International Journal of Industrial Engineering, 15(1), 73-82, 2008.

A Hit-Rate Based Dispatching Rule For Semiconductor Manufacturing

Muh-Cherng Wu and Ting-Uao Hung

Department of Industrial Engineering and Management
National Chiao Tung University,
Hsin-Chu, Taiwan (ROC)

Hit-rate, the percentage of on-time completion, is a very important performance measure in a make-to-order semiconductor fab. This paper presents a dispatching algorithm for such a fab with *machine-dedication* feature. This feature imposes a constraint on the production route due to the advance of manufacturing technology, and has been rarely addressed in previous literature. A dispatching algorithm, called LBSA, was recently developed for a fab with machine-dedication feature. The LBSA algorithm outperformed many other dispatching methods in terms of hit-rate for short-routing products but not so well for long-routing products. This paper develops a dispatching method that shows a high hit-rate performance for both short-routing and long-routing products.

Significance: This paper presents a dispatching algorithm, which outperforms previous methods, in terms of hit-rate of a make-to-order semiconductor fab, with machine dedication feature.

Keywords: Dispatching; Semiconductor manufacturing; Machine-dedication; Line-balanced; Starvation avoidance.

(Received 27 July 2005; Accepted in revised form 19 March 2007)

1. INTRODUCTION

Semiconductor manufacturing is more complicated than most other production processes. The production of a semiconductor wafer requires about 500 operation steps, with reentry characteristics. That is, a wafer has to pass through a tool group several times because the tool group is assigned to perform several operations on the wafer. A wafer fab involves about 100 tool groups, each of which consists of several functionally identical machines and the failure of a machine is unpredictable (Uzsoy *et al*., 1992). Due to the complexity and uncertainty, the completion time of a wafer is quite unpredictable. For a make-to-order fab, this unpredictability consequently leads to volatile *hit-rate*—the percentage of on-time completion. How to develop effective shop floor control methods to improve the hit-rate is therefore very important.

Much literature on the shop floor control of semiconductor manufacturing has been published. These studies focus on two research problems: releasing and dispatching. The releasing problem is to investigate when and which wafer lot to release to the shop floor. Appropriate releasing methods would reduce the total WIP (work-in-process) of a fab and reduce the cycle time. Some of the popularly referred releasing methods involve uniform method, CONWIP (Spearman *et al.,* 1990), workload regulation (Wein, 1988), and starvation avoidance (Glassey and Resende, 1988). To enhance these job releasing methods, some other studies investigated how to determine the threshold WIP level for job releasing (Lin and Lee 2002), and developed methods to estimate cycle time for evaluating the future WIP level (Chung and Huang 1999; Juang et al. 1999; Chang and Hsieh 2003).

The dispatching problem is to determine which wafer lot to process first while a machine is available and a number of lots are waiting to be processed by the machine. Most literature on semiconductor dispatching determines the priority of a lot by considering the lot attributes. Such lot attributes involve the modeling and combination of waiting time, remaining time, and expected lateness. The *lot-attribute approach* deals with each lot individually, while a few other studies proposed a *line-balanced approach*. That is, a production line is partitioned into several segments and the dispatching policy is to smooth the flow rate of each segment and make the production line-balanced (Lee *et al.,* 2002, Wu *et al.,* 2004).

Due to the advance of semiconductor manufacturing technology, a phenomenon known as *machine-dedication* has been widely observed. As stated, a wafer lot has to re-enter a tool group several times with associated operations. For a tool group without machine-dedication characteristics, each of these associated operations can be freely assigned to any machine in the tool group. Conversely, for a tool group with machine-dedication feature, each of these associated operations can be only assigned to a particular machine in the tool group. A typical example with the machine-dedication characteristics is the stepper, a machine located in the photolithography area (Uzsoy *et al.,* 1992).

Previous studies on semiconductor dispatching have established significant milestones. Yet, at their time of investigation, the issue of machine-dedication has not been emphasized. A recent study, emphasizing the machine-dedication feature, proposed a dispatching algorithm (called LBSA) for a semiconductor fab to improve the

hit-rate. The LBSA algorithm has been justified to outperform many other algorithms by extensive simulation experiments for a particular logic product family (1P5M).

A logic semiconductor product is usually represented by 1PXM; for example, 1P5M, 1P6M, and 1P7M. The 'P' means poly-layer that is intended to manufacture transistors. The 'M' means metal-layer on which metals connecting transistors are to be formed. The higher the number of metal layers, the longer the manufacturing route, and the more versatile are the functions of the logic product. That is, newly developed logic products generally have more number of metal layers than existing products.

Further testing the performance of the LBSA algorithm by simulation, we find that the LBSA algorithm only performs well for short-routing products, and not so well for long-routing products such as 1P7M and 1P8M. This paper presents a dispatching method that aims to enhance the LBSA algorithm in order to achieve good performance in both short-routing and long-routing logic products. The proposed algorithm involves the alternative use of *line-balanced* and *lot-attribute* paradigms; which paradigm to use is subject to the present scenario. That is, the lot-attribute approach is applied while there are some lots substantially late in progress. Conversely, the LBSA algorithm is used when no such late lots exist.

The remainder of this paper is organized as follows. Section 2 reviews the literature on semiconductor dispatching. Section 3 presents the proposed dispatching algorithm. Section 4 compares the performance of the proposed algorithm with that of some representative methods in literature. Concluding remarks is presented in Section 5.

2. LITERATURE REVIEW

Much research on the dispatching problems in semiconductor manufacturing has been published. Most studies aim to identify a dispatching rule with good performance. PanWalkar and Iskander (1977) have published a survey paper that provides a list of more than 100 dispatching rules. Each of these dispatching rules was designed for various manufacturing systems and aimed to meet distinct objectives.

The dispatching rules can be classified into two approaches: lot-attribute and line-attribute. The lot-attribute approach prioritizes the dispatching of lots based on the attributes of a lot, such as arrival time, processing time, remaining time, remaining cycle time, due dates, and a combination of these attributes. Examples of the lot-attribute approach include (e.g. Blackstone *et al.,* 1982; Kim *et al.,* 1998; Kim *et al.,* 2001).

Of the lot-attribute studies, some are designed for reducing tardiness or improving on-time delivery. For example, the critical ratio (CR) rule, denoted by the ratio of remaining time divided by remaining processing time, is designed for speeding up the progress of late lots. Lu *et al.* (1994) proposed a modified least-slack rule, modeled by the remaining time subtracted by the remaining cycle time, in order to reduce the variance of tardiness. In terms of hit-rates, these two rules have shown good performance for some manufacturing systems.

The line-attribute approach is designed to prioritize the segments of a production line, where a segment denotes a sequence of operations. A production route in semiconductor manufacturing can be decomposed into a number of segments. Different segments may have to be processed by the same tool group, due to the reentry characteristic. Dabbas and Fowler (2003) proposed a method that first allocates daily capacity to each segment and then dispatches the lot in each segment in a real-time manner based on a combination of lot-attributes. Some other examples of the approach include (Lee *et al.,* 2002).

Yet, the scenarios addressed by most of the dispatching studies on semiconductor manufacturing do not involve the "machine-dedication" feature. The machine-dedication feature imposes a constraint on the production route, which appears recently due to the advance of manufacturing technology. Wu *et al.* (2004) proposed a LBSA (line-balanced and starvation avoidance) dispatching algorithm for a make-to-order fab with the machine-dedication feature. The LBSA algorithm shows a very good performance in terms of hit-rate for short routing products, but not so well for long-routing products. This paper aims to enhance the LBSA algorithm to make it perform well in both short-routing and long-routing products.

3. DISPATCHING ALGORITHMS

Machines in a fab can be classified into two types: series and batch. A series machine processes a wafer at a time until a lot of wafers are completed, while a batch machine (Neuts, 1967) processes several lots of wafers at a time. This research focuses on the dispatching of series machines, which are either with or without the *machine-dedication* feature. The dispatching algorithm for dedicated and non-dedicated machines are respectively described below.

3.1 Dispatching for Dedicated Machines

The dispatching for dedicated machine is developed based on two paradigms: (1) keeping the production line balance and (2) giving higher priority to the lots that tend to be urgently late.

The line-balanced paradigm models a production route by a number of segments. The route is so decomposed that each segment is ended with an operation processed by a dedicated machine. A dedicated machine, mostly referring to a high-resolution stepper, is very expensive and is usually the bottleneck. Due to the reentry characteristics, a dedicated machine has to process the WIPs located in many segments. Appropriately dispatching these WIPs could control the

throughput of each segment.

The idea of the line-balanced paradigm is keeping the throughput of each segment as uniform as possible. Higher throughput on a particular segment tends to output its WIPs earlier than expected. Conversely, lower throughput tends to delay the WIP progress. Consequently, non-uniform throughput would reduce the resulting hit- rate of the production line.

In this research, the throughput of a segment is measured by the formula: $\frac{WIP}{CT}$, where *WIP* denotes the total number of WIPs of the segment and *CT* denotes the cycle time of the segment. The WIPs in higher throughput segments should have higher priority in dispatching in order to lower their WIP levels so that the flow rate of each segment is smoothed. Suppose the highest-priority segment has many lots to be dispatched, CR (critical ratio) is subsequently used to prioritize them.

In summary, the line-balanced paradigm involves two-stage decisions: prioritizing segments followed by ranking the lots in the highest-priority segment. Such a lot prioritization approach may have a drawback. Lots that are substantially delay but located in a low-priority segment have little chance to remedy their progresses and may lead to the decrease of total hit-rate.

The second paradigm is proposed to overcome the drawback by defining an exception to the application of the line-balanced paradigm. The exception is that lots urgently late in progress, regardless of which segments they might stay, always have higher dispatching priority than all other lots. We prioritize these urgently late lots by using CR. While there is no urgently late lot, we apply only the line-balanced paradigm.

Detail steps of the dispatching algorithm are illustrated below. Consider a fab that has a number of dedicated machines and one of which is to be dispatched. The fab produces only one product family that involves many similar products. The route of each product is the same, with *s* segments, but slightly different in operation times. Let *n(i)* denote the number of WIPs in segment *i* and L_{ij} denotes the *j*-th lot where $1 \le j \le n(i)$. Define the processing time of L_{ij} by t_{ij} and its CR value by CR_{ij}. The average cycle time of segment *i* is represented by CT_i, which is obtained by simulation. A parameter γ is so defined that L_{ij} is considered as urgently late when $CR_{ij} \le \gamma$. The procedure for dispatching the dedicated machine, called *Dedicate_Dispatch*, is presented below.

Procedure *Dedicate_Dispatch*

Step1：Compute $CR_{ij}, 1 \le i \le s; 1 \le j \le n(i)$

Step2：Check if there are urgently late lots

Delay_Set = $\{L_{ij} \mid CR_{ij} \le \gamma,\ 1 \le i \le s; 1 \le j \le n(i)\}$

Step 3：Use CR to dispatch if there are urgently delay lots

If Delay_Set $\neq \phi$,

Then $(i^*, j^*) = ArgMin(CR_{ij})$ for all $L_{ij} \in$ Delay_Set

Go To Step 5

Step4：Use average flow rate to dispatch if there are no urgently delay lots

If Delay_Set = ϕ,

Compute the average flow rate of segment *i*, denoted by ν_i.

$$\nu_i = \frac{\sum_{j=1}^{n(i)} t_{ij}}{CT_i},\ 1 \le i \le s$$

Give highest priority to the segment that has maximum ν_i

$$i^* = ArgMax(\nu_i);\ \text{for}\ 1 \le i \le s$$

Prioritize the lots in segment i^*

$$j^* = ArgMin(CR_{i^*j})$$

Step 5: Output the lot $L_{i^*j^*}$, which is the lot to be dispatched. STOP.

3.2 Dispatching for Non-dedicated Machine

A starvation avoidance paradigm is proposed for the dispatching of non-dedicated machine. As stated, the ending operation of a segment is processed by a dedicated machine, which is bottleneck and critical to the fab throughput. Therefore, non-dedicated machine should be so dispatched to keep dedicated machines not starving.

A wafer lot waiting before a non-dedicated machine has three important attributes: (1) the dedicated machine to which the lot is assigned, (2) the segment where the lot stay, and (3) the estimated lateness of the lot. The starvation-avoidance dispatching paradigm first identifies which dedicated machine and at which segment tend to be most-starving. In the most-starving segment, if there are more than one lots to be dispatched, we use CR to prioritize them. The starvation-avoidance algorithm may have the same drawback as the line-balanced paradigm. That is, urgently delay lots may have no chance to remedy their progresses. To overcome this issue, the above methods for dealing with urgently delay lots are also applied here.

Detail steps of the dispatching algorithm for non-dedicated machines are illustrated below. Consider a fab, with *K* dedicated machines, which is having a non-dedicated machine to be dispatched. The fab produces only one product family, consisting of many similar products. The route of each product has *s* segments. The WIPs in segment *i* has *K* types; each type is assigned to a particular dedicated machine. Let the total number of WIPs in segement *i* be represented by $\sum_{k=1}^{K} n(i,k)$, where *n(i,k)* denote the number of lots in segment *i* that has been assigned to the dedicated machine *k*. Define L_{ijk} as the associated *j*-lot $1 \le j \le n(i,k)$. Define the processing time of L_{ijk} by t_{ijk}, the CR value of L_{ijk} by CR_{ijk}, and the mean cycle time of segment *i* by CT_i, which is obtained by simulation.

The algorithm for dispatching a non-dedicated machine is described below. In the presentation, we denote the non-dedicated machine by *Y*, its associated tool group by *T*, and the WIPs waiting before *T* by *WIP(T)*. For a segment *i*, the WIPs that has passed tool group *T* and is leaving for dedicated machine *k* is denoted by WIP(*i* ,*T*, *k*). The mean cycle time for processing WIP(*i*, *T*, *k*) is denoted by CT_{ik}, which is also obtained by simulation. Fig. 1 illustrates the above notations.

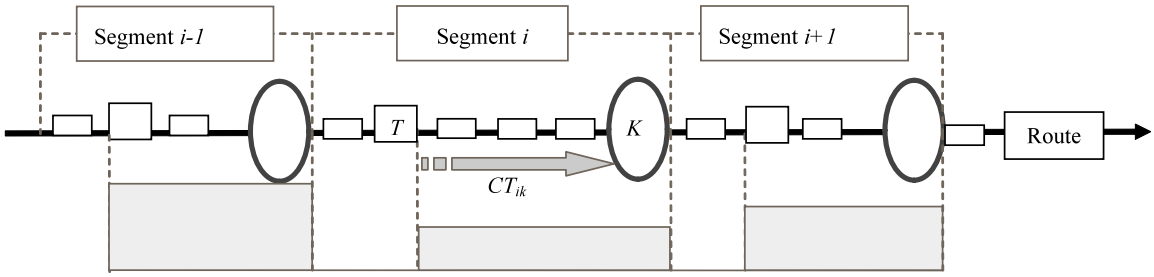

Figure 1 Dispatching for non-dedicated machine

Procedure *Non-dedicated_Dispatch(Y, T)*

Step 1：Check if there are urgently delay lots waiting before tool group *T*

Delay_Set = $\{L_{ijk} \mid CR_{ijk} \le \gamma, L_{ijk} \in WIP(T)\}$

Step 2：Dispatching for cases with urgently delay lots

If Delay_Set $\neq \phi$,

Then $(i^*, j^*, k^*) = ArgMin(CR_{ijk})$ for all $L_{ijk} \in$ Delay_Set

Go To Step 4

Step3：Dispatching for cases without urgently delay lots

If Delay_Set = ϕ,

Compute the average flow rate of the region between *T* and *k* in segment *i*

$$\nu_{ik} = \frac{\sum_{j \in S_{ik}} t_{ijk}}{CT_{ik}} \quad ; \quad S_{ik} = \{ j \mid L_{ijk} \in WIP(i,T,k) \}$$

Determine the highest priority segment

$$(i^*, k^*) = ArgMin(\nu_{ik}) \quad \text{for } 1 \le i \le s,\ 1 \le k \le K$$

Determine the highest priority lot

$$j^* = ArgMin(CR_{i^* j k^*})$$

Step 4: Output the highest priority lot $L_{i^* j^* k^*}$; STOP

3.3 Assumptions and Parameters of the Fab Scenario

The assumptions of the fab scenario are defined below. First, the uniform releasing policy is adopted in the fab; that is, each day a fixed amount of lots is released to the fab. Second, the due date of a lot is so defined: $D = R + Roundup(X \cdot PT)$, where D denotes the due date, PT denotes the total processing time of the lot, and X is a parameter manually given as follows. A simulation program is constructed for a fab that applies the FIFO (first-in-first out) policy in dispatching. The value of X is so given that the hit-rate of the fab is between 30%-75%. A higher value of X, surely leading to a higher hit-rate, yet provides less satisfied service to customers. So, the performance of dispatching is measured by the hit-rate at a particular X value.

The parameter CT_i, the mean cycle time of segment i, is determined by iteratively running a simulation program that includes the proposed dispatching algorithm. We firstly run the simulation by assuming the value of CT_i to be the total processing time in segment i. The simulation results will yield a new CT_i, which is subsequently used in the next simulation. The process is repeated until the newly-generated CT_i is close to the latest assumed one. The value of parameter CT_{ik} is set as that of CT_i. The value of parameter γ is constrained in [0, X], where we use the binary search method to find a γ that maximize the hit-rate.

3.4 Analysis of the Dispatching Method

The proposed dispatching algorithm becomes CR when γ is a very large value. In this case, all lots are regarded as urgently-delay lots; therefore only CR is used in dispatching lots. To the contrary, the proposed dispatching algorithm becomes LBSA when $\gamma = 0$. In summary, the proposed algorithm combines the effects of CR and LBSA, and the value of γ determines the effects of CR and LBSA. The higher the value of γ, the higher is the effect of CR.

4. SIMULATION EXPERIMENTS

By simulation, we compare the performance of the proposed algorithm with that of some representative algorithms in the literature. The main performance criterion is *hit-rate*—the percentage of on-time completion. Four other performance criteria are also measured in the comparison, which includes the mean and variance of cycle time as well as that of tardiness.

4.1 Simulation Scenario

A hypothetic fab is used in the simulation, where the process routes as well as its tool groups are provided by a fab in the real world. The fab includes 60 tool groups, of which 51 are series type and nine are batch type. Dedicated steppers are the bottleneck of the fab. The time between failure and the time to repair of each machine both follow exponential distributions. Setup time of each machine is included in the processing time.

The fab produces just a single product family, which consists of many similar products. These products are with the same process route but slightly different in the processing time. Such a difference in operation time is modeled by $pt \cdot Uniform(0.9, 1.1)$, where pt is the standard operation time and $Uniform(0.9, 1.1)$ is a uniform distribution.

The product family produced in the fab is of logic products. The complexity of a logic product is typically described by 1PXM, where 1P denotes one poly-silicon layer, and XM means X number of metal layers. The more complex a logic product, the higher is the number of X, and the longer is the process route. The simulation tests six product families, which are of 1P3M, 1P4M, 1P5M, 1P6M, 1P7M, and 1P8M. For each product family, Table 1 shows the number of operation steps and segments, the values of X and γ, and the number of lots that should be released per day.

Table 1 Route and relevant information of each product family

	1P3M	1P4M	1P5M	1P6M	1P7M	1P8M
Step number	276	310	344	378	412	446
Segment number	8	10	12	14	16	18
X value	1.50	1.40	1.50	1.40	1.70	1.90
γ value	1.30	1.30	1.40	1.30	1.60	1.90
Released lot per day	36	35	32	29	25	22

The simulation starts with no WIP and runs 270 days. The data of the last 180 days is collected. The performance of each testing scenario is measured by 15 replicates. The simulation program is coded with eM-plant, a commercially available software, and run on a personal computer with 3.0GHz CPU.

4.2 Comparison of Hit-Rate

Four dispatching algorithms are compared in the simulation. They are FIFO (first-in-first-out), CR (critical ratio), a cycle-time based slack method (Lu *et al.,* 1994) and LBSA (Wu *et al.,* 2004), which are selected due to the following reasons. FIFO and CR may be the most two popular methods used in make-to-order fabs. The slack method proposed by Lu *et al*. (1994) was declared to have very good performance in terms of tardiness for scenarios without machine-dedication feature. The LBSA method was claimed to perform very well for fabs with machine-dedication feature.

Table 2 shows the comparison in terms of hit-rate. The proposed algorithm outperforms the other methods in short-routing products (1P3M-1P6M), and ranks the second in long-routing products (1P7M-1P8M) where CR is the best one; however, the difference is statistically not significant. Referring to Appendix A, the P-value of Duncan test for 1P7M with regard to hit rate is 0.11 and 0.08 for 1P8M.

The proposed algorithm is better than the LBSA method in each of the six product families. Notice that the LBSA has a very good performance for short-routing products (1P3M-1P6M), but not so well for long-routing products (1P7M-1P8M), far worse than CR and the proposed algorithm. The proposed algorithm outperforms the LBSA by prioritizing the dispatching of urgently-delay lots.

The reason why CR outperforms the proposed algorithm for long-routing products is analyzed below. The proposed algorithm aims to achieve two targets: (1) keeping line balance and (2) speeding up urgently-delay lots. Keeping line balance is to maintain a constant level of throughput at each segment, which is relatively difficult to achieve as the number of segments increases. Consider the cases of producing 1P3M and 1P7M. Referring to Table 1, the process route of 1P3M has eight segments and that of 1P7M has 16 segments. Each segment of 1P3M requires $\frac{1}{8}$ of the stepper capacity to keep line balance; while that of 1P7M requires $\frac{1}{16}$. Therefore, 1% difference in the stepper capacity between any two segments imposes higher impacts on 1P7M than on 1P3M. Therefore, keeping line balance may not be so effective as CR in dealing with long-routing products, from the perspective of controlling hit-rate.

Table 2 Comparison of hit-rate (unit: %)

Methods	1P3M	1P4M	1P5M	1P6M	1P7M	1P8M
FIFO	66.4	52.1	75.3	67.7	46.2	36.4
CR	67.0	24.8	75.6	55.5	96.7	92.9
Lu-Slack	70.4	21.5	79.4	57.6	54.7	70.6
LBSA	87.2	80.8	88.94	80.6	70.8	54.6
Proposed	95.2	90.8	97.5	91.2	94.7	90.6

4.3 Comparison of Tardiness and Cycle Time

As shown in Tables 3 and 4, the ranking of the proposed algorithm in terms of tardiness is the same as that in terms of hit-rate. That is, the proposed algorithm ranks first in short-routing products (1P3M-1P6M), and ranks second in long-routing products where CR ranks first. However, the difference between the first two ranks is not statistically significant. The P-value of Duncan test for 1P7M with regard to mean tardiness is 0.78, and 0.35 with regard to variance tardiness (see Appendix A). The P-value of Duncan test for 1P8M with regard to mean tardiness is 0.71, and 0.23 with regard to variance tardiness.

Table 3 Comparison of mean tardiness (unit: hour)

Methods	1P3M	1P4M	1P5M	1P6M	1P7M	1P8M
FIFO	11.4	14.3	8.0	9.5	44.2	94.1
CR	7.9	18.7	5.9	10.7	0.8	1.7
Lu-Slack	7.3	19.3	5.2	10.3	20.9	14.2
LBSA	3.4	4.8	3.1	5.0	14.0	40.3
Proposed	1.1	2.2	0.6	2.1	1.3	2.2

Table 4 Comparison of variance tardiness (unit: hour)

Methods	1P3M	1P4M	1P5M	1P6M	1P7M	1P8M
FIFO	17.9	16.6	15.8	14.9	54.8	96.2
CR	11.3	11.1	10.1	11.9	4.2	5.7
Lu-Slack	11.1	10.4	9.9	11.8	27.2	24.9
LBSA	8.6	9.8	8.9	10.4	25.5	54.4
Proposed	4.8	6.8	3.3	6.6	5.5	7.2

Table 5 shows the mean cycle time of each dispatching method. The proposed algorithm outperforms the other methods in almost all products, except at 1P4M where the proposed algorithm ranks second. Table 6 shows the variance of cycle time of each dispatching method. The proposed algorithm ranks third in almost all products, except ranking second at 1P8M. The first two ranks, CR and the slack method, have very small variance of cycle times, which intuitively would lead to higher hit-rates; however, the effects are offset by their longer mean cycle times.

Table 5 Comparison of mean cycle time (unit: hour)

Methods	1P3M	1P4M	1P5M	1P6M	1P7M	1P8M
FIFO	471	504	559	568	781	970
CR	475	510	569	575	745	887
Lu-Slack	469	511	566	574	762	863
LBSA	461	493	552	564	732	897
Proposed	459	498	552	562	715	854

Table 6 Comparison of variance cycle time (unit: hour)

Methods	1P3M	1P4M	1P5M	1P6M	1P7M	1P8M
FIFO	24	18	26	20	76	124
CR	7	8	9	7	13	18
Lu-Slack	9	8	11	8	41	69
LBSA	15	12	18	14	57	96
Proposed	12	9	13	10	43	61

A semiconductor fab in Taiwan, which used the FIFO dispatching rule in an early period, has implemented the LBSA method and later on evolved to the implementation of the proposed dispatching method. However, in their implementations, the estimation of cycle time is based on historical data of the fab, rather than by performing discrete-event simulation. The other data for implementing the dispatching methods is obtainable from the MES (manufacturing execution system). In implementing the dispatching method, a database is developed for accessing from MES the data for dispatching in a near real-time manner. The fab reported that the proposed dispatching method indeed outperforms the LBSA method, and much better than FIFO. Notice that the fab essentially produces long-routing products because most newly developed products are more complicated in their functions and requires long-routing production. Short-routing products have been fading and

accounted for only a small percentage in the fab.

5. CONCLUDING REMARKS

This paper presents a dispatching algorithm for a make-to-order semiconductor fab with machine-dedication feature in order to improve hit-rate. Compared with some representative methods previously published, the algorithm shows highest hit-rate in almost all the tested products, ranging from 1P3M to 1P8M. This algorithm also performs very well in terms of some other performance criteria such as mean cycle time, mean tardiness, and variance of tardiness, but not so well in variance of cycle time.

The idea of this algorithm is two-fold: keeping line-balanced and speeding up urgently-delay lots. That is, the main target of dispatching is keeping the production line-balanced. However, the line-balanced target must be temporarily ignored in case there are urgently-delay lots in the shop floor. The algorithm involves the dispatching of dedicated and non-dedicated series machines. Dedicated machines, the bottleneck of a production line, are dispatched based on the notion of line-balanced. Non-dedicated machines are so dispatched to keep the dedicated machines not starved.

The proposed dispatching method requires the estimation of segment cycle time in advance by performing a discrete-event simulation. In case a well-modeled simulation program is not available, to facilitate the implementation, practitioners may estimate the cycle time by referring to the historical data of the fab.

This algorithm is developed for a fab that manufactures just a single product family, either short-routing or long-routing products. An extended research is being conducted in order to develop a dispatching algorithm for a fab producing both short-routing and long-routing products simultaneously.

ACKNOWLEDGEMENTS

This research is financially support by National Science Council, Taiwan (ROC), under contract number NSC-94-2213-E009-083.We are also grateful to Taiwan Semiconductor Manufacturing Corp. (TSMC) for providing the test data.

6. REFERENCES

1. Blackstone Jr, J. H., Phillips, D. T. and Hogg, G. L. (1982). A state-of-the-art survey of dispatching rules for manufacturing job shop operations. International Journal of Production Research, 20(1): 27-45.
2. Chang, P. C. and Hsieh, J. C. (2003). A neural networks approach for due-date assignment in a wafer fabrication factory. International Journal of Industrial Engineering, 10(1): 55-61.
3. Chung, S. H. and Huang, H. W. (1999) The block-based cycle time estimation algorithm for wafer fabrication facilities. International Journal of Industrial Engineering, 6(4): 307-316.
4. Dabbas, R. M. and Fowler, J. W. (2003). A new scheduling approach using combined dispatching criteria in wafer Fabs. IEEE Transactions on Semiconductor Manufacturing, 16(3): 501-510.
5. Glassey, C. R. and Resende, M. G. C. (1988). Closed-loop jop shop release control for VLSI circuit manufacturing. IEEE Transactions on Semiconductor Manufacturing, 1(1): 36-46.
6. Juang, J. Y., Li, Y. C., and Huang, H. P. (1999). Estimation of lot processing time in an IC fab. International Journal of Industrial Engineering, 6(4): 1999.
7. Kim, Y. D., Kim, J. U., Lim, S. K. and Jun, H. B. (1998). Due-date based scheduling and control policies in a multiproduct semiconductor wafer fabrication facility. IEEE Transactions on Semiconductor Manufacturing, 11(1): 155-164.
8. Kim, Y. D., Kim, J. G., Choi, B. and Kim, H. U. (2001). Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates. IEEE Transactions on Robotics and Automation, 17(5): 589-598.
9. Lee, Y. H., Park, J. and Kim, S. (2002). Experimental study on input and bottleneck scheduling for a semiconductor fabrication line. IIE Transactions, 34(2): 179-190.
10. Lin, Y. S. and Lee, C. E. (2002) A WIP estimation model for wafer fabrication. International Journal of Industrial Engineering, 9(3): 222-237.
11. Lu, S. C. H., Ramaswamy, D. and Kumar, P. R. (1994). Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. IEEE Transactions on Semiconductor Manufacturing, 7(3): 374-388.
12. Neuts, M. F. (1967). A general class of bulk queues with Poisson input. Ann. Math. Stat., 38: 759–770.
13. PanWalkar, S. S. and Iskander, W. (1977). A survey of scheduling rules. Operations Research, 25(1): 45-61.
14. Spearman, M. L., Woodruff, D. L. and Hopp, W. J. (1990). CONWIP: a pull alternative to kanban. International Journal of Production Research, 28(5): 879-894.

15. Uzsoy, R., Lee, C. Y. and Martin-Vega, L. A. (1992). A review of production planning and scheduling models in the semiconductor industry. part I: system characteristics, performance evaluation and production planning. IIE Transactions on Scheduling and Logistics, 24(4): 47-60.
16. Wein, L. M. (1988). Scheduling semiconductor wafer fabrication. IEEE Transactions on Semiconductor Manufacturing, 1(3): 115-130.
17. Wu, M. C., Huang, Y. L., Chang, Y. C. and Yang, K. F. (2004). Dispatching for semiconductor fabs with machine-dedication features. International of Advanced Manufacturing Technology, (accepted for publication, 2004).

APPENDIX A: THE DUNCAN TEST

Table 7. The P-value of Duncan test for 1P7M with regard to hit rate

Methods	FIFO	CR	Lu-Slack	LBSA	Proposed
FIFO		0.000031	0.000116	0.000056	0.000050
CR	0.000031		0.000050	0.000056	**0.112265**
Lu-Slack	0.000116	0.000050		0.000116	0.000056
LBSA	0.000056	0.000056	0.000116		0.000116
Proposed	0.000050	**0.112265**	0.000056	0.000116	

Table 8. The P-value of Duncan test for 1P7M with regard to mean tardiness

Methods	FIFO	CR	Lu-Slack	LBSA	Proposed
FIFO		0.000031	0.000116	0.000056	0.000050
CR	0.000031		0.000050	0.000056	**0.777358**
Lu-Slack	0.000116	0.000050		0.000303	0.000056
LBSA	0.000056	0.000056	0.000303		0.000116
Proposed	0.000050	**0.777358**	0.000056	0.000116	

Table 9. The P-value of Duncan test for 1P7M with regard to variance tardiness

Methods	FIFO	CR	Lu-Slack	LBSA	Proposed
FIFO		0.000031	0.000116	0.000056	0.000050
CR	0.000031		0.000050	0.000056	**0.353471**
Lu-Slack	0.000116	0.000050		**0.199810**	0.000056
LBSA	0.000056	0.000056	**0.199810**		0.000116
Proposed	0.000050	**0.353471**	0.000056	0.000116	

Table 10. The P-value of Duncan test for 1P8M with regard to hit rate

Methods	FIFO	CR	Lu-Slack	LBSA	Proposed
FIFO		0.000031	0.000056	0.000116	0.000050
CR	0.000031		0.000056	0.000050	**0.084203**
Lu-Slack	0.000056	0.000056		0.000116	0.000116
LBSA	0.000116	0.000050	0.000116		0.000056
Proposed	0.000050	**0.084203**	0.000116	0.000056	

Table 11. The P-value of Duncan test for 1P8M with regard to mean tardiness

Methods	FIFO	CR	Lu-Slack	LBSA	Proposed
FIFO		0.000031	0.000056	0.000116	0.000050
CR	0.000031		0.000056	0.000050	**0.709330**
Lu-Slack	0.000056	0.000056		0.000116	0.000116
LBSA	0.000116	0.000050	0.000116		0.000056
Proposed	0.000050	**0.709330**	0.000116	0.000056	

Table 12. The P-value of Duncan test for 1P8M with regard to variance tardiness

Methods	FIFO	CR	Lu-Slack	LBSA	Proposed
FIFO		0.000031	0.000056	0.000116	0.000050
CR	0.000031		0.000056	0.000050	**0.230996**
Lu-Slack	0.000056	0.000056		0.000116	0.000116
LBSA	0.000116	0.000050	0.000116		0.000056
Proposed	0.000050	**0.230996**	0.000116	0.000056	

BIOGRAPHICAL SKETCH

M-C Wu received his MS and PhD in Industrial Engineering from Purdue University. He also has an MBA from National Chen-Chi University. His areas of research include Computer Integrated Manufacturing, Manufacturing Management (in particular for semiconductor industry), Supply Chain Management, and Data Mining.