ADVANCED WIP CONTROL METHOD FOR SEMICONDUCTOR MANUFACTURING

Tsukasa Okazaki
Hitachi East Japan Solutions, Ltd
2-16-10 Honcho, Aoba, Sendai, 980-0014, Japan
okazaki@hitachi-to.co.jp

Kazu Miyaishi
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki 305-8564, Japan
k.miyaishi@aist.go.jp

Hiromi Matsuo
University of Tsukuba
Tsukuba, Ibaraki 305-0006, Japan
matsuo@shako.sk.tsukuba.ac.jp

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 decides adequate work-in-process inventories (WIP) levels to meet demands. And, in the lower layer, dispatching decisions are made at each workstation to keep the fixed WIP level. Computational experiments with wafer fabrication process data show that DISCS, when compared with a traditional control method, succeeds in making less gaps between outputs and demands while keeping lower WIP level. Therefore, we conclude that DISCS is a promising methodology for production planning and control in semiconductor manufacturing.

Keywords: semiconductor manufacturing, distributed control, production planning, inventory control

1. INTRODUCTION
Semiconductor manufacturing process is among the most complicated manufacturing processes. For example, the number of production steps for semiconductor manufacturing are usually not less than a few hundreds with a large number of repetitive reentrant loops, and its Lead-time extends over a couple of months (Atheron and Antheron, 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 must of the time operate early in the learning 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 a large number 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 to keep high output levels are of critical importance in semiconductor manufacturing.

Nevertheless, developing an optimal production plan for the semiconductor manufacturing process in an exact sense is computationally intractable. Hence, many simulation-based methods using various releasing heuristics and dispatching rules have been investigated extensively in the literature, and have been applied to the factories (Wein, 1988; Rose, 2001). However, quality of results obtained by those conventional methods are far from "satisficing" in terms of both backorder cost and inventory cost.

To solve the above-mentioned problems, an advanced production planning and control methodology, which can maintain WIP at the levels indispensable for meeting the demands, is required. And, to develop such methodologies, the following issues need to be addressed; (1) how to determine the adequate WIP level for a given semiconductor manufacturing process, and (2) how to control the manufacturing process to maintain the fixed WIP level. In this paper, we propose a hierarchical distributed planning and control methodology for semiconductor manufacturing process, and
show that the proposed method can reduce 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 shown in comparison with those of the traditional method. The paper concludes with a summary and suggestions of future work.

2. PROPOSED METHODOLOGY

The authors are developing the system called DISCS (Distributed Production Control System) for production planning and control in semiconductor manufacturing. The goal of the system is to achieve robust and efficient production in large-scale 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 decided 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 fixed in the planning layer by appropriately deciding when to release a new lot of what product and which lot to be processed next.

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 should address the problem. Thus, 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.

However, 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 considering current demands and manufacturing environments. And, then each controller in the distributed control layer of DISCS has only to make sure that WIP at the workstation is maintained at the minimum level fixed by the planner. 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 difficulty of distributed control problem.

3. PLANNING TARGET WIP LEVEL

To determine the target WIP levels, DISCS needs to calculate the minimum WIP level enough for satisfying the demands in a given manufacturing environment. For the calculation, using any optimization method such as Genetic Algorithm and Simulated Annealing, many simulation runs of the manufacturing process should be executed in the search of a near optimal solution. Since semiconductor manufacturing process is large-scaled 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 (CONSTant Time Interval method), is developed to be used for determining the target WIP level in DISCS (Miyashita et al., 2003). In CONSTIN, a manufacturing process comprises a sequence of operations that are run in a synchronized manner, and WIP is transferred 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 values every
time when related events occur in the manufacturing process, CONSTIN calculates WIP values 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\(^1\) and produces results of comparable quality (Miyashita et al., 2003).

3.2 Optimizing Target WIP

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 \(\alpha\) 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, throughput should take the same value with a demand rate. But, since a value of lead time is unknown, a target WIP level for the \(i\)-th step of a product \(p\) at the workstation is calculated as

\[ \alpha L_p D_p \]

where \(s^p_i\) is the processing time of the \(i\)-th step of the product \(p\) at the workstation and \(D_p\) is a 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 \(\alpha\). Hence, GA is applied to optimize the value of target WIP control factor \(\alpha\).

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.

And, for further speeding up of the GA calculation, individuals in GA population are evaluated in parallel using the Beowulf cluster computer, which has 8 nodes of 3.06 GHz dual Xeon computers interconnected with Gigabit Ethernet.

As a result, setting GA parameters as 96 individuals and 100 generations, DISCS can calculate the target WIP level in less than 20 minutes, which make 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 2 kinds of control rules to maintain the target WIP level fixed by the DISCS planner at the workstation. One rule is the Release Rule, which controls the timing of job release into the workstation, and the other rule is Dispatching Rule, which controls the sequence of lots to be processed at the 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 decided 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 an identical rule with the above method. The differences of the DISCS methodology from the above methods are: (1) not only the total WIP level but also each WIP level for workstations are 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 by the events in the other workstations. These points are critical in realizing distributed control for semiconductor manufacturing.

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

\[ \max(0, \sum_{j=0}^{n_p} W^j_p - \sum_{j=0}^{n_p} w^j_p) \]

where \(n_p\) is the number of steps for product \(p\), \(W^j_p\) is a target WIP level at \(j\)-th step of product \(p\) and \(w^j_p\) is a current WIP level at \(j\)-th step of product \(p\). \(\sum_{j=0}^{n_p} W^j_p\) and \(\sum_{j=0}^{n_p} w^j_p\) represents 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 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 to keep its WIP as close to the target level as possible.

The controllers in DISCS maintain the WIP level at each process step using the dispatching rule. The dispatching rule sorts lots to be processed in the workstation according to a priority of each jobs. Thus, the basic idea of the dispatching

---

\(^1\)In this experiment, AutoSched AP is used for comparison.
rule in DISCS is to put a higher priority on the lot, which 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 $j$-th step of product $p$, $x_p^j$, is calculated as:

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

And, in semiconductor manufacturing, occasional backorders of 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 promote production of the lot with bigger backorders. Backorders of $j$-th step of product $p$ is calculated as follows:

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

where $d_p^j$ is a total demand for $j$-th step of product $p$ and $z_p^j$ is a total output of $j$-th step of product $p$.

Therefore, In DISCS, the priority of $j$-th step of product $p$ is determined as follows:

$$x_p^j + b_p^j

The priority is normalized by $D_p$ in order to unbiased the effect of different demand quantity for products.

Backorders are caused by the fact that the step has not been processed for a while. And the same fact also results in the large WIP of the step. Hence, the 2 terms of reducing the WIP average 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.

And 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 objectives of minimizing WIP and backorders.

5. COMPUTATIONAL EXPERIMENTS

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

As test data of wafer fabrication processes, we used the MIMAC (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 http://www.eas.asu.edu/masmlab/index.html.

Table 1 shows the properties of the MIMAC test problem. It has basic characteristics of a semiconductor manufacturing process such as lengthy process flow with many repetitive reentrant loops and a couple of bottleneck workstations.

<table>
<thead>
<tr>
<th>Table 1 Test Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of product</td>
</tr>
<tr>
<td>Process flows</td>
</tr>
<tr>
<td>Products</td>
</tr>
<tr>
<td>Workstation groups</td>
</tr>
<tr>
<td>Workstation</td>
</tr>
<tr>
<td>Steps</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Raw processing time</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Demand rate</td>
</tr>
<tr>
<td>(wafer/day)</td>
</tr>
<tr>
<td>Lot size (wafers)</td>
</tr>
</tbody>
</table>

In this experiment, we made the following assumptions to focus our investigative attentions to the basic properties of DISCS: (1) there are 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 through static capacity analysis.

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

(1) A release rule is defined as the constant release rule, which 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) For initial WIP level, WIP control factor $\alpha$ is set as 4.5. These rules and values are often used in practice of real semiconductor manufacturing process.

In the experiments, the simulation of 360 days is executed using 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.3 shows the transitions of the total WIP along the simulation time. From the graph of DISCS results, it is shown that the control system in DISCS smoothly maintains the target WIP level decided by the planning system periodically (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 whole 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 to reduce the WIP in more than 13%.

Figs.4 and 5 show the product inventories and backorders. In the graphs, positive numbers show the amount of product
inventories and negative numbers are 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. From the above results, it is shown that, compared with the traditional method, DISCS can reduce the WIP level as well as backorders.

Figs.6 and 7 show the distribution of average WIP levels among the workstations. The graphs have a marked difference in the WIP value of the workstation No.67. The workstation No.67 is the most significant bottleneck machine in the test problem, which has a high utilization above 93%. The DISCS system succeeds in reducing WIP by eliminating redundant WIP at the bottleneck machine. From these results, it is shown that the dispatching rule used in DISCS is effective for realizing stable and lean production in the face of complexity and variability of semiconductor manufacturing.

Figs.8 and 9 show the inventory and backorder at each step of Product 1. Each line is plotted every month. Again, positive values mean product inventories and negative values show backorders in the graphs.

In the traditional method, inventories exist in the product steps between 30 to 180. And, 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 values added at the previous steps, the traditional method causes higher inventory costs than the DISCS system.

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

5.3 Estimation of Economic Benefits

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

6. Conclusion

This paper shows that the hierarchical distributed planning and control methodology, DISCS, for large-scale and complicated manufacturing processes such as semiconductor fabrication. For manufacturing processes comprising of 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 lengthy lead times and increased obsolescence stocks. DISCS proposed in this paper optimizes WIP level of each step to minimize the inventory cost and the backorder cost, and it controls production sequences at each workstation distributively to keep optimized WIP level. The results of computational experiments show that DISCS can reduce the WIP level and backorders simultaneously.

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

References