Controlling Parallel Batch Processing Machines for Minimizing Tardiness

M.J. Park, P. Mansoer, and Pyung-Hoi Koo

***Abstract*—**Batch process machines (BPMs) process a number of jobs simultaneously as a batch. Since the BPMs require long processing time and increase flow variability, they have great effect on the system performance. When a BPM completes its task and is ready to start a new process, a decision should be made whether to start the process right away or to wait for the upcoming products. This paper presents real-time control procedure for BPMs, where multiple product types are available, with the objective of average tardiness minimization. The proposed procedure takes advantage of due-date information from current products in queue as well as upcoming products. The experimental results show that the proposed strategy give lower average tardiness compared to existing strategies.

***Index Terms*—**Batch processing, Real-Time Loading Decision, Semiconductor Manufacturing, Tardiness Minimization.

I. Introduction

The semiconductor chips are produced in wafer fabrication (aka. wafer fab) in which layers and patterns are built up on wafers for required circuit. The wafers in wafer fab move through the processes in lots each of which generally consists of 20~24 individual wafers. As different layers are added to the wafer surface, wafer lots at different production stages visit (reenter) the same processing equipment several times. In addition to the reentrant flow, the wafer fabs have some distinct characteristics in terms of management perspective including long lead time (more than one month), complex product flows, multiple facility involvement, rapidly changing products, and very large initial investment required. These characteristics make the production scheduling and control problems more important and difficult.

The machines in semiconductor wafer fabrication (wafer fab) may be classified into batch processing machines (BPMs) and discrete processing machines (DPMs). BPMs process a number of wafer lots simultaneously as a batch while DPMs process wafer lots individually. Diffusion and oxidation processes are performed by the BPMs. The products (wafer lots) arriving at the BMPs is formed as a batch before being served by a BPM. Because of the machine or process constraint, there is a limitation to the number of lots that can be included in a batch. Usually, a BPM can accommodate six to eight wafer lots in one batch. The lots in a batch is processed together and released at the same time. Due to the chemical nature of the process, it may be impossible to process jobs with different recipes together in the same batch. The lots with the same recipe can be viewed as a product type and all lots in the same product type have the same processing time. The wafer lots in different reentrant loops can be considered as different product types. Batch processing is one of the major sources of variation in the production flow. Before batch processing, wafer lots should be in the queue waiting more lots to arrive to form a large batch. After batch processing, multiple wafer lots are off-loaded at a time onto machines that are capable of processing only one lot at a time. This leads to the formation of long queues in front of these DPMs and to a non-smooth flow of products. Aside of that, the process time of a BPM is about five to ten times longer than the DPM. Because of these properties, the batch operations have a great effect on system performance in terms of throughput, work-in-process(WIP), cycle time and on-time delivery.

One of the characteristics in wafer fab is high uncertainty due to a variety of processes involved, long lead time and urgent orders, to name a few. The uncertainty reduces the performance of the static scheduling decisions or sometimes makes the schedules infeasible. Therefore, in most real-world semiconductor manufacturing systems, the production line is controlled in realtime by considering current system status. This paper addresses a real-time control problem of batch processing machines in semiconductor manufacturing where parallel BPMs are available for processing multiple product types, with the objective of tardiness minimization

II. Problem Description

Realtime control strategies of BPMs may be classified into two policies, threshold policy and look-ahead policy, according to the use of knowledge on future arrivals of products. The most basic control strategy of the threshold policy is the minimum batch size (MBS) by Neuts [11]. In this control strategy, processing of a batch is started when the number of products waiting in the queue becomes greater than or equal to the predetermined number. There are many studies dedicated to find the optimal MBS size in many different environments, from single-product single-machine environment to multiple-product multiple-machine environment. As this paper focuses on look-ahead control problems, our discussion on previous literature will be mostly on look-ahead policies.

Manuscript received December 4, 2014. This research was supported by Basic Science research program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A4A01014897)

P. Mansoer is with Daechang INC., Busan, Korea (e-mail: mithamansoer@gmail.com).

M.J. Park is with the Department of Systems Management and Engineering, Busan, Korea (e-mail: alswjd6976@gmail.com).

P.H. Koo with the Department of Systems Management and Engineering, Busan, Korea (e-mail: phkoo@pknu.ac.kr).

http://dx.doi.org/10.15242/IAE.IAE0215212

Look-ahead strategy in BPM control decisions considers the near future information such as product arrival and machine status. Glassey and Weng [4] may be among the first to use near-future information for realtime BPM control in semiconductor manufacturing. They present a batching heuristic called DBH (dynamic batching heuristic) that takes future lot information into account for the single-machine single-product type scenario. The policy relies on determining the length of time or the number of incoming lots the furnace should wait for in order to minimize average waiting time. A variation of DBH, called modified DBH (MDBH), is proposed by Kim et al. [7] where the slack time of lots is considered in the decision making. Fowler et al. [2] develop a heuristic for both single-product type and multiple product type case, called the next arrival control heuristic (NACH), which is a modified version of DBH. The difference is that NACH considers only the next arrival lot. The experimental results indicate that NACH is robust in the sense that it performs well even with errors in the prediction of next arrival time. Fowler et al. [3] extend their previous work for multiple batch processing machines. Weng and Leachman [14] propose a control heuristic called MCR (minimum cost rate). The difference between MCR and DBH is the choice of look-ahead horizon. While DBH uses fixed look-ahead horizon, namely process time, MCR uses process time plus prior waiting time. A variation of MCR is dynamic job arrival assignment (DJAH), presented by van der Zee et al. [12] where the structures of MCR and NACH are combined. The performance criterion in DJAH is the minimization of logistic costs per part on a long term. Logistic costs associated within a job consist of linear waiting costs and a fixed amount of setup costs. Later, van der Zee et al. [13] study batch processing system with multiple non-identical machines, and develop a new strategy called DSH (Dynamic Scheduling Heuristic) to choose a machine from different types based on the required processing condition, product characteristics, and operating cost.

The research works discussed above attempt to improve system related performance, lead time or waiting time, to develop their control rules. However, the due-date related performance has received more attention recently because of the characteristic of reentrant product flows. In most wafer fabs, the lithography operation is the bottleneck process so that the lithography equipment should be utilized in full capacity. In order to fully utilize the bottleneck machine, all the other operations including batch machines should feed the products to the bottleneck machine smoothly to prevent the bottleneck machine from being idle due to starvation. In most manufacturing systems, the bottleneck resource has a predefined production schedule. This gives rise to pseudo due dates for jobs in the reentrant loop. The late arrival of lots at subsequent stations may lead to starvation of the system bottleneck and thus lose output. Tardiness is a common measure for the due-date related performance. Recently some researchers deal with the realtime BPM control considering due date and tardiness. Kim et al. [7] present three batching decision making strategy on multiple-products environment, MMBS, MDBH and PUCH. MMBS and MDBH are the modification of MBS and DBH where slack times are additionally included in control decision making. PUCH is a modification of MMBS which gives urgency to each product and selects product family with highest urgency. Kim et al. [8] extend their research by introducing PRALC (Priority Rule-based Algorithm with Look-Ahead Checks) for multiple product type multiple BPMs environment. Gupta and Sivakumar [5] present a control heuristic called look ahead batching (LAB) for single machine single product type cases. The decision of LAB is made considering arrival time and due date of incoming lots. The objective of LAB is to minimize the average tardiness as well as variation of tardiness. Later, Gupta et al. [6] improve LAB by considering both tardiness and earliness. Cerecki and Banerjee [1] present NACH-T, a multiple-product look-ahead strategy where minimum tardiness for each product type at each future arrival is calculated and the effects of a product's tardiness on all the other product types are considered in control decision making. Mansoer and Koo [9] presents a new real-time loading strategy, LBT (Look-ahead batching for Tardiness minimization), for single product type case where the jobs arriving after decision making point are also considered. They show their strategy outperforms the previous models including LAB and NACH-T. Later, Mansoer and Koo [10] extend their previous work by considering multiple product types. This paper extends the works in Mansoer and Koo [9], [10] by considering batch processing station with parallel BPMs with multiple product types. (The term, LBTm will be used for this new control strategy throughout the paper)

III. A New Real-Time Control Strategy

This section presents a new realtime control strategy for parallel BPMs with the objective of tardiness minimization. Some important characteristics of the system under consideration are as follows: (1) There are M identical BPM machines in the system. Each BPM can accommodate a limited number of wafer lots. (2) It can be estimated when the BPMs finish their current jobs and become available. (3) The jobs of different product types cannot be formed as a batch. (4)The jobs arrive to the BPM station dynamically. The arrival time of the jobs can be predicted. Each job has its own due date. (5) The processing time for each product type may be different from each other. The decision time for this strategy occurs when a BPM is available or when a wafer lot arrives to the queue. When a loading decision is to be made, batches of all product types are temporarily constituted and priorities of the batches are calculated based on the total tardiness involved, and then loading time and the jobs to be loaded are selected. A detailed description for LBTm is given as follows.

Step 1: Determine the look-ahead window.

We only consider the upcoming products arriving within a look-ahead time window. The look-ahead time window depends on processing time and machine available time. The look-ahead time window is then obtained by

$$LW = min(t_0 + \bar{T},\ A_{m^*}) \qquad (1)$$

Where t_0 is current time, is average processing time for all product types, A_{m*} is available time of machine with second smallest available time after t_0.

Step 2: At t_0, calculate total tardiness for each product type, then suggest the product type giving minimum tardiness.

Before calculating the total tardiness, set of products of type *j* to be loaded at decision time t_0, is formed based on earliest due date rule. Then, the total tardiness is calculated for each product type. The calculation is adapted from the equation for mean tardiness metric in Cerecki and Banerjee [1]. The product type with the smallest tardiness value is selected as a loading decision alternative at t_0,

Step 3: At t_n, calculate total tardiness of the product type and suggest a loading decision

At a particular decision time, t_n, a product with a specific product type arrives to the system. Therefore, instead of calculating total tardiness value for all possible product types, the calculation of total tardiness should only be done to that specific product type. Then, the set of products of the type to be loaded at decision time t_n is formed. The total tardiness is then calculated by using the same expression in the previous step.

Step 4: Compare all loading decision alternatives and choose the one which creates the smallest total tardiness as the loading decision.

In this step, the result obtained from step 2 and step 3 are compared to find the loading time, t_{n^*}. The loading time is selected by $n^* = arg_n$ min (total travel time at time *n*). If n^* is 0, then the BPM should start right away by loading a batch with product type j^*. Otherwise, the BPM stays idle to wait until one next arrival and then runs the LBTm logic again.

IV. Experimental Result And Analysis

In order to evaluate the performance of the proposed model, LBTm, a series of experiments have been performed. The batch processing station under consideration has three identical BPMs with the same specification. Each BPM can process up to six wafer lots at a time. There are five types of products due to chemical or mechanical requirements for the diffusion process. The manufacturing systems are modeled with ARENA software package and the control logic is written in Visual Basic Application (VBA). For the experiments, the simulation runs for 1,200 hours, in which the first 200 hours are treated as warm-up period. To secure the statistical reliability of the experiments, the simulation runs 100 times. In the experiments, at most five upcoming products are considered in decision making. Each product type may require different processing time. However, for the base scenario, the processing times for all the product types are the same as 6 hours. The products arrive in batch processing station and wait in a temporary queue buffer before being loaded on a BPM. For the base scenario, the traffic intensity is 60%. The traffic intensity (TI) is defined by TI = average processing time/(interarrival time × capacity × number of machines)

The inter-arrival time was adjusted in a way to obtain desirable traffic intensity. To have TI of 60%, the average inter-arrival time of the products should be 0.555 hours. The due-date for each product is set based on the arrival time by using following equation: d_j = arrival time of product *j* + (1+UNIF[0,*k*])*(processing time). Here, due-date value is also dependent on *k*, in which *k* represents the tightness of the due-date. For the base scenario, we assume that *k* is 2. (the effect of varying *k* will be also examined later.)

The proposed strategy, LBTm, is compared with existing strategies which deal with tardiness minimization, MMBS, NACHM and PRALC. MMBS is modified version of MBS in which due dates are involved in decision making. The logic for MMBS is when MBS requirement is fulfilled, the product type with the most products in queue is loaded. If there are ties between product-types, the one with larger tardiness value is loaded. Since MMBS performance for the system with different TIs is dependent on MBS value, the MBS value of two is determined in this case through preliminary experiments. The control procedures of NACHM and PRALC are found in Fowler [3] and Kim [8], respectively. NACHM, which is a control strategy to minimize lead time, is chosen as the benchmark strategy in order to observe the performance of such control strategy in tardiness minimization. Meanwhile, PRALC is chosen as the benchmark strategy because to our best knowledge, it is the only strategy which considers tardiness minimization in multiple product type multiple machine environment. Fig. 1 shows the average tardiness obtained from the different control strategies for the base scenario. It is seen that for the base scenario, LBTm outperforms the other control strategies from the literature. NACHM gives the worst performance which shows that the strategy with lead-time minimization objective does not perform well in tardiness minimization.

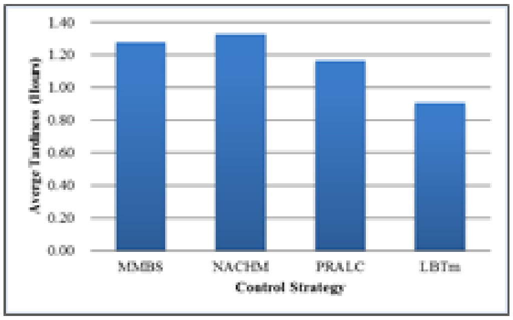
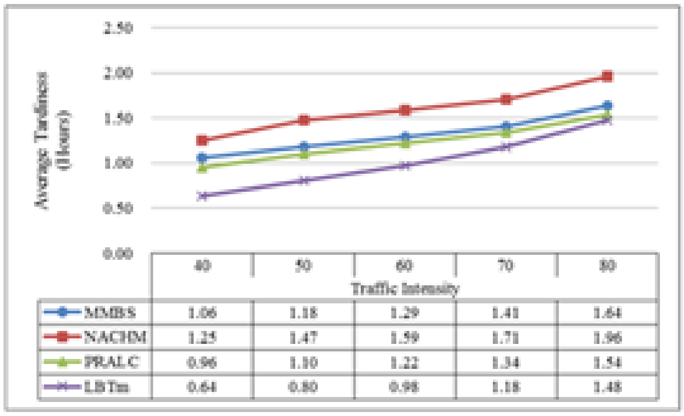


Fig. 1 Performance Of Control Strategies For Base Scenario

We have examined the effect of the traffic intensities on the performance of the control strategies. Experiments are performed with traffic intensities from 40% to 80%. The experimental results are given in Fig. 2. It is seen that the performance of LBTm works well consistently over various traffic intensities. LBTm works especially well in lower and intermediate traffic intensities. When traffic intensity is very high like 80%, then the performance of MMBS, PRALC, and LBTm is very close to each other. This results can be explained as follows: If the traffic intensity is very high, the machines are busy. Then, when a machine becomes idle, it is often better to load the jobs waiting at the queue immediately. In this situation, the decision logic becomes similar to MMBS strategy no matter what strategies are applied.

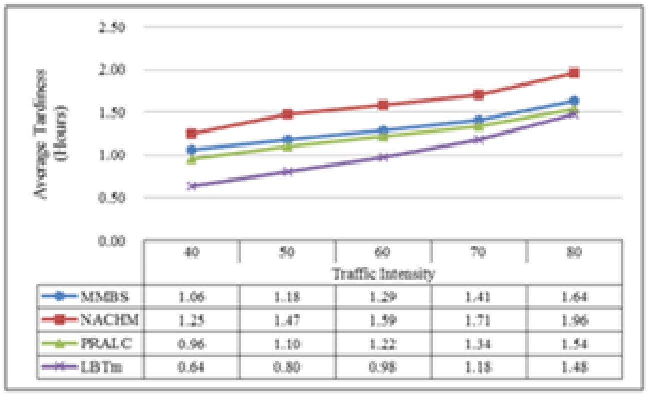
http://dx.doi.org/10.15242/IAE.IAE0215212



	40	50	60	70	80
MMBS	1.06	1.18	1.29	1.41	1.64
NACHM	1.25	1.47	1.59	1.71	1.96
PRALC	0.96	1.10	1.22	1.34	1.54
LBTm	0.64	0.80	0.98	1.18	1.48

Fig. 2 Performance Of Control Strategies Over Various Traffic Intensities

In the base scenario, we assume that the processing times for all the product types are the same. We also have tested the performance with different processing times for different product types. The processing times of the five product types assume to be 4, 5, 6, 7, and 8 hours, respectively. (Their average processing time is still 6 hours.) Fig. 3 shows the experimental results with different processing times. It is seen that even with the different processing times, LBTm shows stable performance.



	40	50	60	70	80
MMBS	1.06	1.18	1.29	1.41	1.64
NACHM	1.25	1.47	1.59	1.71	1.96
PRALC	0.96	1.10	1.22	1.34	1.54
LBTm	0.64	0.80	0.98	1.18	1.48

Fig. 3 Tardiness Over Traffic Intensity In Different Process Time

In the previous section, it is mentioned that the proposed method takes advantage of a certain look-ahead value, which determined as five. In this part of the experiment, the performance of LBTm over different look-ahead values is observed. Fig. 4 shows the average tardiness value when the look-ahead number is varied. A declining curve can be seen from the graph, and at the same time declining rate is decreasing as the look-ahead number increases. Large look-ahead number requires more information and increases the complexity of decision making. Therefore, it is necessary to select an appropriate look-ahead number. We select the look-ahead number of five because there are only little improvements with more than five

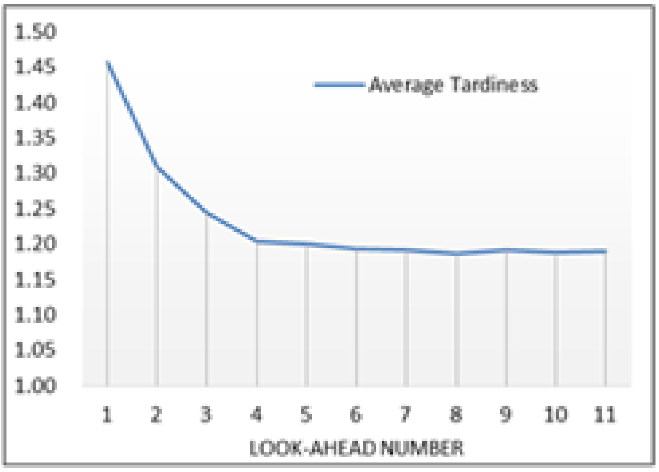


Fig. 4 Average Tardiness On Different Look-Ahead Number

V. CONCLUSION AND FUTURE WORKS

This paper presents a look-ahead control strategy, LBTm, for batch processing stations with multiple machines in semiconductor manufacturing for tardiness minimization on multiple product type environments. LBTm takes advantage of due-date and arrival time information of upcoming products as well as current products in queue, and the the availability information of the machines. Unlike the existing control policies, even when the number of products in queue is greater than capacity, loading can be delayed for more urgent products. In addition, the fixed number of look-ahead in LB Tm reduces the chance of being affected by the arrival of the next product. The experimental results show that the new control strategy provides performance of good quality.

Our work is now ongoing to observe the effect of uncertainty on the performance of the current model. For example, machine breakdown and stochastic processing time are to be included in the simulation experiments. In addition, it will be also interesting to see the relationship between the performances measured at the batch processing machines and the whole system of semiconductor manufacturing. Here, the relationship of product input control, lot scheduling strategy at the bottleneck machine, dispatching decisions at the non-bottleneck machines, and batch loading decisions on batch processing machines are of special interest.

REFERENCES

[1] A. Cerekci, A. Banerjee, "Dynamic control of the batch processor in a serial-batch processor system with mean tardiness performance", *International Journal of Production Research*, vol. 48, no. 5, pp. 1339-1359, 2010.
http://dx.doi.org/10.1080/00207540802641437

[2] J. W. Fowler, D. T. Phillips, G. L. Hogg, "Real-Time Control of Multiproduct Bulk-Service Semiconductor Manufacturing Processes", *IEEE Transactions on Semiconductor Manufacturing*, vol. 5, no.2, pp. 158-163, 1992.
http://dx.doi.org/10.1109/66.136278

[3] J. W. Fowler, G. L. Hogg, D. T. Phillips, "Control of multiproduct bulk server diffusion/oxidation processes. Part 2: multiple servers", *IIE Transactions*, vol. 32, pp. 167-176, 2010.
http://dx.doi.org/10.1080/07408170008963889

[4] C. Glassey, W. Weng, "Dynamic batching heuristic for simultaneous processing", *IEEE Transactions on Semiconductor Manufacturing*, vol. 4, no.2, pp. 77-82, 1991.
http://dx.doi.org/10.1109/66.79719

[5] A. Gupta, A. Sivakumar, "Optimization of due-date objectives in scheduling semiconductor batch manufacturing", *International Journal of Machine Tools and Manufacture*, vol. 46, pp. 1671-1679, 2006.
http://dx.doi.org/10.1016/j.ijmachtools.2005.08.017

[6] A. Gupta, A. Sivakumar, "Controlling delivery performance in semiconductor manufacturing using look ahead batching", *International Journal of Production Research*, vol. 45, no. 3, pp. 591-613, 2007.
http://dx.doi.org/10.1080/00207540600792226

[7] Y. D. Kim, J. G. Kim, B. Choi, and H. U. Kim, "Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates", *IEEE Transaction on Robotics and Automation*, vol. 17, no. 5, pp.589-598, 2001.
http://dx.doi.org/10.1109/70.964660

[8] Y. D. Kim, B. J. Joo, and S. Y. Choi, "Scheduling wafer lots on diffusion machines in a semiconductor wafer fabrication facility", *IEEE Transactions on Semiconductor Manufacturing*, vol.23, no.2, pp. 246-254, 2010.
http://dx.doi.org/10.1109/TSM.2010.2045666

[9] P. Mansoer, and P. H. Koo, "A control strategy of batch processing machines in semiconductor manufacturing", *Proceedings of 17th International Conference on Industrial Engineering*, pp. 830-837, Korea, 2013.

[10] P. Mansoer, and P. H. Koo, "Real-time Control of the Batch Processor for Tardiness Minimization", *ICIC Express Letters,* to be published, 2015.

[11] M. F. Neuts, "A general class of bulk queues with Poisson input", *Annals of Mathematical Statistics*, vol. 50, no. 20, pp. 6022-6035, 1967.
http://dx.doi.org/10.1214/aoms/1177698869

[12] D. Van Der Zee, A. Harten, and P. C. Schuur, "Dynamic job assignment heuristics for multi-server batch operations - A cost-based approach". *International Journal of Production Research*, vol. 35, pp. 3063-3093, 1997.
http://dx.doi.org/10.1080/002075497194291

[13] D. Van Der Zee, A. Harten, and P. C. Schuur, "On-line scheduling of multi-server batch operations, IIE Transactions", vol. 33, pp. 569-586, 2001.
http://dx.doi.org/10.1023/A:1010844500752

[14] W. Weng, and R. C. Leachman, "An improved methodology for real-time production decisions at batch-process work stations", *IEEE Transactions on Semiconductor Manufacturing*, vol. 6, no. 3, pp. 219-225, 1993.
http://dx.doi.org/10.1109/66.238169

Min Jeong Park received Bachelor's degree from Pukyong National University, and currently pursuing her Master degree at System Management and Engineering in Pukyong National University, Korea. Her research interest includes manufacturing logistics, supply chain management and production management. She is a member of Korean Institute of Industrial Engineering.

Paramitha Mansoer received Bachelor's degree from University of Indonesia, and MS in Pukyong National University, Korea. She is currently working on Dae-Chang company. Her research interest includes manufacturing logistics, and supply chain management. She is a member of Korean Institute of Industrial Engineering

Pyung Hoi Koo received Bachelor's degree from Hanyang University, Korea, and MS and Ph.D. in Industrial Engineering from Purdue University, USA. His major research area includes manufacturing logistics, supply chain management and OR applications in industrial problems. Prof. Koo is now a professor in Department of Systems Management and Engineering in Pukyong National University, Korea. He is a member of Korean Institute of Industrial Engineering, Korean SCM Society and Korean Management Science and Operations Research Society.