Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

A BOTTLENECK DETECTION AND DYNAMIC DISPATCHING STRATEGY FOR SEMICONDUCTOR WAFER FABRICATION FACILITIES

Zhugen Zhou
Oliver Rose

Institute of Applied Computer Science
Dresden University of Technology
Dresden, 01187, GERMANY

ABSTRACT

According to the Theory of Constraints (TOC), the performance of a complex manufacturing system such as semiconductor wafer fabrication facility (wafer fab) is mainly determined by its bottleneck. In this paper, we investigate a bottleneck detection and corresponding dynamic dispatching strategy to regulate the workload of the bottleneck and non-bottleneck machines, in order to prevent bottleneck starvation and non-bottleneck with a high work in process (WIP) occurrence. The simulation results indicate that the proposed method achieves improvement with respect to average cycle time, cycle time variance and on time delivery compared with the classical dispatching strategies such as First In First Out (FIFO), Critical Ratio (CR).

1 INTRODUCTION

A typical wafer fabrication facility (wafer fab) contains hundreds of production equipments and dozens kinds of wafer products. Each kind of product has a unique technologic process flow which includes hundreds of processing steps. There are many characteristics of wafer fab, such as re-entrant processing flow, batch tools, sequence dependent setups, unpredictable equipment failure and so on, which differentiate wafer fab from other traditional flow shop or job shop. In general, release strategy and dispatching strategy are two major ways which are applied to control the wafer fab with the purpose of decreasing average cycle time and cycle time variance, achieving on time delivery of the products (Lu, Ramaswamy, and Kumar 1994).

The wafer fab has been intensively studied by academic and industrial researchers, especially applying dispatching strategies to it. Various research works focus on developing different dispatching rules, batching rules, or using combined dispatching rules with the objective of simultaneously optimizing multiple performance measures. The bottleneck approach, originated from Theory of Constraints (TOC), was successfully applied through the use of workload regulation (Wein 1988) and starvation avoidance methods (Glassey and Resende 1988). These methods aim to avoid the bottleneck starvation via carefully controlling lot release, and make sure that the bottleneck output is not intensely reduced due to the wafer fab uncertainty. However, these methods do not consider the working behavior of non-bottleneck. A high work in process (WIP) level in non-bottleneck has a possibility to make bottleneck starved.

Tsai et al. (Tsai, Feng, and Li 2003) proposed four categories of dispatching rules for the bottleneck and non-bottleneck machines. This method uses a pre-defined lower control limit and upper control limit of WIP to determine whether a bottleneck is in a starved state and a non-bottleneck is in a crowded state. The simulation results show that the hybrid dispatching rules achieve improvement of reducing average cycle time and cycle time variance, increasing throughput and decreasing WIP simultaneously. Whereas, defining the pre-defined lower control limit and upper control limit of WIP is not an easy job, and they are based on the capacity loading analysis and the experiences of production manager, which is not always supportable from industry.

In this research, based on Tsai et al.'s method we propose a simple way to define the pre-defined lower control limit to the bottleneck's WIP and upper control limit to the non-bottleneck's WIP. We first focus on identifying the bottleneck and ensure it is not starved, then intend to reduce a high WIP level of non-bottleneck back to a normal level. We compare the results between the proposed method and First In First Out (FIFO), Critical Ratio (CR) with regard to average cycle time, cycle time variance and on time delivery.

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

2 PROPOSED METHOD AND MODEL DESCRIPTION

According to the TOC, the performance of a complex manufacturing system such as the wafer fab is mainly determined by its bottleneck. Therefore, in order to improve its performance, the bottleneck behavior has to be improved. The TOC approach also emphasizes that the non-bottleneck should subordinate the bottleneck, which means the dispatching strategy of non-bottleneck should cooperate with the dispatching strategy of bottleneck. Two types of abnormal situations degrade the wafer fab's performance. The first one is a low WIP level in the bottleneck, another one is a high WIP level in the non-bottleneck (Tsai, Feng, and Li 2003). Based on these theories, we investigate two dynamic dispatching strategies for the bottleneck and non-bottleneck in order to prevent bottleneck starvation and non-bottleneck with a high WIP occurrence.

2.1 Bottleneck Detection and Dynamic Dispatching Strategy

Before describing the bottleneck detection and the dynamic dispatching strategy, some notations are introduced firstly.

- i : lot number.
- j : machine number.
- m : the numbers of machines in wafer fab or model.
- T : past T hours.
- P_i : the priority of lot i.
- Due_i : due date of lot i.
- RPT_i : remain processing time of lot i.
- Now : current time.
- $WT_j(T)$: working time of machine j during the past T hours.
- $OT_j(T)$: offline time of machine j during the past T hours.
- AW_b : the actual WIP level of bottleneck machine.
- DW_b : the predefined Lowest Limited Value of WIP of bottleneck machine.
- AW_{nb} : the actual WIP level of non-bottleneck machine.
- DW_{nb} : the predefined Highest Limited Value of WIP of non-bottleneck machine.

2.1.1 Bottleneck Detection

The utilization method is adopted to detect the bottleneck in this research. We measure the utilizations of different machines. The machine with the highest utilization is considered as the bottleneck. As both working time and offline time of the bottleneck can constrain the wafer fab's performance, the bottleneck is determined as follows.

$$Bottleneck = Machine_j = \max_{1\leq j\leq m}\left(\frac{WT_j(T)+OT_j(T)}{T}\right). \quad (1)$$

Different T is specified to different model in this paper. It is assumed that T equals 8 (one work shift length), 16, 24 or 48 hours respectively. For instance, we test these 4 bottleneck detection periods on MIMAC6 which is a model used in this study case. The experiments tell us that T equals 24 hours could reach a better performance results regarding average cycle time and cycle time variance.

2.1.2 Default Priorities of lots

The default priority of each lot for each step is 1 in the proposed method. If lots have the same priority, the one that enters queue first is favored. If lots have different priority, the one with lower value is chosen for processing. If the priority of lot i is changed in the kth processing step, it will be re-changed to 1 automatically in the $(k+1)$th step.

2.1.3 Dispatching Strategy for Bottleneck Machine

The dispatching objective for the bottleneck machine is to make sure it is not in a starvation state. Therefore, once the WIP level lower than the pre-defined Lowest Limited Value (LLV) takes place in the bottleneck, the upstream machines of bottleneck must be identified first. Then the lots, which are queued in the upstream machines and will be processed next in the bottleneck, will be reset with higher priorities, which means those lots should be sent to the bottleneck as fast as possible. Therefore, the WIP level of bottleneck can be maintained to a normal level. The lots' priorities are calculated as follows.

If $AW_b < DW_b$, then for the lots in the upstream machines of bottleneck, if the next step of lot i will be operated by the bottleneck, the priority of lot i is:

$$P_i = \begin{cases} \frac{1+Due_i-Now}{1+RPT_i} - \frac{DW_b-AW_b}{DW_b} & \text{if } Due_i>\text{Now} \\ \frac{1}{((1+Now-Due_i)\times(1+RPT_i))} - \frac{DW_b-AW_b}{DW_b} & \text{otherwise} \end{cases}. \tag{2}$$

Otherwise for all other lots, just retain their original priority of 1. We will explain the reason why this formula is used in section 2.1.4

2.1.4 Dispatching Strategy for Non-bottleneck Machine

The dispatching objective for the non-bottleneck machine is to prevent a high WIP level occurrence. Therefore, once WIP level higher than the pre-defined Highest Limited Value (HLV) occurs in the non-bottleneck, the upstream machines of non-bottleneck must be identified first. Then the lots, which are queued in the upstream machines and will be processed next in the non-bottleneck, will be given lower priorities, which means those lots should not be sent to the non-bottleneck until its high WIP level is reduced to a normal level. The lots' priorities are calculated as follows.

If $AW_{nb} > DW_{nb}$, then for lots in the upstream machines of non-bottleneck, if lot i will be processed next in the non-bottleneck, the priority of lot i is:

$$P_i = \begin{cases} \frac{1+Due_i-Now}{1+RPT_i} + \frac{AW_{nb}-DW_{nb}}{AW_b} & \text{if } Due_i>\text{Now} \\ \frac{1}{((1+Now-Due_i)\times(1+RPT_i))} + \frac{AW_{nb}-DW_{nb}}{AW_b} & \text{otherwise} \end{cases}. \tag{3}$$

Otherwise for all other lots, just retain their original priority of 1.

Two factors are considered to use formulas (2) and (3). Firstly, We want to introduce due date flow factor (FF) in the proposed method to compare it with due date oriented method such as CR and Earliest Due Date (EDD). Thus, CR is introduced as a part of the formula. Secondly, Changing lots' priorities should be according to the WIP level of the bottleneck and non-bottleneck. Hence, the difference values between the bottleneck's WIP and the LLV, the non-bottleneck's WIP and the HLV are introduced.

The priority from formulas (2) and (3) includes two parts. One is CR value. A value of 1.0 is on schedule, a value less than 1.0 is behind, and larger than 1.0 is ahead of schedule. The other one is $(DW_b\text{-}AW_b)/DW_b$ or $(AW_{nb}\text{-}DW_{nb})/AW_{nb}$ which describes the working states of machines such as hungry, proper and crowded. For $(DW_b\text{-}AW_b)/DW_b$, it is larger than 0, less or equal to 1. The larger value is, the more hungry is the bottleneck. If it is equal to 1, there are no lots in the queue of bottleneck. Therefore, those lots even ahead of schedule have to be sent to the bottleneck as soon as possible. For $(AW_{nb}\text{-}DW_{nb})/AW_{nb}$, it is larger or equal to 0, less than 1. The larger value is, the more crowded is the non-bottleneck. Therefore, those lots even behind schedule can not be sent to the non-bottleneck. For instance, lot i has a CR value of 1.9, it is ahead of its schedule. But the $(DW_b\text{-}AW_b)/DW_b$ value is 1, the bottleneck is extremely starved. According to formula (2), the new priority of lot i is 0.9 (1.9-1). It has a higher priority than those lots with priority 1 which are in the same queue of the bottleneck's upstream machine but not be processed next in the bottleneck, and will be processed first.

2.1.5 Detailed Algorithm

The detailed algorithm of proposed method is described as follows.

- Step1: Apply the utilization method to calculate the utilization of each machine in the model with formula (1), the machine with the highest utilization is considered as bottleneck.
- Step2: Compare the WIP of bottleneck with its Lowest Limited Value (LLV), if the actual WIP is lower than the LLV, then goto step3, otherwise goto step4.
- Step3: Find out the upstream machines of bottleneck, and reset the lots' priorities queued in the upstream machines according to formula (2).
- Step4: Check the WIP of each non-bottleneck machine, and compare it with its Highest Limited Value (HLV), if its actual WIP is higher than the HLV, then goto step5, otherwise goto step6.

- Step5: Find out the upstream machines of non-bottlenecks, and reset the lots' priorities queued in these upstream machines according to formula (3), then goto step7.
- Step6: Remain the lots' original priorities in the upstream machines of non-bottleneck machines.
- Step7: Process lots in the machines according to their priorities;
- Step8: If a high loading machine has a breakdown, handle it as the Step2-Step7 after it is repaired.

The bottleneck constantly changes in a wafer fab mainly due to high variance. Therefore, the bottleneck detection method has to be repeated quite often in order to detect and manage the bottleneck. In this research, the algorithm is repeated according to the reevaluation period of bottleneck detection method, e.g. as we mentioned before, the bottleneck detection period is every 24 hours to MIMAC6, so this algorithm is carried out every 24 hours to MIMAC6.

2.2 Experimental Model and Simulation Software

In order to test the proposed method, we use simulation models from Measurement and Improvement of MAnufacturing Capacities (MIMAC) (Fowler and Robinson 1995) test bed datasets which include 7 models named by MIMAC1, 2, 3, 4, 5, 6, 7. Table 1 shows the basic characteristics of these 7 models.

Table 1: Basic characteristics of MIMAC models.

MIMAC	Products	Tool Groups	Tools	Operator Groups	Operators	Process Flows	Max. Steps
1	2	83	265	32	90	2	245
2	7	97	274	97	274	7	354
3	11	73	354	0	0	11	547
4	7	35	69	0	0	7	92
5	21	85	176	4	24	21	266
6	9	104	228	7	97	9	355
7	1	24	38	0	0	1	172

First of all, MIMAC6 is chosen to be tested with the proposed method, because MIMAC6 is a typical complex wafer fab model including multiple products, various production facilities, and process flows with hundreds of steps, which is rather close to a real 200mm wafer fab. Other MIMAC models are tested with the proposed method as well but with less detail.

The simulation experiments are carried out with Factory Explorer (FX) from WWK. The proposed method is not provided by the FX simulation package, but FX supports customization via a set of user-supplied code and dispatch rules. Here we use a customized FX interface developed by René Wolf (Wolf 2008) to control the operation of FX.

3 EXPERIMENT AND PERFORMANCE ANALYSIS

3.1 Definition of the Lowest Limited Value (LLV) and Highest Limited Value (HLV)

The LLV and HLV are based on the predicted Theoretical Maximum Wafer Processing Rate (TMWPR) which is the maximum feasible rate at which the machine could process wafers, assuming no nonscheduled time, no offline time, no setup, no repair, and full batches (for machines that perform batch processing). In order to get the TMWPR of each machine, one year capacity simulation of MIMAC6 with only one product is carried out by FX each time, then each machine's TMWPR of that product is obtained. Since there are 9 products in MIMAC6, 9 capacity simulation runs are performed. At last an average of each machine's TMWPR of these 9 products is calculated. The bottleneck's LLV and non-bottleneck's HLV can be defined as follows.

$$LLV = X \times TMWPR. \tag{4}$$

$$HLV = Y \times TMWPR. \tag{5}$$

where X>0, Y>0, and X<Y.

Different LLV and HLV combinations should be defined to different models with different loading and due date flow factors. The definition of LLV and HLV is related to the machines' processing rate, lots' inter arrival time for machines, and the online and offline time of machines. Due to hundreds of machines in the model and the complexity, we can not specify an accurate LLV and HLV to each machine. Here a simple method is used to define the range of the LLV and HLV, then the experiment is carried out to confirm which combination can achieve the best performance to the model. First a 18 months capacity simulation of the model is carried out. Then each machine's maximum unit (wafer) WIP is obtained and divided by its TMWPR, each result is assigned to an index of the corresponding machine. We take the 95% utilization case as example. To MIMAC6, most of the high capacity loading machines' indexes range from 10 to 13. Therefore, according to the formula (5), we estimate that the parameter Y ranges from 10 to 13, so the HLV is estimated to range from 10×TMWPR to 13×TMWPR, which means if the WIP of the high capacity loading machine exceeds this estimated HLV value, this machine is highly crowded. Moreover, the parameter X from formula (4) is estimated approximately a half of Y, so the LLV is estimated to range from 4×TMWPR to 7×TMWPR. Finally the experiment confirms that 6×TMWPR for LLV and 12×TMWPR for HLV can achieve the best performance for MIMAC6.

3.2 Performance Analysis

The simulation length of MIMAC6 was carried out for 18 months with 3 dispatching strategies of the proposed method, FIFO and CR. The first 6 months were considered as warm-up period and not taken into account for the statistics. At first we performed parameter studies of the proposed method, and tried to find out how parameters X and Y from formulas (4) and (5) depend on loading and due date flow factor. Then we considered average cycle time (CT), cycle time variance, cycle time Upper Pctile 95% and on time delivery as main performance measures for comparing the proposed method with FIFO and CR. Last we applied the proposed method on different models and compared the results with in the case of FIFO.

3.2.1 Parameter Studies

The Lowest Limited Value (LLV) and the Highest Limited Value (HLV) are the parameters of the proposed method. Different models with different loading and different due date flow factors may need different LLV and HLV combinations. Firstly, we tried to find out the relationship between the LLV, HLV and loading. We considered three utilization cases of 70%, 80% and 95%. To each utilization, we defined 4 different LLV and HLV test values respectively through the method we described in section 3.1, i.e., 16 combinations together. The due date flow factor was set to 2.0. Table 2 lists the average CT and CT variance of different LLV and HLV combinations with 70% utilization, Table 3 with 80% utilization and Table 4 with 95% utilization. We discover that

- With 70% utilization, LLV = 4×TMWPR, HLV = 10×TMWPR,
- With 80% utilization, LLV = 6×TMWPR, HLV = 11×TMWPR,
- With 95% utilization, LLV = 6×TMWPR, HLV = 12×TMWPR,

we obtain the best performance among all tested LLV and HLV combinations. It also tells us that with a higher utilization, higher LLV and HLV values should be defined, because the machines with a high utilization have much work to do and there have to be many lots waiting in the queue. If a lower LLV is defined, the bottleneck may be considered to be not hungry, however, in fact the bottleneck is already in starvation for a long time and could not get any lots. A lower HLV also mistakenly considers a non-bottleneck to be crowded and stops feeding lots to it.

Next, we tried to determine the relationship between the LLV, HLV and due date flow factor. The due date flow factor (FF) is defined as the target cycle time divided by the raw processing time (Rose 2002). To each utilization above, we set 3 different due date flow factors of 1.5, 2.0 and 3.0 to test how the LLV and HLV depend on the due date flow factor. However, the results are almost the same as Table 2, 3 and 4, and the results are not listed here. Under the same loading, the system behavior with different LLV and HLV combinations is not affected by changing the due date flow factors. It indicates that the LLV and HLV do not depend on the due date flow factor as much as the loading in the proposed method. Does it imply that the due date flow factor only has small effect on the average CT and CT variance? What about other performance measures such as percent tardy lots and average time tardy for tardy lots? We will discuss it in section 3.2.3.

Table 2: Average CT and CT variance of different LLV and HLV combinations of MIMAC6 with 70% utilization. The best combination is LLV=4×TMWPR, HLV=10×TMWPR.

HLV	8×TMWPR		9×TMWPR		10×TMWPR		11×TMWPR	
LLV	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)
4× TMWPR	19.4	0.93	19.2	0.92	19.4	0.78	19.4	0.93
5× TMWPR	19.6	0.94	19.3	0.91	19.4	0.85	19.5	0.90
6× TMWPR	19.4	0.92	19.3	0.94	19.5	0.90	19.4	0.96
7× TMWPR	19.3	0.90	19.5	0.95	19.3	0.92	19.4	0.93

Table 3: Average CT and CT variance of different LLV and HLV combinations of MIMAC6 with 80% utilization. The best combination is LLV=6×TMWPR, HLV=11×TMWPR.

HLV	8×TMWPR		9×TMWPR		10×TMWPR		11×TMWPR	
LLV	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)
4× TMWPR	21.6	1.18	21.5	1.20	21.3	1.25	21.3	1.18
5× TMWPR	21.4	1.20	21.4	1.23	21.6	1.20	21.3	1.15
6× TMWPR	21.7	1.24	21.3	1.16	21.5	1.18	21.4	1.09
7× TMWPR	21.5	1.18	21.6	1.20	21.5	1.24	21.6	1.20

Table 4: Average CT and CT variance of different LLV and HLV combinations of MIMAC6 with 95% utilization. The best combination is LLV=6×TMWPR, HLV=12×TMWPR.

HLV	10×TMWPR		11×TMWPR		12×TMWPR		13×TMWPR	
LLV	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)	Ave.CT (days)	CT Var. ($days^2$)
4× TMWPR	29.1	1.68	29.2	1.68	29.1	1.71	29.4	1.75
5× TMWPR	29.0	1.73	29.3	1.70	29.2	1.68	29.2	1.70
6× TMWPR	29.0	1.75	29.1	1.69	29.0	1.65	29.5	1.72
7× TMWPR	29.4	1.71	29.2	1.70	29.0	1.67	29.0	1.71

3.2.2 Comparison between the Proposed Method and FIFO

FIFO always has an excellent average CT. It outperforms those due date oriented dispatching rules such as CR and EDD concerning average CT and CT variance. The proposed method was compared with FIFO with 70%, 80%, and 95% utilization with respect to 3 performance measures of average CT, CT variance, CT upper pctile 95%. The results are listed in Table 5. The results illuminate that the proposed method obtains some improvements of these three performance measures, although the improvements are not significant. We don't expect that the proposed method can exceed FIFO greatly concerning average CT, considering the limitation of FX interface. However, the more important thing is that the CT variance is improved by 17.9%, 9.9% and 1.8% respectively to 70%, 80% and 95% utilization compared with in that case of FIFO. Because a less CT variance stands for a more accurate prediction of production completion time for the wafer fab, achieves greater repeatability and makes the wafer fab stable and predictable, this is a good result.

Table 5: Three performance measures comparison between the proposed method and FIFO for MIMAC6 with different utilizations. The performance measures are average CT, CT variance, CT upper pctile 95%.

Utilization(%)		70	80	95
Ave.CT (days)	Proposed Method	19.4	21.5	29.0
	FIFO	19.5	21.6	29.4
CT var. ($days^2$)	Proposed Method	0.78	1.09	1.65
	FIFO	0.95	1.21	1.68
CT UP 95% (days)	Proposed Method	24.8	27.6	38.2
	FIFO	24.9	27.6	38.8

3.2.3 Comparison between the Proposed Method and CR

It is not a trivial task to set a target due date for CR. With a proper target due date, not only 100% on time delivery, but also a good average CT and CT variance can be achieved. However, if the target due date is set too tight in a high loading of the wafer fab, the average CT and CT variance could be extraordinary larger than in the case of FIFO (Rose 2002). We introduced the due date flow factor in the proposed method in order to test how the system behaves meeting different target due date. First, three due date flow factors of 1.5, 2.0 and 3.0 were defined respectively to three utilizations of 70%, 80% and 95%. Average CT and CT variance are the performance measures comparing with in the case of FIFO. We do not compare with CR because the average CT and CT variance from FIFO outperforms CR in this case. The results are showed in Table 6. As we can see, the due date flow factors do not have any effect on the 70% utilization case. Only small change to the CT variances with 80% and 95% utilization cases. But the results are better than the FIFO. After analysis, we notice that 2777 lots are released in the wafer fab (MIMAC6) each year, and only small part of them need to be changed the priority according to formula (4) and (5) based on the due date flow factor. With different due date flow factors, lots that need to be changed to a new priority could have different priority values. Let us explain with a simple example, if there are 2 lots in the queue, lot *i* needs to be changed to a new priority, lot *j* needs not to be changed and has a default priority 1. With flow factor 1.5, lot *i* has a new priority 0.9, which means lot *i* has a higher priority and is processed first. With flow factor 2.0, lot *i* has a new priority 1, which lot is processed first depends on which lot arrives at the queue first. With flow factor 3.0, lot *i* has a new priority 1.4, so lot *j* is processed first. However, these changes just happen in a small part of lots, they do not affect the wafer fab's performance greatly among all the lots. This is why due date flow factor just has a small effect on the proposed method.

Table 6: Average CT and CT variance comparison between the proposed method and FIFO for MIMAC6 with different utilizations and different due date flow factors.

Utilization(%)		70	80	95
Ave.CT (days)	Proposed Method (1.5FF)	19.4	21.5	29.1
	Proposed Method (2.0FF)	19.4	21.5	29.0
	Proposed Method (3.0FF)	19.4	21.4	29.0
	FIFO	19.5	21.6	29.4
CT var. ($days^2$)	Proposed Method (1.5FF)	0.78	1.11	1.66
	Proposed Method (2.0FF)	0.78	1.09	1.65
	Proposed Method (3.0FF)	0.78	1.09	1.65
	FIFO	0.95	1.21	1.68

In the following, we focused on the 95% utilization case, and compared the proposed method with CR with due date flow factors of 1.5, 2.0 and 3.0, respectively. Table 7 illustrates the performance measures of average CT, CT variance, CT upper pctile 95%, percent tardy lots and average time tardy for tardy lots. As we can see from the results, the proposed method outperforms CR in each performance measure.

Table 7: Five performance measures comparison between the proposed method and CR for MIMAC6 with 95% utilization and different due date flow factors. The performance measures are average CT, CT variance, CT upper pctile 95%, percent tardy lots and average time tardy for tardy lots.

	Ave.CT (days)	CT var. ($days^2$)	CT Upper Pctile 95% (days)	Pct Tardy Lots (%)	Ave. Time Tardy for Tardy lots (days)
Proposed Method (1.5FF)	29.1	1.66	38.3	100	8.9
Proposed Method (2.0FF)	29.0	1.65	38.2	84.1	2.5
Proposed Method (3.0FF)	29.0	1.65	38.0	0.0	0.0
CR (1.5FF)	56.0	4.13	62.3	100.0	35.7
CR (2.0FF)	49.7	1.82	56.8	100.0	22.7
CR (3.0FF)	31.3	2.75	39.8	0.0	0.0
FIFO	29.4	1.68	38.8		

The reason why the cycle time becomes considerably large with a tight target due date of CR is when old lots become late, higher priorities are assigned to them by CR to speed them up, so the fresh lots have to wait and becomes late again, this procedure circulates and the consequence is that the cycle time becomes high. However, the proposed method does not behave like CR. As we can see from Table 7, with a tight target due date 1.5, for CR case, the average CT are 56 days and CT variance are 4.13 $days^2$, which are extremely high. However, for the proposed method the average CT are 29.1 days and CT variance are 1.66 $days^2$, which exceed CR and are still better than FIFO. When the flow factor changes from 2.0 to 3.0, the percent tardy lots changes from 84.1% to 0 of the proposed method, and from 100% to 0 of CR. Does the percent tardy lots change smoothly or suddenly? We were curious about this. So we concentrated on the percent tardy lots and average time tardy for tardy lots with flow factors ranging from 1.5 to 2.8 in steps of 0.2. The results are listed on Table 8 and illustrated by Figure 1.

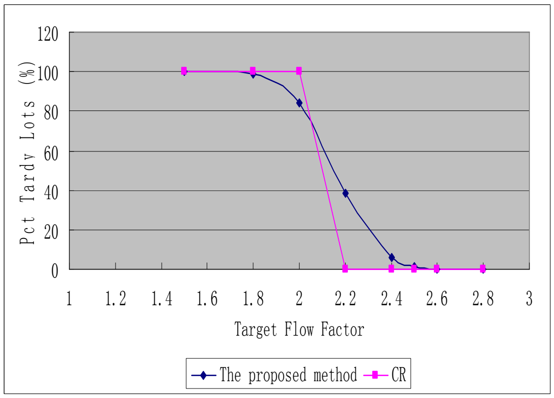

Figure 1: Percent tardy lots comparison between the proposed method and CR for MIMAC6 with 95% utilization and different due date flow factors.

From Table 8, with a tight due date flow factor 1.5, the average time tardy for tardy lots are 8.9 days from the proposed method. It is much smaller than 35.7 days from the CR. From Figure 1, we can see that when due date flow factor changes from 1.8 to 2.6, the percent tardy lots changes smoothly from 99% to 1.3% of the proposed method. Whereas, for the CR, when the due date flow factor changes from 2 to 2.2, the percent tardy lots suddenly jumps from 100% to 0.2%. If the

Table 8: Percent tardy lots and average time tardy for tardy lots comparison between the proposed method and CR for MIMAC6 with 95% utilization and different due date flow factors.

Due Date FF		1.5	1.8	2.0	2.2	2.4	2.5	2.6	2.8
Pct Tardy Lots (%)	Proposed Method	100	99	84.1	38.2	6.1	1.3	0.1	0
	CR	100	100	100	0.2	0	0	0	0
Ave. Time Tardy for Tardy Lots (days)	Proposed Method	8.90	4.70	2.50	1.25	0.59	0.32	0.10	0
	CR	35.70	31.10	22.70	0.06	0	0	0	0

target due date flow factor is set 2.2 in the wafer fab, due to the variability of machines breakdown or customer demand, the flow factor must change to 2.0. For the CR, 100% of the lots become tardy, small change of the due date flow factor is not tolerable, which is a disaster to the fab. However, for the proposed method, the range to tolerate variability is wider than the CR, the flow factor change mentioned above is tolerable. Therefore, the proposed method is more robust than CR.

3.2.4 Further Analysis

In most systems, there is one equipment or part that is the main constraint. This equipment or part is called primary bottleneck. However, there are also a number of other equipments or parts that constrain the system, which are called secondary bottlenecks (Law and Kelton 2000). Similarly, we tried to determine how many machines belong to the secondary bottlenecks of MIMAC6 with 95% utilization, and how they affect the system. We considered the machines with the second, third, fourth, fifth highest utilization as secondary bottlenecks. First we only detected the primary bottleneck. Then we included the second highest utilization machine in the bottleneck detection. Each time we included one more secondary bottleneck in the bottleneck detection, and saw how the average CT and CT variance behaved. Table 9 depicts the results.

Table 9: Simulation results of average CT and CT variance considering secondary bottlenecks of MIMAC6 with 95% utilization . PB = Primary bottleneck (the first highest utilization machine); 2ndB = the second highest utilization machine; 3rdB = the third highest utilization machine; 4thB = the fourth highest utilization machine; 5thB = the fifth highest utilization machine.

		Ave. CT (days)	CT var. ($days^2$)
Proposed Method	PB	29.0	1.65
	PB+2ndB	29.2	1.60
	PB+2ndB+3rdB	29.2	1.58
	PB+2ndB+3rdB+4thB	29.3	1.68
	PB+2ndB+3rdB+4thB+5thB	29.4	1.70
FIFO		29.4	1.68

The best CT variance of 1.58 is achieved by including second and third highest utilization machines in the secondary bottlenecks, then it becomes larger and worse than FIFO when including 4 machines in the secondary bottlenecks. We also considered the sixth, seventh and eighth highest utilization machines as secondary bottlenecks. Some results are approximate to FIFO, some results are worse than FIFO. In most cases the bottleneck is not static but rather shifts between different machines. Including more and more secondary bottlenecks in the bottleneck detection is not a good idea, especially when a secondary bottleneck shifts to another machines, but we still mistakenly consider it as bottleneck, at this situation it would degrade system performance.

3.2.5 Experiments on other Models

The FX interface can not handle the model with rework steps, so MIMAC1 and MIMAC3 could not be tested with the proposed method due to rework steps included in them. MIMAC2 was reported with problem. Thus, we tested MIMAC4, 5 and 7 with the proposed method with 95% utilization. The average CT, CT variance and CT upper pctile 95% are the three performance measures compared with in that case of FIFO. Table 10 lists the results.

Table 10: Simulation results of average CT, CT variance and CT upper pctile 95% of different models with 95% utilization. MIMAC1 is excluded with rework steps.

Models		Ave. CT (days)	CT var. ($days^2$)	CT Upper Pctile (days)
MIMAC4 (7 products)	Proposed Method	8.0	0.17	9.9
	FIFO	8.5	0.25	10.4
MIMAC5 (21 products)	Proposed Method	21.5	0.60	26.2
	FIFO	21.5	0.64	26.4
MIMAC7 (1 product)	Proposed Method	29.9	2.18	32.9
	FIFO	29.9	2.18	32.9
MIMAC1 (2 products)	Proposed Method	34.1	19.6	47.2
	FIFO	33.8	30.2	52.7

We can see that the proposed method is applicable for other models and obtains improvements. However, MIMAC7 is an exception that no matter how we change the LLV and HLV combinations and due date flow factors, the results from the proposed method are the same as FIFO. It is because only one product in MIMAC7. We find out that most of the machines' average queue length is 0, only 0.2 lots in the queue of bottleneck 'PHGCA_LITHOGRAPHY', which indicates that MIMAC7 runs quite smoothly. Most of the machines have a low WIP. Although the bottleneck sometimes starves, the upstream machines of the bottleneck can not feed it because there is no lot or only one lot in the queue. This is why the behavior of MIMAC7 can not be affected by the proposed method. MIMAC1 only has 2 products. It is interesting to test whether it behaves like MIMAC7. Thus, we excluded the rework steps of MIMAC1, the results are also listed in Table 10, but just for a reference.

4 CONCLUSIONS

A bottleneck detection and corresponding dynamic dispatching strategy based on the TOC for the wafer fab was investigated in this research. At first the utilization method was adopted to detect the bottleneck. Then, through comparing the bottleneck's WIP with the pre-defined Lowest Limited Value (LLV) and non-bottleneck's WIP with pre-defined Highest Limited Value (HLV) respectively, the proposed method dynamically changed lots' priorities to regulate the workload of machines in order to prevent bottleneck starvation and non-bottleneck with a high WIP occurrence. Due to the complexity, we suggested a simple way to define the LLV and HLV for different models. The experiment illuminates that with a high loading, higher LLV and HLV values should be defined. We applied the proposed method on MIMAC6 which is a complex wafer fab model, and compared with FIFO with respect to average cycle time, cycle time variance. To 70%, 80% and 95% utilization, the proposed method achieved improvements in these 2 performance measures. Although the average cycle time did not improve greatly, we reduced the cycle time variance. Furthermore, the proposed method was more robust than CR concerning different due date targets change. The proposed method was also applicable for other models in the MIMAC test bed datasets.

5 FURTHER WORK

Based on the present models, more bottleneck detection methods may be introduced and tested, especially considering that the bottleneck shifts between different machines. To obtain an accurate LLV and HLV combination for each machine, an in-depth analysis should be taken to the machines' performance such as processing rate, lots' inter-arrival time for the machines and machines' online and offline time. In order to further reduce the average cycle time and cycle time variance, dispatching strategies for the batch machine should be considered to eliminate unnecessary waiting time to form batches.

REFERENCES

Fowler, J., and J. Robinson. 1995. Measurement and improvement of manufacturing capacities (mimac): Final report. Technical Report 95062861A-TR, SEMATECH, Austin, TX.

Glassey, C. R., and M. G. C. Resende. 1988. A scheduling rule for job release in semiconductor fabrication. *Operations Research Letters* 7:213–217.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling & analysis*. 3rd ed. New York: McGraw-Hill, Inc.

Lu, S. C. H., D. Ramaswamy, and P. R. Kumar. 1994. Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. *IEEE Transactions semiconductor manufacturing* 7:374–376.

Rose, O. 2002. Some issues of the critical ratio dispatch rules in semiconductor manufacturing. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yucesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, 1401–1405. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Tsai, C.-H., Y.-M. Feng, and R.-K. Li. 2003. A hybrid dispatching rules in wafer fabrication factories. *International Journal of The Computer* 11:64–68.

Wein, L. M. 1988. Scheduling semiconductor wafer fabrication. *IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING* 1:1–16.

Wolf, R. 2008. *Entwicklung einer steuerungsschnittstelle für den simulator factory explorer einschließlich ausführlichem test am beispiel der abfertigungsregel "operation due date (odd)"*. M.S. thesis, Department of Computer Science, Dresden University of Technology, Dresend, Germany.

AUTHOR BIOGRAPHIES

ZHUGEN ZHOU is a PhD student at Dresden University of Technology. 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 workcenter modeling for wafer fab. His email address is <zhugen.zhou@tu-dresden.de>.

OLIVER ROSE holds the Chair for Modeling and Simulation at the Institute of Applied Computer Science of the Dresden University of Technology, 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. His web address is <www.simulation-dresden.com>.