Toward a real-time scheduler and controller for semiconductor fabrication systems

Hyeung-Sik Min and Albert T. Jones
Manufacturing Systems Integration Division
Manufacturing Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899
and
Yuehwern Yih
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907

ABSTRACT

The complexity of a semiconductor fabrication system is high because of complicated product flows, uncertain operation times, variable yields, changing products, and evolving technologies. Effective scheduling, which attempts (1) to predict accurately the start and completion times, and (2) to regulate the work-in-process inventory levels, can be quite challenging in this environment. The challenge grows when there are multiple performance objectives that vary over time. In this paper, we describe a competitive, neural-network-based approach that runs in real-time and at predetermined, fixed, time intervals. At each such interval, the network uses the current performance objectives and the current system status to generate a new schedule. We also describe briefly the simulation approach we used to train the network. Finally, we discuss our current efforts to include robustness computations in the decision-making and to use this approach as part of a real-time controller for the system.
Keywords: neural networks; real-time control; robustness; scheduling; wafer fabrication

1. INTRODUCTION

Scheduling of semiconductor wafer fabrication (fab) systems is complicated because of certain distinguishing characteristics such as re-entrant product flows, high uncertainties in operations, and rapidly changing products and technologies. It is thus a significant challenge to develop effective scheduling methods that can successfully control early and late completion of a task, work in process inventory, and frequent production changeovers.

In this paper, we describe a flexible scheduling tool for the fab operator, who can use it to react to various system changes in near real-time. This tool chooses rules to schedule the processing of wafer lots on machines and their movements, via automated material handling systems, into and out of temporary work-in-process storage facilities. Rule selection is complicated because the system configuration changes frequently and the system goal varies among two or more performance criteria.

To address this situation, we chose a competitive neural network, which has been shown to perform well when selecting decision rules for multiple decision variables to satisfy multiple objectives. In addition, a competitive neural network can classify all the information obtained from a simulation model and produce scheduling knowledge. Thus, it can help users extract decision rules for multiple decision variables to achieve the desired system objective in a real time.

2. ISSUES IN THE SCHEDULING OF WAFER FABRICATION

Of the characteristics mentioned above, re-entrant product flow (RPF) has the biggest impact on production planning and scheduling of wafer fabrication. RPF means that wafers at different stages of their fab life must compete with each other for the same machines. The result is that wafers tend to spend a larger amount of their fab life waiting for machines, rather than being processed.

This waiting contributes to long and unstable cycle time. Other contributing factors include long net processing times, system uncertainties, and delays caused by batch processing and machine setups. Net total processing times are fairly long because each wafer requires over 200 operations at a number of workstations.

System uncertainties include machine failures, process yield, and rework. Machines sometimes fail to operate within their design specifications. Duenyas *et al.* (1994) indicate the machine availability ranges from 60% to 75%. This disrupts the flow of materials and causes the cycle time to increase and fluctuate. Wafers are inspected several times in the process. Those that fail inspection are either removed or sent back to an earlier operation for rework. This increases both the mean and variance of the cycle time.

Batch sizes differ by wafer type and process. Therefore, one batch of wafers often waits for another

one to form the right batch size for the next operation. In addition, setup times are required when a machine changes a tool from one type of wafers to another or from one layer to another. Taken together, all of these conditions imply that cycle time has high variability.

Previous researchers have developed numerous fab-scheduling strategies. Two rule-based strategies have been widely used in both practice and academia: dispatching and input-control. Dispatching rules select the particular wafer lot to be scheduled whenever a processing machine becomes available. Input-control rules decide the type, the amount, the time, and the point-of-release whenever new wafer lots enter into the fab. Wein (1988) pointed out that input-control strategies can impact performance more than the dispatching policies. However, later Lu *et al.* (1994), Kumar (1994), and Li *et al.* (1996) showed that a good dispatching policy could also improve the performance.

Therefore, it is natural to hypothesize that significant improvements could be achieved by combining the two. However, there is still very little research focusing on combining them under various situations and finding their interactions and effects. Also, no one has shown that a single input-control strategy combined with a single dispatching strategy consistently dominates others in all situations. Therefore, we believe that it is more meaningful to identify a combination of policies that give good performance under a range of situations.

In addition, the main emphasis of much prior work on scheduling has been on static systems with a single objective at a time. As noted above, wafer fabrication is complex, dynamic, and highly stochastic. Satisfying the multiple objectives might be more important than only optimally meeting a single objective. In this study, we propose an effective approach based on a competitive neural network technique for multi-objective and multi-decision scheduling problems in semiconductor wafer fabrication.

3. SIMULATION MODEL AND SOLUTION METHODOLOGY

3.1 Simulation Model

The semiconductor fab is divided into a number of bays (aisles) that contain a number of similar or identical processing equipment. This configuration creates a large amount of material flow between bays – this is necessary since product flow is highly re-entrant. The transport operations are classified into inter-bay and intra-bay. Inter-bay lot transfers are carried out using an overhead monorail system and stockers. The overhead monorails transfer wafer lots using vehicles and are linked with automated stockers, which are furnished with a device for the lot exchanges with vehicles. The stockers guarantee the continuous control of lot positions and reduce the chance of wafer contamination. Operators within a bay perform intrabay lot transfers between the input-output port of a stocker and workstations for specific process steps or deposits of wafer lot to a stocker. Figure 1 shows the layout of the LG semiconductor fab.

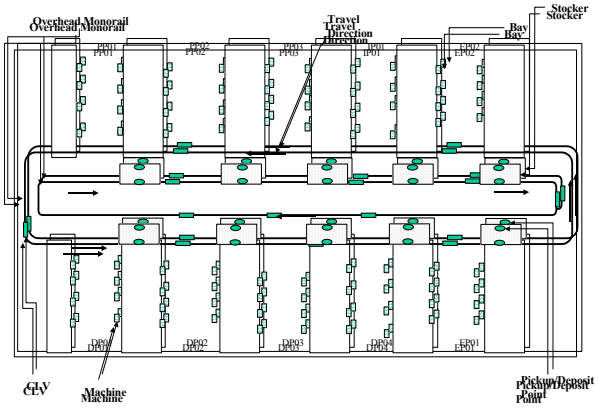

Figure 1. Semiconductor wafer fab model

There are 24 multiserver stations which consist of several identical machines in the simulation model of the semiconductor fab. We assume that all visits by all lots to a specific station have the same processing time distribution, and the lot size is 12 wafers per lot and is held constant through the study. The process flow of wafer lots is presented in table 1, where the number refers to the workstation. In table 1, each lot flows through the photolithography expose station (workstation 14) 12 times. Workstation 14 consists of GCA steppers and is considered as the extreme bottleneck, which is utilized much more than any other workstation. The operation of workstation 14 is referred to as a critical operation in this study. Our simulation model consists of two opposite directional overhead monorails and ten stockers for transportation and intermediate storage of wafer lots. Each monorail contains ten vehicles with a speed of .36 m/s for interbay lot transfers. A bay utilizes one or two stockers and also can share a stocker with an adjacent bay. The size of each stocker is varied and dependent on the workloads of corresponding bays.

Table 1. Process flow

Enter-1-2-13-14-23-15-20-22-23-22-23-22-17-13-14-15-23-16-24-23-22-17-1-8-4-22-22-1-2-8-13-14-18-23-15-16-23-18-22-1-1-13-14-23-15-16-24-23-22-17-1-2-8-9-21-22-1-4-22-22-1-2-13-14-23-15-16-24-24-23-22-17-24-1-2-7-1-3-22-13-15-23-22-22-22-17-13-14-18-23-15-16-20-23-1-17-1-1-3-13-14-16-24-23-22-17-9-21-1-3-13-14-15-23-15-16-24-23-22-17-1-3-10-22-12-6-22-6-1-1-4-10-19-23-1-10-13-14-16-21-12-13-14-18-23-15-15-15-16-19-23-22-17-11-13-14-15-21-23-5-Exit

3.2 Methodology

Our approach integrates a discrete-event simulation and a competitive neural network. This provides a basis for a multi-objective scheduler that controls the behavior of part flows to accomplish multiple objectives by generating the appropriate decision rules on the entire number of decision variables. Moreover, the fab

scheduler has the ability to respond in near real-time -- the scheduler suggests good decision rules in a few seconds whenever the fab manager or operator inputs the proper information. That information includes the detailed definition of decision variables, associated decision rules, and evaluation. These are described presently.

3.2.1 Decision variables and decision rules

Currently, there are five decision variables. Their definition associated rules are given as follows.

(1) Input control

Input-control decision variable determines the time and quantity of raw wafer lots to release into the wafer fab. The associated input-control rules release new lots into the fab whenever the workload level of critical machines falls below a given threshold, which changes over time. The workload level is determined by the sum of remaining processing times at the critical machines for all lots in the fab. The following associated rules determine the critical value.

1. SIWL (Small Increase in Workload): The threshold is set to a 5% increase of the current workload of the critical workstation.
2. SDWR (Small Decrease in Workload): The threshold is set to a 5% decrease of the current workload level of the critical workstation.
3. SWR (Same Workload): The threshold is set to the current workload level of the critical workstation.
4. LIWR (Large Increase in Workload): The threshold is set to a 10% increase of the current workload level of the critical workstation.
5. LDWR (Large Decrease in Workload): The threshold is set to a 10% decrease of the current workload level of the critical workstation.

(2) Selection of a critical machine (input buffer)

If an input buffer of a critical machine is empty and more than one wafer lot is waiting for the machine in stockers, the machine has to select which wafer lot to be processed next. The associated decision rules are

1. FCFS (First Come First Serve): A wafer lot that comes first is processed next.
2. SRPT (Shortest Remaining Processing Time): A wafer lot that has the shortest remaining processing time is processed next.
3. EDD (Earliest Due Date): A wafer lot that has the earliest due date is processed next.
4. CR (Critical Ratio): A wafer lot that has the smallest critical ratio is processed next. The critical ratio is calculated as follows:
   Critical Ratio = (Due date – Current time – Remaining processing time) / (Due date – Current time)

(3) Selection of a non-critical machine (input buffer)

If an input buffer of a non-critical machine is empty and there is more than one waiting wafer lot for the machine in stockers, the machine has to select one wafer lot to be processed next. The associated decision rules are the same as those of selection of wafer lots by critical machines.

(4) Selection of a wafer lot by a stocker

After finishing an operation in a bay or being transferred from another bay for the next operation, a wafer lot is temporarily stored at a corresponding stocker. However, if the corresponding stocker is full, a wafer lot has to wait until a storage position is available. When a stocker has an empty storage position and there is more than one waiting wafer lot, the stocker has to decide which wafer lot to store next.

1. FRFS (First Request First Serve): The stocker selects the wafer lot that requests it first.
2. IBF (In Bay First): The stocker selects a wafer lot that is waiting in a bay (output buffer of machines) for the stocker. This rule is to avoid the deadlock of a machine where holding an output buffer of a machine blocks the process of a next job. If multiple wafer lots are waiting for the stocker in a bay, FRFS is applied to select a lot among them.
3. LRS (Lowest Remaining Spaces In Stocker): The stocker selects a wafer lot in the other stocker that has the lowest remaining storage spaces. This rule is to balance the utilization of stockers. If multiple wafer lots are waiting for the stocker in the other stocker, FCFS is applied to select a lot among them.
4. EDD (Earliest Due Date): Same definition is used as in (2).
5. SRPT (Shortest Remaining Processing Time): Same definition is used as in (2).
6. CR (Critical Ratio): Same definition is used as in (2).

(5) Selection of a wafer lot by a vehicle on a monorail

When a vehicle on the monorail finishes its task and more than one wafer lot requests a vehicle, it has to decide which part will be transported next. The associated decision rules are as follows.

1. FRFS (First Request First Serve)
2. LRS (Lowest Remaining Spaces In Stocker)
3. EDD (Earliest Due Date)
4. SRPT (Shortest Remaining Processing Time)
5. CR (Critical Ratio)

3.2.2 Evaluation Criteria

In a semiconductor manufacturing fab, changing the scheduling rules in real time will have an effect on both system status and performance measure. There are multiple criteria that can be used in evaluating a given rule. These criteria are mainly based on completion times, due-dates, inventory level, and machine utilization. One of the measures we use is projected flow time. Projected flow time is defined as follows.

$$\text{Projected flow time} = \begin{cases} \left[\frac{(t_c - t_i)}{Rp}\right] \times Tp, & \text{if the wafer lot arrives before } t_i, \\ t_c - t_a, & \text{otherwise} \end{cases}$$

Where t_c is the completion time of a wafer lot
t_i is starting time of ith production interval
t_a is arrival time of a wafer lot
Tp is the total processing time of a wafer lot
Rp is the remaining processing time of a wafer lot at t_i

Table 2 shows the other evaluation criteria of this study. Decision rules for decision variables are based on system status in the beginning of each production interval and achieve the desired system performance at the end of each production interval.

Table 2 Evaluation criteria

System performance criteria	System status criteria
- Mean of projected flow time - Standard deviation of projected flow time - Number of tardy jobs	- Total work in process - Total workload of critical machines - Average number of remaining operations of each wafer lot - Mean slack time - Mean remaining processing time

3.2.3 Development of a scheduler

As noted above, we used a competitive neural network (CNN) to perform the scheduling. At the beginning of each production interval, we would like to use the CNN to tell us which decision rules to select for the five decision variables listed above. Furthermore, we would make this selection to optimize the current performance measures, a subset of measures listed in Table 2. To do this, we need to train the CNN.

We had no real training data, so we generated a set from a simulation of the target semiconductor fab. The simulation, which consisted of a lengthy sequence of short production intervals $t_1, \ldots, t_i, \ldots, t_n$, was designed to evaluate all combinations of decision rules against various combinations of performance measures. Given the current system status and performance measures at the end of interval t_{i-1}, a decision rule for each dispatching decision variable for the current production interval t_i was selected randomly. After interval t_i ends, the initial system status, the updated systems status, the updated performance measures, and the selected decision rules are recorded. This procedure continues until the simulation terminates. All of this data is fed into the CNN as an input vector for training.

Following data collection phase, the CNN classifies the simulation output data into instances. Each instance contains the initial system status, the decisions rules selected, and the performance outcomes. Figure 2 shows the framework of data. In stage 1, all instances with similar system status are assigned to the same class. In stage 2, each instance in a class of stage 1 is assigned to a subclass with similar performance measures from training the CNN-P. Therefore, a final class obtained from stage 2 consists of classified instances with similar system status and performance measures, together with the decision rules that generated those measures (Min, 2002).

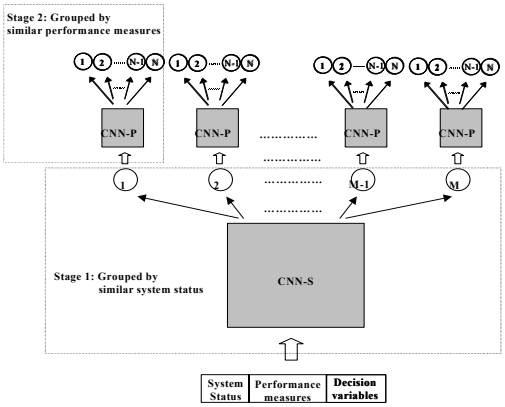

Figure 2. Framework of data classification

When the time comes for the operator to make a new schedule, another CNN is used to match the real systems status and the real performance measures to those instance classes generated from the training data. After this step, the scheduler can obtain the matching decision rules for decision variables for the next production interval.

4. EXPERIMENTAL RESULTS

4.1. Experimental Design

We compared our approach, which we labeled Method 1, against two other methods. Method 2 is controlled by the best rule chosen, based on five simulation runs conducted at the beginning of a production interval, among five randomly selected decision rules for decision variables. Method 3 is regulated by fixed decision rules for decision variables at the start of each production interval. The simulation of fixed decision rules for decision variables uses decision rules [31111]: SWR for input control, FCFS for selection of a wafer lot by a critical machine, FCFS for selection of wafer lots by a non-critical machine, FRFS for selection of a wafer lot by a stocker, and FRFS for selection of a wafer lot by a vehicle on a monorail. These fixed decision rules are used conventionally in a real semiconductor manufacturing industries.

To demonstrate the effectiveness of Method 1, we ran fifteen experimental simulation tests. In each test, all three methods were implemented with the same initial status conditions, desired performance measures, and operational parameters.

4.2. Result and Analysis

From each simulation run, we collect the difference between desired value and actual values for each method. Table 3 shows normalized values of each performance criterion between zero and one, using maximum and minimum values of each performance criterion. For example, for the mean of projected flow

time, the maximum value is 255 and minimum value was 5 among columns |M1-D|, |M2-D| and |M3-D. Then each normalized value is calculated as (p-5)/(255-5), where p is a value among columns |M1-D|, |M2-D| and |M3-D|. Similar calculations are performed for normalization of the other performance measures - standard deviation of projected flow time and number of tardy jobs. Using normalized values of each criterion of performance measure, we obtain overall performance values, which are the average of normalized values of three performance criteria for each method, that is, the three performance criteria are equally weighted. For example, for test *t1*, overall performance of |M1-D| is calculated as (0.008+0.0593+0.278)/3 = 0.115.

In table 3, the average overall performance of method 1 is improved by 45% of the average overall performance of method 2 and 77% of the average overall performance of method 3.

Table 3. Normalized difference between actual and performance measures

	Mean of projected flow time				Standard deviation of projected flow time				Number of tardy jobs				Applied decision rules		
	D	M1	M2	M3	D	M1	M2	M3	D	M1	M2	M3	M1	M2	M3
t1	1630	1637	1568	1529	82	75	135	85	80	132	64	36	44461	22424	31111
t2	1506	1542	1599	1657	102	153	102	107	95	75	99	169	54334	21113	31111
t3	1378	1349	1293	1535	125	138	149	83	63	43	7	24	13152	12242	31111
t4	1551	1562	1507	1621	115	141	120	126	4	4	52	137	12435	32243	31111
t5	1608	1582	1590	1679	112	92	65	120	0	7	10	180	33425	42445	31111
t6	1432	1469	1462	1657	80	161	198	108	25	31	37	193	33432	52321	31111
t7	1505	1446	1492	1665	191	192	196	122	17	14	4	204	31462	21133	31111
t8	1420	1387	1463	1529	101	91	76	85	11	6	12	36	32263	33414	31111
t9	1503	1535	1637	1511	154	132	75	89	49	51	132	31	13244	11331	31111
t10	1438	1415	1451	1401	130	139	102	97	15	21	9	5	14321	43144	31111
t11	1638	1643	1586	1580	88	73	105	89	75	127	53	41	21354	31254	31111
t12	1320	1335	1397	1465	120	126	113	68	17	3	7	4	24433	42143	31111
t13	1578	1530	1564	1602	78	67	96	91	3	8	54	82	52322	21133	31111
t14	1629	1660	1607	1755	134	139	101	114	39	23	59	97	24155	43153	31111
t15	1505	1461	1657	1449	101	86	89	71	3	0	9	0	44341	32425	31111

(D= desired values by an user, M1 = actual value (simulation output) by Method 1, M2 = actual value (simulation output) by Method 2, M3 = actual value (simulation output) by Method 3)

To compare the overall performance of method 1 with that of method 2 and that of method 3, we used Dunnet's one- tailed *t* test. The results indicate that the overall performance of method 1 is significantly superior to that of method 2 and to that of method 3 at the 99% confidence level (Min, 2002).

5. CONCLUSIONS AND FUTURE WORK

Preliminary experimental results strongly indicate that our approach can produce schedules that predict good performance. Our current research focuses on three questions:

1) How likely is it that this prediction will be achieved in the real fabrication system?

2) What impact will this schedule have on the rest of the manufacturing system?

3) What new capability is required to make this scheduler part of a real-time controller?

The first question is equivalent to determining the robustness of the generated schedule. The typical approach to robustness uses discrete event simulations to estimate the impact that random variables - including arrival times, service times, and random events such as priority jobs and machine breakdowns - has on the stated performance measure(s). We plan to augment these simulations with an analysis of the impact that underlying network structures, both physical and informational, have on the performance measure(s).

The second question is equivalent to determining the stability of the entire manufacturing system, should that schedule be implemented. Little prior research exists that provides insight in how to answer this question. We plan to develop an extensive system dynamics simulation that (1) incorporates variables and flows exogenous to the scheduling system, such as raw materials, work-in-process inventory, and production plans and (2) looks at the feedback loops created by scheduling outputs on those variables.

To be part of a real-time controller we must be able to execute the scheduler whenever a new job enters the system or the system performance degrades to an unacceptable level. Only minor modifications are required to integrate it with the production planning system, which controls the release of new jobs. As for performance degradation, Since the scheduler already incorporates the current system status, we need only include a monitoring function. This function will track the actual versus the predicted performance and determine when the difference exceeds a predetermined threshold.

6. REFERENCES

1. Duenyas, I., Fowler, J. W. and Schruben, L. W., "Planning and scheduling in Japanese semiconductor manufacturing," *Journal of Manufacturing Systems*, vol. 13, no. 5, 1994, pp. 323 – 332.
2. Kumar, P. R., "Scheduling semiconductor manufacturing plants," *IEEE Control Systems*, December, 1994, pp. 33 – 40.
3. Li, S., Tang, T., and Collins, D. W., 1996, "Minimum inventory variability schedule with application in semiconductor fabrication," *IEEE Transactions On Semiconductor Manufacturing*, vol. 9, no. 1, 1996, pp. 145 – 149.
4. Lu, S. C. H., Ramaswamy, D. and Kumar, P. R., "Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants," *IEEE Transactions On Semiconductor Manufacturing*, vol. 7, no. 3, 1994, pp. 374 - 388.
5. Min, Hyeung-Sik, "Development of a Real-time Multi-objective Scheduler for Semiconductor Fabrication Systems", PhD Thesis, Purdue University, July, 2002.
6. Wein, Lawrence M., "Scheduling semiconductor wafer fabrication," *IEEE Transactions On Semiconductor Manufacturing*, vol. 1, no. 3, 1988, pp. 115 – 130.