SIMULATION-BASED ADVANCED WIP MANAGEMENT AND CONTROL IN SEMICONDUCTOR MANUFACTURING

Kazuo Miyashita

National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki 305-8564, JAPAN Tsukasa Okazaki

Hitachi East Japan Solutions, Ltd Sendai, Miyagi 980-0014, JAPAN

Hirofumi Matsuo

Kobe University Kobe, Hyogo 657-8501, JAPAN

ABSTRACT

We develop a hierarchical distributed production planning and control methodology, called DISCS, for a large and unstable semiconductor manufacturing process. The upper layer of DISCS periodically optimizes work-in-process inventory (WIP) levels to meet demands and sets a target WIP level for each workstation. One of key technologies required for the purpose is a fast simulation method to make the iterative optimization process tractable. In the lower layer, dispatching decisions are made at each workstation based on its target WIP level. Computational experiments using wafer fabrication process data show that DISCS, when compared with a traditional control method, succeeds in meeting the demand while keeping lower WIP levels. This indicates that DISCS is a promising methodology for production planning and control in semiconductor manufacturing.

1 INTRODUCTION

Semiconductor manufacturing process is among the most complicated manufacturing processes. For example, the number of production steps for semiconductor manufacturing is usually not less than a few hundreds with a large number of repetitive reentrant loops, and its throughput time extends over two months (Atherton and Atherton 1995). Furthermore, the fierce competition in the global market place and short technology life cycles require the semiconductor industry to always deploy state-of-the-art manufacturing technologies. It causes their manufacturing processes to be unstable and unpredictable because they most of the time operate in the early part of the experience curves of manufacturing. In order to hedge disruptions in manufacturing and avoid the ripple effects of exceptional events such as machine breakdowns and yield loss, they tend to keep large amount of work-in-process inventory (WIP) and end up with unnecessarily long production lead times. The costs of having large WIP are high because of rapid product obsolescence associated with short product and technology life cycles. Consumers of semiconductors also demand short lead times and customized products. Hence, a production planning and control methodology that can always maintain the lowest possible levels of WIP while keeping high output levels is of critical importance in semiconductor manufacturing.

Developing an optimal production plan for the semiconductor manufacturing process in an exact sense is computationally intractable. Hence, many simulationbased methods using various releasing heuristics and dispatching rules have been investigated extensively in the literature, and have been applied to the factories (Wein 1988, Leachman, Kang, and Lin 2002, Folwer, Hogg, and Mason 2002). However, quality of results obtained by those conventional methods are far from satisfactory in terms of both backorder and inventory costs.

To solve the above-mentioned problems, an advanced production planning and control methodology, which can maintain the minimum amount of WIP for meeting the demands, is required. To develop such methodologies, the following issues need to be addressed; (1) how to determine the adequate WIP level for a given semiconductor manufacturing line, and (2) how to control the material flow to maintain the fixed WIP level at each workstation. In this paper, we propose a hierarchical distributed planning and control methodology for semiconductor manufacturing process, and show that the proposed method can reduce



Figure 1: The DISCS System

production gaps from the demands and keep reasonably low WIP levels.

This paper is organized as follows. In Section 2, the outline of the proposed system for advanced production planning and control is explained. Section 3 describes the method to determine the minimum WIP level for meeting the demands. Then, Section 4 explains the detailed algorithms of distributed control. In Section 5, the results of the experiments are presented in comparison with those of the traditional method. The paper concludes with a summary and suggestions of future work.

2 PROPOSED METHODOLOGY

We are developing the system called DISCS (<u>Distributed</u> Production <u>Control System</u>) for production planning and control in semiconductor manufacturing. The goal of the system is to achieve robust and efficient production in largescale and complicated manufacturing environments with highly fluctuating factors such as machine failures, yield loss and unpredictable demands.

As shown in Fig. 1, DISCS has a hierarchical architecture that consists of a planning layer and a control layer. In the planning layer, the target WIP level for each workstation is determined periodically to respond to the unsteady manufacturing situations. The target WIP level needs to be optimized based on the current demands and WIP level. Since such optimization for large-scale manufacturing problems needs enormous computation, an efficient optimization method should be developed for DISCS.

The control layer of DISCS makes production decisions in a distributed manner. Each controller in the control layer is in charge of job releasing and sequencing at each workstation. The controller keeps WIP of the workstation at the target level set in the planning layer by determining when and what product to be released to the production line and which lot to be processed next at each workstation.

The distributed production control has the following advantages to be exploited in large-scale and unstable manufacturing environments:

- 1. **Responsiveness to unexpected events:** when an unexpected event occurs at a workstation, the controller for the workstation needs to respond to the problem. That is, each controller has clear responsibility, and therefore can prepare for it in advance by having specific rules and functions. As a result, each controller can respond to the unexpected events quickly and correctly.
- 2. **Robustness against variability:** when a machine failure happens at one workstation, the controllers of the other workstations do not have to interrupt their process. Hence, the overall manufacturing process proceeds without stopping if the failure is resolved timely at the workstation.
- 3. Flexibility in extension: since each controller has distinct responsibility, every controller does not need to be completely developed at the time of deployment. If necessary, any controller can be improved later in the factory floor without losing the overall functionality of the manufacturing system.

It is inherently difficult to understand behaviors of the complicated distributed control system and to improve its performances. Hence, in DISCS, the planner calculates the minimum WIP level of every product for each controller that prevents backorders by considering current demands and manufacturing environments. Then, each controller in the distributed control layer of DISCS has to make sure that WIP at the workstation is maintained at the minimum level set by the planner. Thus, the controllers in DISCS do not need to communicate each other for coordinating their operations. Since each controller can work independently, DISCS does not suffer from the common difficulty in the distributed control.

3 PLANNING TARGET WIP LEVEL

To determine the target WIP levels, DISCS needs to calculate the minimum WIP level that satisfies the demands in a given manufacturing environment. The calculation using any optimization method such as Genetic Algorithm and Simulated Annealing requires many simulation runs of the manufacturing process in search of a near optimal solution. Since semiconductor manufacturing process is large-scale and complicated, traditional simulation methods take enormous time for simulation. Hence, they are unsuitable to be used in DISCS.

3.1 Fast Simulation Method: CONSTIN

To solve the problem, the fast simulation method, called CONSTIN (<u>CONS</u>tant <u>Time IN</u>terval method), is developed to determine the target WIP level in DISCS (Miyashita et al. 2003). In CONSTIN, a manufacturing



Figure 2: Periodic Transition of WIP Level in CONSTIN

process comprises a sequence of operations which are run in a synchronized manner, and WIP is transfered between operations only at the end of a fixed time interval. Hence, the WIP level at each operation changes periodically as shown in Fig. 2. As an additional restriction on WIP movements in CONSTIN, WIP is assumed not to move beyond its next operation at a single time interval. By setting an appropriate time interval (e.g., 60 minutes), compared with a conventional event-based simulator that updates WIP quantities every time related events occur in the manufacturing process, CONSTIN calculates WIP quantities in far less occasions (i.e., once in a fixed time interval). Preliminary experiments show that CONSTIN simulates a semiconductor manufacturing process about 20 to 100 times faster than a commercial event-based simulator (in this experiment, AutoSched AP is used for comparison) and produces results of comparable quality (Miyashita, Ozaki, Senoh, and Matsuo 2003).

3.2 Optimizing Target WIP

In the literature, some theoretical aspects of setting the target WIP levels and reducing the gaps between the target and current WIP levels are discussed in Tang (1990) and Gong and Matsuo (1997). This paper attempts to identify optimal target WIP levels more directly than the previous research incorporating detailed information on manufacturing environments.

In the planning layer of DISCS, the target WIP level is calculated using Genetic Algorithm (GA) and CONSTIN. Each chromosome used in GA represents a distribution of target WIP control factor α for a workstation. Little's Law (Little, 1961) states that, in a steady state, a queue length (i.e, WIP level) should be equal to an arrival rate (i.e., throughput) times a service rate (i.e., lead time). In a steady state, a value of throughput is known to be identical with a demand rate, although a value of lead time is undecidable a priori. Hence, in practice, a target WIP level for the *i*-th



Figure 3: Beowulf Cluster Computer

step of a product p at the workstation is approximately set as

 $\alpha s_p^i D_p$

where s_p^i is the processing time of the *i*-th step of the product *p* at the workstation and D_p is the demand rate of the product *p*. Since the values of s_p^i and D_p are given, the target WIP level is controlled solely with the value of target WIP control factor α . Hence, GA is applied to optimize the value of target WIP control factor α .

The objective function for evaluating fitness of each individual in GA is the sum of resulting WIP and backorders after a 6 months' simulation run. For this simulation, CONSTIN is used to enable real-time re-calculation of the optimized target WIP level corresponding to changes in the manufacturing process.

For further speeding up of the GA calculation, individuals in GA population are evaluated in parallel using the Beowulf cluster computer (see Fig. 3), which has 8 nodes of 3.06 GHz dual Xeon computers interconnected with Gigabit Ethernet. Setting GA parameters as 96 individuals and 100 generations, DISCS can calculate the optimized target WIP level in less than 20 minutes, which makes it possible to update the target WIP for every shift (i.e., every 8 hour) of operations.

4 DISTRIBUTED CONTROL OF WIP

In the control layer of DISCS, each controller exploits two kinds of control rules to maintain the target WIP level set by the DISCS planner at the workstation. One rule is the *Release Rule*, which controls the timing of job release into the production line, and the other rule is *Dispatching Rule*, which controls the sequence of lots to be processed at each workstation when it becomes available.

4.1 Release Rule

In DISCS, the total amount of WIP is controlled to be equal to the sum of the target WIP level for each workstation set by the planner. This type of WIP control is also done in the CONWIP methodology (Hopp and Spearman 2000) and the JIT production system. In terms of job release, DISCS adopts a similar rule to the above method. The differences of the DISCS methodology from the above methods are: (1) not only the total WIP level but also the WIP level at each workstation is planned and controlled in DISCS, and (2) each controller of a workstation works independently from the other controllers and can avoid suffering from the cascading effects of the exceptional events in the other workstations. These points are critically important to realize distributed control in semiconductor manufacturing.

To maintain the target WIP level in DISCS, the amount of job release is determined by the following rule:

$$max(0, \sum_{j=0}^{n_p} W_p^j - \sum_{j=0}^{n_p} w_p^j)$$

where n_p is the number of steps for product p, W_p^j is the target WIP level at the *j*-th step of product p, and w_p^j is the current WIP level at the *j*-th step of product p. $\sum_{j=0}^{n_p} W_p^j$ and $\sum_{j=0}^{n_p} w_p^j$ represent respectively the target WIP level and the current WIP level of product p in the system. When the current WIP level is less than the target WIP level, the shortfall is compensated by releasing new lots. When the current WIP level is more than the target WIP level, no new lot is released. Thus, DISCS maintains the fixed total target WIP level in the system through the above release rule.

4.2 Dispatching Rule

In addition to maintaining the total WIP level constant, DISCS sets a target WIP level for each process step and makes each workstation keep its WIP as close to the target level as possible. The controllers in DISCS attempt to maintain the WIP level at each process step using a dispatching rule. The dispatching rule sorts lots to be processed at the workstation according to a priority of each jobs. That is, the basic idea of the dispatching rule in DISCS is to put a higher priority on the lot that has a larger (positive) gap between the current WIP level and the target WIP level. The gap between the current WIP and the target WIP for the *j*-th step of product p, x_p^j , is calculated as:

$$x_p^j = max(0, w_p^j - W_p^j).$$
 (1)

In semiconductor manufacturing, occasional backorders at the product steps are inevitable due to machine failures and/or yield loss. Hence, to meet the demands from customers, the controller in DISCS needs to expedite production of the lot with larger backorders. Backorders of the j-th step of product p is calculated as follows:

$$b_p^j = max(0, d_p^j - z_p^j)$$
 (2)

where d_p^j is the total demand for the *j*-th step of product p, and z_p^j is the total output of the *j*-th step of product p. Therefore, in DISCS, the priority of the *j*-th step of product p is determined as follows:

$$\frac{x_p^j + b_p^j}{D_p}.$$
 (3)

Here, the priority is normalized by D_p in order to cancel the effect of different demand quantity for products.

Backorders are caused by the situation that the step has not been processed for a while, and the same situation also causes the large WIP for the step. Hence, the two terms of reducing the WIP overage and filling up backorders in the above dispatching rule do not contradict each other. The dispatching rule in DISCS is capable of maintaining the target WIP level as well as reducing the backorders.

Another important point in the above dispatching rule is that a controller does not refer to the information managed by another controller. Hence, each controller can work independently and autonomously in the distributed control architecture, and DISCS as a total system can pursue the global objective of minimizing WIP and backorders.

5 COMPUTATIONAL EXPERIMENTS AND ANALYSIS

We conducted computational experiments to test validity of DISCS and compare its results with the traditional control method using realistic data of semiconductor manufacturing processes.

5.1 Problem Definitions

test data of wafer fabrication processes, As we MIMAC used the (Measurement and Improvement of MAnufacturing Capacities) testbed datasets (Fowler and Robinson 1995), which are now maintained and made public by MASM lab in Arizona State University. For further details and downloads of dataset, see <www.eas.asu.edu/~masmlab/home.htm>.

Table 1 shows the properties of the MIMAC test problem. It has basic characteristics of a semiconductor manufacturing process such as lengthly process flow with many

Non-volatile memory
2
2 (i.e., 1 per process)
83
265
210 (Product 1)
245 (Product 2)
313.4 (Product 1)
358.6 (Product 2)
365.72 (Product 1)
182.86 (Product 2)
48(Product 1,2)

Table 1: Test Problem

repetitive reentrant loops and a couple of bottleneck workstations.

In this experiment, we made the following assumptions to focus our investigative attentions to the basic properties of DISCS: (1) there is no variabilities in processing times of operations, (2) no setup time is considered, (3) operators are not considered in the model, (4) there is neither product rework nor scrap, (5) stochastic machine failure is modeled using exponential distribution, and (6) the demand rates are tuned to realize 100% resource utilization for a bottleneck machine.

To compare with the DISCS method, a traditional control method is designed as follows:

- 1. A release rule to the production line is defined as the constant release rule that releases a new job in a constant time interval based on the demand rate of products.
- 2. As a dispatching rule, the FIFO (First-In-First-Out) rule is used.
- 3. To determine initial WIP level, WIP control factor α is set as 4.5.

These rules and values are often used in practice of real semiconductor manufacturing.

In the experiments, the simulation of 360 days is executed under the above conditions. For DISCS, the target WIP level is updated every 30 days. And, to implement the traditional control method and the controllers in DISCS, the DPSS (Distributed Production Scheduling System) system is developed as an event-based simulation system in a distributed control architecture.

5.2 Experimental Results

Fig.4 shows the transitions of the total WIP over the simulation time. From the graph of DISCS results, it is shown that the control system in DISCS smoothly maintains the target WIP level determined periodically by the planning system



Figure 4: Total WIP level

(i.e., every 30 days for this experiment). The graphs depict that the WIP level of the DISCS system is less than that of the traditional method over the entire simulation periods. The average WIP level of the traditional method is 25,000 wafers, and for DISCS it is 21,700 wafers. That is, DISCS succeeds in reducing the WIP by more than 13 percents.

Figs.5 and 6 show the product inventories and backorders. In the graphs, positive numbers indicate the amount of product inventories and negative numbers indicate backorders. In the traditional method, it is shown that there are many backorders for Product 2. In the DISCS system, however, there are few backorders for both products.



Figure 5: Product Inventory and Backorder (Traditional Method)

The above results show that, as compared with the traditional method, the DISCS system can reduce the WIP level as well as backorders.

Figs.7 and 8 show the inventory and backorder at each step for Product 1. Each line is plotted every month. Again, positive values indicate product inventories and negative values indicate backorders in the graphs.

In the traditional method, inventories exist in the product steps between 30 to 180. In the DISCS system, inventories exist only in the earlier steps. Since inventories at the later product steps are more expensive to hold because of the



Figure 6: Product Inventory and Backorder (DISCS)



Figure 7: Inventory and Backorder Distribution (Traditional Method)

values added at the previous steps, the traditional method causes higher inventory costs than the DISCS system.

These results demonstrate that DISCS succeeds not only reducing WIP but also reducing *expensive* WIP to hold in the manufacturing process.

5.3 Analysis of Target WIP

The experimental results show that the DISCS system is able to reduce WIP while maintaining the stable production by setting and updating the appropriate target WIP levels for each workstation. To understand how the target WIP works for controlling WIP in DISCS, we investigate the relationships among the values of utilization, target WIP and actual WIP for the workstations.

Fig. 9 shows the average utilization of the workstations during 2 years long simulation. It depicts that the workstations No.67, No.76 and No.78 are the bottleneck workstations that have high utilization rates(i.e., 93.9%, 92.7% and 85.8% respectively).

Fig. 10 represents WIP at the workstations. The graph shows that the 3 bottleneck machines have larger WIP than other workstations and the workstation No.67, which has the highest utilization rate, has by far the largest WIP. It is to be noted that the workstation No.9 also has large WIP. But



Figure 8: Inventory and Backorder Distribution (DISCS)



Figure 9: Utilization of Workstation

this is not because the workstation No.9 is highly utilized but because it is a batch production machine that cannot start processing until 4 lots are accumulated in its buffer to make a batch.

Fig. 11 shows the target WIP for each workstation. The black bar in the graph shows the target WIP level of the bottleneck machine, the workstation No.67. Although the bottleneck machine naturally has large WIP as shown in Fig. 10, DISCS sets a small value as its target WIP. This experiment result can be rationally explained as follows.

The dispatching rule in the DISCS system (see Eq.(3)) consists of the two important terms (since the third term D_p in the equation is for normalization); one is the gap between



Figure 10: WIP of Workstation



Figure 11: Target WIP of Workstation

actual WIP and target WIP of the *j*-th step of product *p*, x_p^j in Eq.(1), and the other is the backorder of the *j*-step of product *p*, b_p^j in Eq.(2).

Fig. 12 shows the average value of x_p^j for all the product steps processed at each workstation. Apparently the workstation with large WIP and small target WIP has a large x_p^j value. For the workstations whose values are 0.0 in the graph, the dispatching rule Eq.(3) works actually in a late-lot-first manner. For the workstations whose average x_p^j values are large compared with the corresponding value of b_p^j , the difference of the x_p^j values among the lots determines a priority of being processed next.



Figure 12: Average Gap between Target WIP and Actual WIP for All the Product Steps Processed at Workstation

Fig. 13 shows the standard deviation of the average value of x_p^j for all the product steps processed at each workstation. The graph reveals that the bottleneck machine, workstation No.67, has a much larger deviation in the average value of x_p^j among the product steps than the other workstations.

Since, in the DISCS system, the target WIP is determined to keep backorders (i.e., the value of b_p^j) small, this explains that for the bottleneck machine the value of x_p^j is a dominant factor in the dispatching rule. Therefore, in DISCS, the bottleneck machine urges a lot with larger WIP to be processed earlier, making its WIP balanced among the



Figure 13: Standard Deviation of Average Gap between Target WIP and Actual WIP for All the Product Steps Processed at Workstation

product steps. Unbalancing WIP at the bottleneck machine causes backorders of some product and it takes long time to recover the backorders produced in the complicated manufacturing process. Setting the small target WIP level for the bottleneck machine is considered effective for preventing backorders and realizing stable production.

Since we use the CONWIP-type control rule for lot releasing, total WIP is always maintained to be equal to the sum of target WIP for each workstation. The DISCS determines the minimum total WIP level as the sum of target WIP levels to achieve stable production and distributes the target WIP to each workstation so that it can effectively prevent backorders.

5.4 Estimation of Economic Benefits

The followings are very rough estimation of direct economic benefits of reducing inventory holding costs. By reducing WIP, in addition to reducing inventory holding cost, several secondary effects can be realized such as reducing lead time, decreasing yield loss and improving facility utilization. Here, the only most direct effect is evaluated for reference.

Let us suppose as follows:

- 1. Cost per a chip is 2 dollars.
- 2. One wafer produces 500 chips.
- 3. Annual inventory carrying factor is 20%.
- 4. Annual value loss factor is 50%.

Then, the annual inventory related cost of a wafer is $2 \times 500 \times (0.2 + 0.5) = 700$ dollars.

From the above experiments, the reduced WIP in DISCS is 25, 000-21, 700 = 3, 300 wafers. Therefore, the reduced inventory related cost per year is 3, $300 \times 700 = 2$, 310, 000 dollars. The reduced cost is about 1.1% of the annual revenue (about 260 million dollars) calculated from the demand rates. Although this estimation depends upon several assumptions, the value is calculated in a conservative manner (e.g., it does not include the obsolescence cost of a wafer, nor

the benefits produced from the reduced leadtime). Actual benefits brought by the DISCS system will be severalfold of the estimation.

6 Conclusion

This paper presented the hierarchical distributed planning and control methodology, DISCS, for large-scale and complicated manufacturing processes such as semiconductor fabrication. For manufacturing processes involving stochastic elements such as machine failures and yield loss, it is difficult to keep stable outputs without a high level of WIP. However, unless its size is properly controlled, WIP results in long lead times and increased obsolescence stocks. DISCS proposed in this paper determines the target WIP level of each step to minimize the inventory and backorder costs, and it controls production sequences at each workstation distributively based on the target WIP level. The results of computational experiments show that DISCS can reduce the WIP level and also decrease backorders at the same time.

As a future study, a distributed learning method will be developed to improve coordination among controllers in the DISCS. One of the possible scenarios to be studied is to adjust the sequencing decisions of the controller to reduce set-ups or accelerate batch operations at the succeeding product steps.

REFERENCES

- Atherton, L. F., and R. W. Atherton. 1995. *Wafer fabrication: Factory performance and analysis*. Kluwer Academic Publishers.
- Folwer, J. W., G. L. Hogg, and S. J. Mason. 2002. Workload control in the semiconductor industry. *Production Planning & Control* 13 (7): 568–578.
- Fowler, J., and J. Robinson. 1995. Measurement and improvement of mafucaturing capacities (MIMAC): Final report. Technical Report 95062861A-TR, SEMATECH, Austin, TX.
- Gong, L., and H. Matsuo. 1997. Control policy for manufaturing system with random yield and rework. *Journal of Optimization Theory and Applications* 95 (1): 149–175.
- Hopp, W. J., and M. L. Spearman. 2000. Factory physics. Second ed. McGraw-Hill.
- Leachman, R. C., J. Kang, and V. Lin. 2002. SLIM: Short cycle time and low inventory in manufacturing at Samsung Electronics. *INTERFACES* 32 (1): 61–77.
- Miyashita, K., H. Ozaki, K. Senoh, and H. Matsuo. 2003. Constant time interval simulation for semiconductor manufacturing. In Proc. of 2003 IEEE Conference on Emerging Technologies and Factory Automation, 364– 371. IEEE.

- Miyashita, K., K. Senoh, H. Ozaki, and H. Matsuo. 2003. Constant time interval production planning with application to WIP control in semiconductor fabrication. In *Proceedings of the 2003 Winter Simulation Conference*, 1329–1337. INFORMS.
- Tang, C. 1990. The impact of uncertainty on a production line. *Management Science* 36 (12): 1518–1531.
- Wein, L. M. 1988. Scheduling semiconductor wafer fabrication. *IEEE Transaction on Semiconductor Manufactruing* 1 (3): 115–130.

AUTHOR BIOGRAPHIES

KAZUO MIYASHITA is a Senior Researcher of Intelligent Systems Institute at the National Institute of Industrial Science and Technology, and an Associate Professor of Institute of Computer Science at the University of Tsukuba. Prior to his current position, he spent ten years working for Production Engineering Laboratory of Panasonic. Dr. Miyashita received his B.E. and M.E. from the University of Tokyo, and his Ph.D. from Osaka University, respectively. He is a member of AAAI and IEEE CS. His e-mail address is <k.miyashita@aist.go.jp>, and his web page is <http://staff.aist.go.jp/k.miyashita/>.

TSUKASA OKAZAKI is a Senior Researcher of Research and Develpoment Department at Hitachi East Solutions, Ltd. His e-mail address is <okazaki@hitachi-to.co.jp>.

HIROFUMI MATSUO is a Professor of Graduate School of Business Administration at Kobe University. Dr. Matsuo received his B.E. and M.E. from Kyoto University and hid Ph.D. from Massachusetts Institute of Technology, respectively. He is a member of INFORMS. His e-mail address is <mastuch@kobe-u.ac.jp>.