Proceedings of the 2001 IEEE
International Conference on Robotics & Automation
Seoul, Korea · May 21-26, 2001

Dynamic Scheduling Rule Selection for Semiconductor Wafer Fabrication

Bo-Wei Hsieh
Dept. of Electrical Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
bwhsieh@ac.ee.ntu.edu.tw

Shi-Chung Chang
Dept. of Electrical Engineering &
Grad. Inst. of Industrial Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
scchang@cc.ee.ntu.edu.tw

Chun-Hung Chen
Dept. of Systems Engineering &
Operations Research
George Mason University
Fairfax, VA 22030
cchen9@gmu.edu

Abstract

In this paper, we exploit the speed of an ordinal optimization (OO)-based simulation tool designed by Hsieh et al. to investigate dynamic selection of scheduling rules for semiconductor wafer fabrication (fab). Although a scheduling rule is a combination of loading wafer release and dispatching rules, this paper specifically focuses on dispatching when significant amount of wafers-in-process (WIPs) are held due to engineering causes and when major machine failures occur. Four prominent dispatching rules combined with the wafer release policy of workload regulation constitute a basic set of rule options. The dispatching rule may be weekly selected based on fab states over a four-week horizon. A total of 256 rule options are then evaluated and ranked by the OO-based simulation tool under the performance index of mean cycle time and throughput rate. Results demonstrate the value of dynamic rule selection for uncertainty handling, the insightful selection of good rules and the needs for further research.

1. Introduction

. Major fab scheduling problems include how wafers should be released into a fab and how they should be dispatched among machines for processing. A popular practitioners' approach for scheduling the production in a fab is to select from the many empirical scheduling rules available for IC fabs [6]. To quickly select a good enough scheduling rule from a rule library, Hsieh et al. developed a fast simulation tool (Figure 1) based on the ordinal optimization (OO) and optimal computing budget allocation (OCBA) methods, which will be referred to as the OO-based method hereafter [2].

Operation objectives of a fab change dynamically as well as the machine and inventory states. To achieve competitive fab operations, such a dynamic nature intuitively may lead to the need for dynamic selection of a scheduling rule based on the changes of objectives and states. In addition to the finding of [5], experimental studies of static rule selection by Hsieh et al. have indicated that rule selections vary with factors of initial state, performance index (objective) and time horizon [2]. This motivates our further investigation about how the efficiency of the OO-based simulation may be exploited to facilitate dynamic rule selection.

Dynamic dispatching rule selection is essentially a stochastic optimal control problem. As a closed-loop solution is generally impossible to be obtained for stochastic optimal control of a complex system [1], we consider the open-loop feedback selection (OLFS) instead. At each decision point, OLFS uses whatever available system information to select a good rule for a coming period of time as if no further information will be received in the period. Although not truly dynamic, OLFS exploits feedback information and fast evaluation of rule options to select scheduling rules. In specific, OLFS is applied to selection of dispatching rules upon the occurrence of two significant uncertain events in fab operation: holding of a significant amount of WIPs due to engineering causes and failure of a major machine. The study exploits the speed of the OO-based simulation tool of [2] and adopts a 10-product, 60-step and 12-tool-group fab model, which is extended from the single-product model of Lu et al. [3]. The potential of the OO-based simulation for application to dynamic selection of dispatching rules is also assessed.

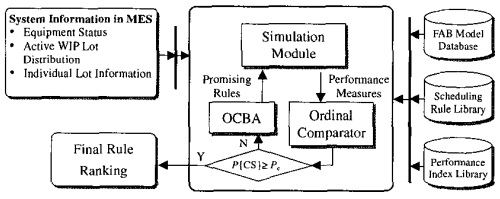

Figure 1 An OO-based Simulation Tool

The remainder of this paper is organized as follows. Section 2 describes the needs for dynamic selection of scheduling rules. The simulation model is described in Section 3. Dynamic selections of

dispatching rules under significant WIP holding and major machine failures are given in Sections 4 and 5 respectively. Section 6 concludes this paper.

2. Dynamic Selection of Scheduling Rules

Reentrant feature of the fabrication line and uncertainties of machines are two critical characteristics that make scheduling problems challenging. As the circuitry is fabricated layer by layer onto a wafer and basic processing steps among layers may be similar, the production flow of each type of product may re-visit the same type of machines, i.e., the same machine group, a few times. Wafers of different product types as well as those of the same type but processed at different layers may compete for the finite capacity of a machine group.

Among the uncertainty factors in fab production management, machine failures and temporary holding of WIPs from processing due to engineering causes are known as the two most prominent perturbations that lead to significant state and/or objective changes. When an engineering hold event occurs, a certain amount of WIPs is held from production until the engineering problem is cleared. Such holding results in a sudden reduction of available WIPs at each stage and may lead to a shortage of WIP for processing at the stage. If a stage in short of WIP requires the processing by a bottleneck machine, WIP holding may then cause bottleneck capacity loss, i.e., the output volume may decrease. When the held WIPs are released back to the production line, they are very often expedited to meet the due dates or the output volume target. When an unscheduled machine failure occurs, the machine group of the failed machine may become a short-term bottleneck. Its loss of capacity may also result in a lower fab output than the original target.

In practice, minimizing the mean cycle time while keeping the output volume per week or month above a target level is usually a fab operation objective. The latter, however, is really the bottom line performance requirement of a fab. Since the occurrence of either of the two aforementioned events may reduce the output volume, fab operation objective might be shifted from cycle time reduction to output volume maximization. In the event of a long period of engineering hold, one may want to adjust the wafer release policy. First, increase wafer release during the period of holding so that there is enough workload in the fab to keep a good utilization of fab capacity and the output volume. Then reduce wafer release after the held WIPs are back to production. And finally put wafer release back to a normal level. On the contrary, wafer release may first need to be reduced under a long time of machine failure to avoid unnecessary increase of WIP levels and cycle times, and then return to the normal level after the machine is repaired.

As for the selection of dispatching rule under a given wafer release policy, there are two intuitive and common strategies in response to the aforementioned two events. One is to feed proper amounts of WIPs to the bottleneck machines so that their available capacity is prevented from starvation. The other is to use machines in processing available WIPs that can be effectively moved to reduce total waiting times incurred by the holding or failure events. The selection is challenging because of the re-entrant nature, where a machine failure may affect the processing of several stages and the productions at one time become the future re-entrant flows to individual machine groups. What the proper amounts of WIPs are to prevent capacity from loss and how to effectively move WIPs and feed the bottleneck machines obviously depend on the state(s) of a re-entrant line.

3. Simulation Model Description

In this paper, a 10-product model (named FAB) extended from the single-product model of Lu et al. [3] is adopted. There are three types of processing technologies: *T1*, *T2* and *T3*, each having a specific sequence of processing stages. Among the ten product types, four product types use technology *T1*, three product types use *T2*, and the other three use *T3*. The model involves 12 failure-prone processing stations, each having one or more identical but independent machines. Wafers are moved among machines in the unit of lot, which consists of 24 wafers. Among the processing stations, Station 8 is modeled as a batch-processing machine group where each batch consists exactly of 6 lots. Processing times, times between failures and times to repair are exponentially distributed. The numbers of operation steps of *T1*, *T2* and *T3* are 60, 41 and 30 respectively. With release rates of 0.3 lots/hour for *T1*, 0.2 lots/hour for *T2* and 0.12 lots/hour for *T3*, the capacity bottleneck machine is Station 6 whose percentage utilization is 95.3%. Detailed model parameters of FAB are given in Table 1.

In our experimental study, two performance indices are considered: per circuit layer mean cycle time (LMCT) and total throughput rate, which are among the most frequently used fab performance indices. There are four prominent dispatching rules and a representative wafer release policy considered in this study as listed in Table 2. Workload regulation release policy proposed by Wein [7], FSVCT dispatching rules proposed by Lu et al. [3], and the OSA rule proposed by Li et al. [4] are known to be good for reducing mean and variance of cycle time. We designed the LDF rule for controlling production smoothness and for tracking production targets [8].

Table 1 Plant Data of FAB

Station	# of Machines	# of Visits (T1)	# of Visits (T2)	# of Visits (T3)	MPT[1]	MTBF[2]	MTTR[3]	% Util
1	4	14	10	8	0.500	150	5	92.7%
2	3	12	9	7	0.375	200	9	82.3%
3	10	7	5	4	2.500	200	5	91.9%
4	1	1	0	1	1.800	200	1	76.1%
5	1	2	1	1	0.900	200	1	83.3%
6	2	3	2	2	1.200	200	6	95.3%
7	1	1	1	0	1.800	200	1	90.5%
8	4	8	6	4	0.800	150	5	84.8%
9	1	3	0	0	1.000	200	5	92.4%
10	9	5	4	1	3.000	130	5	84.4%
11	2	3	2	1	1.200	200	5	87.6%
12	2	1	1	1	2.500	200	5	79.9%

[1] MPT: Mean Processing Time (by hours)
[2] MTBF: Mean Time between Failures (by hours)
[3] MTTR: Mean Time to Repair (by hours)

Table 2 Scheduling Rules

Rule	Symbol	Description
Release policy	WR(C_p)	In a one-bottleneck system, whenever the expected work of type-p products in fab drops below C_p hours for the bottleneck machine, then release a new type-p lot into the fab.
Dispatching rules	FSVCT	Choose the lot with smallest $(a_n + C_p - \zeta_i)$, where p represents the index of product type, a_n is the release time of lot n, C_p is the mean cycle time, and ζ_i is the estimate of the remaining cycle time from buffer i.
	LDF	Let the completion of one wafer processing at a stage be a move. Choose a stage with the largest deviation of completed moves from the desired moves, where the desired number of moves of each product type at each stage is pre-specified. Then choose from the stage a lot which is released into the fab the earliest.
	OSA	Choose a step according to the following priorities: Priority I: step i such that $N_i(t) > \overline{N}_i$ and $N_{i+1}(t) < \overline{N}_{i+1}$; Priority II: step i such that $N_i(t) < \overline{N}_i$ and $N_{i+1}(t) < \overline{N}_{i+1}$; Priority III: step i such that $N_i(t) > \overline{N}_i$ and $N_{i+1}(t) > \overline{N}_{i+1}$; Priority IV: step i such that $N_i(t) < \overline{N}_i$ and $N_{i+1}(t) > \overline{N}_{i+1}$, where $N_i(t)$ is the WIP at time t at step i, $\overline{N}_i$ is the average WIP at step i. Choose a lot with the same priority using FSVCT.
	FIFO	Select the lot which arrived at the station the earliest.

4. Rule Selection under Engineering Holds

Consider the operations of FAB for the coming four weeks. The fab has been operated under the scheduling rule of the workload regulation release policy combined with FSVCT dispatching rule (WR-FSVCT) for one year. Now suppose that over the whole line, half of the technology-*T1* WIP belongs to one customer order and that customer orders an engineering hold for one week. The WR wafer release policy remains unchanged due to its capability of regulating the workload of the production line. Under the WR policy, the workload for the bottleneck machine group, Station 6, is set at a level of 106 hours, which is the long-term average workload to achieve a throughput rate of 0.62 lots/hr. At this throughput rate, the utilization of Station 6 is 95.31%. Recall that the machine of Station 9, whose utilization is 92%, is only used in processing products of technology T1. A little calculation reveals that the engineering hold may lead to 12% capacity loss of Station 9 over the week of holding.

As has been discussed at the beginning of this section, proper amounts of available WIPs should be supplied in priority to both the bottleneck station and Station 9 to prevent their capacity from loss. When the held wafers are released back, the utilization of Station 9 needs to be raised to about 95% and Station 6 to 97% for the later three weeks in order to catch up with the delayed work of the first week. Since FSVCT determines lot priority based on individual lot information rather than station information, continual application of FSVCT as the dispatching rule does not match the needs when a significant holding/release event occurs. The question is then how dispatching rules should be dynamically selected so that the performance requirements for LMCT and/or throughput rate can be well achieved.

The dispatching rule library now consists of only four rules: FSVCT, FIFO, LDF, and OSA, which will be referred to as rules A, B, C, and D respectively. By means of OLFS, weekly change of dispatching rules is investigated. Over a four-week horizon, there are therefore 256 (4x4x4x4) options. The OO-based simulations are then conducted to find good combinations of the four dispatching rules from the 256 options.

Simulation results listed in Tables 3 and 4 clearly indicate that throughput close to 0.62 lots/hr can be achieved by many options, and that to minimize LMCT, the dispatching rule should be changed from WR-FSVCT to WR-LDF. The former observation is due to the fact that the capacity loss of Station 9 is only 12% in the first week and it can be made up by a higher utilization of Station 9 for the rest three weeks. The best rule listed in Table 4, A-C-D-D, achieves the maximum throughput rate, which clearly shows the transient of dispatching rule selection over a four-week horizon from the originally used rule A. In the latter observation, although not the best in throughput rate performance, the option C-C-C-C obtains a throughput rate of 0.618 lots/hr, which is only 0.418% lower than

that of rule option A-C-D-D. But The LMCT performance of C-C-C-C, 12.388 hour, is 10% shorter than the 13.690 hour of A-A-A-A. Computation time required for this rule selection experiment by using the OO-based simulation tool is about three hours. Our study shows that evaluation of the 256 rule combinations by using a regular simulation may take 150 to 300 hours of computation time, which is infeasible for such an application.

Table 3 Dynamic Rule Selection under Engineering Holds (ranked by LMCT)

Rank	Rule[4]	LMCT	%	Throughput
1	C-C-C-C	12.388	-	0.618
2	C-C-C-B	12.395	0.06%	0.616
3	C-C-C-D	12.492	0.84%	0.617
4	B-C-C-B	12.644	2.07%	0.602
5	B-C-C-D	12.653	2.14%	0.607
6	B-C-C-C	12.705	2.56%	0.600
7	A-C-C-C	12.733	2.78%	0.608
8	A-C-C-B	12.740	2.84%	0.607
9	D-C-C-B	12.759	2.99%	0.601
10	C-D-C-C	12.781	3.17%	0.608
131	D-D-D-D	13.628	10.01%	0.610
145	A-A-A-A	13.690	10.51%	0.612

[4] Rule-A: WR-FSVCT; Rule-B: WR-FIFO; Rule-C: WR-LDF; Rule-D: WR-OSA

Table 4 Dynamic Rule Selection under Engineering Holds (ranked by throughout)

Rank	Rule	Throughput	%	LMCT
1	A-C-D-D	0.621	-	13.514
2	A-A-B-A	0.621	0.00%	13.970
3	A-A-D-A	0.620	0.16%	13.995
4	A-C-C-A	0.619	0.32%	13.279
5	A-B-C-B	0.619	0.32%	14.043
6	A-B-B-D	0.619	0.32%	14.170
7	C-C-C-C	0.618	0.48%	12.388
8	D-D-C-C	0.618	0.48%	13.106
9	D-D-C-A	0.618	0.48%	13.605
10	D-B-A-D	0.618	0.48%	13.642
46	A-A-A-A	0.612	1.45%	13.690

It can be concluded from Table 3 that when engineering hold occurs, the switching from FSVCT to LDF rule leads to a superior LMCT while maintaining reasonable throughput rates. Conceptually, this is no surprise because under holding, the actual production of wafers of *T1* technology largely deviates from the desired targets over the whole line. LDF then gives a higher priority to available WIPs of *T1* technology than WIPs of other types. In so doing, available WIPs of *T1* move faster than ordinary to Station 9, which can reduce the capacity loss of Station 9 and the cycle times of available WIPs of *T1* technology. Such a gain compensates the LMCT increase for the held WIPs of *T1* technology. Similarly, at the bottleneck station, available WIPs of *T1* technology are given a higher priority of processing and re-enter the station faster. When the held WIPs are released, they are still given a higher priority and are expedited until the actual production of wafers of *T1* technology catches up its desired targets of individual steps. In contrast, the FSVCT rule prioritize WIPs of all types by their slack times; the held WIPs of *T1* technology get a high priority after being released because of resultant short slack times. But the available WIPs of *T1* technology does not get a higher priority during the holding period. So, when WIPs of *T1* technology are rushed to Station 9 in the later three weeks, congestion occurs and the FSVCT rule leads to a longer LMCT.

5. Rule Selection after Unusual Machine Failure

Again, consider the four-week operations of FAB model, where one machine of Station 9 goes into an unusual down situation at the beginning of a week and will need five days to be repaired. Although not frequently happened in the fab, this unusual event may lead to significant impact on fab performance. To complete the target 4-week workload of Station 9 in the remaining 23 days, the average utilization of Station 9 has to be more than 100%, which means that the capacity of Station 9 is not enough to complete the 4-week target after a 5-day failure. Station 9 then becomes a short-term capacity bottleneck instead of Station 6. Even 100% utilized for the remaining 23 days, Station 9 can only complete 89% of the target throughput which leads to at least 5% decrease in total throughput rate, i.e. the throughput rate of FAB over the four-week horizon is not greater than 95% of the target throughput. Under such a failure, WIPs belonging to technology *T1* cumulate at the steps processed by Station 9 and machines for processing downstream steps are starved. Due to the reentrant feature of IC fabrications, the shortage of *T1* products at downstream steps of Station 9 will propagate to the original bottleneck station (Station 6). Station 6 then allocates more resource to products of *T2* and *T3* technologies, which occupy 57% of capacity of Station 6 under normal states, to prevent capacity from loss and results in shorter LMCT of products of *T2* and *T3*, which are not processed by Station 9.

Under the failure of Station 9, throughput rate of the fab will decrease by a certain amount due to capacity loss of Station 9. For such an unusual event, the most important of all might be maintaining the throughput, which is again set as 0.62 lots/hour. Dispatching rule of LDF is supposed to be good in the aspects of maintaining throughput rate and reducing LMCT. LDF gives a higher priority to WIPs of *T1* technology after Station 9 is repaired to prevent available capacity of Station 9 from further loss and to reduce the LMCT of *T1* WIPs by expedition.

Table 5 Dynamic Rule Selection under Unusual Machine Failure (ranked by LMCT)

Rank	Rule[5]	LMCT	%	Throughput
1	C-C-C-C	14.527	-	0.553
2	C-C-C-D	14.702	1.20%	0.549
3	C-C-C-B	14.795	1.84%	0.546
4	B-C-C-C	14.833	2.11%	0.538
5	D-C-C-B	14.845	2.19%	0.543
6	D-C-C-C	14.850	2.22%	0.537
7	C-C-A-D	14.912	2.65%	0.546
8	B-C-C-D	14.919	2.70%	0.543
9	C-A-C-C	14.940	2.84%	0.543
10	A-C-C-C	14.966	3.02%	0.545
69	A-A-A-A	15.547	7.02%	0.545

5 Rule-A: WR-FSVCT; Rule-B: WR-FIFO; Rule-C: WR-LDF; Rule-D: WR-OSA

Table 6 Dynamic Rule Selection under Unusual Machine Failure (ranked by throughout)

Rank	Rule	Throughput	%	LMCT
1	A-B-A-A	0.555	-	15.889
2	D-B-D-D	0.555	0.00%	16.419
3	C-C-C-C	0.553	0.36%	14.527
4	C-B-D-D	0.553	0.36%	15.946
5	D-A-A-A	0.552	0.54%	15.163
6	D-A-D-A	0.552	0.54%	15.865
7	D-A-A-B	0.551	0.72%	15.243
8	C-D-A-B	0.551	0.72%	15.843
9	B-D-A-A	0.551	0.72%	15.936
10	B-D-D-A	0.551	0.72%	16.092
59	A-A-A-A	0.545	1.80%	15.547

Simulation experiments similar to those of Section 4 are conducted to select a good combination of dispatching rules. Only the initial states are different. The initial state of this experiment is obtained by running the FAB simulation with a machine of Station 9 set to be down and to be repaired after five days. Simulation results listed in Tables 5 and 6 indicate that throughput rate decreases by more than 10% under the unusual machine failure for all rules, and that to minimize LMCT, dispatching rule should be changed from WR-FSVCT to WR-LDF. It is our expectation that the throughput rate decreases due to capacity loss of Station 9. If the operation objective is to maximize throughput rate, rule option A-B-A-A results in the best throughput rate of 0.555 lots/hr, shown in Table 6. However, the throughput rates of the top-ranking rules are not significantly different in this machine failure case as well as in the previous engineering hold case. Such observations imply that throughput rate is insensitive to dispatching rules under the WR release policy. Since we only calculate $P\{CS\}$ for the top-ranking option, the relative ranking among other options does not really have a significant statistical support under the insensitivity of throughput. Namely, the differences between the top-10 rule options for maximizing throughput and the top-10 rule options for minimizing LMCT (Tables 3-6) are not so big as they appear to be.

In Table 5, the best selection for LMCT performance is rule option C-C-C-C, whose LMCT is 14.527 hours, 7% shorter than the 15.547 hours of A-A-A-A, and throughput rate is 0.553 lots/hr. Therefore, under a capacity loss situation, rule option C-C-C-C not only performs the best in LMCT performance but also obtains a good throughput rate. Dispatching rule should be changed from WR-FSVCT to WR-LDF for the coming four weeks when the failure event occurs. Under both engineering holds and unusual machine failure events, CCCC (WR-LDF) is the only option that commonly appears in the top-10 rule combinations across Tables 3-6. This reveals a strong appeal of CCCC for handling these two unusual events in the fab. Note that such a conclusion may not be applicable to other problems. The main objective of the study here is to demonstrate that our OO-based simulation tool can determine a good dispatching rule very efficiently when any unexpected event occurs.

6. Conclusions

In this paper, selections of dispatching rules at the occurrence of significant WIP holding and major machine failure were investigated. We exploited the speed of the OO-based simulation tool to select a good scheduling rule for the coming four weeks, where rule changes weekly. Simulations yielded insightful results that dispatching rule should be switched from the slack time-based FSVCT to the deviation-from-target-based LDF to handle these unusual events. These observations justified that dispatching rule should be changed dynamically to handle these unusual events. Simulation time saving up to 50 times can be achieved by the OO-based simulation. However, the number of rule combinations grows combinatorially over time and the number of tool groups. Further research on option search method exploiting the OO-based simulation is thus needed for large problems with combinatorial complexity.

Acknowledgement

This work was supported in part by the National Science Council of the Republic of China under Grants NSC88-2212-E-002-065 and NSC 89-2212-E-002-040, and by NSF Grant DMI-9732173 and DMI-0002900, Sandia National Laboratory Grant BD-0618, and the George Mason University Research Foundation.

References

[1] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*. New York: Academic Press, 1976.

[2] B. W. Hsieh, C. H. Chen, and S. C. Chang, "Fast Fab Scheduling Rule Selection by Ordinal Comparison-based Simulation," in *Proc. 1999 Int. Symposium*

Semicond. Manuf., pp.53-56, Oct. 1999.

[3] S. H. Lu, D. Ramaswamy, and P. R. Kumar, "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants," *IEEE Trans. Semicond. Manuf.*, vol. 7, no. 3, pp. 374-388, Aug. 1994.

[4] S. Li, T. Tang, and D. W. Collins, "Minimum Inventory Variability Schedule with Applications in Semiconductor Fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 9, no. 1, pp. 145-149, Feb. 1996.

[5] S. C. Park, N. Raman, and M. J. Shaw, "Adaptive Scheduling in Dynamic Flexible Manufacturing System: A Dynamic Rule Selection Approach," *IEEE Trans. Robot. Automat.*, vol. 13, no. 4, pp. 486-502, Aug. 1997.

[6] M. Thompson, "Using Simulation-Based Finite Capacity Planning and Scheduling Software to Improve Cycle Time in Front End Operations," in *Proc. 1995 IEEE/SEMI Advanced Semicond. Manuf. Conf. and Workshop*, pp. 131-135, 1995.

[7] L. M. Wein, "Scheduling Semiconductor Wafer Fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 1, no. 3, pp. 115-130, Aug. 1988.

[8] G. L. Wu, K. Wei, C. Y. Tsai, S. C. Chang, N. J. Wang, R. L. Tsai, and H. P. Liu, "TSS: a Daily Production Target Setting System for Foundry Fabs," in *Proc. 1998 Int. Symposium Semicond. Manuf.*, pp. 75-78, Oct. 1998.