and the cumulative extent of over etch of WSix, poly2 and ONO OE has been normalized with for the various wafer regions. (a) Center. The curve for devices at the center, i.e., curve (a), to the right for the mediocre results, e.g., footings in poly2 owing to inadequate layer plasma etch and radiation damage, have been identified. A first-order relation between t_{ed} and the WSiX, poly2 and ONO OE time has been empirically derived to comprehend the interplay relation between them and to optimize the etch recipe. With optimization, t_{ed} has improved by over four-fold, from the range of 41–68 s (average ~50 s) using the baseline process conditions to the range of 200–210 s (average ~208 s). More significantly, the spread in t_{ed} has also been greatly minimized, from ~54% down to ~5%.

Fig. 3. Empirically derived first-order approximate relation between ONO breakdown time t_{ed} and the cumulative extent of over etch of WSiX, poly2 and ONO layers Σ (AOE)_{t=τ,p} for the various wafer regions. (a) Center. (b) Top/bottom. (c) Left/right. Σ(AOE)_{t=τ,p} has been normalized with respect to that at the center wafer region using Control processing conditions. The t_{ed} curve for devices in the top/bottom region is obtained by shifting the t_{ed} curve for devices at the center, i.e., curve (a), to the right by an additional extent of OE of (1/0.9737 × 1) = 2.7% to compensate for the lower etch rates there of approximately 0.9737[±1/3 × 0.975] for WSiX at 0.977 for poly2 + 0.960 for ONO] times that at the center, as shown by curve (b). Curve (c) is correspondingly derived for the left/right region. The t_{ed} ranges for Control and Process X are also shown.

for the mediocre t_{ed} results, e.g., footings in poly2 owing to inadequate layer plasma etch and radiation damage, have been identified. A first-order relation between t_{ed} and the WSiX, poly2 and ONO OE time has been empirically derived to comprehend the interplay relation between them and to optimize the etch recipe. With optimization, t_{ed} has improved by over four-fold, from the range of 41–68 s (average ~50 s) using the baseline process conditions to the range of 200–210 s (average ~208 s). More significantly, the spread in t_{ed} has also been greatly minimized, from ~54% down to ~5%.

REFERENCES


A Control Method to Reduce the Standard Deviation of Flow Time in Wafer Fabrication

Hyun Joong Yoon and Doo Yong Lee

Abstract—This paper proposes a control method for reducing flow time variability in a wafer fabrication facility with multiple wafer types. We employ stochastic petri nets to model and analyze the machine module, and define operation due dates using a novel utilization index metric. An operation due date (OPNDD) rule for lot dispatch is proposed and evaluated against other lot dispatch policies.

Index Terms—Flow time (cycle time), scheduling, stochastic petri nets (SPN), wafer fabrication.

I. INTRODUCTION

Semiconductor manufacturing has four basic steps: wafer fabrication, wafer probe, assembly, and final testing. Wafer fabrication is the most technologically complex and capital-intensive of the four steps [7]. The wafer fabrication is the process to form integrated circuits (IC) on the wafer. An integrated circuit is composed of several layers, and hundreds of operations are required to complete it. The wafers are fabricated in the facility called a workstation, and there are one or more identical parallel machines in each workstation. Due to the process characteristics and the complexity of process flow, it is difficult to control the wafer fabrication using the methods applicable to other manufacturing systems.

Related research on control schemes in the semiconductor manufacturing system has been reported in [2]–[4], [8]. In particular, Wein [8] simulates factory performance using data from a development wafer fabrication facility; mean flow time is evaluated under six different input regulation rules and 12 different sequencing rules are evaluated. Lu et al. [4] propose a fluctuation smoothing policy to reduce the mean and the variance of the flow time for single cassette type.

This paper proposes the operation due date (OPNDD) rule to reduce the standard deviation of the flow time in the wafer fabrication with multiple wafer types. To compute the operation due date, we define the utilization index that can be obtained by analyzing the stochastic petri nets (SPN) model of the machine module.

Manuscript received ; revised .

The authors are with the Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Taejon, Korea (e-mail: leedy@kaist.ac.kr; yunkii@discrete.kaist.ac.kr).

Publisher Item Identifier S 0894-6507(00)06372-7.
II. SPN MODELING AND UTILIZATION INDEX

For modeling and analysis of the machine module, we employ SPN, a modeling and analysis tool for discrete event systems. Because it is graphical in nature and capable of performance analysis, SPN is suitable for modeling and analysis of discrete-event systems such as manufacturing systems [1], [5].

Fig. 1 and Table I illustrate the SPN model of the machine module that is composed of four places and four transitions. If mean processing time (MPT), mean time between failure (MTBF), and mean time to repair (MTTR) are exponentially distributed, the transition rates of these events, \( \lambda_1, \lambda_2, \lambda_3, \) and \( \lambda_4 \) are equal to \( 2/\text{MPT}, 2/\text{MPT}, 1/\text{MTBF}, \) and \( 1/\text{MTTR} \), respectively.

Table II and Fig. 2 are the reachable marking and reachability graph of the SPN model shown in Fig. 1.

From the reachability graph, we can compute the transition rate matrix as follows:

\[
C = \begin{bmatrix}
-\lambda_1 & \lambda_1 & 0 \\
\lambda_2 & -\lambda_2 - \lambda_3 & \lambda_3 \\
0 & \lambda_4 & -\lambda_4
\end{bmatrix}
\]  

(1)

The probability of a particular marking in the steady state \( \pi_i \) \( (i = 0, 1, 2) \) can be computed using the following two equations:

\[
\begin{bmatrix}
\pi_0 & \pi_1 & \pi_2
\end{bmatrix}C = 0
\]  

(2)

\[
\pi_0 + \pi_1 + \pi_2 = 1.
\]  

(3)

From (2) and (3), we can compute the mean throughput rate of the machine module as follows:

\[
TR = \pi_1 \lambda_2 \frac{\lambda_1 \lambda_2 \lambda_4}{\lambda_2 \lambda_4 + \lambda_1 \lambda_4 + \lambda_1 \lambda_3}.
\]  

(4)

The utilization index is defined and can be obtained using the following procedure. Suppose that cassette type \( \tau \) \( (\tau = 1, 2, \ldots, T) \) is produced with a production ratio \( \mu_\tau (\mu_1 + \cdots + \mu_\tau + \cdots + \mu_T = 1) \). If a workstation \( s \) includes \( M^{\tau} \) machines whose throughput rate is \( TR^{\tau} \) and the number of visits of cassette type \( \tau \) to the workstation \( s \) is \( N^{\tau} \), the mean of the sojourn time per cassette in the workstation \( s \) is proportional to

\[
U^{\tau} = \frac{\sum_{\tau=1}^{T} \mu_\tau N^{\tau}}{M^{\tau} \cdot TR^{\tau}}.
\]  

(5)

If there are \( n \) workstations in the fabrication, the utilization index of the workstation \( s \) is defined as

\[
U^{\tau} = \frac{\sum_{s=1}^{n} U^{\tau}}{n}
\]  

(6)

which indicates the rate of congestion in the workstation \( s \) for processing a single cassette.

III. OPNDD RULE

To reduce the standard deviation of the flow time, we propose the OPNDD rule, where the operation due date is the due date assigned to each operation to satisfy the final due date. The OPNDD rule prioritizes cassettes according to operation due date, from earliest to latest. Assigning a higher priority to the cassette that is relatively delayed in its process can effectively reduce the standard deviation of the flow time. The system time, the sum of queuing time and processing time in a workstation, becomes longer as the throughput rate of the workstation becomes lower and/or the arrival rate of cassettes becomes higher. Also, because the arrival rate of cassettes is proportional to the number of visits to the workstation per cassette, we let the operation due date be proportional to the utilization index defined in the previous section. If we let the operation due date of the \( l \)th operation of the cassette \( \pi \) be \( \text{opndd}_{[\pi]}^{[\tau]} \), the difference between the \( l \)th and the \( (l-1) \)th operation due date is

\[
\Delta \text{opndd}_{[\pi]}^{[\tau]} = \text{opndd}_{[\pi]}^{[\tau]} - \text{opndd}_{[\pi]}^{[\tau]} = \alpha^{[\tau]} \cdot LT_{[\tau]}^{[\pi]}
\]  

(7)

where \( LT_{[\tau]}^{[\pi]} \) is the utilization index of the workstation where \( l \)th operation is to be processed and \( \alpha^{[\tau]} \) is a proportional constant. The sum for all the operations is equal to the lead time \( LT^{[\tau]} \) of the cassette type \( \tau \). Thus

\[
\sum_{l=1}^{\ell_{[\pi]}} \Delta \text{opndd}_{[\pi]}^{[\tau]} = \alpha^{[\tau]} \sum_{l=1}^{\ell_{[\pi]}} U^{[\tau]} = LT^{[\tau]}
\]  

(8)
where $I_{i}^{[r]}$ is the number of the total operations of the cassette type $r$.

The lead time can be determined by using the desired flow time or the due date of each cassette type. If it is not possible, merely the sum of processing time of all types of operations can be used as the lead time for the second best. Solving for (7) and (8), we obtain

$$\Delta_{opndd}^{[r]}[x]_{[i]} = \alpha^{[r]} \cdot U I_i^{[r]}$$

$$= \frac{LT^{[r]}_{i}}{I_i^{[r]}} \cdot U I_i^{[r]} \quad (l = 1, 2, \ldots, L^{[r]}).$$

(9)

Therefore, we can get the operation due date of the $l$th operation of the cassette $\pi$ as follows:

$$opndd^{[r]}[x]_{[i]} = opndd^{[r]}[x]_{[i-1]} + \Delta_{opndd}^{[r]}[x]_{[i]};$$

$$opndd^{[r]}[x]_{[0]} = r^{[r]};$$

(10)

where $r^{[r]}$ is the arrival time of the cassette $\pi$ of type $r$.

### IV. Simulation

The OPNDD rule is evaluated through simulation using the Hewlett-Packard Technology Research Center Silicon fab (TRC fab). Plant data for the TRC fab can be found in [8]. The TRC fab has three fab models: Fab1, Fab2, and Fab3. The three fab models are different in the number of machines in several workstations, which leads to a difference in the number of bottleneck workstations. The simulations have been performed for both single and multiple cassette types. In the simulation experiments for the single process, cassettes have 172 process steps. In the simulation for the multiple processes, it is assumed that cassette type A, B, and C are produced with production ratio of 0.2 : 0.3 : 0.5. The total number of operations of each cassette type are assumed to be 172, 154, and 139, respectively. Simulation with 10,000 produced cassettes each has been carried out five times each. The fabrication is assumed to be empty in the initial state. When a machine fails during the processing, the remaining work is processed after the machine is repaired. To eliminate the transient effect, the first 10% and the final 10% of the simulation are removed from the analysis.

The OPNDD rule is compared with fluctuation smoothing policy for variance of cycle time (FSVCT) [4] and five sequencing rules out of the 12 sequencing rules presented in [8], that are applicable to multiple process flows. However, since the FSVCT is developed only for the single process, the FSVCT cannot be included in the simulation results for the multiple processes. We compared the following sequencing rules.

- **OPNDD**: Select the cassette with the earliest operation due date.
- **FSVCT**: Select the cassette with the smallest $\alpha (r) - \zeta_i$, where $\alpha (r)$ is the release time of the cassette and $\zeta_i$ is the estimate of remaining flow time.
- **FIFO**: Select the cassette that arrives at the queue at the earliest time.
- **FIFO+**: If there are cassettes going next to a workstation which has a queue of size four or smaller, select among these using FIFO. If not, use FIFO.
- **SRPT**: Select the cassette that has the shortest expected remaining processing time until it exits the fab.
- **SRPT+**: If there are cassettes going next to a workstation which has a queue of size four or smaller, select among these using SRPT. If not, use FIFO.
- **LWNQM**: Select the cassette going next to the workstation with the least amount of expected work per machine.

Tables III and IV show the mean and the standard deviation of the flow time and the throughput rate for the three fab models under each sequencing rule. In the simulation of the OPNDD rule, CONWIP [6] is used as an input regulation rule. In the simulation of the other sequencing rules, the cassettes are assumed to be released into fabrication with uniform input rate (DETERMIN [8]), which shows better performance among the input regulation rules in the simulation performed by Wein [8].

For the single process, the simulation result for Fab1 shows that the standard deviation of the flow time is 64.47 under the OPNDD rule, 69.24 under the FSVCT, and 111.91 under the FIFO rule. The OPNDD rule reduces the standard deviation of the flow time by 6.89% compared to the FSVCT and 42.39% compared to the FIFO rule. In the simulations for the Fab2 and the Fab3, the OPNDD rule reduces the standard deviation of the flow time by 26.80% and 30.02% compared to the FSVCT and 54.47% and 52.55% compared to the FIFO rule, respectively.

To compare OPNDD with the FSVCT, one needs to fix the estimate of the remaining flow time ($\zeta_i$). In this simulation experiment, the value obtained from simulation has been used as recommended in [4]. Although the FSVCT effectively reduces the standard deviation of the flow time in most cases, it has a drawback of difficulty in selecting the estimate of the remaining time.

In the simulation experiments for the multiple processes, the simulation result for the Fab1 shows that the OPNDD rule reduces the standard deviation of the flow time of the cassette type A by 44.25%.

### TABLE III

**Simulation Results for Single Process**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Fab1</th>
<th>Fab2</th>
<th>Fab3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flow Time (hr)</td>
<td>Through. Rate (1/hr)</td>
<td>Flow Time (hr)</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Devn.</td>
<td>Mean</td>
</tr>
<tr>
<td>OPNDD</td>
<td>956.31</td>
<td>64.47</td>
<td>0.0234</td>
</tr>
<tr>
<td>FSVCT</td>
<td>953.51</td>
<td>69.24</td>
<td>0.0237</td>
</tr>
<tr>
<td>FIFO</td>
<td>972.69</td>
<td>111.91</td>
<td>0.0236</td>
</tr>
<tr>
<td>FIFO+</td>
<td>972.94</td>
<td>110.72</td>
<td>0.0236</td>
</tr>
<tr>
<td>SRPT</td>
<td>989.58</td>
<td>118.16</td>
<td>0.0236</td>
</tr>
<tr>
<td>SRPT+</td>
<td>984.76</td>
<td>117.00</td>
<td>0.0236</td>
</tr>
<tr>
<td>LWNQM</td>
<td>985.34</td>
<td>116.94</td>
<td>0.0236</td>
</tr>
</tbody>
</table>

**TABLE IV**

**Simulation Results for Multiple Processes**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Fab1</th>
<th>Fab2</th>
<th>Fab3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flow Time (hr)</td>
<td>Through. Rate (1/hr)</td>
<td>Flow Time (hr)</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Devn.</td>
<td>Mean</td>
</tr>
<tr>
<td>OPNDD</td>
<td>2516.72</td>
<td>144.01</td>
<td>0.0235</td>
</tr>
<tr>
<td>FSVCT</td>
<td>2123.91</td>
<td>160.47</td>
<td>0.0235</td>
</tr>
<tr>
<td>FIFO</td>
<td>2381.50</td>
<td>306.00</td>
<td>0.0235</td>
</tr>
<tr>
<td>FIFO+</td>
<td>2601.22</td>
<td>278.20</td>
<td>0.0235</td>
</tr>
<tr>
<td>SRPT</td>
<td>2196.35</td>
<td>502.61</td>
<td>0.0235</td>
</tr>
<tr>
<td>LWNQM</td>
<td>2196.35</td>
<td>502.61</td>
<td>0.0235</td>
</tr>
</tbody>
</table>
TABLE IV
SIMULATION RESULTS FOR MULTIPLE PROCESSES

<table>
<thead>
<tr>
<th>Rule</th>
<th>Wafer Type</th>
<th>FAB1</th>
<th>FAB2</th>
<th>FAB3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Flow Time (hr)</td>
<td>Through. Rate (1/hr)</td>
<td>Flow Time (hr)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Mean</td>
</tr>
<tr>
<td>OPNDD</td>
<td>TYPE 1</td>
<td>981.31</td>
<td>63.58</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>TYPE 2</td>
<td>876.53</td>
<td>58.17</td>
<td>0.0079</td>
</tr>
<tr>
<td></td>
<td>TYPE 3</td>
<td>785.00</td>
<td>57.93</td>
<td>0.0131</td>
</tr>
<tr>
<td>FIFO</td>
<td>TYPE 1</td>
<td>975.43</td>
<td>114.04</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>TYPE 2</td>
<td>903.01</td>
<td>108.25</td>
<td>0.0080</td>
</tr>
<tr>
<td></td>
<td>TYPE 3</td>
<td>811.25</td>
<td>101.18</td>
<td>0.0133</td>
</tr>
<tr>
<td>FIFO+</td>
<td>TYPE 1</td>
<td>1032.30</td>
<td>144.16</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>TYPE 2</td>
<td>947.57</td>
<td>138.28</td>
<td>0.0080</td>
</tr>
<tr>
<td></td>
<td>TYPE 3</td>
<td>783.06</td>
<td>109.79</td>
<td>0.0133</td>
</tr>
<tr>
<td>SRPT</td>
<td>TYPE 1</td>
<td>1079.90</td>
<td>178.24</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>TYPE 2</td>
<td>961.38</td>
<td>149.30</td>
<td>0.0080</td>
</tr>
<tr>
<td></td>
<td>TYPE 3</td>
<td>790.14</td>
<td>105.89</td>
<td>0.0133</td>
</tr>
<tr>
<td>SRPT+</td>
<td>TYPE 1</td>
<td>1070.83</td>
<td>169.85</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>TYPE 2</td>
<td>952.39</td>
<td>142.28</td>
<td>0.0080</td>
</tr>
<tr>
<td></td>
<td>TYPE 3</td>
<td>788.52</td>
<td>105.02</td>
<td>0.0133</td>
</tr>
<tr>
<td>LWNQ/M</td>
<td>TYPE 1</td>
<td>807.79</td>
<td>88.68</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>TYPE 2</td>
<td>802.38</td>
<td>95.15</td>
<td>0.0080</td>
</tr>
<tr>
<td></td>
<td>TYPE 3</td>
<td>941.56</td>
<td>195.91</td>
<td>0.0133</td>
</tr>
</tbody>
</table>

compared to the FIFO rule. With the cassette type B and C, the OPNDD rule reduces the standard deviation of the flow time by 46.26% and 47.25%, respectively. In the simulation for the FAB2, the OPNDD rule reduces the standard deviation of the flow time of the cassette type A, B, and C by 54.22%, 55.10%, and 47.30%, respectively, compared to the FIFO rule. In the simulation for the FAB3, the OPNDD rule reduces the standard deviation of the flow time of the cassette type A, B, and C, by 54.22%, 55.10%, and 47.30%, respectively, compared to the FIFO rule. In the simulation for the FAB3, the OPNDD rule reduces the standard deviation of the flow time of the three cassette types by 54.22%, 55.10%, and 47.30%, respectively.

In the simulation results of FAB3, SRPT, SRPT+, and LWNQ/M show poor performance for some cassette types. That is related to the fact that there is only one machine in most workstations except two workstations (workstations 13 and 14) in FAB3. Because there is no alternative machine in most workstations, machine failures severely affect the flow times of the cassettes. The sequencing rules of OPNDD, FIFO, and FIFO+ give higher priority to the cassettes that are delayed by machine failure, which is helpful to reduce the standard deviation of the flow time. On the other hand, SRPT, SRPT+, and LWNQ/M give higher priority to the cassettes that have shortest remaining time (SRPT and SRPT+) or are going to the workstation with the shortest queue for the next operation (LWNQ/M). They do not take the delayed cassettes into consideration. Therefore, the standard deviation of the flow time can become large under these sequencing rules.

V. CONCLUSION

The simulation result shows that, as a sequence control method, the OPNDD rule is effective in reducing the standard deviation of the flow time. A manufacturing supervisor can easily implement the OPNDD rule. If one can obtain data for MPT, MTBF, and MTTR, the utilization index of machines and the operation due date of cassettes can be computed easily using the equations (6) and (11). Then the OPNDD rule gives higher priority to the cassette with the earliest operation due date when selecting a cassette to be processed in an available machine in a workstation.

A proper selection of the input regulation rule is important for the better performance. From the simulation experiments, the CONWIP shows the best performance when it is applied with the OPNDD rule. For future research, input regulation rules to improve the performance of the OPNDD rule will be studied.

REFERENCES


