determine the precision of the point estimators for the performance measures under consideration.

5.2 Experiment 1: Loading Policies

In our first experiment we examine the effects of different loading policies. We test threshold levels of t = 2, 3, ..., 8 for a system where FTP is in effect. We also test systems where VTPP with t_l = 1, 2, 3 and t_h = 4 are in effect as well as VTRP(***t***), where ***t*** is determined according to the heuristic presented in Section 2. The estimates of the average flow time and average tardiness are given in Tables 3 and 4. The numbers in parentheses are the standard errors associated with each of the estimates. For confidentiality reasons, all summarized output has been scaled by a constant.

We do not report the results for the instances where threshold is 2 or 3, since they yield unstable systems. We conclude that for the planned increased output, the fab can no longer use a threshold of 2 as the minimum batch size, as it will decrease the furnace utilization significantly and increase product cycle-times drastically. If the fab operates under a fixed threshold policy, the average flow time and average tardiness are at their lowest levels when the threshold is 4. FTP(5) is also a competitive policy. However, as can be seen from Table 4, the cycle time per mask layer for the low volume product is higher than desired. VTPPs, however, are better as they decrease both the average tardiness and the cycle time per mask layer for the low volume product. Finally, although VTRP(***t***) has the lowest average flow time, the cycle time for the low volume product is relatively high and the average tardiness is significantly higher.

5.3 Experiment 2: Dispatch Policies

In this experiment we examine the effects of five different dispatching policies that have been explained in detail in Section 2. In this experiment, single lot starts are made. The estimates of the average flow time and average tardiness are given in Table 5. LLCR is the current policy of the plant and none of the dispatching policies is significantly superior to others in terms of its average flow time performance. However, if LNGQ is used in conjunction with the critical ratio prioritization, the average tardiness improves slightly. Results of this experiment show that the average flow time through diffusion workcenter is not very sensitive to the dispatching policy in effect.

Table 3: Average Diffusion Flow Time and Average Tardiness in Hours under Different Furnace Loading Policies

Policy	Average Flow Time	Average Tardiness
FTP(4)	12.67 (*0.02*)	42.52 (*0.89*)
FTP(5)	14.10 (*0.03*)	43.30 (*1.11*)
FTP(6)	17.82(*0.00*)	99.82 (*2.75*)
FTP(7)	20.24 (*0.05*)	109.77 (*1.55*)
FTP(8)	20.65 (*0.01*)	134.13 (*1.28*)
VTPP(4, 1)	12.42(*0.01*)	29.41 (*1.44*)
VTPP(4, 2)	12.19 (*0.01*)	34.40 (*0.62*)
VTPP(4, 3)	12.29 (*0.00*)	36.50 (*0.92*)
VTRP(***t***)	12.12 (*0.01*)	51.17 (*0.12*)

Table 4: Average Cycle Time per Mask Layer/Theoretical Cycle Time per Mask Layer Ratios for Products under Different Furnace Loading Policies

Policy	P1	P2	P3	P4	P5
FTP(4)	2.27 (*0.01*)	2.11 (*0.01*)	2.08 (*0.00*)	2.16 (*0.00*)	2.88 (*0.04*)
FTP(5)	2.47 (*0.02*)	2.33 (*0.01*)	2.23 (*0.01*)	2.29 (*0.02*)	3.41 (*0.03*)
FTP(6)	3.18 (*0.00*)	2.65 (*0.00*)	2.51 (*0.01*)	2.57 (*0.00*)	18.27 (*0.13*)
FTP(7)	3.51 (*0.02*)	2.88 (*0.01*)	2.73 (*0.00*)	2.84 (*0.01*)	22.17 (*0.09*)
FTP(8)	3.72 (*0.01*)	3.09 (*0.00*)	2.89 (*0.00*)	2.85 (*0.01*)	26.29 (*0.55*)
VTPP(4, 1)	2.20 (*0.00*)	2.07 (*0.01*)	2.02 (*0.01*)	2.06 (*0.01*)	1.57 (*0.01*)
VTPP(4, 2)	2.25 (*0.00*)	2.07 (*0.00*)	2.03 (*0.01*)	2.06 (*0.00*)	1.99 (*0.00*)
VTPP(4, 3)	2.24 (*0.02*)	2.12 (*0.01*)	2.03 (*0.01*)	2.09 (*0.01*)	2.46 (*0.01*)
VTRP(***t***)	2.19 (*0.01*)	2.05 (*0.01*)	2.03 (*0.00*)	2.02 (*0.00*)	3.50 (*0.01*)

Table 5: Average Diffusion Flow Time and Average Tardiness in Hours under Different Dispatching Policies

Policy	Average Flow Time	Average Tardiness
LLCR	12.67 (*0.02*)	42.52 (*0.89*)
LTCR	12.18 (*0.01*)	43.46 (*0.42*)
LACR	12.18 (*0.00*)	41.48 (*1.45*)
LNGQ-CR	12.22 (*0.01*)	36.25 (*0.11*)
LNGQ-FCFS	12.13 (*0.00*)	45.67 (*0.95*)
LPCR	13.31 (*0.10*)	60.47 (*2.20*)

5.4 Experiment 3: Batch Starts

In order to alleviate the adverse effect of high threshold values on the cycle time of the low volume products, we consider starting those products in batches so as to coordinate the lot starts with the loading policy on the furnaces.

The first scenario is the original setting of the system where single lot release is in effect and FTP(4) is in used for furnace loading. Under the second scenario, we start the high volume products in single lots and the low volume ones in batches of four and use FTP(4) for furnace loading. The third scenario is again an instance from Experiment 2 where single lot release is in effect and VTPP(4, 2) is in use. In the fourth scenario, we start the low volume product in batches of two and use VTPP(4, 2) for furnace loading. Finally in order to see if there is any benefit to using batch starts for all products, we start all the products in batches of four and use FTP(4) for furnace loading. The experimental conditions for the scenarios are summarized in Table 6.

Table 6: Experimental Conditions for Scenarios

Scenario	High Volume Product Starts	Low Volume Product Starts	Loading Policy
I	Single lot	Single lot	FTP(4)
II	Single lot	Batches of 4	FTP(4)
III	Single lot	Single lot	VTPP(4, 2)
IV	Single lot	Batches of 2	VTPP(4, 2)
V	Batches of 4	Batches of 4	FTP(4)

Results from the simulation runs are summarized in Tables 7 and 8. Scenarios II and IV, where low volume is released in batches, improve the due date performance decreasing the average tardiness. In comparing the performance of the scenarios that have batch starts to the ones with single lot releases in terms of cycle time metrics, we see the decrease in cycle time per mask layer for low volume product. For Scenario V, where all products are started in batches, both the average flow time and average tardiness decrease. Therefore, we can conclude that there is benefit in coordinating lot releases with the loading policy of the batch processing machines.

Table 7: Average Diffusion Flow Time and Average Tardiness in Hours under Different Lot Start Policies

Scenario	Average Flow Time	Average Tardiness
I	12.67 (*0.02*)	42.52 (*0.89*)
II	12.23 (*0.00*)	29.97 (*0.46*)
III	12.19 (*0.02*)	34.40 (*0.62*)
IV	12.15 (*0.03*)	27.41 (*1.07*)
V	11.84 (*0.02*)	38.36 (*1.67*)

Table 8: Average Cycle Time per Mask Layer/Theoretical Cycle Time per Mask Layer Ratios for Products under Different Lot Start Policies

Policy	P1	P2	P3	P4	P5
I	2.27 (*0.01*)	2.11 (*0.01*)	2.08 (*0.00*)	2.16 (*0.00*)	2.88 (*0.04*)
II	2.23 (*0.00*)	2.09 (*0.01*)	2.03 (*0.01*)	2.08 (*0.00*)	2.01 (*0.00*)
III	2.25 (*0.00*)	2.07 (*0.00*)	2.03 (*0.01*)	2.06 (*0.00*)	1.99 (*0.00*)
IV	2.17 (*0.02*)	2.04 (*0.01*)	2.01 (*0.00*)	2.06 (*0.01*)	1.74 (*0.02*)
V	2.21 (*0.01*)	2.05 (*0.00*)	1.98 (*0.01*)	2.07 (*0.01*)	1.93 (*0.01*)

6 CONCLUSIONS AND RECOMMENDATIONS FOR IMPLEMENTATION

In this paper we have examined the performance of different loading and dispatching policies for batch processing operations of a semiconductor wafer fab. Our study complements previous analytical work in that it

provides empirical evidence for the good performance of threshold policies in multiple batch processing machine environments where processing times of the operations and machine capabilities vary significantly. We also have shown that the loading policy has a more significant effect than the dispatching policy on the average diffusion flow time and due date performance.

The average flow time through diffusion appears to be a convex function of the threshold value of the loading policy in effect. A threshold level too low decreases the furnace utilization and increases the queue time dramatically, which adversely affects the due date performance of the fab. On the other extreme, a threshold level that is too high increases the waiting time significantly increasing the cycle times for the lots. The degree of this effect, however, is dependent on the production volume of the product. A low volume product may experience very long queue times if the threshold is high compared to its work-in-process inventory level. In order to mitigate the adverse effects of high thresholds on the low volume products, semiconductor manufacturers may choose from two remedies. Either a hybrid loading policy can be implemented, under which the threshold value for the low volume products is set to a lower value so that these products do not experience long queue times. Another approach for the low volume products is to start them in batches so that their work-in-process is aligned with the loading policy of the batch processing machines.

The threshold policies we study consider the work-in-process inventory at the work center to make a dispatching decision. However, by using the information on the future arrivals to the workcenter the performance can be improved. One direction for future research is to test the performance of heuristic dispatch rules that consider the incoming work-in-process from upstream operations.

ACKNOWLEDGMENTS

This research was supported by an applied research grant from Intersil Corporation. We would like to thank Roxie Karg (currently at Bridgestone, Upper Sandusky, OH), Ken Park, Dave Hronek, Dennis Benson, Kevin Hoopingarner, Suzi Stanton, Neil Stanton, and Jeff Short of Intersil Corporation, Findlay, OH for their dedicated help during data collection and their useful insights for model development. We also would like to thank Prof. Bruce Schmeiser of Purdue University for the helpful discussion on how the threshold should be set.

REFERENCES

Avramidis, A. N., K. J. Healy, and R. Uzsoy. 1998. Control of a batch processing machine: a computational approach. *International Journal of Production Research*, 36: 3167-3181.

Deb, R. K., and R. F. Serfozo. 1973. Optimal control of batch service queues. *Applied Probability*, 5: 340-361.

Glassey, C. R., and W. Weng. 1991. Dynamic batching heuristic for simultaneous processing. *IEEE Transactions on Semiconductor Manufacturing*, 4: 77-82.

Ikura, Y., and M. Gimple. 1986. Efficient scheduling algorithms for a single batch processing machine. *Operations Research Letters*, 5: 61-65.

Fowler, J. W., D. T. Phillips, and G. L. Hogg. 1992. Real-time control of multiproduct bulk service semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, 5: 158-163.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*. 2nd ed. New York: Mc Graw Hill Inc.

Pegden, C. D., D. R. Shannon, and R. P. Sadowski. 1995. *Introduction to Simulation Using SIMAN*. 2nd ed. New York: McGraw-Hill Inc.

Song, W. T. and B. W. Schmeiser. 1994. Reporting the precision of simulation experiments. In *New Directions in Simulation for Manufacturing and Communications*, ed. S. Morito, H. Sakasegawa, K. Yoneda, M. Fushimi, and K. Nakano, 402-407.

Systems Modeling Corporation. 1994. *ARENA: Variables guide*.

Uzsoy, R. 1995. Scheduling batch-processing machines with incompatible job families. *International Journal of Production Research*, 33: 2605-2708.

Weng, W., and R. C. Leachmann. 1993. An improved methodology for real-time production decisions at batch-process work stations. *IEEE Transactions on Semiconductor Manufacturing*, 6: 219-225.

AUTHOR BIOGRAPHIES

ELİF AKÇALI received the B.S. degree from Middle East Technical University, Ankara, Turkey and the M.S. degree from Purdue University, West Lafayette, IN in Industrial Engineering. She is currently pursuing the Ph.D. degree in the School of Industrial Engineering at Purdue University. Her research interests include simulation modeling, discrete and heuristic optimization with application to production planning and control methods for semiconductor industry. She is a member of INFORMS and IIE and is one of the IIE Gilbreth Memorial Fellowship recipients. Her email address is <akcali@purdue.edu>.

REHA UZSOY received the B.S. degrees in Industrial Engineering and Mathematics, from Boğaziçi University, Istanbul, Turkey, and the Ph.D. degree in Industrial Engineering from the University of Florida, Gainesville, FL in 1990. He is now a Professor in the School of Industrial Engineering at Purdue University, West

Lafayette, IN. His research and teaching interests are in production planning and scheduling, particularly in semiconductor manufacturing. He also worked as a visiting researcher at Intel Corporation and IC Delco. Prof. Uzsoy received the Outstanding Young Industrial Engineer in Education Award from the Institute of Industrial Engineers in 1997. His email address is <uzsoy@ecn.purdue.edu>.

DAVID HISCOCK has received the B.S. degree in Mechanical Engineering from Ohio State University, Columbus, OH in 1984. He joined RCA Solid State and held various engineering positions in Hi-Rel assembly and test for Military and Space applications. Under GE Solid State, in 1989 he lead the ion implantation engineering sustaining and development projects in Findlay, OH. In 1992 he worked the installation of the first cell control system in Harris Semiconductor lowering wrong recipe scrap levels. Since 1994 he has been the Senior Manager of Yield Enhancement and MIS at Intersil, Findlay, OH. His email address is <dhiscock@intersil.com>.

ANNE MOSER is a Staff Engineer for diffusion operations at Intersil, Findlay, OH. She received the B.A. degrees in Mathematics and Philosophy from College of Wooster, Wooster, OH in 1978. She got a M.A. degree in Philosophy in 1980 and a B.S. in Electrical Engineering in 1985 from the University of Toledo, Toledo, OH. Her email address is <amoser@intersil.com>.

TIM TEYNER is the Manager of the Continuous Improvement Group at Intersil, Findlay, OH. He received a B.S. degree in Electrical Engineering Technology from Purdue University in 1976. After working in R&D for RCA on the videodisc technology, he worked in semiconductor manufacturing as a Probe Engineer and Probe Engineering Manager before taking his current position, which involves extensive liaison with other Intersil sites in the United States and the Far East. His email address is <tteyner@intersil.com>.