Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds

CYCLE TIME VARIANCE MINIMIZATION FOR WIP BALANCE APPROACHES IN WAFER FABS

Zhugen Zhou
Oliver Rose

Computer Science Department
University of Federal Armed Forces Munich
85577 Neubiberg, GERMANY

ABSTRACT

Although work-in-process (WIP) balance approaches can achieve average cycle time reduction, due to the characteristics of wafer fabrication facilities (wafer fabs), e.g., re-entrant flow, setup time and batch processing, a lack of effective mechanism to ensure lot movement at the right pace results in degraded cycle time variance, which might be a potential problem when due date is concerned. This paper attempts to solve this problem. Firstly four cycle time variance minimization rules which utilize waiting time, cycle time and due date information of lot are investigated. Then they are incorporated into two WIP balance approaches in literature to figure out whether they can overcome the drawback arising from WIP balance. In the end the benefit of cycle time variance minimization is illustrated by one example to address an improved ability to meet due date reliably.

1 INTRODUCTION

Work-in-process (WIP) is an important performance indicator in wafer fabrication facilities (wafer fabs). According to the Little's Law, a reduction of WIP leads to a production cycle time reduction given the same throughput, which has great economic benefit to the semiconductor manufacturer. WIP balance is a way to achieve average cycle time reduction by reducing WIP, for the reason that WIP balance regulates the workload of work-centers or operations to prevent overload and starvation, which directly reduces the WIP variability and smoothes the process flow thus reducing the queue time of lots (Li et al. 1996; Fowler et al. 2002; Zhou and Rose 2012).

WIP balance approaches have been well studied by academic and industrial researchers. In general, they can be classified into operation oriented and work-center oriented based on different viewpoints of WIP flow. Minimum Inventory Variability Scheduling (MIVS) (Li et al. 1996) and Line Balance algorithm (LB) (Dabbas and Fowler 2003) are the representatives of operation oriented approaches. Some researchers have different opinions and believe that managing WIP from viewpoint of work-center is beneficial (Ham and Fowler 2007, Zhou and Rose 2010). In our previous study (Zhou and Rose 2010), we proposed a WIP Control Table (WIPCT) to manage WIP for work-centers. Each upstream work-center maintains a WIPCT which contains current WIP information of downstream work-centers such as target WIP, actual WIP and WIP difference. In case lot moves in/out and machine status changes, the WIPCT is updated. Therefore, the upstream work-center is able to detect WIP distribution and measure the pull request of downstream work-centers dynamically. We also reported about that due to short of consideration of lot status like processing time, waiting time or due date information, the cycle time variance improvement is moderate by the WIPCT.

No matter which manner to achieve WIP balance is used, they both have the same objective to reduce average cycle time. However, as far as cycle time variance is concerned, the effects of both WIP balance

scenarios are modest. Because of re-entrant flow, setup time and batch processing, the lots chase after each other constantly, for instance, an early-arrival lot can be bypassed by a late-arrival lot because the late-arrival lot fulfills the setup or batch requirement. Even if WIP balance is applied, the lots go through the fab out of synchronization which represents as some lots are processed rapidly while some lots are delayed. As a consequence, the cycle time distribution spreads out causing a degraded variance. The WIP balance approaches in literature only focus on average cycle time reduction and throughput increase but seldom address the importance of cycle time variance minimization. Obviously, cycle time variance minimization is substantial as well. The reason is a low cycle time variance indicates a precise prediction of production completion time which allows more relaxed coordination with downstream operation on products like in assembly lines (Lu et al. 1994). In particular, this is critical to customer oriented companies since a low cycle time variance achieves a greater repeatability and quality to meet due date reliably. Thus, they are able to provide an accurate lead time commitment to customers.

Therefore, it gives us a potential research direction which is unsettled by MIVS and WIPCT. In this paper, we explore four rules aiming at minimizing cycle time variance. Then we combine MIVS and WIPCT with the proposed four rules, to figure out whether they can overcome the drawback arising from WIP balance when due date is introduced.

This paper is organized as follows. In Section 2, we introduce MIVS and WIPCT briefly. Then we describe the proposed cycle time variance minimization rules in detail in Section 3. Section 4 gives the simulation results and performance analysis. Section 5 is the conclusion and future work.

2 WIP BALANCE APPROACHES – MIVS AND WIPCT

2.1 Minimum Inventory Variability Scheduling (MIVS)

MIVS is a representative approach to balance the WIP of operations. It utilizes global WIP information about the current operation and downstream operation, and tries to keep the WIP of each operation close to the average target WIP level. Intuitively, the current operation with a high-than-average WIP level should have a higher priority than the one with low-than-average WIP level. In the meantime, the downstream operation with a low-than-average WIP level should have a higher priority than the one with high-than-average WIP level. Consequently, it leads to a combination of four different priorities described in Table 1.

Table 1: The principle of MIVS.

		Downstream Operation	
		Actual WIP < Target WIP	Actual WIP >= Target WIP
Current Operation	Actual WIP >= Target WIP	Priority 1	Priority 2
	Actual WIP < Target WIP	Priority 3	Priority 4

2.1 WIP Control Table (WIPCT)

The objectives of WIPCT are: (i) Evaluating the pull requests of downstream work-centers; (ii) Minimizing the deviation of actual WIP to target WIP of downstream work-centers. Each upstream work-center maintains a WIPCT which contains current WIP information of all its downstream work-centers e.g., tar-

get WIP level, actual WIP level, WIP difference and utilization. Suppose work-center 1 has three downstream work-centers 2, 3 and 4. Table 2 describes an example of WIPCT of work-center 1.

Table 2: An example of WIPCT.

Downstream Work-center	Target WIP (lots)	Actual WIP (lots)	WIP Difference (%)	Utilization (%)
2	12	6	-50	65
3	20	10	-50	80
4	8	12	50	70

Where:
Target WIP: The desired WIP level of the work-center has to be maintained. The target WIP levels used in MIVS and WIPCT are from the simulation model running with FIFO dispatching;
Actual WIP: The current WIP level of work-center including lots in queue and in process;
Utilization: Work-center utilization from lot release to current time;
WIP Difference: The deviation of actual WIP to target WIP, *(Actual WIP - Target WIP) / Target WIP*; The negative value means that the work-center is running out of WIP. The smaller the difference, the stronger pull request the work-center has;
The Actual WIP, WIP Difference and Utilization will be updated when lot moves in/out and machine status changes in work-center 1.

At time *t*, when a machine in work-center 1 is available for processing, work-center 1 checks the WIPCT. The downstream work-centers are ranked in descending order according to the *WIP difference*, the smaller the WIP difference is, the higher the rank. If the work-centers have the same WIP difference, the one with higher utilization has a higher rank. Based on this algorithm, in Table 2 work-center 3 has the strongest pull request, the next is work-center 2 and work-center 3 is the last. Accordingly, the lots in the queue of work-center 1 are divided into three priority categories. The lots heading towards work-center 3 obtain the highest priority, next priority level is for the lots for work-center 2 and the last one is for work-center 4.

3 RULES TO MINIMIZE CYCLE TIME VARIANCE

In both MIVS and WIPCT approaches, more than one lot might have the same priority. In this case, first-in-first-out (FIFO) is used to distinguish the urgency of the lots. Actually, it explains the reason why the lots go through the fab out of the right pace, since FIFO does not utilize any information about processing time, waiting time or due date etc. to process lots harmonically. In literature Hsieh et al. (2003) suggest to apply specified rules to better distinguish the urgency of the lots. If we consider two layers priority for hierarchical dispatching, the first layer priority tells us which lots fulfill the WIP balance requirement. While the second layer priority tells us the urgency (which lot is the optimal to optimize a certain target) among the lots selected from the first layer, when other performance indicators such as cycle time variance and on-time delivery have to be optimized. Therefore, it is our hypothesis that there will be potential room for improvement if we can utilize more information when replacing FIFO.

Before presenting the rules to minimize cycle time variance, we introduce the following notations:

P_i : Processing time of lot *i*;

Q_i : Queue time of lot *i*;

R_i : Release time of lot *i* to enter the fab;

F_i : Finish time of lot *i* to leave the fab;

D_i : Due date of lot *i*;

L_i : Lateness of lot *i*;

Now : Current time t;
CT_i : Cycle time of lot i when it finishes process and leaves the fab;
$AccCT_{i,t}$: Accumulated cycle time of lot i at time t (still in the fab);
RPT_i : Raw processing time of lot i;
$AccRPT_i$: Accumulated raw processing time of lot i;
$AvgCT_{i,j}$: Average cycle time of lot i and j if lot i is selected for processing ahead lot j;
$CTVar_{i,j}$: Cycle time variance of lot i and j if lot i is selected for processing ahead lot j;
$AvgCT_{j,i}$: Average cycle time of lot i and j if lot j is selected for processing ahead lot i;
$CTVar_{j,i}$: Cycle time variance of lot i and j if lot j is selected for processing ahead lot i.

3.1 Selection of the Lot with the Longest Queue Time or the Longest Queue Time plus Accumulated Cycle Time

Apparently, the queue time accounts for a large proportion of cycle time, particularly when the fab runs under a high loading case. It is not difficult to prove that choosing the lot with the longest queue time can minimize cycle time variance in a single machine case (Gupta et al. 2009). Assume there are two lots – lot i and lot j in the queue. At time t, a machine in work-center is available to process. If lot i is chosen for processing first, the average cycle time and cycle time variance are expressed as follows:

$$AvgCT_{i,j} = [(Q_i + P_i) + (Q_j + P_i + P_j)]/2\,, \tag{1}$$

$$CTVar_{i,j} = [(Q_i + P_i - AvgCT_{i,j})^2 + (Q_j + P_i + P_j - AvgCT_{i,j})^2]/2\,. \tag{2}$$

Similarly, if lot j is chosen for processing first, the average cycle time and cycle time variance are expressed as follows:

$$AvgCT_{j,i} = [(Q_j + P_j) + (Q_i + P_i + P_j)]/2\,, \tag{3}$$

$$CTVar_{j,i} = [(Q_j + P_j - AvgCT_{j,i})^2 + (Q_i + P_i + P_j - AvgCT_{j,i})^2]/2\,. \tag{4}$$

The target is to minimize cycle time variance of lot i and j. Thus, if lot i is chosen for processing ahead lot j, the following condition holds:

$$CTVar_{i,j} < CTVar_{j,i}\,. \tag{5}$$

We can deduce the following equation from Equation (5) by using Equations (1), (2), (3) and (4):

$$\begin{aligned}
&[(Q_i + P_i - AvgCT_{i,j})^2 + (Q_j + P_i + P_j - AvgCT_{i,j})^2]/2 \;<\\
&\qquad [(Q_j + P_j - AvgCT_{j,i})^2 + (Q_i + P_i + P_j - AvgCT_{j,i})^2]/2\\
&\Rightarrow [P_i + (Q_i - Q_j)]^2 > [P_j - (Q_i - Q_j)]^2\\
&\Rightarrow (P_i + P_j)(2Q_i - 2Q_j + P_i - P_j) > 0\,,
\end{aligned}$$

$$\because (P_i + P_j) > 0,$$
$$\therefore (2Q_i - 2Q_j + P_i - P_j) > 0$$
$$=> (P_i + 2Q_i) > (P_j + 2Q_j). \quad (6)$$

By extending the two lots case to *n* lots case in the queue and assuming the single machine has a high utilization, we can approximately derive the following:

$$\because Q_i >> P_i \,\&\&\, Q_j >> P_j,$$
$$\therefore Q_i > Q_j. \quad (7)$$

Equations (6) and (7) tell us that selection of the lot with the longest queue time can minimize cycle time variance for the single machine case. However, the wafer fab is a highly dynamic job shop containing hundreds of work-centers. When we extend the single machine case to the whole fab case, Equation (7) still holds if the queue time Q_i is extended to include two parts which are the real queue time that lot *i* spends in the current work-center and the accumulated cycle time that lot *i* spends in the fab. Thus, the following is derived from Equation (7):

$$Q_i + AccCT_{i,t} > Q_j + AccCT_{j,t}. \quad (8)$$

Equation (8) tells us that selection of the lot with the longest queue time plus accumulated cycle time leads to cycle time variance minimization to the whole fab case.

3.2 Due Date Oriented Dispatching Rule - Operation Due Date Rule

Due date oriented dispatching rules target at due date control to improve on-time delivery and minimize tardiness. Although the performances of due date rules are affected by the shop utilization and the tightness of target due date, due date rules inherently minimize the lateness variance. Lateness is defined as the difference between the finish time and the due date. Suppose lot *i* entering the fab is assigned a due date D_i and the finish time is F_i, thus, the lateness L_i is:

$$L_i = F_i - D_i. \quad (9)$$

The due date oriented rules – no matter whether Earliest Due Date (EDD), Least Slack Time (LST), Critical Ratio (CR) or Operation Due Date (ODD), they all attempt to minimize the variance of lateness by different but similar manners (Baker and Trietsch 2009). The due date plays the dominating role in setting a target which forces the lots to catch up with. As long as due date is established, the due date rules keep lots going through the fab as close to their due date as possible. This inherent advantage can directly lead to cycle time variance minimization, which turns out to be able to minimize cycle time variance. Suppose at time *t*, lot *i* is released into the fab and has a release time R_i. We simply define that the due date D_i of lot *i* is its release time R_i and assume lot *i* is tardy. The following follows from Equation (9) (Lu et al. 1994):

$$L_i = F_i - D_i = F_i - R_i = CT_i. \quad (10)$$

The lateness is the same as the cycle time in Equation (10). Consequently, from the above deduction the due date rules should lead to cycle time variance reduction.

We choose ODD rule in this study since ODD shows a powerful strength in minimizing cycle time variance in our previous study (Rose 2003, Zhou and Rose 2011). ODD rule breaks up the slack time into as many segments as the number of operations of a lot. It strictly keeps lots at the right pace to meet the operation due dates through the fab. ODD of lot *i* at operation *p* (*ODD(i,p)*) is defined as follows:

$$ODD(i,p) = R_i + RPT(p) \times DDFF \tag{11}$$

where *RPT(p)* denotes the raw processing time for a sequence of operations from operation *1* to operation *p* (including operation *p*). *DDFF* is due date flow factor and defined as the target cycle times divided by the raw processing time.

In this study, the lot with the smallest ODD obtains the highest priority.

3.3 Flow Factor Rule

There is a performance indicator called Flow Factor (FF) that is used to describe the relationship between the cycle time and the raw processing time. It is expressed as follows:

$$FF_i = CT_i / RPT_i . \tag{12}$$

Actually, the FF tells us how much time a lot spends in waiting, transporting, etc., besides raw processing time. Obviously, FF is expected to be minimized. Therefore, the FF is extended as a dispatching rule. At time *t*, a machine in work-center selects a lot via calculating the FF by accumulated cycle time divided by accumulated raw processing time, as described in Equation (13):

$$FF_i = AccCT_{i,t} / AccRPT_i . \tag{13}$$

At first glance, the FF does not seem to include due date information, but it becomes clear if we modify Equation (13) in the following way:

$$\begin{aligned} & AccCT_{i,t} = AccRPT_i \times FF_i \\ \Rightarrow\ & Now - R_i = AccRPT_i \times FF_i \\ \Rightarrow\ & Now = R_i + AccRPT_i \times FF_i . \end{aligned} \tag{14}$$

Equation (14) has a similar expression as the ODD rule, which indicates that the FF is expected to minimize cycle time variance as well, because it attempts keep the lots going through the fab with a given flow factor. As we mentioned above, the performance of ODD rule is influenced by the tightness of due date. Thus, we need to choose an appropriate target due date for the ODD rule carefully. The advantage of the FF rule is that it requires no effort to choose target due date but the effect of cycle time variance minimization is also expected. In contrast to the ODD rule, the lot with the largest FF is selected to minimize cycle time variance .

3.4 Simulation Model

The small wafer fab dataset MIMAC6 from Measurement and Improvement of MAnufacturing Capacities (MIMAC) is used to test our ideas. We refer the interested reader to Fowler and Robinson (1995) for details. MIMAC6 is a typical complex wafer fab model including:

- 9 products, 9 process flows, maximum 355 process steps. (Table 3 lists the basic information of the products.)
- 24 wafers in a lot. 2777 lots are released per year under fab loading of 100%.
- 104 tool groups (work-centers), 228 tools (machines). 46 single processing tool groups, 58 batching processing tool groups.
- Sequence dependent setup, rework, MTTR (mean time to repair), and MTBF (mean time between failures) of tool group.

Table 3: Raw processing time and release time of the products in MIMAC6 model.

Products	Raw Processing Time (days)	Time until Next Release (hours)
B5C	17.6	30.4762
B6HF	16.6	92.9782
C4PH	10.9	43.9225
C5F	15.1	36.4234
C5P	11.8	10.9271
C5PA	13.5	17.2316
C6N3	14.9	47.6584
C6N2	13.2	41.1018
OX2	12.8	35.2768
The due date of a lot is calculated by 'Release date + RPT*DDFF'; RPT is the raw processing time, DDFF is the due date flow factor.		

4 SIMULATION RESULTS AND PERFORMANCE ANALYSIS

The simulation length of MIMAC6 was set to 18 months. The first 6 months were considered as warm-up period, and not taken into account for statistics.

4.1 Incorporating Cycle Time Variance Minimization Rules into MIVS and WIPCT

As we mentioned above, we need to distinguish the urgency of lots that fall into the same priority for MIVS and WIPCT, with the objective to minimize the cycle time variance. Therefore, we incorporate the proposed four cycle time variance minimization rules into MIVS and WIPCT. If the lots obtain the same priority from MIVS or WIPCT, the proposed rules are applied to distinguish them. The average cycle time (Avg. CT), cycle time variance (CT Var.) and cycle time upper percentile 95% (CT Upper Pct. 95%) are considered as performance measures. The cycle time upper percentile 95% is the value below which 95% of the lots' cycle times fall. The fab loadings are divided into three levels which are 95% (high), 85% (medium) and 75% (low). Tables 4, 5 and 6 show the one year simulation results of the whole fab.

First of all, we take a close look at the results of 95% fab loading case in Table 4. Obviously, MIVS and WIPCT achieve average cycle time reduction compared to FIFO. The default rule to distinguish the lots for MIVS and WIPCT is FIFO, which leads to the absolute variance 1.8 and 4.2 for MIVS and WIPCT respectively, although the cycle time upper percentile 95% is 37.0 and 38.4 days respectively that is better than FIFO. When the proposed rules are incorporated into MIVS and WIPCT to better differentiate the lots, except for the Q rule, other three rules achieve promising performance. For the average cycle time, no matter for MIVS or WIPCT, the rules ODD and FF result in considerable average cycle time reduction that is more than one day in comparison with using FIFO as default rule. While the Q+Acc.CT rule shows limited improvement. With respect to the cycle time variance and cycle time upper percentile 95%, the ODD and FF rules absolutely dominate over the Q+Acc.CT rule. We observe that besides the Q

rule, the other three rules can minimize cycle time variance for certain. However, it is surprising that they have additional positive effect to achieve average cycle time reduction as well, in particular, the excellent improvement due to the ODD and FF rules. Actually, it can be explained by Lu et al. (1994) that reduction of the suddenness of lot arrival can reduce the delay in queue thus reducing cycle time. When the fab is running under a high loading, the ODD and FF rules exactly play the role to progress lots at the right pace to avoid fluctuation. Thus, they can diminish the suddenness of lot arrival. In this study, the ODD and FF rules are equally effective since there is no significant difference from the results.

Table 4: Three performance measures comparison between MIVS and WIPCT combining with cycle time minimization rules under 95% fab loading.

		95% Fab Loading		
		Avg. CT (days)	CT Var. (days^2)	CT Upper Pct. 95% (days)
FIFO		29.6	1.8	39.1
MIVS+	FIFO	28.7	1.8	37.0
	Q	28.5	2.9	37.0
	Q+Acc.CT	28.4	1.5	36.6
	ODD (DDFF 2.2)	27.4	0.4	35.2
	FF	26.9	0.8	34.6
WIPCT+	FIFO	28.9	4.2	37.9
	Q	28.8	10.2	38.9
	Q+Acc.CT	28.6	2.4	36.8
	ODD (DDFF 2.2)	27.1	1.1	34.6
	FF	27.3	1.3	35.1

Where:
FIFO: First-in-first-out; Q: Longest queue time; Q+Acc.CT: Longest queue time plus accumulated cycle time; ODD: Operation due date; DDFF: Due date flow factor; FF: Flow factor.
The same abbreviations are applied in Tables 4 and 5.

Table 5: Three performance measures comparison between MIVS and WIPCT combining with cycle time minimization rules under 85% fab loading.

		85% Fab Loading		
		Avg. CT (days)	CT Var. (days^2)	CT Upper Pct. 95% (days)
FIFO		23.3	1.2	29.8
MIVS+	FIFO	23.1	2.6	29.8
	Q	23.6	4.0	31.0
	Q+Acc.CT	23.4	2.0	28.8
	ODD (DDFF 2.0)	22.6	0.5	29.3
	FF	22.8	0.8	29.2
WIPCT+	FIFO	23.3	2.7	29.3
	Q	23.3	3.6	30.1
	Q+Acc.CT	23.5	2.3	28.8
	ODD (DDFF 2.0)	22.8	0.9	29.0
	FF	22.8	1.0	29.2

Table 6: Three performance measures comparison between MIVS and WIPCT combining with cycle time minimization rules under 75% fab loading.

		75% Fab Loading		
		Avg. CT (days)	CT Var. (days^2)	CT Upper Pct. 95% (days)
FIFO		20.5	0.9	26.2
MIVS+	FIFO	20.4	1.3	25.4
	Q	20.7	1.9	26.2
	Q+Acc.CT	20.8	1.0	24.7
	ODD (DDFF 1.8)	20.4	0.4	26.4
	FF	20.3	0.5	26.2
WIPCT+	FIFO	20.5	1.5	25.9
	Q	20.7	2.1	26.6
	Q+Acc.CT	20.6	1.2	25.5
	ODD (DDFF 1.8)	20.3	0.5	26.0
	FF	20.4	0.6	26.2

Secondly, we see similar performance but along with slight difference for the medium (85%) and low (75%) fab loading cases in Tables 5 and 6. The ODD and FF rules still achieve significant cycle time variance minimization as well as slight average cycle time reduction, whereas the rules Q and Q+Acc.CT degrade the average cycle time for MIVS and WIPCT. It is interesting to see that the Q+Acc.CT rule outperforms the ODD and FF rules with regard to cycle time upper percentile 95%. When the fab is running under low loading, the variability is not as serious as the high fab loading case. Hence, the fab is running in a smoother way which indicates that the positive effect of the ODD and FF rules are smaller than the high fab loading case. The Q+Acc.CT intends to choose the lot with the longest accumulated cycle time to process which turns out to finish the lots as fast as possible without strict pace of lot movement. Although it degrades the average cycle time, it manages to reduce the cycle time upper percentile 95% performance.

4.2 Benefit of Cycle Time Variance Minimization

In this section, we consider two more performance measures which are percent tardy lots (Pct. Tardy Lots) and average tardiness for tardy lots (Avg. Tardiness for Tardy Lots) to further understand the importance of cycle time variance minimization. To do this, we modify the second layer priority for the WIPCT by specifying a simple batch rule for the batch processing work-centers. If the lots have the same priority, the one that leads to the largest batch size is selected for processing. For the single lot processing work-centers, FIFO is utilized to distinguish the lots. It is represented as WIPCT+(Batch+FIFO) in Table 6. The cycle time variance is assumed to degrade by this rule since the batch rule results in serious overtaking movement of the lots. In order to solve this problem, the FIFO rule is replaced by the ODD rule which brings in the WIPCT+(Batch+ODD) in Table 7. The ODD rule is used for the single lot processing work-centers. For the batch processing work-centers, it remains the same as the WIPCT+(Batch+FIFO).

Firstly the Batch+FIFO rule is applied to the WIPCT, and the due date flow factor is set to 2.2. There is no doubt that the average cycle time is reduced since the batch size optimization leads to batches which are as full as possible to save capacity loss and speed up the lot movement. Apparently, the percent tardy lots and average tardiness performances are superior to the WIPCT+FIFO. However, there is the problem that the cycle time variance becomes large because batch optimization allows some lots to become tardy. If the customer requires to receive the products earlier, we have no choice but only to change the due date flow factor from 2.2 to 2.0. The problem arising from this change is, due to the degraded cycle time variance, the cycle time upper percentile 95% is not improved sufficiently. The percent tardy lots increases

from 25.5% to 78.7%, and the average tardiness of tardy lots increases from 0.6 to 1.8 days, if the Batch+FIFO rule is still utilized for the WIPCT. Because of the strength of the ODD rule, the Batch+ODD is applied to the WIPCT. The ODD rule overcomes the drawback arising from the Batch+FIFO rule perfectly by reducing the cycle time variance. Therefore, the percent tardy lots decreases from 78.7% to 20.6%, and average tardiness for tardy lots decreases from 1.8 to 0.5 days. It demonstrates that cycle time variance minimization allows an improved ability to minimize the tardiness and meet the due date reliably.

Table 7: Five performance measures comparison to figure out the benefit of cycle time variance minimization.

	95% Fab Loading				
	Avg. CT (days)	CT Var. (days^2)	CT Upper Pct. 95% (days)	Pct. Tardy Lots (%)	Avg. Tardiness for Tardy Lots (days)
FIFO (DDFF 2.2)	29.6	1.8	39.1	62.4	1.8
MIVS+FIFO (DDFF 2.2)	28.7	1.8	37.0	44.4	1.0
WIPCT+FIFO (DDFF 2.2)	28.9	4.2	37.9	48.3	1.1
WIPCT+(Batch+FIFO) (DDFF 2.2)	27.2	7.8	36.5	25.5	0.6
WIPCT+(Batch+FIFO) (DDFF 2.0)	27.2	7.8	36.5	78.7	1.8
WIPCT+(Batch+ODD) (DDFF 2.0)	26.8	1.8	34.2	20.6	0.5

Where:
FIFO: First-in-first-out; DDFF: Due date flow factor;
Batch+FIFO: For the batch processing work-centers, if more than one lot has the same priority, the one that can achieve the largest batch size obtains the highest priority;
For the single lot processing work-centers, if more than one lot has the same priority, the one that enters the queue first obtains the highest priority;
Batch+ODD: For the batch processing work-centers, if more than one lot has the same priority, the one that can achieve the largest batch size obtains the highest priority;
For the single lot processing work-centers, if more than one lot has the same priority, the one that has the smallest ODD obtains the highest priority;

5 CONCLUSION AND FUTURE WORK

In this paper, we explored four rules to minimize cycle time variance with the objective to solve the problem arising from WIP balancing for wafer fabs. There is no doubt that WIP balance leads to average cycle time reduction. However, due to the poor pace of lot movements, WIP balance allows some lots to accelerate while some lots to delay, thus degrading cycle time variance. The proposed four rules used to minimize cycle time variance were motivated by mathematical reasoning and validated by simulation. The simulation results demonstrated that except for the Q rule, other three rules overcome the drawbacks of WIP balance to different extents. We highlight the ODD rule among the proposed rules to address the importance that cycle time minimization can improve the ability to meet the due date reliably.

We investigated the proposed cycle time variance minimization rules under different fab capacity loading cases. The Q rule that is reported a good performance for single machine case in literature turns out to be ineffective for wafer fabs. This is the reason we proposed other three rules utilizing more infor-

mation other than queue time. For the high loading case, the ODD and FF rules dominate over the Q+Acc.CT rule and show their capability of keeping lots at a strict pace to minimize cycle time variance. It turns out that the significant variance minimization effect of the ODD and FF rules have a positive effect on reducing average cycle times as well. For the medium and low loading cases, the ODD and FF rules still outperform the Q+Acc.CT rule with regard to the variance performance, while the Q+Acc.CT is superior over the ODD and FF rules for the cycle time upper percentile 95%. There is no significant difference between the ODD and FF rules since they are both equally effective from the simulation results. For a customer oriented company, the ODD rule might be a good choice since the ODD rule strictly moves lots toward on-time completion by utilizing due date information, which is fairly comprehensive from the viewpoint of operational control. The FF rule provides another option since it does not involve any due date information while the ODD rule is affected by the tightness of due date.

The promising performances of the WIPCT combined with the ODD rule provide us with a new approach to deal with WIP balance and due date control. WIP balance does not always lead to good on-time delivery performance, and due date control does not primarily lead to low inventory. When short cycle times as well as good on-time delivery are desired simultaneously, the priority-based two-layer hierarchical dispatching turns out to be effective. The WIP balance in the top layer guarantees that the lots can balance the workload of work-centers or operations, the due date control in the bottom layer ensures that the we can choose the optimal lot among the lots fulfilling the WIP balance requirements to optimize cycle time variance. Consequently, both targets are taken into consideration, such that the lots progress smoothly without serious fluctuation to achieve WIP balance. Furthermore, the narrowed cycle time distributions ensures that as many lots as possible complete before reaching their due dates. For the future study, we need to investigate the reliability of the MIVS and WIPCT combining ODD rule because ODD rule is affected by the tightness of due date and the fab loading.

REFERENCES

Baker, K. R., and D. Trietsch. 2009. *Principle of Sequencing and Scheduling*. Wiley Publishing:John Wiley & Sons, Inc.

Dabbas, R. M., and J. W. Fowler. 2003. "A New Scheduling Approach Using Combined Dispatching Criteria in Wafer Fab." *IEEE Transactions on Semiconductor Manufacturing* 16:501-510.

Fowler, J. W., and J. Robinson. 1995. "Measurement and Improvement of Manufacturing Capacities (MIMAC): Final report." Technical Report 95062861A-TR, SEMATECH, Austin.

Fowler, J. W., G. L. Hogg, and S. J. Mason. 2002 "Workload Control in the Semiconductor Industry." *Production Planning & Control* 13:568-578.

Gupta, A. K., V. K. Ganesan, and A. I. Sivakumar. 2009. "Cycle Time Variance Minimization in Dynamic Scheduling of Single Machine Systems." *The International Journal of Advance Manufacturing Technology* 42:544-552.

Ham, M., and J. W. Fowler. 2007. "Balanced Machine Workload Dispatching Scheme for Wafer Fab." *Advanced Semiconductor Manufacturing Conference* 390-395.

Hsieh, B. W., S. C. Chang, C. H. Chen, and M. C. Chang. 2003. "Efficient Composition of Good Enough Dispatching Policies for Semiconductor Manufacturing." *IEEE International Symposium Semiconductor Manufacturing* 67-70.

Li, S., T. Tang, and D. W. Collins. 1996. "Minimum Inventory Variability Schedule with Applications in Semiconductor Fabrication." *IEEE Transactions on Semiconductor Manufacturing* 9:1-5.

Lu, S. C., D. Ramaswamy, and P. R. Kumar. 1994. "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-time in Semiconductor Manufacturing Plants." *IEEE Transactions on Semiconductor Manufacturing* 7:374-388.

Rose, O. 2003. "Comparison of Due-date Oriented Dispatch Rules in Semiconductor Manufacturing." In *Proceedings of the 2003 Industrial Engineering Research Conference* 18-20.

Zhou, Z., and O. Rose. 2010. "A Pull/Push Concept for Tool Group Workload Balance in a Wafer Fab." In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2512-2516. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhou, Z., and O. Rose. 2011. "A Composite Rule Combining Due Date Control and WIP Balance in a Wafer Fab." In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, 2085-2092. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhou, Z., and O. Rose. 2012. "WIP Control and Calibration in a Wafer Fab." In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, 177. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

ZHUGEN ZHOU is a PhD student at University of the Federal Armed Forces Munich, Germany. He is a member of the scientific staff of Prof. Dr. Oliver Rose at the Chair of Modeling and Simulation. He received his M.S. degree in Computational Engineering from Dresden University of Technology. His research interests include dispatching concepts for complex production facilities and work center modeling for wafer fab. His email address is zhugen.zhou@unibw.de.

OLIVER ROSE is the professor for Modeling and Simulation at the Department of Computer Science, University of the Federal Armed Forces Munich, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities, in particular, semiconductor factories. He is a member of IEEE, INFORMS Simulation Society, ASIM, and GI, and General Chair of WSC 2012.