Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

A GLOBAL WIP ORIENTED DISPATCHING SCHEME: WORK-CENTER WORKLOAD BALANCE WITHOUT RELYING ON TARGET WIP

Zhugen Zhou
Oliver Rose

Department of Computer Science
Bundeswehr University Munich
Neubiberg, 85577, GERMANY

ABSTRACT

Work-in-process (WIP) oriented dispatching rules have been widely applied to balance workload for work-center in wafer fabs. The performances of WIP oriented rules, e.g., minimum inventory variability scheduling (MIVS) and WIP control table (WIPCT), highly rely on the accuracy of target WIP, as the target WIP plays a major role in measuring the pull request of work-center. In this paper, to replace the target WIP, we propose a workload indicator (WI), which utilizes global fab information like dynamic workload of work-center, batch size, setup requirement and lot status, to measure the pull request of work-center. Furthermore, the proposed WI is applied in a global fab dispatching scheme which considers *K*-machine look ahead and *J*-machine look back. Simulation results show its significant improvement versus the use of the target WIP oriented rules.

1 INTRODUCTION

Work-in-process (WIP) is an important performance indicator in 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 oriented (also called workload oriented) dispatching rules, e.g., minimum inventory variability scheduling (MIVS) (Li et al. 1996) and WIP control table (WIPCT) (Zhou and Rose 2010), are popular ways to regulate the workload of work-center to prevent overload and starvation (Fowler et al. 2002). However, the performances of the WIP oriented rules highly rely on the accuracy of the target WIP. In practice, the target WIP is determined by sophisticated approach like neural network (Chambers and Mount-Campbell 2002) or trial-and-error approach (Zhou and Rose 2015), which indeed is a challenging task. Several potential drawbacks arise when the target WIP is applied to balance workload of work-center:

- Misleading target WIP results in overload or starvation to the bottleneck work-center, which has significant impact on throughput of the fabs.
- Uncertainty of product volume mix and daily changing lot release rate cause the necessity to update target WIP weekly, daily, or even hourly accordingly.
- Due to hundreds of work-centers in wafer fabs, the explosion of the parameter space is hard to handle for simulation experiment if the target WIP of each work-center is considered as simulation parameter (Zhou 2014).

Therefore, in this paper we seek to develop a dispatching scheme which can achieve workload balance for work-center without the use of target WIP. To achieve that, firstly, we need to understand what kind of role the target WIP plays. Basically, WIP control philosophies used in wafer fabs can be classified into push and pull (Fowler et al. 2002). Simply speaking, the push approach is about what

upstream work-center desires to produce, and the pull approach is about what downstream work-center is capable of producing. The pull approach is proven to produce less WIP congestion than the push approach by means of accurate measurement of downstream status (Spearman and Zazanis 1992). The pull approaches are combinations of order release like starvation avoidance (Glassey and Resende 1988) and constant WIP (Spearman et al. 1990) and dispatching rules like MIVS and WIPCT. The MIVS is a representative pull approach which has successful applications in industry. The MIVS (also called one-step ahead MIVS) attempts to keep the WIP of each operation close to the average target WIP level. It assigns 4 different priorities to operations so as to minimize the deviation of instantaneous WIP from the average WIP. Taking upstream and downstream operations into consideration, the one-step ahead MIVS is extended to *K*-steps ahead and *J*-steps back MIVS which has a more global outlook of the wafer fabs (Li et al. 1996). Collins and Palmeri (1997) conduct a simulation study by comparing one-step ahead MIVS with three-steps ahead MIVS. It is unclear that three-steps ahead MIVS is superior to one-step ahead MIVS. Different from MIVS looking WIP flow from the viewpoint of operation, the WIPCT intends to manage WIP from the viewpoint of work-center (Zhou and Rose 2010). In the WIPCT, each upstream work-center has a WIP control table that 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 WIP control table is updated. Therefore, the upstream work-center is able to detect WIP distribution and measure the pull request of downstream work-centers dynamically. Although the MIVS and WIPCT can balance workload by producing fast lot movement, they have a disadvantage of poor pace of lot movement.

As a matter of fact, the target WIP plays an important role in measuring the pull request of downstream in the MIVS and WIPCT. The balance machine workload (BMW) approach (Ham and Fowler 2007) shows as an example of work-center workload balance without relying on the target WIP. Inspired by *K*-steps-ahead and *J*-steps-back MIVS, the BMW approach considers *K*-work-centers ahead and *J*-work-centers back. Instead of the target WIP, the BMW adopts a real workload (how many hours of production work) to measure the pull request of work-center. The workload is calculated dynamically based on the global fab status. Although the simulation results indicate that the BMW outperforms the MIVS regarding workload balance, three issues of the BMW approach are debatable. (1) The calculated priority for operation is a ratio between actual workload and raw processing time, which causes misleading priority for operation in some cases; (2) The misleading priority reduces the chance to foam batch as full as possible. Consequently, it leads to a long queue in front of the batch processing work-center (this issue will be presented in Section 3); (3) The BMW is applied to two small datasets in which some fields including load/unload time, operator, delay, scrap and rework are removed, which makes it hard to judge its superiority over the MIVS.

According to the theory that an optimal schedule can be achieved, if the information set based on which decision is made is large enough (Baker and Trietsch 2009). We believe that the pull request can be measured without relying on the target WIP, if upstream and downstream information of the wafer fab can be utilized in an appropriate way. In this study, similar to the BMW approach we look at the WIP flow from the viewpoint of work-centers. In a work-center, at first we calculate the workload of each operation and the total workload of work-center. By utilizing the concept of *K*-work-centers ahead and *J*-work-centers back, we draw a WIP flow diagram which demonstrates real time WIP distribution in work-centers.

Figure 1 presents the workload information at the viewpoint of *k*-work-centers ahead based on wafer fab dataset MIMAC6 from Measurement and Improvement of MAnufacturing Capacities (MIMAC). In work-center 'AUTO-CL_dot, operations 4, 21, 105 and 136 have the workload 4 hours, 2 hours, 14 hours and 6 hours, respectively. When a machine is available for processing, a decision on which lot at which operation needs to be made. If we only take local work-center 'AUTO-CL_dot' into consideration, the lot at operation 105 should be processed at first, because it has the heaviest workload in the queue. This dispatching is similar to longest queue first rule. In fact, different dispatching decision can be made according to different viewpoint and information. If we consider 1-work-center ahead, there are four

downstream work-centers 'ASM_B2', 'ASM_C2_H1', 'ASM_B3_B4_D4' and 'POSI_GP' which have total workload 12 hours, 14 hours, 14 hours and 14 hours, respectively. We realize that the 'ASM_B2' has the minimum total workload of 12 hours. Therefore, the lot at operation 4 should be processed in work-center 'AUTO-CL_dot' at first and sent to 'ASM_B2'. It is similar to least work at next queue rule. The dispatching decision is different if we extend it to 2-work-centers ahead. The lot at operation 4 recommended in 1-work-center ahead is not an ideal choice anymore, because the downstream work-center 'CTS_2' which can process the lot at operation 6 is high loaded. If we still send the lot to it, it will only cause heavier workload on downstream work-centers. In this case, the lot at operation 136 is recommended to be sent to work-center 'POSI_GP' because its downstream work-center 'SH' is low loaded, although it is high loaded. This example indicates that local information is too limited to make an effective dispatching decision. In order to obtain the global fab information, we can extend the 1-work-center ahead to the *k*-work-centers ahead.

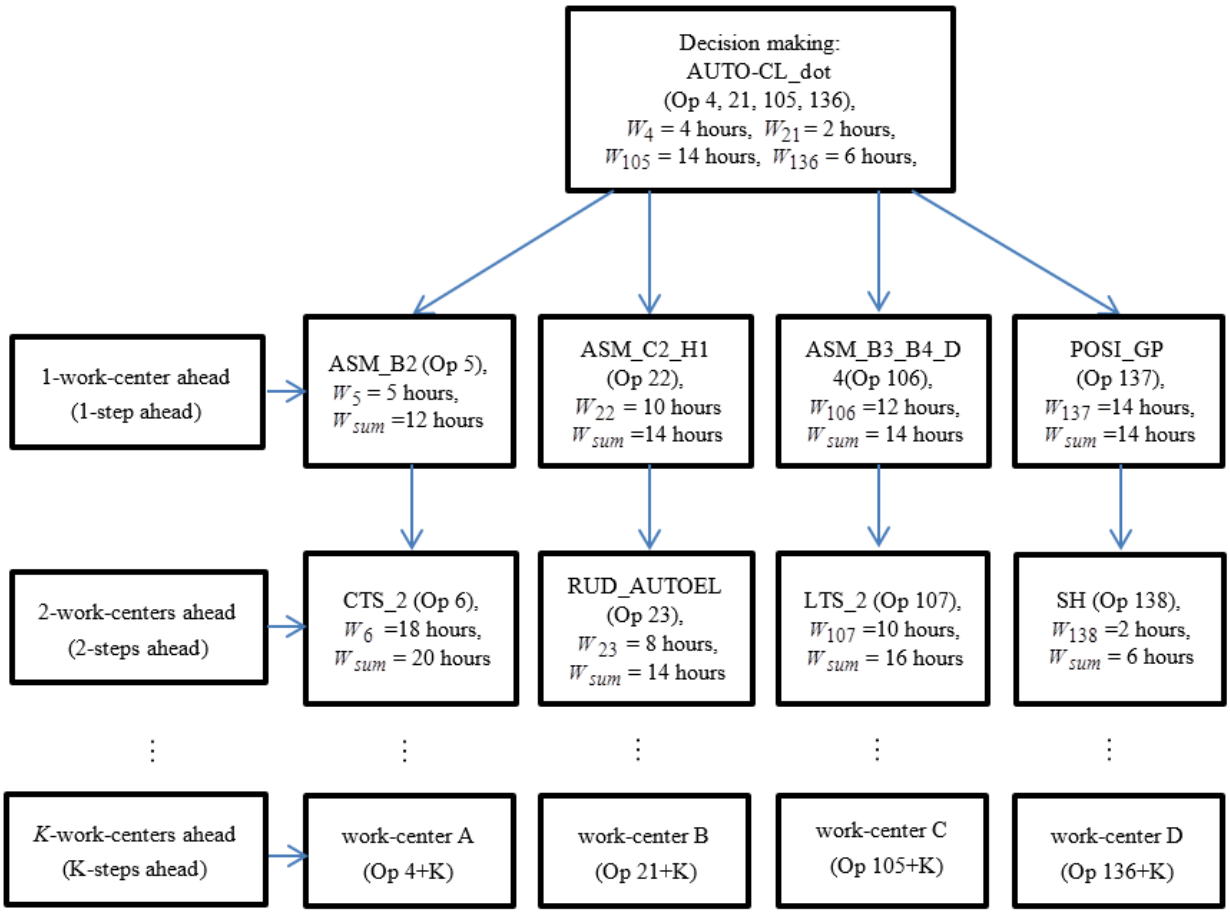

Figure 1: Workload information of work-centers at the viewpoint of K-work-centers ahead.

Similarly, we can apply the same idea to draw the workload information of work-center from the viewpoint of *J*-work-centers back. The combination of the *K*-work-centers ahead and *J*-work-centers back is a global dispatching scheme that is large enough to include the whole wafer fabs. Based on this scheme, in order to overcome the weaknesses of the BMW approach we propose a new workload indicator which employs global fab information as much as possible to measure accurate pull request of work-center. Furthermore, we introduce a mechanism to achieve global fab balance with harmonious lot movement that is unsuccessfully achieved by the MIVS and WIPCT.

This paper is organized as follows. In Section 2, we introduce the workload indicator (WI) which is used to measure the pull request of work-center. Then we address one special case to modify WI by taking batch and setup strategies into account. After that we describe the global dispatching scheme integrated with WI in detail and the simulation model. Section 3 gives the simulation results and performance analysis. Section 4 is the conclusion and future work.

2 WORKLOAD INDICATOR (WI) TO MEASURE PULL REQUEST OF WORK-CENTER

2.1 Methodology

Firstly we will introduce the notations used in the following sections:

- workload: how many hours of production left at a work-center which includes load/unload time and raw processing time of operations;
- *i*: work-center identifier;
- *t*: operation identifier;
- *j*: candidate operation (operation at local work-center to be dispatched) identifier;
- *J*, *K*: number of steps look back/ahead;
- *J', K'*: number of steps between the first downstream/upstream operation and the candidate operation at the bottleneck work-center;
- $W_{i(up),t}$, $W_{i(down),t}$: workload of operation *t* at upstream/downstream work-center *i*;
- $W_{i(local),j}$: workload of candidate operation j at local work-center *i*;
- b_u : batch utilization;
- Local work-center: when a machine at a work-center is available for processing. The lot with candidate operation j needs to be assigned a priority for dispatching.

2.2 Workload Indicator for Local/Upstream/Downstream Work-centers

Based on the concept of *K*-work-centers ahead and *J*-work-centers back, we can obtain the global and dynamic workload information. To precisely reflect the pull request of work-center, we divide the workload information into three components which are workload for local work-center, upstream *J*-work-centers and downstream *K*-work-centers.

For the local work-center *i*, the workload of operation *j* is defined as:

$$Local = W_{i(local),j}. \tag{1}$$

The further away the up/downstream work-centers from the local work-center, the less influence the workload has on the dispatching decision. For instance, for the downstream work-center case the workload of 10-work-centers ahead might be less important than the 1-work-center ahead, because the 1-work-center ahead has more immediate needs than the 10-work-centers ahead does. Thus, to reflect this fact we propose a workload indicator for the upstream and downstream work-centers as follows:

$$Upstream_{J-back} = \sum_{t=j-J}^{j-1} \frac{W_{i(up),t}}{2^{j-t}}, \tag{2}$$

$$Downstream_{K-ahead} = \sum_{t=j+1}^{K} \frac{W_{i(down),t}}{2^{t-j}}. \tag{3}$$

Consequently, we obtain the final WI for the lot at operation *j* at local work-center *i* as follows:

$$WI_j = Upstream_{J-back} + Local + Downstream_{K-ahead} = \sum_{t=j-J}^{j-1} \frac{W_{i(up),t}}{2^{j-t}} + W_{i(local),j} - \sum_{t=j+1}^{K} \frac{W_{i(down),t}}{2^{t-j}} \quad . \quad (4)$$

The local and upstream workload indicators are positive while the downstream is negative. The reason is for the local and upstream work-centers, a high workload leads to a high priority, while a high workload for downstream leads to a low priority.

2.3 Incorporate Batch and Setup Strategies into Workload Indicator for Batch Processing Work-center

One benefit of workload balance from the viewpoint of work-center is that batch and setup strategies can be incorporated into dispatching decision. Batch and setup are actually two significant factors in wafer fabs, especially, increasing batch size and reducing setup time show great appeal for batch processing work-center in literature (Kim et al. 2008, Robinson 2002).

The WI in Equation (4) demonstrates that an operation has a high priority if its upstream is high loaded and its downstream is low loaded. When batch and setup are involved, the situation becomes different. The WI has a potential flaw in foaming batch as full as possible. In Figure 2, work-center *A* performs batch processing, and the maximum batch size is 4 lots. Both operation 4 and 8 have 2 lots, and operation 16 has 4 lots. According to the WI, apparently work-center *A* prefers to process lots at operation 4 because it has high workload at upstream and low workload at downstream. However, it leads to half batch size and capacity loss. If work-center *A* turns out to be a bottleneck, it causes a long queue impacting on the throughput. In this case, lots at operation 16 would be a better choice since it can form a full batch to reduce the queue length.

Based on this fact, a modification factor (mod) is introduced to the WI to avoid the situation mentioned above. The mod is used to weight the influence of batch and setup for batching processing work-center, as show in Equation (5). In principle, if a candidate operation (lot) leads to a low batch size and needs a new setup, the mod has negative effect on the WI to decrease its priority, and vice versa. Thus, the mod is calculated according to the batch utilization and new setup requirement, as presented in Equations (6) and (7). With regard to single processing work-center, the mod is not applied in this study, and the WI for single processing work-center is calculated according to Equation (4).

$$WI_j = mod * Upstream_{J-back} + mod * Local + mod^{-1} * Downstream_{K-ahead} = \sum_{t=j-J}^{j-1} mod * \frac{W_{i(up),t}}{2^{j-t}} + mod * W_{i(local),j} - \sum_{t=j+1}^{K} mod^{-1} * \frac{W_{i(down),t}}{2^{t-j}} \quad . \quad (5)$$

At first, in Equation (6) there is a factor called batch utilization (b_u) which is the percentage of lots that can form a batch. The batch utilization is divided into 3 levels. Less than half batch is considered as a low batch utilization which has a low influence on the WI. In this case, the mod should be a value between 0 and 0.5 represented by '*X*'. Similarly, full batch is considered as a high batch utilization which has a strong contribution the WI, and it has a value of 1. In between the low batch and the full batch, the mod between 0.5 and 1 is determined by '*X*' and '*Y*'. Then, if the candidate operation (lot) requires a new setup, the mod has an impact on the WI by reducing a value of '*Z*', with the purpose to reduce the priority of the candidate operation. For target WIP oriented rules, the simulation experiment is hard to handle due to hundreds of target WIP levels for work-centers. However, in this study '*X*', '*Y*' and '*Z*' are designed as three parameters that influence on the performance of the WI, which is considered as an advantage in

reducing parameter space. As a preliminary study, we apply trial-and-error approach to determine the values of '*X*', '*Y*' and '*Z*'. For future study, in order to obtain accurate values of these three parameters, sophisticated approach like heuristic optimization should be applied.

The values of '*X*', '*Y*' and '*Z*' in Equations (6) and (7) are derived from the following procedures. (1). The value of *mod* is between 0 and 1, the batch utilizations are divided into three levels as shown in Equation (6); (2). For batch utilization less than 50%, obviously, we need to lower its influence on the WI. Thus, we design three levels of the parameter '*X*' which are 0.2, 0.3 and 0.4; (3). For the batch utilization between 50% to 100%, the parameter '*Y*' has three levels that are 1.6, 1.4 and 1.2 corresponding to '*X*'; (4). The parameter '*Z*' depends on the minimum value of '*X*'. If '*X*' is 0.2, '*Z*' is designed to 0.1; If '*X*' is 0.3, '*Z*' is designed to two levels which are 0.1 and 0.2; If '*X*' is 0.4, '*Z*' is designed to three levels which are 0.1, 0.2 and 0.3; (5). The combinations of ('*X*', '*Y*' and '*Z*') are (0.2, 1.6, 0.1), (0.3, 1.4, 0.1), (0.3, 1.4, 0.2), (0.4, 1.2, 0.1), (0.4, 1.2, 0.2) and (0.4, 1.2, 0.3); (6). Design of experiment is carried out and the simulation results tell us that the combination (0.3, 1.4, 0.2) achieves the best performance for the MIMAC6 model (in Section 3). Therefore, it leads to the Equations (8) and (9).

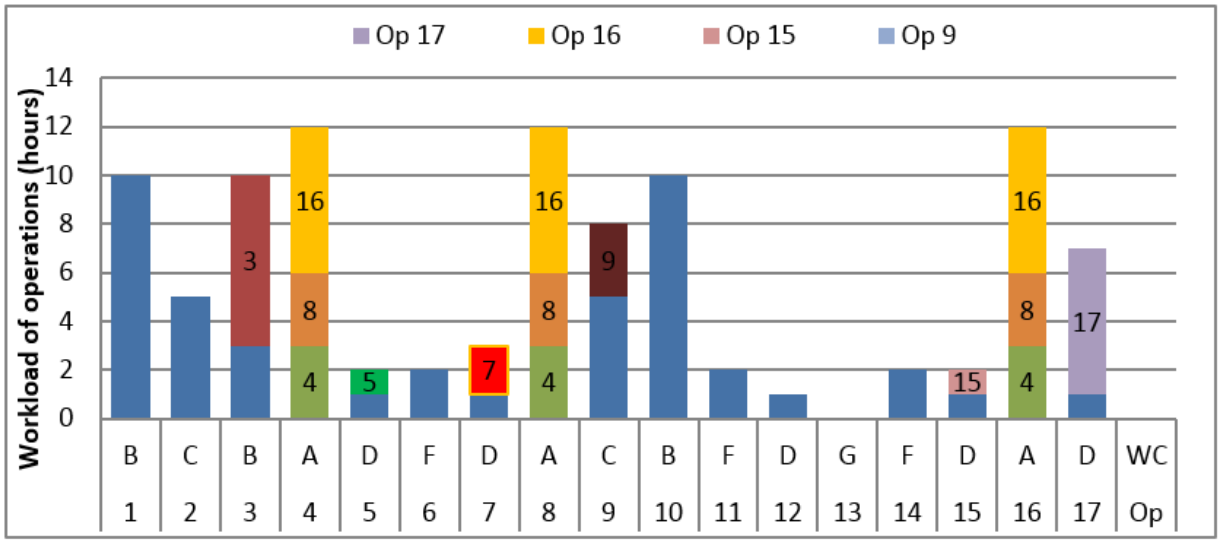

Figure 2: Influence of batch processing to the WI.

$$mod = \begin{cases} X & b_u < 50\%; \\ X + (b_u - 0.5) * Y & 50\% \le b_u < 100\%; \\ 1.0 & b_u = 100\%; \end{cases} \tag{6}$$

$$\begin{aligned} &If\ new\ setup\ is\ required: mod = mod - Z; \\ &else: mod\ remains\ unchanged. \end{aligned} \tag{7}$$

$$mod = \begin{cases} 0.3 & b_u < 50\%; \\ 0.3 + (b_u - 0.5) * 1.4 & 50\% \le b_u < 100\%; \\ 1.0 & b_u = 100\%; \end{cases} \tag{8}$$

$$If\ new\ setup\ is\ required: mod = mod - 0.2;$$
$$else: mod\ remains\ unchanged. \tag{9}$$

2.4 Determine Look Ahead and Look Back Distance (Sizes of *K* and *J*)

In order to make optimal dispatching decisions, information set should not be only restricted to 1-work-center ahead and 1-work-center back. On one hand, the values of *K* and *J* can be large enough to include the entire line. On the other hand, there is no conclusion that what sizes of *K* and *J* are enough and appropriate (Collins and Palmeri 1997). Hence, the sizes of *K* and *J* should be determined dynamically according to the line situation. According to the Theory of Constrains, the bottleneck is the foremost constraint in the fab and the non-bottlenecks should subordinate to the bottleneck to make sure to achieve the best performance. Based on this reason, the seizes of *K* and *J* should reflect the distance between the local work-center and the bottleneck.

In this paper, we apply the method from Ham and Fowler (2007) to determine *K* and *J*. First of all, the work-center with the highest workload is considered as bottleneck. (If more than one work-center have the same highest workload, the one with higher utilization is considered as bottleneck.) Then in the bottleneck work-center, the operations called bottleneck operations are sorted by workload in descending order and marked as the upstream or downstream of the candidate operation *t*. Finally, the bottleneck operations are searched in descending order starting from the candidate operation *t*. When the first downstream bottleneck operation is found, it stops searching and assigns the first downstream bottleneck operation (the number of operation) to *K'*. The first upstream bottleneck operation is found and assigned to *J'* in the same way. If all bottleneck operations are the upstream of candidate operation *t*, *K'* is 0. If all bottleneck operations are the downstream of candidate operation *t*, the *J'* is 0. The sizes of *K* and *J* for candidate operation *t* are determined as follows (Ham and Fowler 2007):

$$\begin{aligned} &\text{K = MAX\{MIN\{number of downstream operations, 5\}, K'\},} \\ &\text{and J = MAX\{MIN\{number of upstream operations, 5\}, J'\}.} \end{aligned} \tag{10}$$

If the candidate operation is close to the end or at the beginning of process flow, *K* and *J* can be less than 5. If the bottleneck operation is close to the candidate operation, the Equation (10) still guarantees that the sizes of *K* and *J* are at least 5.

2.5 Strong Pull at the End/Weak Pull at the Beginning of Process Flow and Strong Pull for Low Utilized Work-center

Based on our observation from the fabs, two unusual events have potential impact on the WI. When the WI is used to calculate the priority of candidate operations, a high workload of upstream increases the priority while a high workload of downstream decreases the priority. Operations at the end of process flow have high possibility to obtain higher priority from WI compared to the operations at the beginning or middle of process flow. The other problem is that some work-centers are designed to be low utilized. In this case, the WI becomes high to force the upstream work-centers to send lots to the low utilized work-centers intentionally.

Consequently, these two unusual events bring poor pace of lot movement, i.e., some lots spend considerable queue time while some lots move fast through the fabs. To improve this situation we intend to break the dominance of the WI by introducing the operation due date (ODD) dispatching rule. The ODD rule is a simple rule where each operation has its own due date that is defined as the release time plus the sum of processing times up to this operation times the target flow factor. The ODD rule strictly keep the lots at the right pace to meet their operation due dates. Thus, it is able to speed up the lots which cannot be processed in time caused by the WI. The target flow factor for the ODD is set to 2.2

which is considered as a medium flow factor. For more information regarding the ODD, the interested readers are referred to Zhou and Rose (2013).

2.6 Detailed Algorithm of Global WIP Oriented Dispatching Scheme

When a machine of a work-center is available for processing, the following algorithm is carried out to calculate the WI for each candidate operation:

Step 1: Check whether a candidate operation is late for its operation due date. If it is true and the local work-center is not the bottleneck, the ODD rules is used to assigned priorities to candidate operations. Otherwise continue to Step 2.
Step 2: Find out the bottleneck work-center with the highest workload, and determine the look ahead/back distances K and J according to Equation (10).
Step 3: Calculate the modification factor *mod* according to Equations (8) and (9) for batch processing work-center.
Step 4: For single processing work-center, calculate the WI for each candidate operation according to Equation (4). For batch processing work-center, calculate the WI for each candidate operation according to Equation (5).
Step 5: Assign WI to each candidate operation as priority for dispatching decision.

2.7 Simulation Model

The wafer fab dataset MIMAC6 from Measurement and Improvement of MAnufacturing Capacities (MIMAC) is used to test the proposed global WIP oriented dispatching scheme. 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.
- 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) for tool group.

3 SIMULATION RESULTS AND PERFORMANCE ANALYSIS

Several factors influencing the pull request of work-center are proposed and integrated into the WI in Section 2. In order to understand their positive/negative effects, we carry out an experiment in which four approaches ranging from a simple one that only considers the workload of local work-center to a sophisticated one that combines all factors.

The first approach represented as '*WI(1):L*' only considers the workload at the local work-center as described in Equation (1). The operation with the highest workload obtains the highest priority, which is similar to the rule called Longest Queue First (LQF). This approach may achieve local balance for some work-centers, but it is too local to achieve balance for the whole wafer fabs. Starting from the second approach represented by '*WI(2):U+L+D*', the upstream and downstream workload information are taken into account and combined with the first approach, which is expressed in Equations (1), (2), (3) and (4). The sizes of K and J are dynamically determined according to the bottleneck detection and described in Equation (10). The performance of '*WI(2):U+L+D*' is expected to outperform '*WI(1):L*' because more information sets are considered. In the third approach represented by '*WI(3):Mod(U+L+D)*', the modification factor for the batch and setup are included into the second approach as shown in Equations (5), (8) and (9). In the fourth approach represented by '*WI(4):Mod(U+L+D)+ODD*', we apply ODD rule

to replace WI to overcome the poor pace of lot movement caused by some unusual events like unidirectional preferences. We also expect that the cycle time variance is improved compared to the third approach. The fourth approach is the detailed algorithm presented in Section 2.6.

Average cycle time, cycle time variance and cycle time upper 95% percentile are considered as performance measures under 95% fab loading. We compare the results to the cases of FIFO, MIVS, WIPCT and BMW. The results are shown in Table 1.

Table 1: Three performance measures comparison among four variations of WI, FIFO, MIVS and WIPCT under 95% fab loading.

Approach	Avg. Cycle Time (days)	Cycle Time Variance (days^2)	Cycle Time Upper 95% Percentile (days)
FIFO	29.6	1.7	39.1
MIVS	28.5	1.8	37.0
WIPCT	28.2	6.7	37.8
BMW	55.2	20.4	80.0
WI(1): L	45.4	20.6	58.3
WI(2): U+L+D	28.5	8.2	38.4
WI(3): Mod(U+L+D)	27.0	7.4	35.5
WI(4): Mod(U+L+D)+ODD	26.2	2.0	34.0

Where: MIVS is Minimum Inventory Variability Scheduling;
WIPCT is WIP Control Table with target WIP level,;
WI(1): L is the first approach which only considers the workload of local work-center;
WI(2): U+L+D is the second approach which extends the first approach by considering the workload of upstream and downstream work-centers;
WI(3): Mod(U+L+D) is the third approach which incorporates the modification factor into the second approach;
WI(4): Mod(U+L+D)+ODD is the fourth approach which combines ODD rule to the third approach to replace the WI if needed.

Next we focus on the results of four variations of the WI. The '*WI(1):L*' only considering the workload of local work-center gives rise to a huge average WIP level, which leads to an enormous average cycle time and variance. Actually, we are not surprised to see this result, as we mentioned before the information set used for the dispatching decision is too limited to achieve a global balance. For the second approach '*WI(2):U+L+D*', which takes the upstream and downstream workload information into consideration and determines the sizes of J and K dynamically, achieves a huge improvement. The average cycle time is reduced from 42.4 to 28.5 days, which becomes competitive with the MIVS and WIPCT. Besides that, the cycle time variance and cycle time upper 95% percentile are improved considerably. It validates the assumption that a global standpoint of workload balance is superior over a local standpoint. Actually, the '*WI(2):U+L+D*' approach is similar to the BMW. However, the BMW produces a considerable amount of average cycle time. The WI in Equation (4) is considered to be more effective in measuring the pull request of work-center than the BMW. In addition, the modification factor reflecting the requirement for batch processing work-center brings positive effects. '*WI(3):Mod(U+L+D)*' results in 2 days of average cycle time reduction compared with '*WI(2):U+L+D*'. Once again, batch and setup are two important factors in wafer fab when workload balance is desired. Because increasing batch utilization and avoiding constant new setup bring faster lot movement to achieve cycle time reduction. Although the modification factor has positive effect on average cycle times, with respect to the cycle time variance it shows otherwise. The intention to introduce ODD rule, which is '*WI(4):Mod(U+L+D)+ODD*', is to overcome the potential weakness of WI and improve the pace of lot movement. The simulation results tell us the cycle time variance is improved significantly after ODD is

applied, meanwhile, the average cycle time and cycle time upper 95% percentile still outperform '*WI(3):Mod(U+L+D)*' approach. So far, we can conclude that the fourth approach incorporating all factors into WI shows significant improvements compared with MIVS, WIPCT and BMW which achieve workload balance with the usage of target WIP. Obviously, the proposed global WIP oriented dispatching scheme shows promising results which support our assumption that WIP balance can be achieved without relying on the target WIP.

Table 2 shows the average batch size comparison of 17 batch processing work-centers among MIVS, WIPCT, BWM and WI(4):Mod(U+L+D)+ODD. We can see that our global WIP oriented dispatching scheme has the advantage in forming larger batches in comparison with the MIVS, WIPCT and BMW.

Table 2: Average batch size comparison of 17 batch processing work-centers among MIVS, WIPCT and WI(4): Mod(U+L+D)+ODD under 95% fab loading.

Work-center	Average Batch Size (wafers per batch)			
	MIVS	WIPCT	BMW	WI(4):Mod(U+L+D)
11022_ASM_A2	24.1	24.5	17.3	26.6
11021_ASM_A1_A3_G1	38.2	39.6	34.6	41.5
11026_ASM_B2	89.4	91.3	82.4	92.2
11027_ASM_B3_B4_D4	40.7	41.6	35.2	43.1
11029_ASM_C1_D1	50.8	51.7	46.5	53.1
11030_ASM_C2_H1	42.2	43.4	38.2	45.5
11032_ASM_C4	26.4	26.8	22.1	29.3
11122_ASM_D2	26.9	27.7	23.7	29.2
11128_AMS_E4	48.5	49.2	45.5	50.6
11524_MAX1+2_AL-TEMP	50.7	51.5	47.8	53.1
12021_AUTO-CL_undot	54.4	54.6	51.6	55.4
12022_AUTO-CL_dot	53.6	53.7	50.8	55.2
14131_AMT-PREC_1+3	52.2	53.9	50.4	54.6
14137_AMT-PREC_7	56.6	58.4	52.3	59.9
14821_DNS-SOG_1	62.4	66.3	58.7	68.9
12221_HF-DIP-5_B	91.8	91.9	90.6	92.2
11132_ASM_F4_D3	33.0	34.5	30.1	35.8

4 CONCLUSION AND FUTURE WORK

Two WIP oriented pull rules, i.e., MIVS and WIPCT, were well studied and applied in literature. However, the successes of these rules highly relied on the accuracy of target WIP. It is difficult and challenging to determine appropriate target WIP. Although the BMW approach shown as an example of work-center workload balance without relying on the target WIP, it suffered from several shortcomings. Thus, in this paper we sought to develop a dispatching scheme to achieve global fab balance without relying on the target WIP, in particular, fast and good pace of lot movement that is unsuccessfully achieved by the MIVS, WIPCT and BMW.

Inspired by the MIVS and BMW rules, we developed a WIP flow of work-centers in which we can observe the dynamic workload information of the fabs. Based on that, we proposed a Workload Indicator (WI) to measure the pull request of work-centers. To replace target WIP, we considered not only the WI of local work-center, but also the WI of upstream and downstream work-centers. Besides that, we incorporated several factors which may affect the workload balance into the WI, based on the theory that the larger information set is used for decision making, the better schedule can be achieved. We conducted

simulation experiments using four WI scenarios from simple one to sophisticated one. The simulation results demonstrated that the proposed global WIP oriented dispatching scheme achieved better workload balance for work-centers in comparison with the MIVS, WIPCT and BMW. Furthermore, compared to the MIVS and WIPCT which have hundreds of target WIP levels, the proposed dispatching scheme only has three parameters which is an advantage of the size of parameter space.

For the future study, we need to investigate the influence of machine breakdown and maintenance on the WI. In order to obtain accurate values of the parameters for the WI, sophisticated approach like heuristic optimization should be applied.

REFERENCES

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

Chambers M. and C.A. Mount-Campbell. 2002. "Process Optimization via Neural Network Metamodeling". *International Journal of Production Economics* 79(2):93-100.

Collins, D. W. and V. Palmeri. 1997. "An Analysis of the "K-step Ahead" Minimum Inventory Variability Policy Using Sematech Semiconductor Manufacturing Data in a Discrete-event Simulation Model". In *Proceedings of the 6th International Conference on Emerging Technologies and Factory Automation*, September 9th-12th, Los Angeles, 520-527.

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(7):568-578.

Glassey, C.R. and M.G.C. Resende. 1988. "Closed-Loop Job Release Control for VLSI Circuit Manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 1(1):36-46.

Ham, M. and J. W. Fowler. 2007. "Balanced Machine Workload Dispatching Scheme for Wafer Fab". In *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, June 11th-12th, Stresa, Italy, 390-395.

Kim, S., YH. Lee, and T. Yang. 2008. "Robust Production Control Policies Considering WIP Balance and Setup Time in a Semiconductor Fabrication Line". *International Journal of Advanced Manufacturing Technology* 39(3-4):333-343.

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):1-5.

Robinson, J.K. 2002. "Understanding and Improving Wafer Fab Cycle Times". Technical Report, FabTime Inc., Menlo Park, California.

Spearman, M.L. and M.A. Zazanis. 1992. "Push and Pull Production Systems: Issues and Comparisons". *Operations Research* 40(3):521-532.

Spearman, M.L., D. Woodruff, and W. Hopp. 1990. "CONWIP: a Pull Alternative to Kanban". *International Journal of Production Research* 28(5): 879.894.

Zhou, Z. 2014. *WIP Balance and Due Date Control for Complex Job Shops (Wafer Fabs)*. Ph.D. thesis, Department of Computer Science, die Universität der Bundeswehr München, Munich, Germany. http://atheneforschung.rz.unibw-muenchen.de/node?id=97064, accessed 17th March 2015.

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. 2013. "Cycle Time Variance Minimization for WIP Balance Approaches in Wafer Fabs". In *Proceedings of the 2013 Winter Simulation Conference,* edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 3777-3788. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhou, Z. and O. Rose. 2015. "A Framework for Effective Shop Floor Control in Wafer Fabs". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossett, 3001-3012. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

ZHUGEN ZHOU is a postdoctoral researcher at Bundeswehr University 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 and Ph.D degree in Computer Science from Bundeswehr University Munich. His research interests include dispatching concepts for complex production facilities, work center modeling for wafer fabs and heuristic optimization. His email address is zhugen.zhou@unibw.de.

OLIVER ROSE is the professor for Modeling and Simulation at the Department of Computer Science, Bundeswehr University

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. His email address is oliver.rose@unibw.de.