Universität der Bundeswehr München

WIP BALANCE AND DUE DATE CONTROL FOR COMPLEX JOB SHOPS (WAFER FABS)

A Dissertation Presented
by
Zhugen Zhou

March, 2014

Submitted to die Universität der Bundeswehr München
in conformity with the requirements for the degree of
doctor rerum naturalium (Dr. rer. nat.)

Department of Computer Science
Institute of Computer Engineering
1. Supervisor: Prof. Dr. rer. nat. Oliver Rose
2. Supervisor: Prof. Dr. Cathal Heavey

ACKNOWLEDGEMENT

First of all I want to express my gratitude to Prof. Oliver Rose. He gave me the opportunity to do my research in my desired field of interest. He provided excellent research conditions and a high degree of freedom during my research. This dissertation could not have been completed without his guidance and assistance.

I want to thank Sebastian Werner and Frank Lehmann for their great support and fruitful discussion during my stay at Infineon Dresden. Their valuable industrial experiences are big help to this dissertation.

I also want to thank my colleagues in the group of Modeling and Simulation. It is a great pleasure to work with them.

Last I am grateful to my parents and my wife. They gave me support selflessly. I might never have the chance to carry out my study in Germany without their encouragement, patience and tolerance.

Abstract:

Nowadays to survive in the global market with increasing and fierce competition, the keys to success for the companies are fast product and reliable delivery that are the challenges for shop floor control. As one of the most important key performance indicators (KPI), work-in-process (WIP) attracts more and more attention since it has a major influence on overall manufacturing costs. According to Little's Law, a lower WIP level leads to a shorter production cycle time given the same throughput, which has significant economic importance. Besides that, due date commitment is another critical factor, especially for customer oriented companies to achieve customer satisfaction. A missed due date causes not only penalty, but also confidence lost to the customers.

In order to gain a competitive position within industry, on the shop floor enormous efforts have been spent in developing different kinds of operational control strategies relating to WIP and due date. On one hand, there are a number of operational control strategies which target the control of the flow of lots through wafer fab to achieve balanced WIP like CONstant WIP (CONWIP), Starvation Avoidance (SA), or Minimum Inventory Variability Scheduling (MIVS). These WIP oriented rules attempt to avoid starvation and congestion of work-center or operation, thus reducing WIP variability and cycle time. On the other hand, there are also a number of dispatching rules targeting due date control like Earliest Due Date (EDD), Critical Ratio (CR) and Operation Due Date (ODD). These due date oriented rules focus on progressing lot toward on-time completion based on lot status. As WIP and due date have two different goals, even conflicting goals under certain circumstances, the first set of WIP oriented rules do not always lead to good on-time delivery performance, the

latter due date oriented rules do not primarily lead to low WIP level. Both WIP oriented and due date oriented rules turn out to be insufficient when both targets, i.e., lower WIP level and lower cycle time, better on-time delivery and less tardiness, are desired simultaneously.

As a matter of fact, on the shop floor the challenges to apply WIP oriented or due date oriented rules are way beyond our anticipation. We encounter plenty of questions that cannot be answered with satisfaction from existing literature, when we manage to apply WIP oriented or due date oriented rules. The motivation of this dissertation is to find out the answers for the concerned issues relating to WIP and due date from industry. Particularly, we have a stronger interest in WIP related issues and intend to carry out a comprehensive study about WIP, for the reason that for instance low WIP level in combination with low variance can make sure the lots finish before their due dates as much as possible. Naturally, the due date related issues (on-time delivery and tardiness) can be solved perfectly. We will address the following eight issues and attempt to find out the answers in this dissertation.

The first one is **work-center oriented WIP balance**. The classic WIP balance rules like MIVS and Line Balance algorithm (LB) are operation oriented. Some researchers claim that managing WIP from the viewpoint of operations is beneficial because the WIP flow histogram intuitively tells us that we should push WIP from high WIP operation to low WIP operation. Nevertheless, the disadvantage of operation orientation that is to disregard the workload status of work-center is also obvious. In particular, it tends to cause congestion when work-centers have breakdowns. Therefore, some engineers would prefer to look at WIP flow at the viewpoint of work-center, which brings forward the first issue that is work-center oriented WIP balance.

The second issue, **fast but poor pace lot movement**, arises from the first issue. It is no doubt that work-center oriented WIP balance can achieve low WIP levels and low average cycle times for the fab. Whereas, without consideration of the lot status, i.e., whether a lot is ahead of/on/behind schedule (also expressed as due date information), some lots are accelerated while some are waiting long time in the queue. This poor pace movement causes a high cycle time variance which becomes a potential problem if good on-time delivery is desired. How to improve the lot movement based on WIP balance is particularly important to due date control.

No matter whether applying operation oriented or work-center oriented WIP balance, both can lead to two research directions which are to apply them with or without target WIP levels. The third issue is about **how to determine the target WIP level for work-centers**. The reason why the classic rule MIVS is successful is due to the assistance of target WIP. The target WIP regulates the WIP flow to avoid starvation and congestion. How to set an adequate WIP level is a challenging task because the performance of such a WIP balance approach is sensitive and highly relies on the target WIP level. The target WIP level usually has to be set appropriately by means of pilot studies or educated guessing. Therefore, an adaptive procedure or sophisticated approach to determine the adequate WIP level should be considered like applying queuing model or neural network based on the historical data.

The fourth issue that is **work-center oriented balance without the need of target WIP** is an extension of the third issue. From operational control viewpoint, there are many reasons to abandon the application of target WIP despite the fact that target WIP is helpful and effective. For instance, uncertainty of product volume mix and almost daily changing lot release rates due to frequent changes of customer orders cause the necessity to update the target WIP daily, or even hourly. Not to mention that huge parameter sets for the simulation experiment imposes additional challenges to apply target WIP.

The fifth issue is an extension to the first and second issues. The reason why WIP oriented and due date oriented rules harm each other in some cases is because they only employ their favorable information. It tells us that if we expect to make an optimal dispatching decision, we need to take the workload information of work-centers as well as the lot status information (due date information) into consideration. As a result, **a new dispatching scenario including WIP information and due date information** is desired.

No matter how much effort we spend to achieve WIP balance, WIP imbalance can still occur anytime and anywhere in the fab, since it is time dependent. The sixth issue is about **WIP imbalance monitoring and calibration**. Even small WIP imbalances can grow to serious problems if they could not be restrained in time. Therefore, an effective mechanism to monitor and detect WIP imbalances is necessary. In the literature, researchers have proposed WIP monitoring and calibration approaches which utilize the target WIP as trigger event and the MIVS rule as a calibration method. Once again, we propose a new WIP imbalance detection and calibration approach to differentiate from the one using target WIP. The reason is obvious and analogous to the fourth issue.

The seventh issue is **the performances of due date oriented rules** in literature. This issue looks independently from the above issues at first glance, actually, they connect to each other. On one hand, we spend much effort to figure out the cause of WIP imbalance, and we realize that the due date rules have a common symptom that WIP imbalance occurs under tight due dates and

high capacity loading. It demonstrates that we should pay more attention when the fab runs products with tight due dates and under high fab loading. On the other hand, the inherent characteristic of due date rules is to reduce lateness variance, thus reducing cycle time variance, which can exactly overcome the drawback arising from WIP balance for work-centers. Furthermore, the variants of due date rules - composite rules, e.g., modified operation due date (MOD), solve the WIP imbalance problem under tight due dates successfully, which gives us a hint to deal with the confliction between WIP balance and due date control.

When we obtain an insight into WIP balance and due date control, the eight issue is about **how to combine both ideas, i.e., keeping a low WIP level, avoiding bottleneck starvation and meeting due dates**. In reality, the fact is sometimes that we cannot achieve both targets, and the question which one is more important is controversial and has been raised by academic and industrial researchers. The answer is that it depends on the objective and situation in the fab. For example, in a customer oriented wafer fab there are some low volume products like hot lots, engineering lots and customer sample are expected to leave the fab as fast as possible. Normally, they will be assigned tight due dates to be accelerated. However, the introduction of WIP balance to this kind of wafer fabs seems to reduce the weight of due date control to low volume products. In order to make better trade-off between WIP balance and due date control, we have to figure out the interaction between them first. We intend to carry out preliminary study about this issue in this dissertation.

Key words: Wafer fabs, dispatching rule, WIP balance, cycle time reduction, due date control, on-time delivery

Chapter

CHAPTER 1

INTRODUCTION

1.1 Challenges and Motivation

Nowadays, more and more electronic products such as cell phones, computers and car devices, have expanded rapidly into daily life, which brings enormous opportunities to the semiconductor manufacturing. To survive in the global market with increasing and fierce competition, semiconductor manufacturers have to explore the state-of-the-art manufacturing technologies to launch new products along with minimized cost, shorten production cycle time, increase throughput, machine utilization and on-time delivery, and so on. However, it is not easy to achieve these targets since semiconductor manufacturing is considered as one of the most complex manufacturing processes. The manufacturing process is extremely unpredictable and unstable, and not easy to be traced. The reasons are as follows:

- High investment cost;
- Diverse product mix;
- Large number of process flows and hundreds of process steps (operations);
- Large degree of uncertainty of manufacturing resources like unpredictable machine breakdowns;
- Re-entrant flows;
- Setup and batch requirements.

To handle these, on the shop floor of a wafer fabrication facility (wafer fab) a large variety of operational control policies, in particular work-in-process (WIP) oriented policies including lot release and dispatching rules (also called workload control or WIP balance) [Fowler et al. 2002, Fredendall et al. 2010, Strum et al. 1999, Wein 1988], have been investigated and presented by academic and industrial researchers. The motivation to utilize WIP oriented policies is to control the flow of lots to achieve balanced WIP to reduce variability, thus achieving cycle time reduction that brings significant economic benefit. Nevertheless, cycle time reduction is not a trivial task. Indeed, many WIP related issues are involved, some even have gone so far as to confuse the engineers and been considered as constraints in the fab:

- Whether being efficient to the critical work-center means high risk of high WIP in the fab?
- Effectiveness comparison: release rules vs. dispatching rules?
- Lot flow comparison: WIP balance for work-centers vs. operation?
- What is an acceptable WIP level and how to determine it for work-centers or operation (bottleneck and non-bottleneck), even for the whole fab?
- Which manufacturing area has too much or too little WIP? How to monitor and control?

As many enterprises move from mass production to mass customization to satisfy their customers, for example Application Specific Integrated Circuits (ASIC) production, the superior flexibility dealing with changeable customer orders and the reliable on-time delivery performance are of particular concern. As a result, the complexity of semiconductor manufacturing is even increased. For instance, due to customer unique requirements and changeable orders, the product volume mix is uncertain and the lot release rate is changing daily, weekly and monthly. The due date performance imposes additional challenges to shop floor control, especially challenges to the predominance of WIP oriented issues

Those enterprises which apply WIP oriented policies on the shop floor always consider machine utilization, bottleneck starvation avoidance, etc. as the first priority, as long as no due date performance is involved. The authors of [Chung et al. 2009, Dabbas and Fowler 2003, Glassey and Resende 1988, Li et al. 1996, Lu et al. 1994, Spearman et al. 1990] demonstrated the excellent performance achieved by WIP oriented policies in the fab. However, when due dates are introduced, the situation becomes critical. Due dates seem to reduce the weight of WIP balance. For instance, the upstream machine would rather send a delayed lot to a highly loaded downstream machine instead of sending an on-schedule lot to a lowly loaded downstream machine. In contrast, those who apply due date oriented rules always prefer due date performance instead of WIP performance. Because a lack of global WIP information, especially when the machine has a breakdown, WIP imbalance occurs from time to time. In some cases the wafer fab runs with excessive WIP, which leads to long production cycle times and bad on-time delivery [Rose 2003]. This tells us that both WIP oriented and due date oriented policies have advantages and disadvantages. They turn out to be sophisticated but insufficient in today’s advanced wafer fabs where both targets - low WIP level and good due date performances - are desired simultaneously.

1.2 Problem Definition

1.2.1 What is WIP Balance?

WIP balance has been widely studied in the literature. Although there are various WIP oriented strategies, there is no exact, explicit and unified definition of WIP balance. The term “Workload Control” proposed by Fowler et al. [2002] summarizes WIP oriented strategies as the combination of lot release and dispatching strategies used to control the flow of lots through a wafer fab. In fact, WIP balance is the goal of ‘Workload Control’ and ‘Workload Control’ implicitly tells us the way to achieve WIP balance via lot release and dispatching strategies.

Here is one simple example to illustrate the term ‘Workload Control’. Suppose machine M_0 can process two different products P_1 and P_2, and has two downstream machines M_1 and M_2. P_1 is processed by M_0 and M_1, P_2 is processed by M_0 and M_2, respectively. Lot L_1 and L_2 belonging to P_1 and P_2 respectively are available to be processed by M_0 at a given time. L_1 is ahead of L_2 in the queue. In the meantime, M_1 has a breakdown and M_2 is available. M_1 has other lots of P_1 in the queue, therefore, it makes sense that M_0 chooses L_2 to process although L_1 arrived first. The consequence of such a workload control is to avoid capacity loss of M_2 and long queue in M_1, which is exactly what WIP balance is.

To further understand WIP balance, we need to figure out the opposite side of WIP balance that is WIP imbalance. We sum up three phenomena of WIP imbalance observed in a real wafer fab of Infineon AG, Dresden Germany. (1): From operation (process step) viewpoint, WIP imbalance means WIP piles up in one or some operations as shown in Figure 1.2.1.(a). It is dangerous if the high WIP operations are only performed by one work-center, and when the work-center is down, the process flow is suspended; (2): From work-center (machine) viewpoint, WIP imbalance represents that some work-centers are overloaded, while some are starved, as the example described above. Some lots experience long queue times in the overloaded work-centers, while the capacity is lost to the starved work-centers; (3): From macroscopic (whole fab) viewpoint, one direct symptom of WIP imbalance is the degradation of throughput. In case WIP imbalance occurs, the process flow must be affected and blocked somewhere in the fab. Consequently, WIP accumulates in the fab and throughput decreases.

No matter which way to cure WIP imbalance, e.g., preventing operations or work-centers from being overloaded and starved, and increasing the throughput by fast lot movement, the ultimate goal of WIP balance is to speed up lot movements to achieve low WIP and cycle time reduction. Moreover, since we address the importance of cycle time variance performance considered as pace issue, we raise the WIP balance to a higher level in comparison with the one in literature, and define it as follows:

WIP balance means lots go through the wafer fab in a fast and smooth way by means of workload control strategies. On one hand, ‘fast’ means cycle time is reduced because lots spend less queue time and throughput is increased. On the other hand, ‘smooth’ means that as lots go through the fab with better rhythm and pace, which results in fewer fluctuations in the WIP evolution curve, well-balanced WIP distributions in operations, starvation and congestion avoidance for work-centers .

We should notice that WIP balance is a relative term and there is no absolute balance as it is time dependent. In a given time, from a global viewpoint, WIP is in a balanced state for the whole fab. However, WIP can also be imbalanced for some work-centers from a local viewpoint, and vice versa. Generally speaking, two ways are used to determine if WIP is balanced or not. (1): A target WIP level is predefined to the operation or work-center. Starvation or congestion can be concluded by means of comparison between the actual WIP and target WIP; (2): The improvement of cycle time can be viewed as WIP balance achievement.

1.2.2 What is Due Date Control?

Due date control, also called due date management [Keskinocak and Tayur 2004, Wein 1991], consists of due date assignment policies and due date dispatching policies. In most cases due dates are determined by negotiations with customers or planning decisions that are not discussed operationally. We pay less attention on due date assignment and focus on due date dispatching policies. In contrast to most of the literature which considers, e.g., on-time delivery and tardiness as objectives, we propose the cycle time variance as one performance measure for due date control, because most of the due date dispatching policies intend to minimize the variance between lot finish time and due date, which turns out to minimize cycle time variance.

With regard to the objectives, such as on-time delivery and average tardiness of tardy lots, due date assignment policies have direct and indirect effects on performance [Baker and Trietsch 2009, Keskinocak and Tayur 2004]. The direct effect arises from the due date tightness. The indirect effect results from due date being a parameter of some due date dispatching policies. The due date dispatching policies have direct influence on cycle time and variance performances, whereas due date assignment policies only have an indirect influence.

It is obvious that due date assignment and dispatching policies have conflicting objectives. A tight due date is always preferential to a loose due date. However, a tight due date is more difficult to achieve than loose due date. A tight due date tends to create more tardiness which conflicts and disturbs the scheduling objective. Hence, in order to solve this conflict we consider minimized average tardiness of tardy lots and on-time delivery as objectives subject to a constraint to the due date tightness represented by due date flow factor (DDFF) as discussed in the following chapters.

Based on the observed facts, in this dissertation due date control is defined as:

Due date control employs due date dispatching policies to maximize the on-time delivery, and minimize the average tardiness of tardy lots and lot cycle time variance based on given operation due dates and lot final due dates.

1.2.3 What is WIP Balance Combining with Due Date Control?

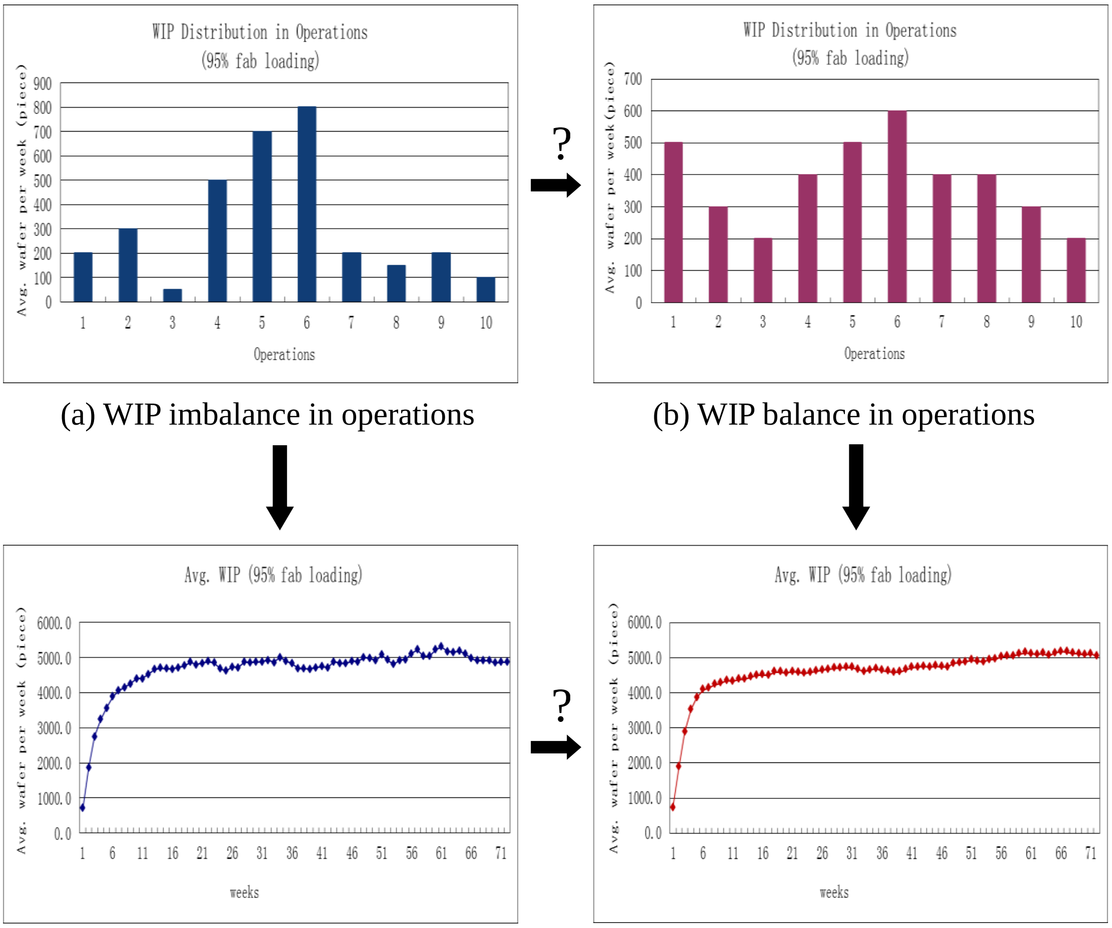
WIP balance and due date control have different objectives, as a result, WIP balance does not always lead to good on-time delivery performance and due date control does not primarily lead to low WIP levels. Nevertheless, WIP balance and due date control can be complementary and overcome the weaknesses of each other. We are aware that both of them are equally important, the situation that one dominating the other depends on the objective desired to achieve. In other words, a trade-off has to be made between WIP balance and due date control, e.g., when target due date is tight and there are a large amount of lots being tardy in the fab, overemphasizing due date control only brings tardiness to the fresh lots. In this case, it would be a good idea that through WIP balance we make sure some lots go through the fab as fast as possible, at the cost of due date performance of tardy lots.

In this dissertation, we intend to incorporate both ideas together. By way of combining ***WIP balance*** and ***due date control***, shortened cycle time, lowered cycle time variance, increased on-time delivery and minimized average tardiness are achieved, simultaneously.

1.3 Objectives

This dissertation deals with issues related to WIP balance and due date control in wafer fabs. First of all, to better understand why WIP balance (workload control) is so critical, we need to know that as one of the most important key performance indicators (KPI), WIP represents the average number of wafers (also referred to lots) in the fab. It is obvious that WIP costs money to produce, and excessive WIP means excessive resources and capital are wasted without adding any value, e.g., floor space utilization, handling systems. According to Little's Law [Little 1992], a lower WIP level leads to a shorter production cycle time given the same throughput, which has significant economic importance to respond to today's quick market change fashion. Besides that, less WIP means lots spend less queue time, thus cycle time predictability increases. This directly increases on-time delivery, because it is easier to predict the exact production cycle time and confirm to customer order accordingly. Additionally, more than 70% cycle time of a lot is consumed in the wafer fabrication process, and the last part of manufacturing process is test and inspection. If a lot fails in the early operation like in wafer fabrication, high WIP causes a long time between the early operation and final inspection. It can be difficult to detect and correct the root cause of the problem since so much time has already passed. Thus, the lower the WIP is, the easier it is to detect and correct failure problems to improve the quality. Actually, WIP balance is a way used to reduce WIP in the fab, for the reason that WIP balance smoothes the manufacturing flow and speeds up lot movement by means of regulating the workload of machines or operations to avoid starvation and congestion [Fowler et al. 2002, Li et al. 1996].

Figure 1.3.1 demonstrates an example of conventional WIP imbalance/balance of operations. (The data of this example is from simulation, but the relationship between WIP imbalance and WIP fluctuation presented is approved by engineers in industry.) If we look at the WIP flow on the basis of operations, Figure 1.3.1 (a) represents WIP imbalance by the fact that WIP distributes unevenly in operations. This WIP imbalance is driven by different events like hot lot, setup, batching and inappropriate dispatching, et al. Consequently, WIP fluctuation occurs oftentimes as shown in Figure 1.3.1 (c), which causes trouble in cycle time predictability. Conventional WIP balance only focuses on WIP and cycle time reduction meaning fast lot movement. In this dissertation, WIP balance is raised to a higher level. Except for fast lot movement, WIP balance means less WIP fluctuation and fast lot movement with better pace as well which are presented in Figure 1.3.1 (b) and (d). This is the first objective in this dissertation.



(a) WIP imbalance in operations

(b) WIP balance in operations

(c) WIP curve with serious fluctuation

(d) A relatively balanced WIP curve

Figure 1.3.1: WIP imbalance vs. WIP balance

As a matter of fact, we encounter different kinds of issues during the exploration of our first objective. These issues help us to further understand the challenges in applying WIP balance.

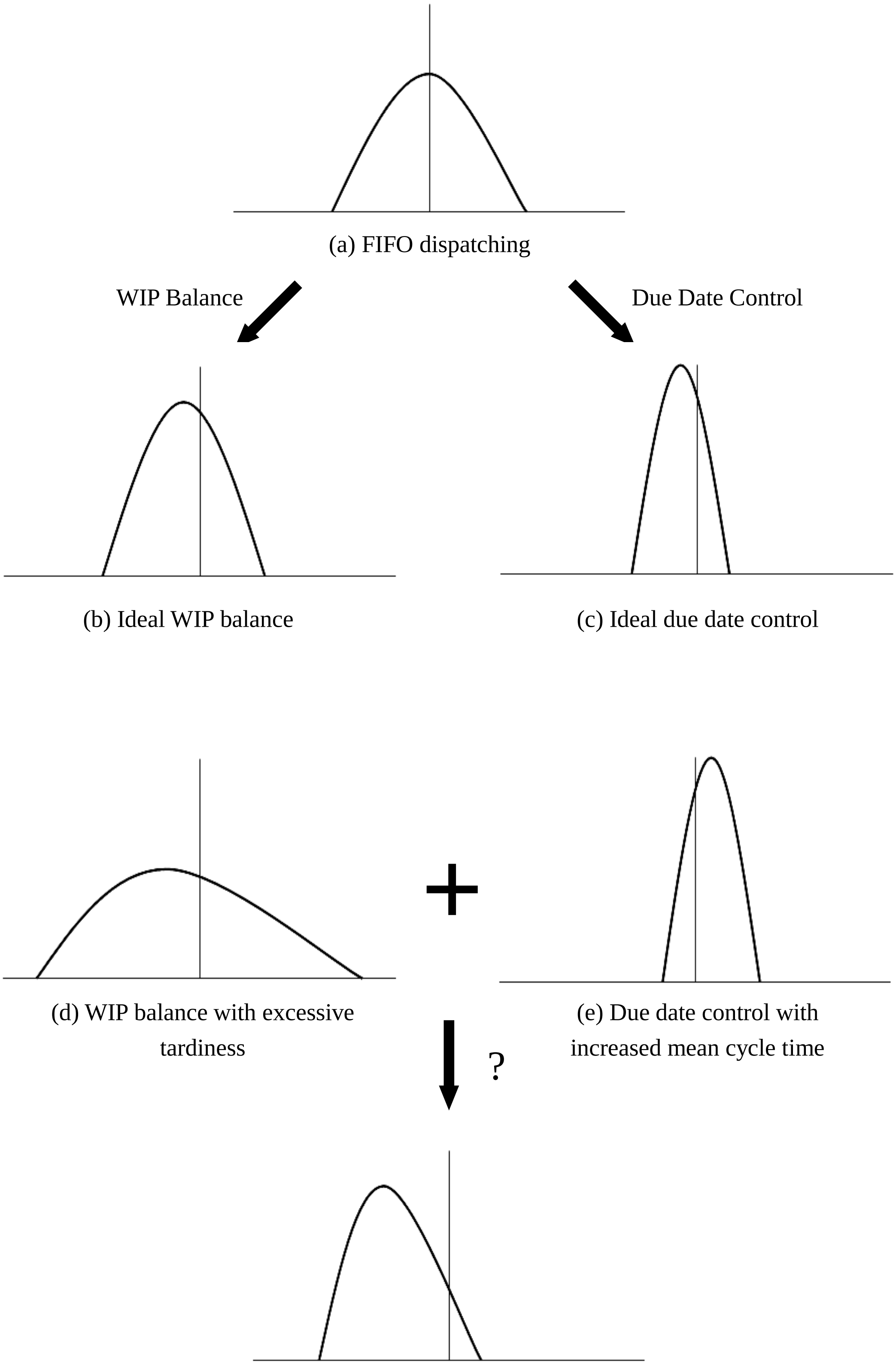
- In contrast to conventional WIP balance for operations presented in Figure 1.3.1, WIP balance for work-centers is considered as a potential beneficial approach to achieve cycle time reduction. To develop a WIP oriented dispatching scheme for work-centers is the first step towards WIP balance objective.
- As discussed above, not only fast lot movement, but also smooth pace lot movement are the inherent requirements of WIP balance. Typically,

WIP balance for work-centers does not take lot status into consideration, which means it is highly possible work-center oriented WIP balance achieves fast lot movement at the cost of losing pace. Thus, as an extension of the first step, how to improve lot movement for work-center oriented WIP balance is of concern as well.

- The reason, why we are aware that an operation is starved or overloaded in Figure 1.3.1, is because a target WIP level is pre-specified for an operation. Actually, target WIP plays an important role for the success of WIP balance policies like CONWIP and MIVS. Similarly, it is very natural to ask whether we can apply target WIP to achieve work-center oriented WIP balance. More importantly, how to determine appropriate target WIP is extremely difficult and costing.
- It is true that target WIP can achieve excellent performance, whereas, when we realize that the practical drawbacks are as obvious as its superiorities, it drives us to develop an alternative to replace target WIP for work-center oriented WIP balance.
- Last but not the least, we find out that WIP imbalance still appears after huge effort spent in achieving WIP balance purposefully, as WIP balance is relative and time dependent, and some events like unpredictable machine breakdowns are unavoidable in the fab. An effective detection and calibration method for WIP imbalance is vital to prevent small WIP imbalances from accumulating and becoming a serious problem. This WIP calibration procedure can enhance the intelligence of automatic manufacturing because it can be adapted and integrated into the current manufacturing systems.

Besides WIP, due date is another important KPI in wafer fab. For customer oriented enterprises, due date control is their major concern. On the shop floor, plenty of due date oriented rules have been utilized to increase on-time delivery and minimize tardiness. As due date control only focuses on processing lots toward on-time completion based on lot status, it tends to ignore the WIP situation in the fab, which turns out to cause excessive WIP and long production cycle times. On one hand, promising short lead time (cycle time) and delivery reliability, in reality, can attract more customers. On the other hand, delivery reliability cannot be guaranteed since due date control cannot always achieve short lead times for all products. We realize that as long as we want to gain a competitive position in the market by providing short lead times to customers, it seems due date control might be not fully satisfactory. If we seek help from WIP balance, it might lead to low WIP for the fab but without on-time completion pace.

In order to understand the relationship between WIP balance and due date control, Figure 1.3.2 shows hypothetical cycle time distribution of lot with the due date (zero tardiness) represented by the vertical axis. Figure 1.3.2 (a) represents the cycle time distribution of a dispatching methodology that ignores both WIP balance and due date control like FIFO. As we mentioned above, WIP balance can achieve cycle time reduction which has two cases. Figure 1.3.2 (b) shows that WIP balance results in low mean cycle time and variance which attempt to reduce tardiness. On the contrary, Figure 1.3.2 (d) presents an intention to minimize mean cycle time while allowing some lots to become quite tardy, which is mentioned above already that WIP balance sometimes sacrifices due date control. In figure 1.3.2 (c), when due date control is applied, it tends to finish the lots as close to the due date as possible, which results in a low mean cycle time and variance. Nevertheless, in Figure 1.3.2 (e), a low cycle time variance is achieved at the cost of an increased mean cycle time, which turns out that still a proportion of lots become tardy. It also tells us that due date



(a) FIFO dispatching

(b) Ideal WIP balance

(c) Ideal due date control

(d) WIP balance with excessive tardiness

(e) Due date control with increased mean cycle time

(f) WIP balance combined with due date control

Figure 1.3.2: Hypothetical cycle time distribution of lot for WIP balance and due date control.

control does not primary lead to low WIP levels. It would be perfect if the cycle time distribution can be acquired by the way shown in Figure 1.3.2 (b) and (c). Nevertheless, in reality, we always obtain the cycle time distribution presented in 1.3.2 (d) and (e) instead due to lot movement in poor pace. Since WIP balance and due date control have their pros and cons, the question comes naturally that whether WIP balance and due date control can be integrated to show complementary strengths, as shown in Figure 1.3.2 (f): a lower mean cycle time can be achieved by WIP balance while the tardiness is reduced as much as possible because of a low variance achieved by due date control. This is the second objective expected to be achieved in this dissertation.

1.4 Wafer Fabrication Facilities (Wafer Fabs)

In this section we will present a short introduction to wafer fabrication facilities (wafer fabs).

Semiconductor manufacturing is considered as one of the most complicated manufacturing processes. It consists of four basic phases: wafer fabrication, wafer probe, assembly and final testing [Fowler et al. 2002, Strum et al. 1999, Wein 1988]. The most technologically complex and expensive stage is the wafer fabrication, in which hundreds of circuits are built up through hundreds of operations on a silicon wafer to provide the required circuitry. In wafer probe, individual circuits are tested electrically by way of thin probes. Then the wafers are cut up into individual circuits and the failure circuits are discarded. In assembly, the circuits are mounted in plastic or ceramic packages to be protected from environment. The final test is used to detect whether the circuits is functional according to the required specification before shipping to the customer. In general wafer fabrication and probe are referred to front-end operations, while the assembly and final test are referred to back-end operations. One characteristic of the semiconductor manufacturing is the manufacturing cycle time is relatively longer than other manufacturing processes, for instance, many products need more than one month to be produced. Among those four process stages, wafer fabrication is the most complex and time consuming. Therefore, wafer fabrication is the first stage needed to reduce manufacturing time and improve performance, which exactly our research dedicates to.

There are many complexities that differentiate wafer fabs from traditional flow shops or job shops. We highlight the following two characteristics that make the production planning and scheduling difficult, especially for WIP balance and due date control.

(1). Re-entrant flow

There are many kinds of products in wafer fab. Each product has a unique process flow to follow until it is finished. Normally one process flow has hundreds of steps. A number of steps are repeated at the same production equipment. This is because wafer is manufactured layer by layer, some layers are produced in the same manner with variations like temperature, accuracy. Additionally, high capital investment requires some expensive machines like photolithography to perform different process steps. Therefore, products at different process stages visit the same machines many times, which is known as re-entrant flows. The following Figure 1.4.1 shows a typical re-entrant flow. Wafers are processed repeatedly from operation 2 to 15 for each layer. Re-entrant flow creates the need for WIP balance, because it brings a loop to the manufacturing line. Different products type and identical product type at different process stages compete for machines in the loop so that machines may be shared unequally and lose capacity.

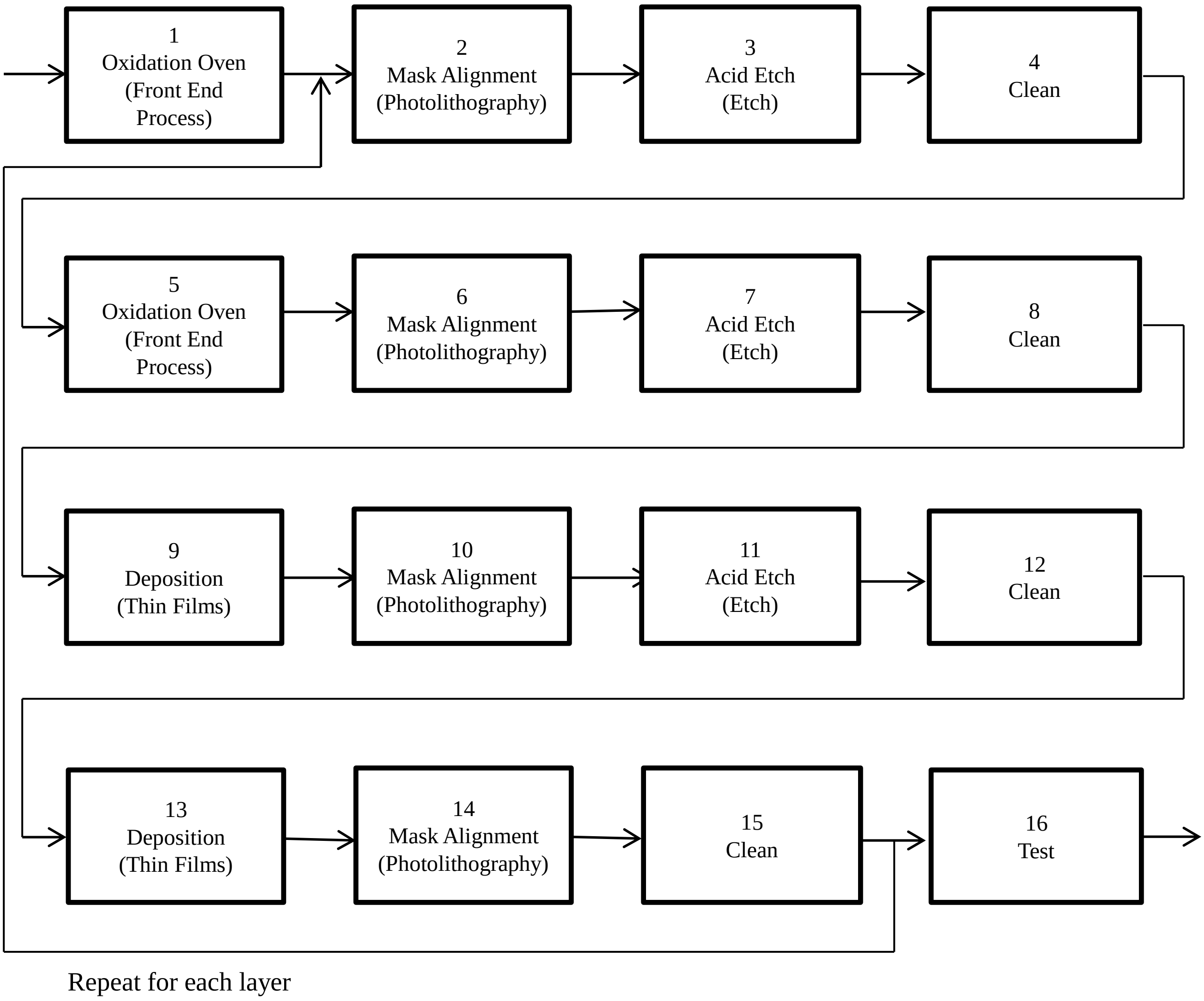


Figure 1.4.1: Typical re-entrant flow of wafer fab

(2). Diverse machine types and characteristics

There are different kinds of machines in wafer fabs. Some machines performing on single wafer or lot are referred to single processing machines, while some machines performing on groups of lots are referred to batch processing machines. There is also one kind of machine called cluster tool which is a subgroup of single processing machines allowing to process more than one wafer at a time at different chambers. Identical or similar machines are grouped together, which is normally called work-center (work station), to share interchangeable operations. The characteristics of machines differ widely. Some machines have sequence-dependent setup times, while some machines have sequence-independent setup times. The batch processing machines have different batch criteria, e.g., product type based, setup time based and process recipe based. The collection of lots to fulfill setup time requirement or form a batch leads to a non-smooth product flow causing WIP fluctuation. Besides that, the machines in wafer fabs are technologically extremely sophisticated and require preventive maintenance. They are all subject to unpredictable failures which is considered as the main cause of uncertainty in wafer fab. The breakdown of a critical machine like a bottleneck could result in an excessive WIP level and tardiness of lots.

Furthermore, as many enterprises change the manufacturing style from make-to-stock to make-to-order, the complexity of wafer fabs is even increased for the following reasons:

- Hundreds of products and corresponding process flows;
- Due to customer unique requirements and changeable orders, the product volume mix is uncertain and the lot release rate changes daily, weekly and monthly;
- Bottlenecks are changing frequently because of frequent product changes;
- Manufacturing processes are disturbed constantly by prioritized lots like engineering lots, qualification lots, test samples, especially delayed lots due to due date commitments.

Consequently, researchers have concluded that wafer fab is the most complex of all manufacturing environment [Fowler et al. 2002, Glassey and Resende 1988, Kumar 1993, Sze and Lee 1985] in which it is not easy to achieve WIP balance and due date control.

1.5 Structure of this Dissertation

This dissertation is organized as followings.

In Chapter 2, methodologies used to study wafer fabs are presented. Firstly some important performance indicators utilized to evaluate the performance of wafer fabs are presented. Secondly the current state of the art of operational control regarding to WIP and due date is available. It provides detailed literature review of WIP oriented rules and due date oriented rules, individually. Then it highlights the deficiencies in the current literature and demonstrates how exactly the work of this dissertation is able to address those deficiencies. Thirdly an introduction to the simulation model and software is available as well.

Chapter 3 attempts to give an insight into work-center oriented WIP balance and solve the related issues arising from it. Firstly, Section 3.1 presents a detailed WIP imbalance study of simulation model. Then work-center oriented WIP dispatching policies are developed to solve this WIP imbalance. Secondly, Section 3.2 shows how to improve the degraded cycle time variance arising from the WIP balance approaches in Section 3.1. Thirdly, Section 3.3 reports on the challenges of obtaining estimates of target WIP for WIP balance approaches. Fourthly, a full scale study of due date oriented rules in Section 3.4 is carried out to address the possible WIP imbalance caused by tight due dates and the corresponding strategies to handle.

Chapter 4 presents two achievements of this dissertation. The first one WIP balance for work-center without the need of target WIP is described in Section 4.1. It intends to use look-ahead and look-back strategies to replace target WIP by large set of information. Section 4.2 addresses the importance of WIP control by means of combining workload information of work-center with lot status information. Then a WIP detection and calibration approach without setting target WIP level is developed to smooth the material flow and prevent WIP curve from increasing.

Chapter 5 provides three simulation studies that combine WIP balance and due date control to show their complementary strengths and how to deal with the trade-off between them. This is another achievement of this dissertation.

The conclusion in Chapter 6 highlights the achievements of this dissertation and the contribution to science area. It contains future research ideas as well.

CHAPTER 2

METHODOLOGIES FOR WAFER FABS

Firstly several performance indicators considered as performance measures for simulation in the following chapters are introduced in Section 2.1. A detailed literature review regarding WIP and due date oriented rules is presented in Section 2.2. The simulation model and software used for simulation study are available in Section 2.3.

2.1 Important Performance Indicators in Wafer Fabs

The main objectives of production control in wafer fabs are on achieving shorter production cycle times and minimizing production costs while improving on-time delivery performance. They can be classified into two categories. The first one is WIP oriented performance measures such as short cycle time with low WIP level, WIP balance to achieve high utilization of work-center and decrease waiting time of lots. The latter one is due date oriented performance measures such as minimizing tardiness and increasing on-time delivery. Both operational control objectives are equally important, since low WIP levels can avoid to waste excessive resources and capital and short cycle time is a critical factor to respond the need of market, while on-time delivery is also a crucial factor to capture the market with due date commitment. We will describe the most important performance measures in the following.

2.1.1 Cycle Time

Cycle time (CT) is the total time required to produce a lot (wafer), from entering the fab to leaving the fab. From operation viewpoint, cycle time is the sum of time spent in at each operation (process step) c_i which is called operation cycle time.

$$CT = \sum_{i=1}^{n} c_i \qquad (2.1.1)$$

Each operation cycle time c_i includes the following components presented in Table 2.1.1.

Operation Cycle Time						
Transport Time	Queue Time			Process Time		
Transport Time	Batching Time	Queue Waiting Time	Setup Time	Load Time	Raw Processing Time	Unload Time

Table 2.1.1: Components of operation cycle time

Since the processing time is related to a physical or chemical process, it is the domain of process engineering. Thus, it is less important from the operational control consideration. As in most cases, the largest contributor to cycle time is the queue time which is the time a lot is waiting to be processed. The major contribution of WIP balance is to smooth the manufacturing process to reduce the queue time.

2.1.2 Cycle Time Variance

Cycle time variance is a measurement of how far a set of cycle time spreads out. For a set of cycle time *CT*, there is a mean value $\mu = E[CT]$, the variance of *CT* is given by:

$$Var(CT) = E[(CT-\mu)^2] \tag{2.1.2}$$

Cycle time variance is important as well because it tells us how cycle time distributes. A low variance indicates a precise prediction of production completion time. In particular, this is critical to customer oriented companies because they are able to provide an accurate lead time commitment to customers. As we mentioned above, WIP balance leads to cycle time reduction. However, sometimes WIP balance might bring a poor variance which results in excessive tardiness of some lots. This is the reason why due date control needs to be taken into consideration since due date control provides a mechanism to minimize variance.

2.1.3 Cycle Time Upper 95% Percentile

This performance measure provides a cycle time value below which 95% of the lots' cycle times fall. It is another important indicator for cycle time distribution.

2.1.4 Work-in-Process (WIP) and Throughput

WIP is the average number of wafers (lots) in the fab. The WIP includes wafers (lots) being processed in a work-center, as well as being transported or waiting in queue. Throughput is the average wafers (lots) can be manufactured per time unit in the fab.

According to Little’s Law [Little 1992] there is a relationship between the average of cycle time and WIP, as shown below:

$$Avg.Throughput = \frac{Avg.WIP}{Avg.CycleTime} \tag{2.1.3}$$

In other words, if the throughput is maintained to be constant, a reduction of WIP results in a reduction of cycle time.

$$Avg.CycleTime = \frac{Avg.WIP}{Avg.Throughput} \tag{2.1.4}$$

Although it is a mathematical formula, it tells us the way to reduce cycle time. In reality, it is difficult to maintain a constant throughput. However, maintaining a constant WIP level is a popular way for operational control. Using dispatching and scheduling strategies like WIP balance effectively improve the manufacturing process to increase throughput, so as to reduce the cycle time.

2.1.5 Due Date, Tardiness and On-time Delivery

In general, due date is the promised date to deliver the order to the customer. From operation control viewpoint, for those due date oriented rules due date means the date a lot has to finish processing and leave the fab. In this dissertation, due date of a lot is calculated as the release date of the lot plus the target cycle time. The target cycle time is calculated as the raw processing time

(RPT) multiplied by target due date flow factor (DDFF). Thus, the due date of a lot *i* is shown as follows:

$$D_i = R_i + RPT_i * DDFF \tag{2.1.5}$$

Where D_i is the due date of lot *i*, R_i is the release date of lot *i*, RPT_i is the raw processing time of lot *i*, *DDFF* is the target due date flow factor.

Along with development of due date control, a new due date concept called operation due date is raised [Bertrand 1983]. The operation due date is determined by dividing the interval between the lot final due date and its release date into as many segments as the number of operations. The operation due date of the final operation is equivalent to the lot due date. More detailed information about due date can be found in Section 3.4 of Chapter 3.

Once the due date is determined, we can define the average tardiness performance of tardy lots as follows.

$$Avg(Tar) = \frac{\sum_{i=1}^{n}(T_i - D_i)}{N} \tag{2.1.6}$$

Where T_i is the finish time of lot *i*, D_i is the due date of lot *i*, *N* is the number of tardy lots.

The on-time delivery performance also described as percent tardy lots is defined as follows.

$$OTD = \frac{N}{M} \tag{2.1.7}$$

Where *N* is the number of tardy lots, *M* is the number of finished lots.

2.2 Literature Review

During the past 30 years, a number of researchers have investigated the performance of various operational control policies for complex manufacturing facilities like the ones found in the semiconductor industry. We refer the interested reader to [Atherton and Atherton 1995, Panwalker and Iskandar 1977] for details. Different control policies have different performance objectives. Some rules target WIP balance for operations or work-centers which can lead to cycle time reduction. Some rules target due date control to achieve on-time delivery or at least minimal tardiness. While the WIP oriented rules do not always guarantee a good on-time delivery performance, the due date oriented rules do not primarily lead to low WIP levels.

2.2.1 WIP Oriented Release Rules

In contrast to due date oriented rules, WIP oriented rules focus on workload control [Fowler et al. 2002] which is a combination of lot release rules and dispatching/scheduling rules used to control how lots flow through work-centers to achieve WIP balance in the line. WIP oriented rules are typically global rules which utilize information not only from the local work-center where the dispatching decision is made, but also from upstream and downstream work-centers. Push and pull philosophies are two classical lot release rules for workload control. On one hand, the push rule is a make-to-order approach and originated from Material Requirements Planning (MRP) in the early 1970s

[Spearman and Zazanis 1992, Wight 1970]. The product (lot) release is based on shop throughput targets. The weakness of the push philosophy is that excessive WIP will cause considerable cycle times. On the other hand, the appearance of Japanese manufacturing techniques such as Just-In-Time (JIT) etc. supported the introduction of the pull philosophy in the early 1980s. With the pull philosophy product (lot) releases are based on the downstream shop status. A downstream work-center tries to pull a lot from an upstream work-center. The pull philosophy has been proven to lead to less WIP congestion and easier inventory control than the push philosophy [Spearman and Zazanis 1992]. Kanban and CONWIP (CONstant Work In Process) are two popular representatives of the pull philosophy. For the Kanban approach [Marek et al. 2001, Monden and Yasuhiro 1981], there is a card set between each pair of work-centers, and the total system WIP level is limited to the sum of the numbers of cards in all card sets. A lot is pulled by each work-center from the previous work-center only if the lot receives a card authorization. Kanban controls the WIP at the individual work-center level. In contrast to Kanban, CONWIP [Marek et al 2001, Spearman et al. 1990] only uses a single global set of cards to control the WIP level of the whole shop. Every lot seizes a card when it is released to the system for the first time. If all cards are taken by lots, a fresh lot expecting to enter the system has to wait until a lot leaves the system and the corresponding card is released. Kanban pulls lots between each pair of work-centers, while CONWIP pulls lots only at the beginning of the line. Recently there is a strong interest in CONWIP. Firstly, CONWIP is similar to an input/output control rule. It is easy to understand and robust to control only requiring understanding the relationship between WIP and throughput [Fowler et al. 2002]; Secondary, due to product mix changes, the bottleneck may shift over time. The Kanban approach needs to adjust the number of cards in each card set to avoid bottleneck starvation and make sure throughput. Therefore, the CONWIP approach is easier to manage because there is no tight WIP control between each pair of work-centers [Kalisch et al. 2008].

Due to the success of Kanban and the appearance of Theory of Constraint (TOC) [Goldratt 1984], the bottleneck oriented pull approach was developed. Wein [1988] introduced a Workload Regulation (WR) input approach for lot releases to the shop. For WR a target workload of the bottleneck has to be defined. If the actual workload of the bottleneck drops to the target workload, a new lot is released into the shop. Wein carried out a design of experiments which combines four lot release approaches (Poisson arrival, Constant arrival, CONWIP, and WR) with several dispatching rules. He found out that the effects of specific dispatching rules rely considerably on both the type of lot release approach and the number of bottlenecks in the shop. The WR approach is quite intuitive and only requires understanding the relationship between the target workload of the bottleneck and the system throughput. Therefore, it has been already widely adapted in real factory environments. However, setting the appropriate target workload is the core issue of WR.

Glassey and Resende [1988] presented another well-known bottleneck oriented lot release approach called Starvation Avoidance (SA). They defined a virtual inventory of the bottleneck which is used as a measure to keep a proper inventory level at the bottleneck. The virtual inventory includes the total bottleneck processing time of the next operations of all lots which reach the bottleneck work-center within a given lead time plus the expected time to repair the bottleneck machines which are currently broken down. The lead time is the sum of the processing times of all lots required to arrive at the bottleneck the first time after their release. Glassey and Resende compared the SA rule with three other lot release approaches (Uniform arrival, WR, and CONWIP). They concluded that SA is more effective than the other lot release approaches concerning near-capacity throughput while maintaining lower average lot delays. However, compared to the WR approach, the SA approach requires more conceptual understanding and considerable implementation effort because it requires global inventory information about the whole shop.

Rose [1999] developed another promising bottleneck oriented lot release rule called CONstant Load (CONLOAD). This rule aims at overcoming some performance problems of traditional lot release lot rules like CONWIP and WR during product mix changes. In contrast to CONWIP and WR, CONLOAD takes into consideration how much load is added to a single machine or a group of machines by a particular lot to decide on releasing this lot into the fab or not. Rose concluded that CONLOAD outperforms CONWIP and WR with regard to keeping the bottleneck utilization at a desired level and providing a smooth WIP evolution curve.

2.2.2 WIP Oriented Dispatching Rules

Although some researchers claimed that the lot release approach is more important regarding workload control than dispatching [Glassey and Resende 1988, Wein 1988], there is no doubt that dispatching is still a powerful way to assist or improve the workload control, because dispatching approaches have low computational requirements and an intuitive appeal. In addition, they can be used to avoid machine starvation and they can handle re-entrant flows to effectively balance the line.

A promising WIP oriented dispatching rule named Minimum Inventory Variability Scheduling (MIVS) was proposed by Li et al. [1996]. MIVS considers both upstream and downstream operations, and tries to keep the WIP of each operation close to an average target WIP level. It gives higher priority to an operation which has a high WIP level while its downstream operation has a low WIP level to avoid starvation at downstream operations. In contrast, it gives the lower priority to an operation which has a low WIP level while its downstream operation has a high WIP level. MIVS succeeds in adapting to the nature of re-entrant flows and in reducing the WIP imbalance through pulling lots into low WIP operations, the results are reduced WIP variability and reduced cycle times. Similarly, *K*-step ahead and *J*-step back MIVS was also developed. In a real application, the *K* and *J* are not fixed and depend on fab status like machine availabilities. Collins and Palmeri [1997] compared 1-step ahead MIVS with *K*-steps ahead MIVS and concluded that there is no obvious evidence that 3-step ahead MIVS outperforms 1-step ahead MIVS.

Similar to MIVS, Dabbas and Fowler [2003] proposed a global Line Balance (LB) algorithm with the objective of minimizing the deviations of actual WIP to target WIP for each operation. Through calculating throughput signals, cumulative signals, and unconstrained quantities, LB determines portions of WIP at all operation stages required to be pushed forward to balance the downstream operations. The main novelty and contribution of this approach is that the authors considered LB as a global dispatching approach combined with several local dispatching rules such as CR, Flow Control (FC) and Throughput (TP) into a single rule, with the objective of optimizing different performance measures simultaneously. Defining an appropriate target WIP level is the key issue of applying MIVS and LB.

Unlike MIVS and LB considering WIP balance from operation viewpoint, there is another research direction which considers WIP balance for work-centers. Ham and Fowler [2007] proposed a Balanced Machine Workload (BMW) dispatching scheme in order to overcome the potential weakness of MIVS. By calculating the workload of each machine and corresponding workload of *K*-downstream and *J*-upstream operations, the BMW is able to choose the best operation to maintain the machine workload in a balanced state. Zhou and Rose [2010] proposed a WIP Control Table for each work-center to keep the actual WIP level close to target WIP level. They addressed that WIP balance for work-center lead to cycle time reduction, however, with regard to lower cycle time variance taking WIP balance for operation into account is very necessary. Leachman et al. [2002] introduced a comprehensive WIP management project called SLIM for whole wafer fabs in Samsung Electronics Corp., Ltd. SLIM is a set of methodologies and scheduling application for managing WIP and cycle time. In particular SLIM schedules non-bottleneck, bottleneck and diffusion furnaces work-centers considering upstream and downstream information.

Perdean et al. [2008] proposed an interesting dispatching concept combining a push policy (first in first out) and a pull policy (shortest remaining processing time) together via a push-pull point (PPP) to control a typical re-entrant manufacturing line, with the objective to reduce the mismatch between the daily output and demand. The novelty of this approach which has not been considered in the literature before is to introduce the PPP to divide the line where push policy is applied in the upstream of PPP and pull policy is employed in the downstream of PPP. Through simulation experiments, they found out that when the PPP control works together with CONWIP release policy, significant improvements were obtained for high demands with high variances compared to pure pull policy or pure push policy, or CONWIP combined with pure pull policy.

2.2.3 Other WIP Related Research Directions

Several WIP related research directions have emerged along with the development of workload control. The first one, most of the workload control rules above only focus on reducing cycle time and increasing throughput but seldom address the importance of cycle time variance minimization. Some researchers only concentrate on the cycle time variance minimization of single machine cases [Gupta et al. 2009, Ventura and Weng 1972], only a few papers apply it in job shops like wafer fabs [Kuo et al. 2008, Lu et al. 1994].

The second one, as a critical factor to WR, MIVS and LB mentioned above, is the target WIP level determination. Since the performances of those WIP oriented rules highly rely on target WIP, how to acquire appropriate target WIP levels attracts more and more attentions. In general, as shown in previous studies [Burman et al. 1986, Dabbas and Fowler 2003, Lee and Kim 2002, Li et al. 1996, Lin and Lee 2001, Pai 2004], using simulation models or queuing models is a popular way to estimate target WIP levels. Due to the difficulties in developing algorithms and statistical models, neural networks trained with observed data attracted attention recently [Huang et al. 1999]. Kuo et al. [2008] proposed a back-propagation neural network model to determine the target WIP level for the bottleneck and non-bottleneck work-centers instead of a queuing model with the purpose to guarantee a maximum throughput of the bottleneck while achieving a minimum WIP level.

When setups, batches or unusual events caused by machine breakdowns and hot lots etc. are involved, additional challenges of workload control are imposed. Some high level workload control rules that consider upstream and downstream shop status information are the examples to deal with setups and batches in wafer fabs [Glassey and Weng 1991, Kim et al. 2008, Robinson et al. 1995]. Some approaches focus on smoothing manufacturing processes by means of WIP imbalance monitor and correction to minimize the impact of the production variations, with the objective to enhance the intelligence of automatic manufacturing [Guo et al. 2007, Yeh et al. 2008].

2.2.4 Single Due Date Oriented Dispatching Rules

When the performance objective involves meeting a given due date, due date oriented dispatching rules are generally employed to minimize the proportion of tardy lots, mean tardiness of tardy lots, and the like. They can be categorized into static rules like Earliest Due Date (EDD) and dynamic rules like Least Slack Time (LST), Critical Ratio (CR), Operation Due Date (ODD). The EDD rule aims at meeting the due date, and gives the highest priority to the lot which has the earliest due date. LST and CR are variants of EDD. Besides the due date information, LST and CR consider the remaining raw processing time of a lot as well. The LST rule calculates the slack for each lot as: *Slack = Due – Now – RemainingRPT*, where Due is the due date of a lot, Now is the current time, and RemainingRPT denotes the remaining raw processing time. The lot with the smallest slack is favored. LST is an extension to EDD for the reason that it tells us if two lots have the same due date, the lot with longer remaining raw processing time is more urgent because its due date allows less delay. The CR rule distinguishes lot urgency by a ratio between remaining time to the due date and remaining raw processing time, *Critical Ratio = (Due - Now) /*

RemainingRPT, instead of computing a difference like LST. A CR value of less than 1 denotes a lot which falls behind schedule; a CR value equivalent to 1 means that a lot is on schedule, a CR value of greater than 1 represents a lot which is ahead of schedule and has slack time left. CR assigns the highest priority to the lot with the smallest CR value. Baker and Bertrand [Baker and Bertrand 1981] presented a simulation study of combining due date assignment rules with due date oriented dispatching rules. Considering minimizing Mean Tardiness as performance measure, they concluded that compared with SLT and CR, SPT (Shortest Processing Time) is effective with tight target due dates and EDD is superior with loose target due dates. While considering Conditional Mean Tardiness as performance measure, Muhlemann et al. [1982] found out that CR outperforms EDD and LST. Rose [2002] presented a detailed study of the CR rule, and showed that CR leads to sudden performance degradation when the target due date is too tight. This issue arises because CR only focuses on the final due date and speeds up lots which are close to due date or already late. In contrast, the fresh lots run out of their slack time and have to wait in early operations.

The ODD rule [Bertrand 1983, Rose 2003] succeeds to avoid the above problems of CR. ODD breaks up the slack time into as many segments as the number of operations of a lot, which means ODD considers due dates for all intermediate operations, unlike CR which only considers due date of the final processing operation. The ODD value of operation *i* is defined as: *ODD* = *Release Time* + *RPT(i)* * *DDFF* where RPT(*i*) denotes the RPT for a sequence of processing steps or operations from operation 1 to operation *i* (including operation *i*) and DDFF denotes the target due date flow factor which is the ratio of target cycle time and raw processing time of a lot. The ODD rule gives priority to the lot with the smallest ODD value. For the final operation of a lot the ODD is equal to the classical due date as it is used for CR. Because slack times for young lots assigned by the ODD rule are smaller than in the CR case. Therefore they do not have to let old lots pass before they are processed. As a consequence, it is not possible with the ODD rule that problems at operations at the end of the processing sequence propagate back to the operations at the beginning. Once the operation due date have been established, the lots are strictly kept at the right pace to meet their due date through the factory from the early operations on. Thus, the ODD rule is able to minimize the variance of lot lateness relative to the due date and typically also leads to a low cycle time variance.

Most of these due date oriented dispatching rules above are local rules. They only focus on the information of lots which wait in the local work-center buffer instead of taking into account information from elsewhere in the shop, e.g., machines failures, machine utilizations, etc. Furthermore, they only focus on on-time delivery and tardiness performances. Thereby, sometimes they are incapable to handle WIP imbalances because of multiple re-entrant flows, machines breakdowns, etc. Consequently, the shop runs at a high WIP level with considerable cycle times.

2.2.5 Composite Due Date Oriented Dispatching Rules

The performances of these due date oriented dispatching rules are mainly affected by how tight or loose the due date is set [Elvers 1973]. Some rules perform better with tight target due dates like SPT, although SPT does not use any due date information, while some rules perform better with loose target due dates like CR and ODD. By noticing the complementary strengths of different rules working with different target due dates, Baker and Bertrant [1981] proposed a composite rule called Modified Due Date (MDD) which is combination of EDD and Least Work Remaining (LWKR). The MDD is defined as: *MDD = Max (Due, t + RemainingRPT)* where Due is the lot due date, t is current time and RemainingRPT is the remaining raw processing time of the lot. As an extension, Baker and Kanet [1983] presented another composite MOD rule which is a combination of SPT and ODD. It performs like SPT if the target due date is tight and like ODD if the target due date is loose. They assumed that the MOD (operation-based due date and SPT) is more effective than MDD (lot-based due date and LWKR). For each lot in the queue of a work-center at time t MOD is calculated in the following way: *MOD = Max (ODD, t + PT)* where ODD is the operation due date of the lot at work-center, t is current time and PT is the processing time of the lot at the work-center. The MOD rule gives priority to the lot with the smallest value of MOD. It tends to combine the advantages of SPT and ODD and provides short cycle times and minimizes cycle time variance while working with different target due dates simultaneously.

There is another composite rule called Apparent Tardiness Cost heuristic (ATC) [Vepsalainen and Morton 1987]. The ATC rule combines the Weighted Shortest Processing Time (WSPT) rule and the LST rule. There are two characteristics of this rule. Firstly, apart from processing time, the ATC rule utilizes a look-ahead strategy and takes waiting time estimates of lots on downstream work-centers into consideration to calculate the slack time of each operation. Secondly, the ATC rule uses an exponential decay function to calculate the weight/processing time to allocate priorities to lots. The simulation results demonstrate that the ATC rule outperforms other due date oriented dispatching rules with regard to minimization of weighted tardiness penalties.

However, there are several user-defined parameters in the ATC rule. The application and accuracy of ATC rule depend considerably on defining appropriate parameters. The Apparent Tardiness Cost with Setups (ATCS) [Lee et al. 1997] is an extension of ATC rule. ATCS rule includes setup avoidance into ATC rule to evaluate the tradeoffs that exist when trying to sequence lots that have due dates and priorities on machines that require sequence-dependent setups.

2.2.6 Conclusion

In this section, we draw a conclusion regarding to the pros and cons of WIP oriented rules and due date oriented rules in literature, as described in Table 2.2.1.

We realize that WIP oriented rules show weaknesses when due date is involved, and due date oriented rules show deficiencies when low WIP level is desired. The root cause is WIP and due date are two different goals. No literature addresses the possibility of taking these two goals into consideration. The main focus of this dissertation is to demonstrate possible ways to achieve both goals concurrently.

In addition, as an important factor ‘target WIP’ attracts more and more attention in shop floor control. The main controversial point is a misleading target WIP applied in the SA, MIVS and LB approaches causes performance degradation to wafer fabs. Thus, another focus of this dissertation is to search alternative ways to replace the target WIP, i.e. achieving WIP balance without the need of target WIP.

Operational Control Rule		Feature	Disadvantage
WIP oriented	Operation oriented (MIVS, LB)	● Reduce WIP variability ● Reduce cycle time ● Increase throughput	● Lot movement with poor pace ● No due date involved
	Work-center oriented (SA, BMW)	● Avoid congestion and starvation ● Increase utilization	
Due date oriented	EDD, CR, ODD…	● Achieve right pace of lot movement toward on-time completion	● No low WIP level guaranteed ● Sacrifice machine utilization

Table 2.2.1 The pros and cons of WIP oriented and due date oriented rules

Literature also demonstrated that target WIP can be utilized to monitor and correct WIP imbalance to reduce the impact on material flow. Actually, analyzing WIP positions dynamically can help us to smooth the manufacturing flow even without the need of target WIP, which will be addressed in this dissertation as well.

2.3 Simulation Model and Software

In this dissertation the wafer fabs dataset MIMAC6 (Measurement and Improvement of MAnufacturing Capacities) is used for simulation. MIMAC6 is a typical 200mm wafer fabs model of ASIC. It includes actual manufacturing data from real wafer fabs that were organized into a standard format. It contains minimum information necessary to model a fab, including product information, process flow, process time, rework routing, machine availability and so on. The following list gives an overview of the main characteristics of MIMAC6 model. This model is available via anonymous ftp from 'ftp.sematech.org' in directory '/pub/datasets'. For further detail about this model, the interested readers are referred to [Fowler and Robinson 1995].

- Product profile:
  - 9 products, 24 wafers of one lot size;
  - Avg. Steps/mask layers: 30;
  - Max. static capacity: 2777 lots released per year (approx. 5,554 wafers per month);
  - Table 2.3.1 shows the raw processing time of each product.
- Process flow:
  - 9 process flows, max. 355 process steps;
  - Avg. line yield: 93%.
- Tool group and operator:
  - 104 tool groups (work-centers), 228 tools (machines);
  - 46 single processing tool groups, 58 batch processing tool groups;
  - Each tool group have different dispatching rules;
  - 9 operator groups.
- Availability:
  - Failures: information about mean time between failures (MTBF) and mean time to repair (MTTR), clock-time-based exponential distribution is modeled;
  - Avg. downtime per tool: 13.6%;

  - All downtimes are modeled as non-preemptive.

- Process time:
  - Constant per wafer, per lot or per batch process times;
  - Load and unload times;
  - Transport times are not modeled;
  - Extra delay times are not modeled.

- Setup and batch:
  - 6 tool groups have setup requirements;
  - 86 different setup IDs to model setup group;
  - Setup avoidance is modeled; (For tool groups with a priority-based dispatch rule, This strategy first finds the highest priority class of lots waiting for service among which there is at least one lot ready to run. Within this highest priority class with ready lots, this strategy selects the lot that minimizes setup time. If more than one lot minimizes setup, the lots are ordered by the rest of their nominal dispatch rule, and the first lot is selected.)
  - 197 Different batch IDs to form batches, using minimum and maximum batch size.

Product	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

Table 2.3.1: Basic information of the products in MIMAC6 model

We run the simulation with FX (Factory eXplorer, version 2.10.0.4) from WWK [www1], a commercial simulation package for factory models. The proposed WIP balance and due date control approaches are not provided by the FX simulation package, but FX supports customization via a set of user-supplied code and dispatch rules. We use a customized interface developed by René Wolf (2008) to develop our approaches and control the operation of FX. This interface implemented in C++ language consists of four modules which are User, Data Model, Utilities and Dispatching Rules, as presented in Figure 2.3.1. The following is the overview of this interface.

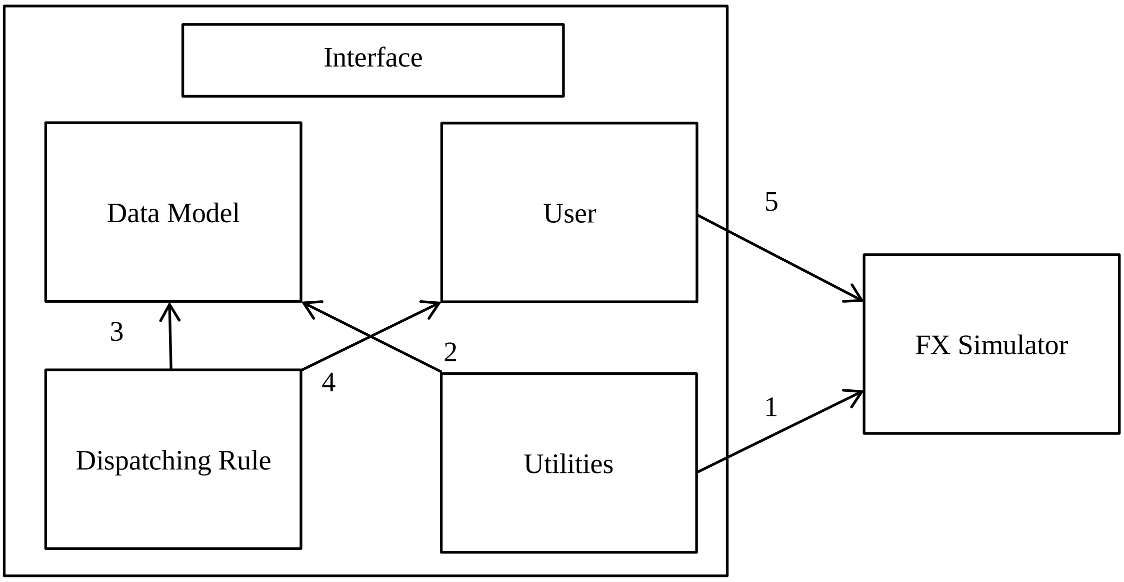
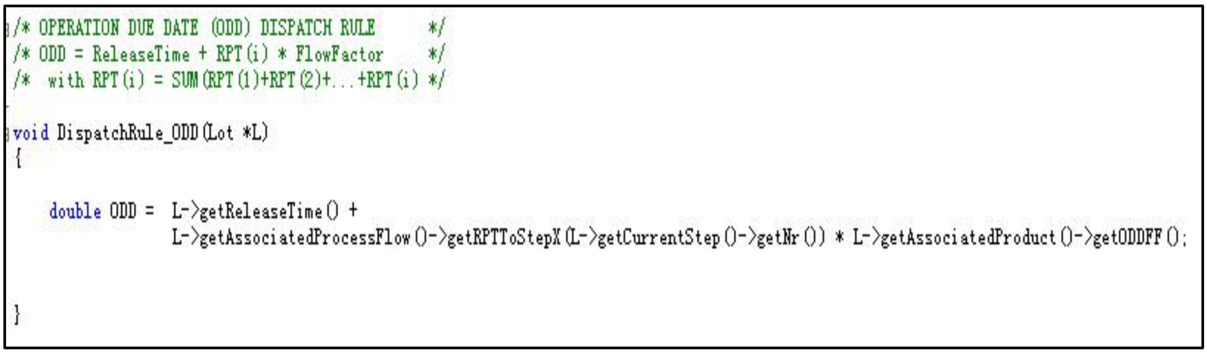


Figure 2.3.1: Basic communication between the interface and FX simulator

- User: is the user-supplied dynamic link library routines. It provides critical functions like loading the simulation model, selecting lot for tool group and so on. It also provides a number of discrete events to interrupt and interact with FX.

- Data Model: is used to store the simulation model during simulation. All the information regarding to product, lot, tool group, process flow and process step, is accessible via this data model.
- Utilities: provides a number of useful functions, e.g. obtaining a reference to the simulation model, and the parameters in the command windows of FX and so on.
- Dispatching Rule: is the module where we develop the operational control policies (WIP balance and due date control rules).

An example about the operation due date (ODD, in Section 3.4 of Chapter 3) implementation and the interaction between this interface and FX is presented as follows. First of all, we implement the ODD rule as Equation in the Dispatching Rule module as illustrated in Figure 2.3.2. During simulation, when a tool becomes available and needs a lot to process, FX will carry out the following procedures to calculate the ODD priority of lot.



```
/* OPERATION DUE DATE (ODD) DISPATCH RULE      */
/* ODD = ReleaseTime + RPT(i) * FlowFactor     */
/*  with RPT(i) = SUM(RPT(1)+RPT(2)+...+RPT(i) */

void DispatchRule_ODD(Lot *L)
{

    double ODD =  L->getReleaseTime() +
                  L->getAssociatedProcessFlow()->getRPTToStepX(L->getCurrentStep()->getNr()) * L->getAssociatedProduct()->getODDFF();


}
```

Figure 2.3.2 ODD implementation in C++ language in Dispatching Rule module

- The Utilities module makes a reference to the simulation model and store it in the Data Model module via steps 1 and 2 in Figure 2.3.1;

- Then the Dispatching Rule module will get access to the Data Model module, and obtain the necessary information e.g. the release date, the raw processing time and due date flow factor, to calculate the ODD value (shown in Figure 2.3.2) via step 3;
- Finally, the Dispatching Rule module will pass the ODD value to the FX simulator via User module, as depicted in steps 4 and 5. FX uses this ODD value as the priority of lot and chooses the lot with highest priority to the available tool.

In this dissertation, all the simulations of MIMAC6 are carried out for 18 months with 3 replications. (Due to large amount of data transfer between the interface and FX simulator, each replication takes approximately 20~25 minutes under 95% fab loading case (computing environment: Windows 7, Intel core 2 Duo CPU 2.4 GHz, 2G RAM)). The first 6 months were considered as warm-up periods, and were not taken into account for statistic.

CHAPTER 3

WORK-CENTER ORIENTED WIP BALANCE

Firstly Section 3.1 gives an in-depth analysis of the symptoms in a wafer fab (MIMAC6 model) when WIP imbalance occurs, so as to address the necessity to balance WIP for operations or work-centers. Then we introduce the MIVS rule in detail and address the key issues applying MIVS. Based on the observation of MIVS, we propose an alternative for operation oriented WIP balance that is work-center oriented WIP balance.

We intend to solve two issues arising from the proposed work-center oriented WIP balance. The first one is fast but poor pace lot movement, in Section 3.2 we propose three methods to improve the pace of lot movement to reduce cycle time variance. The second issue, how to determine the target WIP for work-center, will be answered in Section 3.3. One fast and effective way is to use the average WIP of work-centers from simulation experiments with FIFO dispatching. However, a standard procedure like applying queuing models and neural networks should also be taken into account.

Besides that, in Section 3.4 we conduct a comprehensive study of due date oriented rules. Typically, we focus on the cycle time variance minimization effect and the WIP imbalance phenomenon caused by tight due dates, as we expect to find out the complementary effect of due date control to WIP balance.

3.1 Work-center Oriented WIP Balance

3.1.1 WIP Imbalance Symptoms in Wafer Fabs (MIMAC 6 Model)

To begin this chapter, we investigate the performances of the MIMAC6 model under 95% fab capacity loading applying different dispatching rules which utilize different types of information to assign priorities to lots. We intend to find out whether the WIP imbalances caused by the investigated rules have something in common. Four rules, which are first in first out (FIFO) using lot arrival time information, shortest processing time (SPT) using lot processing time information, critical ratio (CR with a tight due date flow factor of 1.5) using lot due date information and least work at next queue (LWNQ) using workload information of work-centers, are applied.

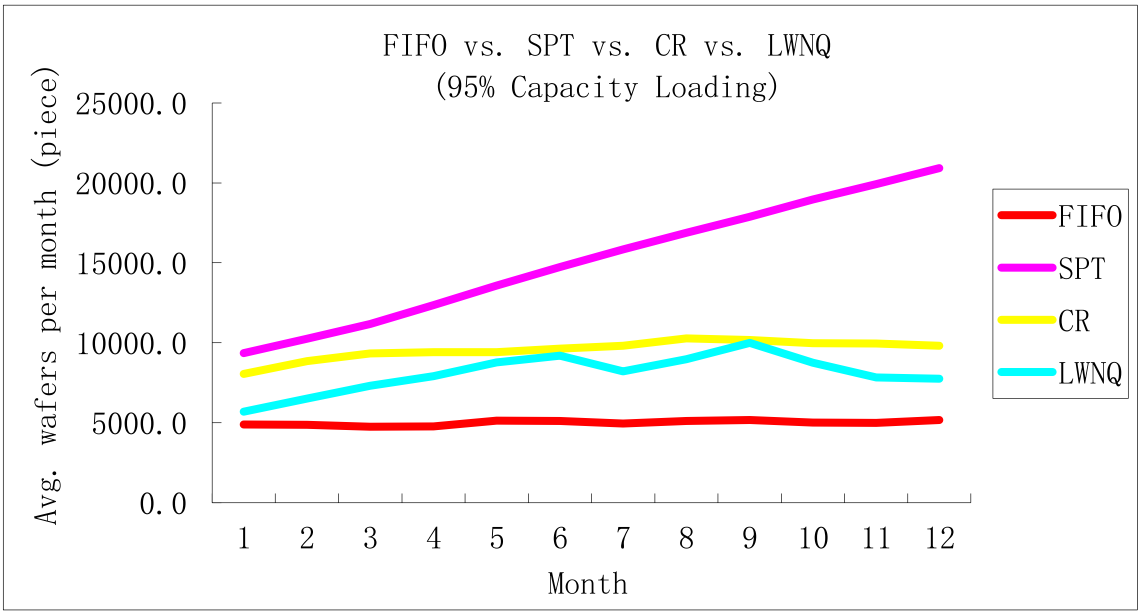


Figure 3.1.1: Average WIP comparison among four rules

From Figure 3.1.1, we notice that the WIP curve of SPT rule is increasing over time, and the ones of CR and LWNQ rules are relative higher than in the case of FIFO rule. It is no doubt that SPT, CR and LWNQ achieve relative higher average cycle time than FIFO. We only have a global picture that the WIP coming from SPT, CR and LWNQ is relatively imbalanced than FIFO. What happened inside the fab? According to the Theory of Constraints (TOC) [Goldratt 1984], the critical resources suffer more from WIP imbalance than the non-critical resources. Thus, we look deeper inside the WIP situation of the first three highest utilized work-centers in MIMAC6 model in Table 3.1.1. For the work-centers ‘20540_CAN_0.43_MII’ and ‘12553_POSI_GP’, these four rules have different performances. Whereas, we can see a clear picture that SPT, CR and LWNQ rules cause extremely long queues and high queue times to work-center ‘11026_ASM_B2’, which is the biggest problem leading to considerable WIP levels in Figure 3.1.1.

As a work-center can perform different operations, we are interested in finding out what happened inside ‘11026_ASM_B2’ by further detailed analysis. ‘11026_ASM_B2’ can process 8 operations with 4 batch IDs. Table 3.1.2 lists the average WIP and average queue delayed of 8 operations in ‘11026_ASM_B2’. We realize that the WIP of the SPT, CR and LWNQ rules is very high in operation 10701 and 12701. CR and LWNQ rules behave slightly different from SPT and produce relative high WIP to other operations as well, e.g., operation 13711. As a consequence, from the operation viewpoint, the long queue of ‘11026_ASM_B2’ can be represented as the high WIP of operations 10701, 12701 and 13711 caused by SPT, CR and LWNQ rules.

Although SPT, CR and LWNQ rules dispatch lots via different mechanisms, which means the reason causing WIP imbalance might be different, they show the common effect that they cause congestion for critical work-centers or operations, specifically for the ‘11026_ASM_B2’ or operations 10701, 12701

Work-center		***Avg. WIP (wafers)***	***Avg. queue delayed (hours)***
20540_CAN_0.43_MII	***FIFO***	256.5	1.7
	SPT	132.1	0.4
	CR	311.1	2.2
	LWNQ	146.2	0.4
12553_POSI_GP	***FIFO***	149.4	8.1
	SPT	76.0	4.5
	CR	184.3	9.9
	LWNQ	95.2	5.3
11026_ASM_B2	***FIFO***	641.6	27.4
	SPT	12517.3	582.9
	CR	5098.1	240.6
	LWNQ	3442.0	165.4

Table 3.1.1: WIP comparison of the first three highest utilized work-centers among four rules

and 13711. Here, it is of less interest to explore the detailed reason why they show such a behavior, instead, we prefer to seek a solution for it. Simply speaking, CR, SPT and LWNQ rules dispatch lots regardless of the workload situation of downstream (although LWNQ dispatches based on the workload of work-center, this effect is rather limited). If we do not consider the characteristics of '11026_ASM_B2', for example, there is only one machine and the processing time is long, one basic fact observed above is although '11026_ASM_B2' is already highly loaded with a long queue, these three rules still send lots to it, which results in an extremely long queue time.

- The situation would be different if these rules "knew" that they should stop sending lots to '11026_ASM_B2' because they are aware that '11026_ASM_B2' is already overloaded. Another alternative is that they stop sending lots to operations 10701, 12701 and 13711, because these three operations already have high WIP levels. The challenge is by what criteria they can judge that either ‘11026_ASM_B2’ or operations 10701, 12701 and 13711 have high WIP levels.

Our assumption brings in two different WIP balance scenarios that are work-center oriented and operation oriented, but with the same mechanism of applying target WIP for work-centers and operations, which are described in the following sections.

Operations		***Avg. WIP (wafers)***	***Avg. queue delayed (hours)***
10701	***FIFO***	22.4	1.2
	SPT	1158.8	85.2
	CR	537.9	30.5
	LWNQ	422.4	26.7
12701	***FIFO***	26.0	1.2
	SPT	1174.8	72.7
	CR	558.6	25.8
	LWNQ	487.5	21.6
12811	***FIFO***	33.4	1.2
	SPT	7.4	0.3
	CR	5.8	0.2
	LWNQ	45.3	1.4
13711	***FIFO***	16.6	1.0
	SPT	1.6	0.1
	CR	387.9	25.5
	LWNQ	326.2	21.3
16001	***FIFO***	20.0	0.9
	SPT	4.0	0.2
	CR	2.3	0.1
	LWNQ	45.4	2.1
21801	***FIFO***	3.0	0.2
	SPT	0.2	0.02

	CR	23.7	1.9
	LWNQ	12.2	0.8
25561	***FIFO***	16.8	0.9
	SPT	1.4	0.1
	CR	2.5	0.1
	LWNQ	25.9	1.8
32601	***FIFO***	32.3	1.0
	SPT	2.9	0.1
	CR	13.8	0.4
	LWNQ	66.7	2.1

Table 3.1.2: WIP comparison of 9 operations among four rules in work-center ‘11026_ASM_B2’

3.1.2 Minimum Inventory Variability Scheduling (MIVS) - Operation Oriented WIP Balance

Before starting with WIP balance for work-centers, the classic and famous rule, minimum inventory variability scheduling (MIVS), is introduced. MIVS considers the WIP flow from the standpoint of operations and a number of researchers claim that managing WIP from operation viewpoint is beneficial since it is quite straight forward to know the distribution of WIP in operations which is represented by WIP distribution histogram illustrated in Figure 3.1.2. One benefit gained from the WIP distribution histogram is, we are aware that in order to balance WIP pulling WIP from high WIP operation to low WIP operation is very necessary, which is easy to understand and causes no complicated computation effort.

The MIVS rule is a representative approach to balance the WIP of operations. MIVS considers both upstream operation and downstream operation, and tries to keep the WIP of each operation close to the average target WIP level. Intuitively, the upstream operation with a higher-than-average WIP level should have a higher priority than the one with lower-than-average WIP level. In the meantime, the downstream operation with a lower-than-average WIP level should have a higher priority than the one with higher-than-average WIP level. As a consequence, it leads to a combination of four different priorities described in Table 3.1.3. Actually, the target WIP in Figure 3.1.2 plays a critical role in regulating the WIP flow. The dynamic WIP will keep close to the long term average WIP (target WIP), which results in WIP variability reduction, e.g., avoiding starvation or congestion at some specific operations [Li et al. 1996]. Thus, the serious problem presented by SPT, CR and LWNQ rules in the previous section can be avoided.

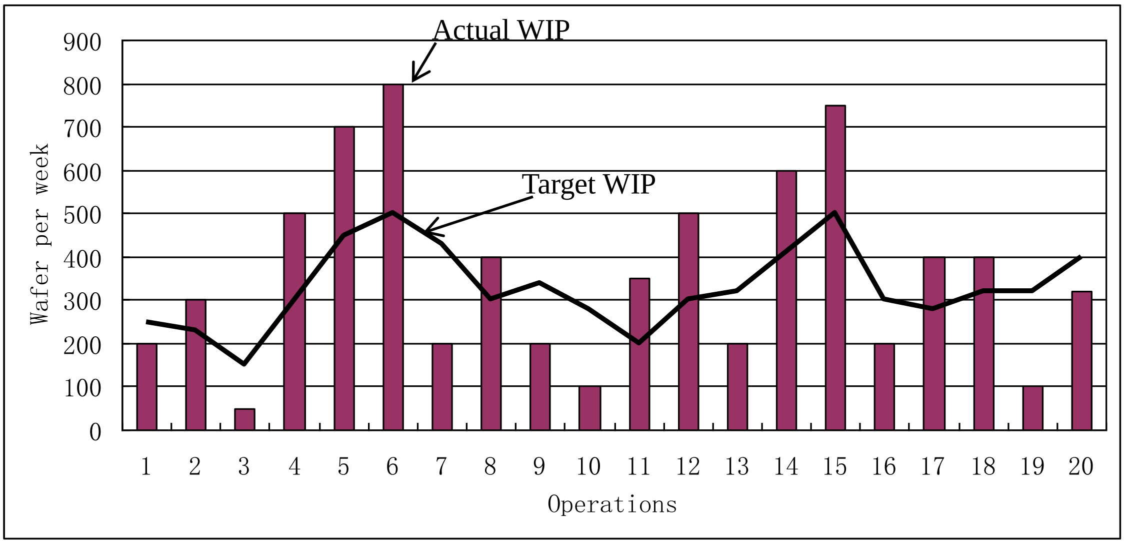


Figure 3.1.2. WIP distribution histogram in operations

In Figure 3.1.2, suppose a work-center can process lots in operations 2, 9, 13 and 17 and the downstream operations are 3, 10, 14 and 18, respectively. According to MIVS rule, operation 2 has the highest priority 1, operation 17 has priority 2, operation 9 has priority 3 and operation 13 has the lowest priority 4.

		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

Table 3.1.3: The principle of MIVS rule

We simply apply the MIVS rule to the MIMAC6 model and consider average cycle time, cycle time variance and cycle time upper 95% percentile as performance measures to compare with FIFO rule. The results are presented in Table 3.1.4. The average WIP level for each operation from FIFO is used as the target average WIP levels for MIVS.

		Fab loading (%)		
		75	***85***	***95***
Average Cycle Time (days)	***MIVS***	20.4	23.1	28.5
	FIFO	20.5	23.3	29.6
Cycle Time Variance (days^2)	***MIVS***	1.3	2.0	1.8
	FIFO	0.9	1.2	1.6
Cycle Time Upper 95% Percentile (days)	***MIVS***	25.4	29.8	37.2
	FIFO	26.2	29.8	39.1

Table 3.1.4: Performance measures comparison between MIVS and FIFO

We can see the benefit of WIP balance that MIVS rule is superior to FIFO with regard to average cycle time and cycle time upper 95% percentile for different loading cases. The MIVS rule succeeds in reducing line imbalance by pulling WIP from high WIP operations to low WIP operations, thus reducing WIP variability to achieve an average cycle time reduction. With respect to the cycle time variance, MIVS provides no significant improvement in comparison to FIFO. This is understandable because MIVS rule does not provide a mechanism to keep lots going through the fab at the right pace. This is one of the potential drawbacks of WIP balance (as described in Figure 1.3.2 (d), P.13) which can result in excessive tardiness for some lots and needs to be overcome in this dissertation.

3.1.3 Minimum Workload Variability Scheduling (MWVS) – Work-center Oriented WIP Balance

In Section 3.1.2 MIVS employs upstream and downstream WIP information (actual WIP and target WIP) to balance WIP effectively. We realize that WIP variability reduction for operations is the core concern of MIVS. As a matter of fact, reducing variability is a crucial factor affecting a manufacturing system [Hopp and Spearman 2011]. What we observed in Section 3.1.1 is long queues in front of some critical work-centers caused by irregular WIP flow, some work-centers are overloaded while some are starved. The average cycle time increases due to the congestion and starvation of work-center even though the WIP remains approximately at the same level [Hopp and Spearman 2011, Li 1991, Li 1993, Tang 1993]. Thus, if we consider the WIP flow from the viewpoint of work-centers, similar to the MIVS rule, we can apply the same idea to reduce WIP variability for work-centers. Inspired by that, we propose a work-center oriented WIP balance approach named Minimum Workload

Variability Scheduling (MWVS) that intends to avoid starvation and congestion of work-centers, which directly avoids capacity losses and reduces queue times.

Another issue dealt with in this section is the estimation of target WIP levels. The MIVS rule tries to minimize the deviation between the actual WIP and target WIP to reduce WIP variation. Similarly, MWVS rule also predefines target WIP levels for work-centers. A work-center is detected as starved or overloaded by means of its target WIP level. However, setting appropriate target WIP levels is a challenging task, which motivates us to develop an alternative to avoid the need of target WIP levels. For this reason, a variant of MWVS rule called one-step-ahead and one-step-back MWVS rule is proposed. Through only comparing the actual WIP levels between upstream and downstream work-centers, we can still detect the workload status of work-centers. This is a preliminary study and we will carry out a detailed study of WIP balancing without target WIP levels in Section 4.1.

3.1.3.1 MWVS with Target WIP Level

As we mentioned before, WIP distribution histograms for operations is an intuitive representation of the dynamic WIP status and the relationship between upstream and downstream operations. Whereas, because wafer fab is a dynamic job shop, normally an upstream work-center has more than one downstream work-center. The upstream and downstream work-centers have no one-to-one relationship anymore, as described in Figure 3.1.3. Under such circumstances, if we only compare the actual WIP with the target WIP, this only results in a one-dimensional priority matrix in Table 3.1.5 that is not feasible. For instance, suppose the actual WIP of work-center 7 is higher than target level, but for work-center 9 it is lower than target level. Therefore, the lots in operation 6 and

15 get higher priority than lots in operation 3 and 10. The problem arising here is how to distinguish the urgency between operation 6 and 15, or operation 3 and 10.

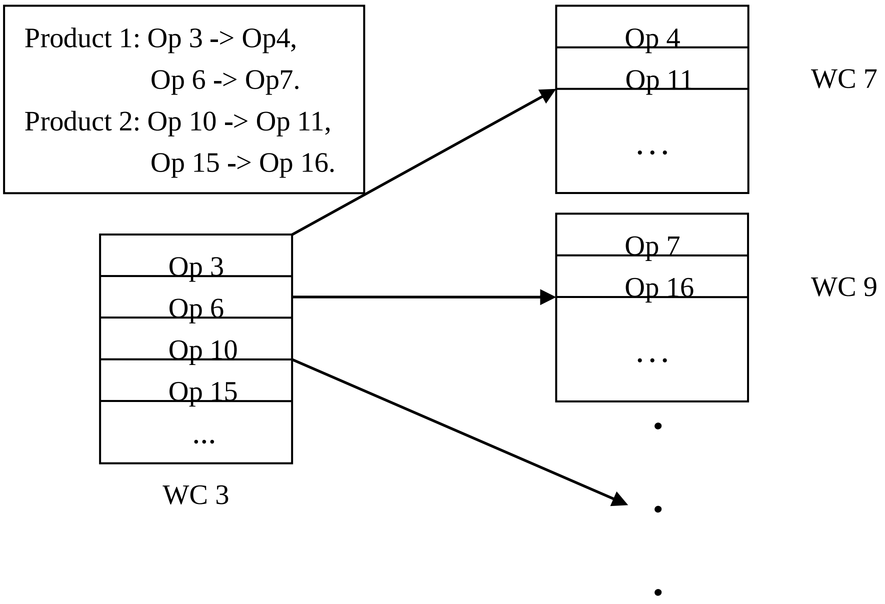


Figure 4.1.3: A simple example of the relationship between current and downstream work-centers

	Downstream Work-center	
	Actual WIP < Target WIP	***Actual WIP >= Target WIP***
Current Work-center	Priority 1	Priority 2

Table 3.1.5: Because of one-to-n relationship, only considering the actual and target WIP level of work-center results in one dimension priority matrix

In order to solve this problem, we introduce Product Weights for work-centers. It is defined as the ratio describing the contribution that a work-center *WC* dedicates to a product *P*. It is expressed as follows:

$$PW_{work-center}^{product} = \frac{\text{the number of operations of P performed by WC}}{\text{the number of operations of all products performed by WC}} \qquad (3.1.1)$$

Then the target WIP level of product *P* for work-center *WC* - $TarWIP_{WC_i}^{P_i}$ is obtained via the Product Weight multiplied by the target WIP level of the work-center. Correspondingly, there is also an actual WIP level of product *P* for work-center *WC* - $ActWIP_{WC_i}^{P_i}$. Now the $TarWIP_{WC_i}^{P_i}$ and $ActWIP_{WC_i}^{P_i}$ lead to a two dimensional priority matrix similar to MIVS rule as shown in Table 3.1.6.

		Downstream Work-center	
		$ActWIP_{WC_d}^{P_i}$ < $TarWIP_{WC_d}^{P_i}$	$ActWIP_{WC_d}^{P_i}$ >= $TarWIP_{WC_d}^{P_i}$
Current Work-center	$ActWIP_{WC_c}^{P_i}$ >= $TarWIP_{WC_c}^{P_i}$	Priority 1	Priority 2
	$ActWIP_{WC_c}^{P_i}$ < $TarWIP_{WC_c}^{P_i}$	Priority 3	Priority 4

Table 3.1.6: Introduction of Product Weight results in two dimension priorities matrix

The detailed algorithm of MWVS rule is described as follows:

(1). Firstly, we compare the actual WIP with the target WIP of downstream work-centers to find out whether it is starved or overloaded. The priorities of lots are assigned according to Table 3.1.5;

(2). Secondly, if two lots (or more) have the same priority from Table 3.1.5, Table 3.1.6 is applied to distinguish them. The purpose is to make sure that the high WIP product in the current work-center is pushed to the starved downstream work-centers, and the low WIP product waits for a while to avoid congestion for the overloaded downstream work-centers.

Next we continue the example of Figure 3.1.3. Assume for operation 6, its actual WIP is higher-than-target and its downstream has the actual WIP lower-than-target. Operation 15 has the actual WIP higher-than-target and its downstream has the actual WIP higher-than-target. Therefore, according to Table 3.1.6, operation 6 has a higher priority than operation 15. The same procedure can be applied to operation 3 and 6.

The target WIP level for work-centers used in Table 3.1.5 and 3.1.6 is from the average WIP level of each work-center of MIMAC6 applying FIFO as dispatching. The Product Weight of each product for all work-centers is listed in Table 3.1.12 in the Appendix.

3.1.3.2 One-step-ahead and One-step-back MWVS without Target WIP Level

The previous section introduces the MWVS rule using target WIP for work-center WIP balance. The target WIP explicitly tells us that under which circumstances the work-center is starved or overloaded, which is fairly understandable from operational control standpoint. Nevertheless, it is a big challenge to determine and apply appropriate target WIP levels in reality. The practical drawbacks and challenges to apply target WIP levels are also obvious and will be discussed in detail in Section 4.1 of Chapter 4. Therefore, we intend to explore the feasibility of WIP balance without the help of target WIP. Another work-center oriented approach, which is a variant of MWVS rule and called one-step-ahead and one-step-back MWVS, is proposed in this section.

The larger the information utilized to make a decision, the better the schedule can be achieved. In order to replace the role of target WIP level, we have to use more information from upstream and downstream work-centers. By extending MWVS, we consider WIP flow for work-center one-step-back. The relationship among upstream, current and downstream work-centers becomes n-1-n and is illustrated in Figure 3.1.4.

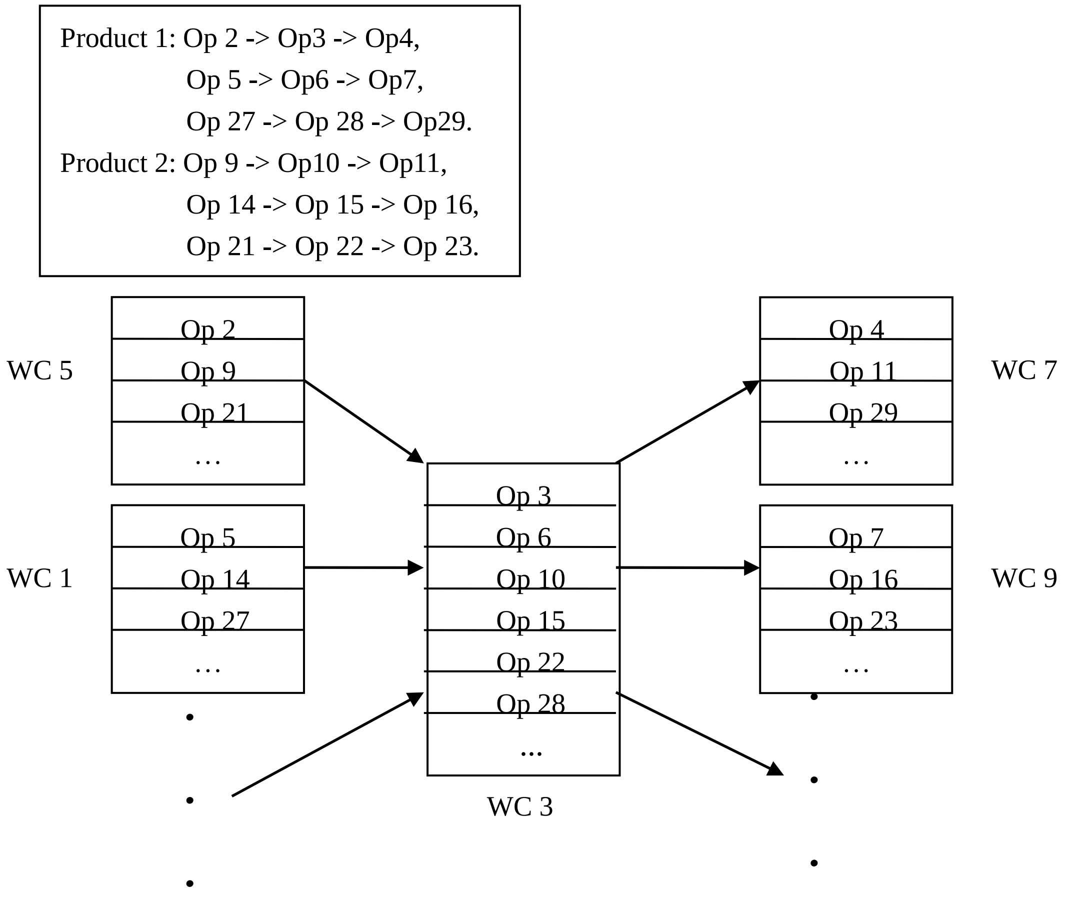


Figure 3.1.4: A simple example of the relationship among upstream, current and downstream work-centers

In fact, except for target WIP levels, without additional information like lot status (due date information) the desired dispatching effect is rather limited in this case. However, Figure 3.1.4 illustrates two facts that are useful and non-negligible. (1): Whether there is high WIP coming from upstream work-centers in future; (2): Whether there is starvation at downstream work-centers. For instance, (1): Suppose there is high WIP at operation 2 of work-center 5, which indicates that work-center 3 will receive those lots soon. In order to avoid congestion of operation 2, work-center 3 should process the lots in operation 2 as fast as possible; (2): Suppose there is a high WIP at operation 6 of work-center 3 and a low WIP at operation 7 of work-center 9. Work-center 3 should push WIP to work-center 9.

Based on the above, the one-step-ahead and one-step-back MWVS with only using actual WIP level information is described in Table 3.1.7. $ActWIP^{P_i}_{WC_u}$, $ActWIP^{P_i}_{WC_c}$ and $ActWIP^{P_i}_{WC_d}$ represent the actual WIP of product *i* in upstream, current and downstream work-center, respectively. We have to notice that dispatching via Table 3.1.7 may contradict the one via Table 3.1.5 and 3.1.6, because they have different viewpoints to balance WIP. We assume one-step-ahead and one-step-back MWVS without target WIP can achieve WIP balance to a certain extent in comparison with MWVS with target WIP level. But without target WIP level, the WIP variance cannot be reduced as much as MWVS. The simulation results in next section will tell us exactly about these effects.

	$(ActWIP^{P_i}_{WC_u} + ActWIP^{P_i}_{WC_c})$ $>=\ ActWIP^{P_i}_{WC_d}$	$(ActWIP^{P_i}_{WC_u} + ActWIP^{P_i}_{WC_c})$ $<\ ActWIP^{P_i}_{WC_d}$
$(ActWIP^{P_i}_{WC_u} >= ActWIP^{P_i}_{WC_c})$ && $(ActWIP^{P_i}_{WC_c} >=\ ActWIP^{P_i}_{WC_d})$	Priority 1	None
$(ActWIP^{P_i}_{WC_u} >= ActWIP^{P_i}_{WC_c})$ && $(ActWIP^{P_i}_{WC_c} <\ ActWIP^{P_i}_{WC_d})$	Priority 3	Priority 5
$(ActWIP^{P_i}_{WC_u} <\ ActWIP^{P_i}_{WC_c})$ && $(ActWIP^{P_i}_{WC_c} >=\ ActWIP^{P_i}_{WC_d})$	Priority 2	None

$(ActWIP^{P_i}_{WC_u} < ActWIP^{P_i}_{WC_c})$ && $(ActWIP^{P_i}_{WC_c} < ActWIP^{P_i}_{WC_d})$	Priority 4	Priority 6

Table 3.1.7: Priority setting of one-step-ahead and one-step-back MWVS

3.1.3.3 Simulation Results and Performance Analysis

First of all, we consider one year average cycle time of the whole fab, cycle time variance and cycle time upper 95% percentile as performance measures under different fab loadings. The simulation results are shown in Table 3.1.8. Here MWVS_1 stands for MWVS with target WIP, and MWVS_2 represents one-step-ahead and one-step-back MWVS without target WIP.

		Fab Loading (%)				
		75	***80***	***85***	***90***	***95***
Average Cycle Time (days)	***FIFO***	20.5	21.6	23.3	25.7	29.6
	MIVS	20.4	21.4	23.1	25.2	28.5
	MWVS_1	20.4	21.6	23.0	25.0	28.2
	MWVS_2	20.5	21.8	23.2	25.3	29.0
Cycle Time Variance (days^2)	***FIFO***	0.9	1.2	1.2	1.4	1.6
	MIVS	1.3	1.4	2.6	2.8	1.8
	MWVS_1	1.4	2.0	2.9	4.4	6.7
	MWVS_2	1.4	1.9	2.5	3.2	4.8
Cycle Time Upper 95% Percentile (days)	***FIFO***	26.2	27.6	29.8	33.1	39.1
	MIVS	25.4	26.8	29.8	32.6	37.0
	MWVS_1	25.7	27.4	29.5	32.1	37.8
	MWVS_2	25.9	27.8	29.5	32.4	38.3
Where MWVS_1 is MWVS with target WIP level, MWVS_2 is one-step-ahead and one-step-back MWVS without target WIP level.						

Table 3.1.8: Performance measures comparison among MIVS, MWVS (with target WIP) and one-step-ahead and one-step-back MMVS (without target WIP)

For the 75% and 80% fab loading cases, these three performance measures of MWVS_1 and MWVS_2 are outperformed by MIVS. It seems that considering WIP balance for operations (MIVS) is more effective than that for work-centers (MWVS_1 and MWVS_2) under a low fab loading. Because most of the work-centers have a low WIP, which implies that lots flow quite smoothly through the work-centers and WIP does not need to be balanced in work-centers. In this case, WIP balance for operations is more efficient, which directly speeds up lot movement and contributes to average cycle time and cycle time variance improvement. However, if the fab runs under a high loading, more and more lots enter the fab, due to random machine failures, the situation becomes more complex. Lots piling up in front of the critical work-centers occurs permanently. In this case, reducing WIP variability for work-centers becomes more effective than for operation. As long as high WIP taking place in one downstream work-center, the upstream work-center is aware to stop feeding lots to it and deliver lots to other downstream work-centers with low WIP in time, which can avoid longer queue times at the high WIP downstream work-centers and starvation in the low WIP downstream work-centers. As we can see from the results, MWVS_1 (with target WIP) outperforms MIVS with regard to these three performance measures in the 85%, 90% and 95% fab loading cases. With regard to MWVS_2 (without target WIP), on one hand, it always underperforms in these three measures under different loadings compared with MWVS_1. This makes sense because MWVS_2 intends to balance WIP without the assistance of target WIP, it is difficult to detect whether a work-center is starved or overloaded as precise as the one applying target WIP (MIVS and MWVS_1). On the other hand, MWVS_2 outperforms FIFO under high fab loading cases, which gives us confidence that we can achieve WIP balance without target WIP if more information can be utilized for dispatching decisions. We also notice that these three WIP balance rules cannot improve cycle time variance. Typically for MWVS_1 and MWVS_2, the WIP balance for work-center causes fast but out of pace lot movement, which is a potential problem that needs to be solved.

Secondly, we take a close look at work-center behavior under a fab loading of 95%. Table 3.1.9 lists the cycle time contribution by top 10 work-centers of product ‘B5C’. For MIVS, MWVS_1 and MWVS_2, the furnace work-center ‘11026_ASM_B2’ contributes most of the cycle time of this product. For the MWVS_2 case, 13.7% of cycle time is spent in this work-center. MIVS is more balanced than MWVS_2, because this critical work-center contributes less cycle time. MWVS_1 is the best, this critical work-center contributes 3.19 days which is almost one day less than MIVS. For other major contributors, although they contribute a little more compared with MIVS and MWVS_2 cases, MWVS_1 successfully avoids high WIP in ‘11026_ASM_B2’ and shifts a certain amount of WIP from ‘11026_ASM_B2’ to other work-centers. Therefore, lots do not experience huge queue times in ‘11026_ASM_B2’. In other words, MWVS_1 is able to balance the WIP among different work-centers which results in a more balanced line than MIVS.

MIVS			***MWVS_1***			***MWVS_2***		
WC	***CTC (days)***	***PoT (%)***	***WC***	***CTC (days)***	***PoT (%)***	***WC***	***CTC (days)***	***PoT (%)***
11026_ASM_B2	4.17	11.6	***11026_ASM_B2***	3.19	9.1	***11026_ASM_B2***	4.98	13.7
20540_CAN_0.43_MII	2.65	7.3	***20540_CAN_0.43_MII***	2.79	8.0	***20540_CAN_0.43_MII***	2.56	6.7

12553_POSI_GP	2.16	6.0	***12553_POSI_GP***	2.15	6.2	***12553_POSI_GP***	2.44	6.1
13024_AME_4+5+7+8	1.47	4.1	***13024_AME_4+5+7+8***	1.56	4.3	***13024_AME_4+5+7+8***	1.50	3.7
15121_LTS_3	1.29	3.6	***15121_LTS_3***	1.34	3.8	***11024_ASM_A4_G3_G4***	1.35	3.5
17421_HOTIN	1.21	3.3	***17421_HOTIN***	1.24	3.4	***15121_LTS_3***	1.28	3.2
11024_ASM_A4_G3_G4	1.20	3.3	***11024_ASM_A4_G3_G4***	1.25	3.6	***17421_HOTIN***	1.22	3.1
16221_IMP- MC _1+2	1.05	2.9	***16221_IMP-MC _1+2***	1.17	3.3	***11027_ASM_B3_B4_D4***	1.04	2.3
15627_HIT_S6000	1.02	2.8	***15627_HIT_S6000***	1.05	2.9	***16221_IMP-MC _1+2***	1.08	2.6
17221_K_SMU236	1	2;7	***17221_K_SMU236***	0.99	2.9	***15627_HIT_S6000***	1	2.4

Where WC stands for work-center, CTC represents cycle time contribution, PoT means percent of total.

Table 3.1.9: Cycle time contribution by top 10 work-centers of product B5C, 95% fab loading

Thirdly, in order to further understand the behavior of work-center oriented WIP balance, we investigate three bottleneck work-centers that are under high capacity loading (fab loading is 95%). Table 3.1.10 lists the basic information of these three work-centers and the simulation results are listed in Table 3.1.11. Sufficient WIP to achieve high utilization of the bottleneck is important. However, if the WIP exceeds the required level to protect bottleneck from starvation, the cycle time increases because lots experience long waiting times in queue. Although the work-center ‘20540_CAN_0.43_MII’ has the highest utilization, the work-center ‘11026_ASM_B2’ seems the one that constrains the whole fab because it only has one machine and long batch processing time. From Table 3.1.11, we can see that the average queue delay of lots in ‘11026_ASM_B2’ is high and up to 27.4 hours for FIFO. When ‘11026_ASM_B2’ has a breakdown, the WIP flow is blocked. In such circumstances, it makes no sense that the upstream sends more lots to it since it is only increasing queue time. The major contribution of WIP balance is to shift WIP between ‘11026_ASM_B2’ and other downstream work-centers. Obviously, MIVS, MWVS_1 and MWVS_2 succeed in reducing average WIP and queue delay for ‘11026_ASM_B2’ in comparison to FIFO.

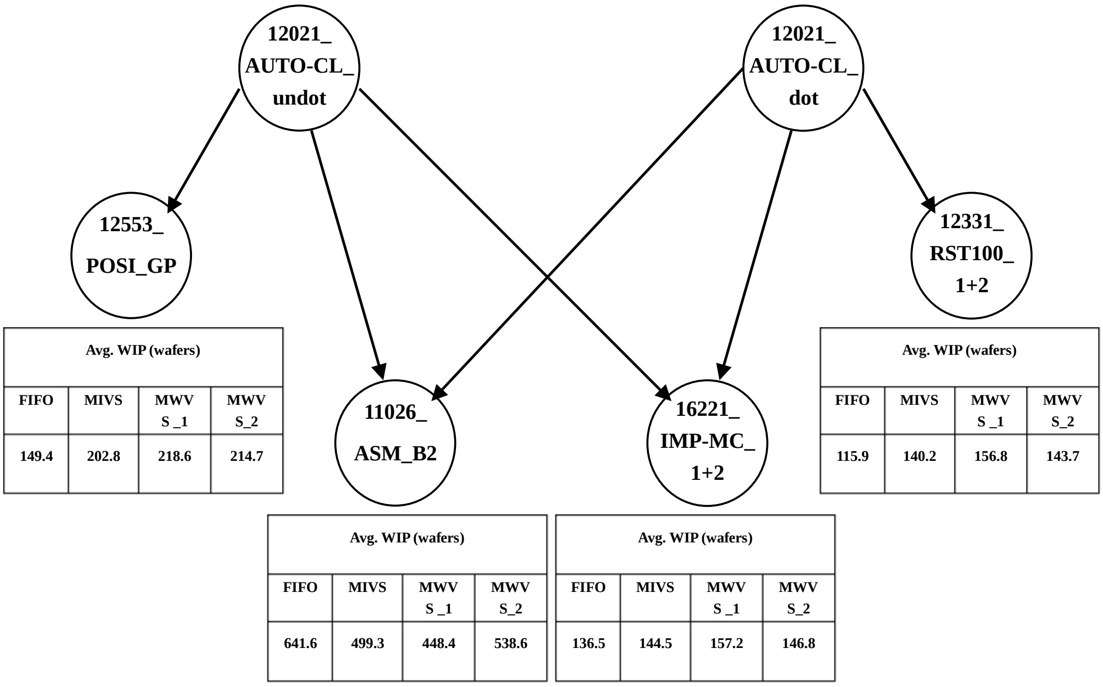
Work-center	*Processing type*	*Number of machines*	*Lot processing time*	*Capacity loading*
20540_CAN_0.43_MII	Single, Photo	7	0.12 hours/lot	95%
12553_POSI_GP	Batch, Wet Etch	1	0.5 hours/batch	93.54%
11026_ASM_B2	Batch, Furnace	1	4 hours/batch	90.56%

Table 3.1.10: Basic information of three high capacity loading work-centers

Figure 3.1.5 demonstrates the WIP shift among ‘11026_ASM_B2’ and other work-centers. ‘11026_ASM_B2’ has two upstream work-centers which are ‘12021_AUTO-CL undo’ and ‘12022_AUTO-CL dot’. These two upstream work-centers have more than 10 downstream work-centers. Here we only list the three major downstream work-centers which have the most WIP shift. Once detecting a high WIP taking place in ‘11026_ASM_B2’, the WIP balance approaches stop sending lots to it. In contrast, the lots are sent to other downstream work-centers, e.g., ‘12553_POSI_GP’ and ‘16221_IMP-MC_1+2’. This is the reason why the average WIP levels of ‘12553 POSI GP’ and ‘16221_IMP-MC_1+2’ of MIVS, MWVS_1 and MWVS_2 are higher than in the case of FIFO. MWVS_1 achieves the best WIP shift compared with MIVS and MWVS_2, which is expected and shows us once again the benefits of applying target WIP for work-center oriented WIP balance.

Work-center		***Avg. WIP (wafers)***	***Avg. queue delayed (hours)***
20540_CAN_0.43_MII	***FIFO***	256.5	1.7
	MIVS	299.7	2.0
	MWVS_1	318.2	2.2
	MWVS_2	284.6	1.8
12553_POSI_GP	***FIFO***	149.4	8.1
	MIVS	218.8	9.8
	MWVS_1	202.6	9.3
	MWVS_2	214.7	10.3
11026_ASM_B2	***FIFO***	641.6	27.4
	MIVS	499.3	20.2
	MWVS_1	448.4	19.3
	MWVS_2	538.6	22.5

Table 3.1.11: Average WIP and queue delayed comparison among three high capacity loading work-centers under 95% fab loading



Avg. WIP (wafers)			
FIFO	MIVS	MWVS _1	MWVS_2
149.4	202.8	218.6	214.7

Avg. WIP (wafers)			
FIFO	MIVS	MWVS _1	MWVS_2
115.9	140.2	156.8	143.7

Avg. WIP (wafers)			
FIFO	MIVS	MWVS _1	MWVS_2
641.6	499.3	448.4	538.6

Avg. WIP (wafers)			
FIFO	MIVS	MWVS _1	MWVS_2
136.5	144.5	157.2	146.8

Figure 3.1.5: WIP shifts among different downstream work-centers under 95% fab loading

3.1.4 Conclusions

In this section, firstly we described the symptoms of WIP imbalance for the MIMAC6 model to introduce the importance of WIP balance for operations or work-centers. Inspired by the MIVS rule, we developed a work-center oriented WIP balance approach named Minimum Workload Variability Scheduling (MWVS). Similar to MIVS, MWVS used target WIP to identify whether a work-center is starved or overloaded. In order to figure out the possibility of avoiding target WIP, we proposed one-step-ahead and one-step-back MWVS that only used actual WIP of upstream, current and downstream work-centers for decision making. We conclude the following from the performance analysis of the simulation results:

- Under the low fab loading case, since lots flow smoothly through the work-centers, the work-center oriented approach shows no significant performance improvements compared to the operation oriented approach.

- However, for the case of high fab loading, because a large number of lots queue in front of work-centers, WIP variability reduction for work-centers is more effective than for operations. The work-center oriented WIP balance can detect whether a work-center is starved or overloaded. That is the information missing in MIVS. It shows the advantage of controlling WIP to flow to low WIP work-center instead of high WIP work-centers. Indeed, as the simulation result told us, it is of particular importance to avoid long queues in front of critical work-center under high fab loading. Essentially, as exhibited in Section 3.1, a work-center can perform different operations, once we reduce the WIP variability for work-centers, the WIP variability of operation is solved naturally. This is the reason why MWVS with target WIP is superior over MIVS.

There are three issues arising from work-center oriented WIP balance.

(1). The first one is the unimproved cycle time variance. It is understandable that WIP balance addresses starvation and congestion avoidance to speed up lot movement. Therefore, it pays less attention to the issue of good pace movement. Indeed, cycle time variance minimization is critical as well, in particular, when due date is taken into account. Therefore, it is very necessary to explore a way to reduce cycle time variance for WIP balance approaches, which will be presented in Section 3.2.

(2). Another issue, the target WIP used in MWVS is from the average WIP of work-centers of MIMAC6 with FIFO dispatching. This method is fast and without much complex computation effort. Since the fab performance highly relies on target WIP, sophisticated approaches considered as standard procedure should be taken into account to create accurate target WIP. In Section 3.3, we will discuss this issue in detail.

(3). MWVS with target WIP always outperforms one-step-ahead and one-step-back MWVS without target WIP. It is true that WIP balance with target WIP is more accurate and effective than the one without target WIP, because the target WIP plays an important role to determine the status of work-centers. As a preliminary study, it is too early to conclude that it is impossible to achieve WIP balance without the need of target WIP, since information used for decision making is rather limited in this case. For this reason, larger information sets are needed to replace the target WIP information, i.e., we need to employ lot information like due dates. The feasibility of achieving WIP balance without target WIP will be further explored and discussed in Section 4.1 of Chapter 4.

3.2 Cycle Time Variance Reduction

3.2.1 Why Cycle Time Variance Reduction is Necessary

From the simulation results in Section 3.1.3 (Table 3.1.8, P.58), we notice that both operation oriented and work-center oriented WIP balance achieve average cycle time reduction in comparison with FIFO. As far as cycle time variance is concerned, the effects of both approaches are modest. Because of re-entrant flows, rework, setups and batch formation, lot overtaking takes place oftentimes, e.g., an early-arrival lot can be bypassed by a late-arrival lot because the late-arrival lot fulfills the batch requirements. Earlier when WIP balance approaches (MIVS and MWVS) are applied, the cycle time variance shows no sign of improvement, sometimes even an increase. The inherent characteristic of WIP balance with only focusing on starvation and congestion avoidance is the root cause. For instance, in order to prevent downstream starvation, some lots are processed rapidly to the downstream work-centers, while some lots still need to wait in the queue although they arrived at the queue first. As a consequence, the poor pace lot movement causes degraded cycle time variance.

There are a large number of articles [Demeester and Tang 1996, Domaschke et al. 1998, Ho et al. 2000, Leachman et al. 2002, Nemoto et al. 1996, Spearman et al. 1990] that focus on reducing cycle time without consideration of cycle time variance minimization. Whereas, in today’s advanced semiconductor manufacturing, cycle time variance is too important to be ignored. The reasons are:

- A low cycle time variance indicates a precise prediction of production completion time which allows more relaxed coordination with downstream operations on wafers like assembly [Lu et al. 1994];
- The strict pace movement is of great significance to WIP forecast and control. It creates a buffer allowing to absorb a fair amount of variability and further enhance the ability to handle unusual events like random machine failures;
- In particular, it is critical to customer oriented companies because a low cycle time variance leads to a greater repeatability and quality to meet due date reliably. Thus, they are able to provide an accurate lead time commitment to customers.

For the MIVS and MWVS approaches, in most cases more than one lot obtains the same priority, and FIFO is used for final dispatching to distinguish the urgency of lots. This is the reason why the cycle time variance cannot be improved, as FIFO does not use any information, e.g., processing time, waiting time and due date to distinguish the urgency of lots. If we change the perspective of ‘urgency’ to ‘better pace lot movement’, naturally, reducing cycle time variance for the lots which obtain the same priority from MIVS and MWVS is the ‘urgency’ issue. Hsieh et al. [2003] also suggest to apply specified rules to better distinguish the urgency of the lots.

- Thereby, we believe there is potential room for cycle time variance improvement as long as more information can be utilized to replace FIFO as final dispatching rule to better distinguish the urgency of lots.

3.2.2 Rules to Minimize Cycle Time Variance

Before introducing the dispatching rules used to replace FIFO for MIVS and MWVS, we define 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*;

$O_{i,j}$: Operation due date of lot *i* for operation *j*;

L_i : Lateness of lot *i*;

Now : Current time *t*;

CT_i : Cycle time of lot *i* when it finishes processing and leaves the fab;

$AccumulatedCT_{i,t}$: Accumulated cycle time of lot *i* at time *t* (still in the fab);

RPT_i : Raw processing time of lot *i*;

$AccumulatedRPT_i$: Accumulated raw processing time of lot *i*;

$RemainingRPT_i$: Remaining 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*;

1. Select the lot with longest queue time plus accumulated cycle time

As we mentioned above, queue time accounts for a large proportion of cycle time, particularly, under high fab loading case. Since the raw processing time is constant in the MIMAC6 model, one basic fact is that cycle time distributions can be narrowed if the queue time can be kept in a fixed range. In other words, dispatching lots on the basis of queue time information can lead to cycle time variance reduction. In the following we prove this assumption.

Suppose a work-center case. There are two lots - lot *i* and lot *j* in the queue. At time *t* the work-center is available to select a lot to process. If lot *i* is chosen for processing first, the average cycle time and cycle time variance are expressed as follows [Gupta et al. 2009]:

$$AvgCT_{i,j} = \frac{(Q_i + P_i) + (Q_j + P_i + P_j)}{2} \tag{3.2.1}$$

$$CTVar_{i,j} = \frac{(Q_i + P_i - AvgCT_{i,j})^2 + (Q_j + P_i + P_j - AvgCT_{i,j})^2}{2} \tag{3.2.2}$$

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

$$AvgCT_{j,i} = \frac{(Q_j + P_j) + (Q_i + P_j + P_i)}{2} \tag{3.2.3}$$

$$C\,T\,V\,a_{j\,,} = \frac{(Q_j + P_j - AvgCT_{j,i})^2 + (Q_i + P_j + P_i - AvgCT_{j,i})^2}{2} \tag{3.2.4}$$

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:

$$CTVar_{i,j} < CTVar_{j,i} \tag{3.2.5}$$

We can deduce the following equation from Equation (3.2.5) by using

Equation (3.2.1), (3.2.2), (3.2.3) and (3.2.4):

$$[(Q_i + P_i - AvgCT_{i,j})^2 + (Q_j + P_i + P_j - AvgCT_{i,j})^2]/2$$
$$< [(Q_j + P_j - AvgCT_{j,i})^2 + (Q_i + P_j + P_i - AvgCT_{j,i})^2]/2$$

$$=> [P_i + (Q_i - Q_j)]^2 > [P_j - (Q_i - Q_j)]^2$$

$$=> (P_i + P_j)(2Q_i - 2Q_j + P_i - P_j) > 0$$

$$\because (P_i + P_j) > 0$$

$$\therefore (2Q_i - 2Q_j + P_i - P_j) > 0$$

$$=> (P_i + 2Q_i) > (P_j + 2Q_j) \qquad (3.2.6)$$

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

$$\because Q_i >> P_i \ \&\& \ Q_j >> P_j$$

$$\therefore Q_i > Q_j \qquad (3.2.7)$$

The above equations tell us that selecting the lot with the longer queue time can minimize cycle time variance in a single work-center case.

However, the above conclusion is based only on one work-center. In reality, there are hundreds of work-centers in the fab. When extending the one work-center case to the whole fab case, it is obvious that selection of the lot with the longest queue time is not sufficient to make sure that Equation (3.2.5) holds. We notice that the queue time Q_i can be extended to be the queue time

Q_i lot spending in the current work-center and the accumulated cycle time $CT_{i,t}$ lot spending in the fab, if the one work-center case is extended to the fab case. Therefore, the following is derived from Equation (3.2.7):

$$Q_i + AccumulatedCT_{i,t} > Q_j + AccumulatedCT_{j,t} \qquad (3.2.8)$$

Equation (3.2.8) tells us that selection of the lot with longer queue time plus accumulated cycle time can minimize cycle time variance in the whole fab case.

2. Due date oriented rule - Operation Due Date

Suppose each 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:

$$L_i = F_i - D_i \qquad (3.2.9)$$

The due date oriented rules - such as Earliest Due Date (EDD), Least Slack Time (LST), Critical Ratio (CR) or Operation Due Date (ODD) - all attempt to minimize the variance of lateness in different but similar manners. Let us have look at these rules in detail:

EDD: The lot with earliest due date is chosen to be processed because it is the most urgent. The due date performs as a milestone, EDD tries to make every lot finish before its due date, or close to due date if not possible. The lateness variance is minimized if all lots are finished around their due dates.

LST: The slack of lot *i* is calculated by ‘D_i- *Now* - $RemainingRPT_i$’, and the lot with the least slack time is favored. The slack is used to measure the urgency of lots. In fact, LST is more ‘fair’ than EDD because only using due date information is too optimistic to keep lots toward completion at the right pace. LST utilizes slack time to replace due date and intends to make every lot equally early or equally tardy. As a result, the lateness variance is even smaller than EDD if all lots are equally early or equally tardy.

CR: The critical ratio is calculated by ‘(D_i - Now) / $RemainingRPT_i$’, and the lot with smaller ratio is favored. CR has a similar mechanism like LST, therefore, it can also provide a reduced variance of lateness.

ODD: Different from EDD, ODD assigns a due date to each operation. It breaks up the slack time into as many segments as the number of operations of a lot. Therefore, ODD is even more ‘fair’ than LST with regard to keep lots going towards operation due date equally early or equally tardy. Once the operation due dates have been established, the lots are strictly kept at the right pace to meet their due date through the fab from the early operations on.

We are aware that once due date is set, these four typical due date oriented rules manage to finish lot as close to due date as possible, so as to minimize the variance of lateness. It is not difficult to find out that the cycle time becomes the same as the lateness if due date D_i is replaced by release time R_i in Equation (3.2.9). Consequently, from the above deduction the due date oriented rule should lead to cycle time variance reduction. (In Section 3.4, we will explain the cycle time variance reduction mechanism of due date oriented rules in detail.)

In this study, ODD is chosen among these four rules as a representative of due date oriented rules. The lot with smaller ODD value is preferred.

3. Flow Factor (FF)

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

$$FF_i = \frac{CT_i}{RPT_i} \tag{3.2.10}$$

Actually, the FF tells us besides raw processing time, how much time a lot spends in waiting, transporting and so on. Obviously, FF is expected to be minimized. By noticing this, the FF can extend to a dispatching rule. At time *t*, a work-center is free to select a lot via calculating the current FF by accumulated cycle time divided by accumulated raw processing time, as described in Equation (3.2.11).

$$FF_i = \frac{AccumulatedCT_{i,t}}{AccmulatedRPT_i} \tag{3.2.11}$$

At first glance, the FF does not seem to include due date information, but it becomes clear if we modify Equation (3.2.11) in the following way:

$$AccumulatedCT_{i,t} = AccumulatedRPT_i \ * \ FF_i$$

$$=> \quad Now - R_i = AccumulatedRPT_i \ * \ FF_i$$

$$=> \quad Now = R_i + AccumulatedRPT_i \ * \ FF_i \tag{3.2.12}$$

Equation (3.2.12) is similar to the ODD rule, which indicates that the FF is expected to minimize cycle time variance as well, because the FF rule attempts to keep each lot going through the fab at constant FF. But unlike ODD, the lot with larger FF is selected to minimize cycle time variance.

3.2.3 Simulation Results and Performance Analysis

As we mentioned above, we need to distinguish the urgency of lots that fall into the same priority for MIVS and MWVS (with target WIP, expressed as MWVS_1), with the objective to minimize cycle time variance. Therefore, we incorporate the proposed three cycle time variance minimization rules into MIVS and MWVS_1. If the lots obtain the same priority from MIVS or MWVS_1, the proposed rules are applied to distinguish them. The average cycle time, cycle time variance and cycle time upper 95% percentile are considered as performance measures. The fab loadings are divided into three levels which are 95% (high), 85% (medium) and 75% (low). Table 3.2.1, 3.2.2 and 3.2.3 show the one year simulation results of the whole fab.

First of all, we focus on the results of 95% fab loading case, as MWVS_1 shows a seriously degraded cycle time variance performance under this loading. From Table 3.2.1, the default rule to distinguish the lots for MIVS and MWVS_1 is FIFO, it leads to cycle time variance 1.8 and 6.7 for MIVS and MWVS_1 respectively, and the cycle time upper 95% percentile is 37.0 and 37.8 days for MIVS and MWVS_1 respectively. When the proposed three rules are incorporated into MIVS and MWVS_1 to better differentiate the urgency of lots, the improvements are promising. For the average cycle time, no matter whether MIVS or MWVS_1 is used, the ODD and FF rules result in considerable average cycle time reduction which is more than one day in comparison with using FIFO as default rule, while the Q+Acc.CT rule shows limited improvement.

		95% Fab Loading		
		Average Cycle Time (days)	***Cycle Time Variance (days^2)***	***Cycle Time Upper 95% Percentile (days)***
FIFO		29.6	1.6	39.1
MIVS+	***FIFO***	28.7	1.8	37.0
	Q + Acc.CT	28.4	1.5	36.6
	ODD (DDFF 2.0)	27.4	0.4	35.2
	FF	26.9	0.8	34.6
MWVS_1+	***FIFO***	28.2	6.7	37.8
	Q + Acc.CT	28.4	4.3	36.9
	ODD (DDFF 2.0)	26.9	1.0	34.4
	FF	27.2	1.4	34.9

Where MIVS: Minimum Inventory Variability Scheduling;
MWVS_1: Minimum Workload Variability Scheduling with target WIP level;
FIFO: First in first out; Q + Acc.CT: Longest queue time plus accumulated cycle time;
ODD: Operation due date; FF: Flow factor; DDFF: Due date flow factor.				

Table 3.2.1: Four cycle time variance reduction methods comparison for MIVS and MWVS under 95% fab loading

With respect to the cycle time variance and cycle time upper 95% percentile, the ODD and FF rules absolutely dominate over the Q+Acc.CT rule. Specifically, ODD and FF rules significantly reduce cycle time variance, from 6.7 to 1.0 and 1.4 respectively, for MWVS_1. Furthermore, the cycle time upper 95% percentile also indicates the remarkable improvement, 95% of lots’ cycle times are reduced considerably to 34.4 days by ODD and 34.9 days by FF in comparison with 37.8 days by FIFO. It is obvious that these three rules can minimize cycle time variance for certain. However, it is surprising that they have an additional positive effect on achieving average cycle time reductions 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 in progressing lots at the right pace to avoid fluctuation. Thus, they can diminish the suddenness of lot arrivals. In this study, the ODD and FF rules are equally effective since there is no significant difference from the results.

		85% Fab Loading		
		Average Cycle Time (days)	***Cycle Time Variance (days^2)***	***Cycle Time Upper 95% Percentile(days)***
FIF0		23.3	1.2	29.8
MIVS+	***FIFO***	23.1	2.6	29.8
	Q + Acc.CT	23.4	2.0	28.8
	ODD (DDFF 1.8)	22.6	0.5	29.3
	FF	22.8	0.8	29.2
MWVS_1+	***FIFO***	23.0	2.9	29.5
	Q + Acc.CT	23.1	1.8	28.7
	ODD (DDFF 1.8)	22.7	0.8	29.0
	FF	22.8	1.0	29.2

Where MIVS: Minimum Inventory Variability Scheduling;
MWVS_1: Minimum Workload Variability Scheduling with target WIP level;
FIFO: First in first out; Q + Acc.CT: Longest queue time plus accumulated cycle time;
ODD: Operation due date; FF: Flow factor; DDFF: Due date flow factor.

Table 3.2.2: Four cycle time variance reduction methods comparison for MIVS and MWVS under 85% fab loading

		75% Fab Loading		
		Average Cycle Time (days)	***Cycle Time Variance (days^2)***	***Cycle Time Upper 95% Percentile (days)***
FIFO		20.5	0.9	26.2
MIVS+	***FIFO***	20.3	1.3	25.5
	Q + Acc.CT	20.8	1.0	24.7
	ODD (DDFF 1.6)	20.4	0.4	26.4
	FF	20.3	0.5	26.2
MWVS_1+	***FIFO***	20.4	1.4	25.7
	Q + Acc.CT	20.6	1.0	25.5
	ODD (DDFF 1.6)	20.3	0.5	25.9
	FF	20.3	0.6	26.0

Where MIVS: Minimum Inventory Variability Scheduling;
MWVS_1: Minimum Workload Variability Scheduling with target WIP level;
FIFO: First in first out; Q + Acc.CT: Longest queue time plus accumulated cycle time;
ODD: Operation due date; FF: Flow factor; DDFF: Due date flow factor.				

Table 3.2.3: Four cycle time variance reduction methods comparison for MIVS and MWVS under 75% fab loading

Secondly, we see similar performances with slight differences for the medium (85%) and low (75%) fab loading cases in Table 3.2.2 and 3.2.3. The ODD and FF rules still achieve significant cycle time variance minimization as well as slight average cycle time reduction, whereas, the Q+Acc.CT rule degrades the average cycle time for MIVS and MWVS_1. It is interesting to see that the Q+Acc.CT rule outperforms the ODD and FF rules with regard to cycle time upper 95% percentile. 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 for 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 a strict pace of lot movements. Although it degrades the average cycle time, it manages to reduce the cycle time upper 95% percentile performance.

3.2.4 Conclusions

In this section, we explored three rules to minimize cycle time variance with the objective to solve the problem arising from WIP balancing shown in Section 3.1.3. 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 three rules used to minimize cycle time variance were motivated by mathematical reasoning and validated by simulation. The simulation results demonstrated that they can overcome the drawbacks of WIP balance to different extents.

We investigated the proposed cycle time variance minimization rules under different fab capacity loading cases.

- 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. This unexpected finding is a strong argument that it is very necessary to make sure that lots are progressed with good pace to achieve balanced WIP.

- 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 95% percentile.
- 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 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 and the ODD rule is affected by the tightness of due dates.

Last the promising performance coming from ODD rule drives us to carry out a comprehensive study of due date control rules in Section 3.4. The benefit of cycle time variance minimization for WIP balance will be addressed in detail in Section 5.1 of Chapter 5 when due date performance is involved.

3.3 Using Queuing Models and Neural Networks to Determine Target WIP Levels for Work-centers for MWVS Approach

3.3.1 Introduction to Queuing Models and Neural Networks

In Section 3.1 the simulation results demonstrated that the target WIP plays a crucial role when MWVS_1 is applied to achieve WIP balance for work-centers. The target WIP levels of work-centers used in MWVS_1 derived from the average WIP levels of work-centers of MIMAC6 with FIFO dispatching (Actually we conducted a parameter setting for the target WIP from FIFO-based-simulation that is explained in Section 3.3.5). This method is fast and without much complex computation effort. However, some doubts regarding the accuracy of target WIP of this method are raised from industry. As the fab performance highly relies on target WIP, the target WIP should be derived from sophisticated approaches that are reasonable and precise.

With regard to the sophisticated approaches to determine target WIP, there are two major ways in literature namely queuing models and neural networks. The GI/G/m queuing model proposed by Whitt [1993] has a single service facility with *m* identical servers and unlimited buffer with first-in-first-out queue discipline, as illustrated in Equation (3.3.1). Queuing models are easily applied because it ignores certain details and responds very quickly under certain conditions [Burman et al. 1986, Connors et al. 1996, Lin and Lee 2001].

$$WIPL_i^{QM} = \frac{c_a^2 + c_s^2}{2} \times \frac{\rho^{\sqrt{2(m+1)}-1}}{1-\rho} + \rho m \qquad (3.3.1)$$

Where $WIPL_i^{QM}$: average WIP level for work-center *i*;

c_a : the coefficient of variation (CV) of inter-arrival time;

c_s : the coefficient of variation (CV) of service time;

m : the number of machines in work-center;

ρ : the utilization of work-center.

Compared to queuing models, neural networks have the advantage that is not necessary to develop complex algorithms and statistical models since it can be trained with observed data to figure out the hidden relationship between the input data and expected output data [Chambers and Mount-Campbell 2002, Narendra 1996]. Neural networks are the most commonly used mathematical and computational model which intends to mimic the behavior of biological neurons [Narendra 1996]. It can be used to predict almost any functions and the trend of data based on historical data. There are two main categories of neural networks architectures which are feed-forward neural network and feed-back neural network. In this study, we choose feed-forward architecture because it performs better on function recognition than feed-back architecture, which is needed to approximate the target WIP for work-centers based on training set data [Narendra 1996]. Besides that, with respect to the learning algorithm, the back-propagation is selected because it is the most commonly used algorithm for the supervised training of feed-forward neural networks. Figure 3.3.1 illustrates an example of the back-propagation neural network (BPNN). When training the back-propagation feed-forward neural network, the data flow transmits one-way from input layer to output layer to acquire the actual output. After each run the error, which is the difference between the actual output and the target output, is back-propagated to the previous hidden layer. The connection weights of the hidden nodes will be altered according to the error using the delta rule. This back-propagated process goes on until the input layer is reached. Finally, the objective of this BPNN is to minimize the error to an acceptable level. Otherwise, the training does not stop unless time over or a predefined maximal number of iterations has been reached. For more details about neural network and its application on semiconductor manufacturing, we refer the interested readers to [Chamber and Mount-Campbell 2002, Huang et al. 1999, Kuo et al. 2008, Narendra 1996, Yu and Huang 2002].

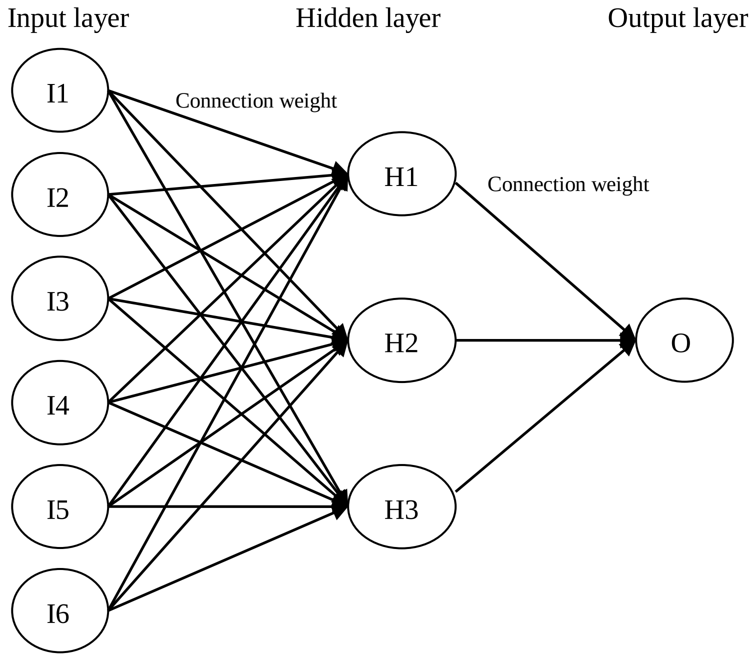


Figure 3.3.1: An example of back-propagation of neural network

In this study, we use both ways to determine the target WIP. The first way: the GI/G/m queuing model is applied to obtain the target WIP level for each work-center. The second way: according to [Kuo et al. 2008] only applying queuing model probably overestimates or underestimates the target WIP level. The overestimation or underestimation effect may amplify due to hundreds of work-centers in the fab. Thus, firstly we make use of BPNN to determine the total WIP level for the whole fab. Then using this whole fab WIP level to adjust the WIP level of non-bottleneck work-centers derived from the queuing model.

All historical data used for training and testing BPNN and queuing model comes from MIMAC6 with 18 months run and FIFO dispatching.

3.3.2 Estimating Target WIP Levels for Bottleneck Work-centers Using Queuing Models under 75%, 85% and 95% Fab Capacity Loadings

According to Little's Law, the bottleneck work-center determines the throughput of the wafer fab, thus, the WIP of fab increases while the WIP of bottleneck increases. In order to obtain the total WIP level of the fab under 75%, 85% and 95% capacity loadings, first of all the target WIP level of bottleneck work-center is calculated by queuing models.

In MIMAC6 model, the work-center '20540_CAN_0.43_MII' is considered as the bottleneck from static viewpoint since it has the highest utilization. The parameters of the bottleneck applied to Equation (3.3.1) under fab loading 75%, 85% and 95% are listed in Table 3.3.1.

	Fab loading (%)		
	75	***85***	***95***
c_a	0.97	0.97	0.96
c_s	0.04	0.04	0.04

m	7	7	7
ρ	0.75	0.85	0.95

Table 3.3.1: The parameters of the bottleneck ‘20540_CAN_0.43_MII’ under different fab loadings used in queuing model

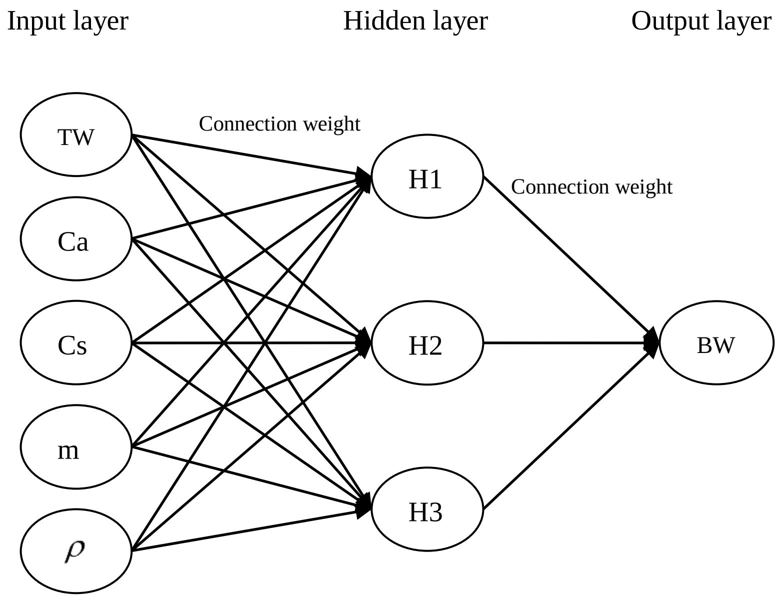
As a consequence, the target WIP levels of bottleneck ‘20540_CAN_0.43_MII’ under 75%, 85% and 95% fab loadings are 6.0 lots, 7.9 lots and 14.6 lots, respectively. For calculation convenience and a lot contains 24 wafers, we determine the average target WIP level for ‘20540_CAN_0.43_MII’ described in Table 3.3.2.

	Fab loading (%)		
‘20540_CAN_0.43_MII’	***75***	***85***	***95***
Avg. target WIP level (wafers)	144	190	350

Table 3.3.2: Average target WIP level for bottleneck ‘20540_CAN_0.43_MII’ under different fab loadings

3.3.3 Estimating Total Target WIP Levels for Wafer Fabs under 75%, 85% and 95% Fab Loadings Using Feed-forward Back-propagation Neural Networks

Our objective is to find out the total target WIP level for the whole wafer fab under 75%, 85% and 95% fab loadings, respectively. Then the total target WIP level is utilized to adjust the target WIP level of non-bottleneck work-centers



Input:

TW: the total WIP level for the fab;

c_a : the average CV of inter-arrival times of the work-center;

c_s : the average CV of service times of the work-center;

m : the number of machines in work-center;

ρ : the utilization of the work-center;

Output:

BW: the average WIP level of the bottleneck.

Figure 3.3.2: A BPNN used to estimate the WIP relationship between the bottleneck and the whole wafer fab

derived from queuing models. For this reason, first of all a feed-forward BPNN is developed to figure out the WIP relationship between the bottleneck and the fab. The total WIP of the fab is set as input and the average WIP of the bottleneck is set as output for BPNN. According to the queuing model and Kuo et al. [2008], the input parameters are total WIP of the fab and four performance measures for each work-center which are c_a, c_s, m and ρ. The BPNN is designed as shown in Figure 3.3.2. We use the Encog workbench to build the

BPNN. For more information about Encog, we refer the interested reader to [www2].

The ‘historical’ data used to train and test the BPNN were collected from the MIMAC6 model carried out by Factory eXplorer (FX) with FIFO dispatching for 18 months, and under different fab capacity loadings ranging from 70% to 100% with 1% step increments. For each fab loading, there are 40 sets of data, 20 sets of which are selected as training data and 20 sets of which are selected as testing data. Thus, there are together 620 sets of data for training and 620 sets of data for testing. When training the BPNN, the learning rate and momentum are adjusted until the root-mean-squared-error (RMSE) and mean-absolute-percentage-error (MAPE) are small enough to be acceptable. Through experiments of training BPNN, we summarize the parameters which yield the acceptable results (activation function = sigmoid activation, learning rate = 0.7, momentum = 0.3, maximum error = 0.01, number of hidden nodes = 6). Table 3.3.3 shows the RMSE and MAPE for training and testing data.

	Training data	***Testing data***
Examples	620	620
RMSE	0.028	0.030
MAPE	3.84%	4.40%

Table 3.3.3: RMSE and MAPE for BPNN

The MIMAC6 model includes 104 work-centers. As described in Figure 3.3.2, there are 4 performance measures of each work-center provided to the BPNN. As a result, including the total WIP of the fab there are total 417 performance measures that may influence the relationship between the average

WIP of the bottleneck and the total average WIP of the fab. In order to find out the most critical performance measures to reduce the input complexity of the BPNN, we figure out three relationships as follows:

(1). The relationship between the average WIP of bottleneck work-center ‘20540_CAN_0.43_MII’ and the total WIP of the fab.

(2). The relationship between the bottleneck and its direct upstream work-centers. Bottleneck ‘20540_CAN_0.43_MII’ has two upstream work-centers which are ‘10123_DNS-3’ and ‘10151_DNS-1’. Actually, the average WIP of the bottleneck is influenced directly and strongly by these two upstream work-centers.

(3). The relationship between bottleneck and non-bottlenecks. Among the non-bottleneck work-centers, work-center ‘11026_ASM_B2’ is highly utilized and has the longest queue length and queue delays. Besides that, it only has one machine, which causes WIP increase during failure period. Consequently, it can represent the non-bottlenecks.

Based on these observations, we select the c_a , c_s , m and ρ of work-centers ‘20540_CAN_0.43_MII’, ‘10123_DNS-3’, ‘10151_DNS-1’ and ‘11026_ASM_B2’ plus the total WIP level of the fab, together 17 performance measures out of 417 as inputs for the BPNN. As we mentioned above, for each fab loading there are 20 sets data used for testing. Therefore, the output data has overlapping problem, e.g., the average WIP of bottleneck of 71% fab loading is lower than the one of 70% fab loading. We select the output data based on the criteria which are (1): the maximum WIP data of current fab loading is lower than the minimum WIP data of next high fab loading; (2): The average WIP level of bottleneck increases gradually. As illustrated in Table 3.3.4, from 106.4

to 200.2 it increases approximately in steps of 2; From 200.2 to 495.3 it increases approximately in steps of 5; From 495.3 to 595.6 it increases approximately in steps of 20; From 595.6 to 1000.8 it increases approximately in steps of 50. As a result, 106 sets of data are selected from the output data. Table 3.3.4 and Figure 3.3.3 show the average WIP level of the bottleneck corresponding to the total WIP level of the fab.

Avg. WIP level of bottleneck ***(Wafers)***	***Total WIP level of fab*** ***(Wafers)***	***Avg. WIP level of bottleneck*** ***(Wafers)***	***Total WIP level of fab*** ***(Wafers)***	***Avg. WIP level of bottleneck*** ***(Wafers)***	***Total WIP level of fab*** ***(Wafers)***	***Avg. WIP level of bottleneck*** ***(Wafers)***	***Total WIP level of fab*** ***(Wafers)***
106.4	1896.4	164.5	2367.4	254.8	3653.3	415.5	5398.6
108.7	1835.1	166.8	2234.4	260.4	3728.2	420.6	5446.9
111.1	1884.6	168.1	2271.1	265.6	3896.8	425.5	5521.8
112.9	1947.6	170.6	2469.6	270.3	3941.4	431.3	5589
114.8	1953.2	172.2	2481.1	275.9	3984.2	435.8	5543.6
116.1	1923.4	174.4	2492.1	280.8	4069.4	440.2	5565.7
118.7	2073.4	176.7	2576.6	285.5	4089.4	445.4	5645.2
121.1	1925.9	178.9	2528.9	290.1	4179.6	450.7	5774.5
122.8	1945.2	180.5	2516.8	294.9	4173.7	455.4	5834.6
124.4	1959.8	182.8	2654.3	301.3	4198.2	460.3	5899.9
126.6	2014.3	184.2	2588.3	305.5	4265.3	465.5	5953.3
128.9	2047.4	185.8	2747.5	310.2	4247.8	470.2	6021.2
130.2	2023.1	188.3	2804.3	315.6	4396.4	475.4	6034.4
132.8	2049.8	190.2	3108.4.	320.3	4423.7	485.3	6167.8
135.2	2116.3	192.7	3145.3	325.4	4463.4	490.9	6256.8
136.8	2138.2	194.4	3247.2	331.3	4436.3	495.3	6399.9
138.3	2163.1	196.6	3224.5	335.8	4447.8	515.5	6564.3
140.9	2158.8	198.5	3365	340.3	4436.2	535.6	6890.3
142	2174.9	200.2	3269.6	345.6	4420.5	557.1	6857.7
144.2	2167.6	205.8	3361.6	350.5	4480.6	575.3	6887.2
146.2	2289	210.3	3293.5	356.1	4634.5	595.6	6944.3
148	2247.5	215.5	3416.7	365.7	4823.9	644.2	6949.6
149.6	2223.9	220.3	3403.4	375.9	5004.5	695.2	6972.1
152.3	2139.2	225.2	3420.8	380.4	5086.5	745.7	6972.3
154.3	2193.4	230.7	3447.3	385.4	5173.4	795.8	6972.3
156.2	2236.8	235.5	3558.2	390.2	5243.6	845.7	6972.4
158.6	2243.6	240	3598.3	394.7	5276.1	895.1	6972.4
161.1	2267.5	245.3	3643.3	404.9	5326.7	945.7	6972.6
162.9	2203.4	250.3	3667.8	410.1	5397.7	1000.8	6972.8

Table 3.3.4: The relationship between the average WIP level of bottleneck and the total WIP of wafer fab under fab loadings from 70% to 100%

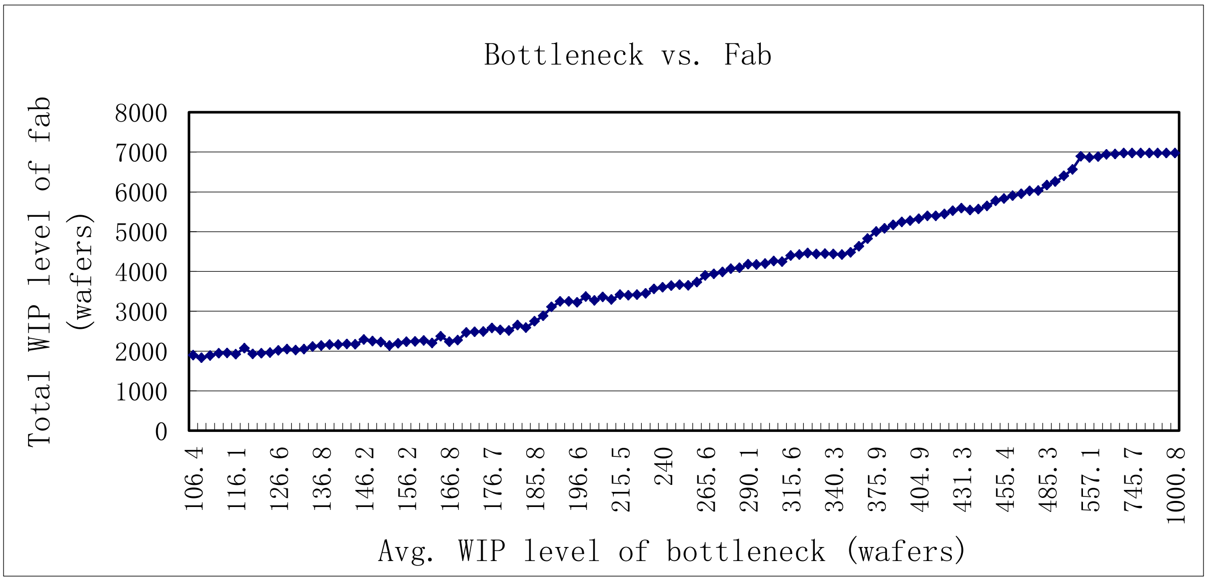


Figure 3.3.3: The relationship between the average WIP level of bottleneck and the total WIP level of wafer fab under fab loadings from 70% to 100%

We can conclude the following according to Table 3.3.4 and Figure 3.3.3

(1). The total WIP of the fab has a trend to increase as the average WIP of the bottleneck increases.

(2). When the average WIP level of the bottleneck is higher than 535.6 wafers, the total WIP of the fab remains the same level approximately, which is in the case of 100% fab loading.

Considering the average WIP level of the bottleneck listed in Table 3.3.2:

(1). For 75% fab loading case, the average WIP level of the bottleneck is 146 wafers. Based on Table 3.3.3, the total WIP level of the fab is 2289.0 wafers corresponding to the average WIP level of the bottleneck 146.2 wafers. Therefore, approximately we consider 2289.0 wafers as the total WIP of the fab under 75% loading. The total WIP level for all non-bottleneck work-centers are 2143.0 (2289.0 - 146) wafers.

(2). For 85% fab loading case, the average WIP level of the bottleneck is 190 wafers. Based on Table 3.3.3, the total WIP level of the fab is 3108.4 wafers corresponding to the average WIP level of the bottleneck 190.2 wafers. Therefore, approximately we consider 3108.4 wafers as the total WIP of the fab under 85% loading. The total WIP level for all non-bottleneck work-centers are 2918.4 (3108.4 - 190) wafers.

(3). For 95% fab loading case, the average WIP level of the bottleneck is 350 wafers. Based on Table 3.3.3, the total WIP level of the fab is 4480.6 wafers corresponding to the average WIP level of the bottleneck 350.5 wafers. Therefore, approximately we consider 4480.6 wafers as the total WIP of the fab under 95% loading. The total WIP level for all non-bottleneck work-centers are 4130.6 (4480.6 - 350) wafers.

3.3.4 Allocating Average WIP Level to Each Non-bottleneck Work-center from the Total WIP Level of the Fab under 75%, 85% and 95% Fab Loadings

In fact, the sum of average WIP level of all work-centers derived from the queuing model is not equivalent to the total WIP level of the fab derived from BPNN. Queuing model may overestimate or underestimate the average WIP level for work-centers. Since there are 104 work-centers in the MIMAC6 model, the overestimation or underestimation effect may become large. Therefore, the total WIP level of the fab acquired from BPNN is used to adjust the target average WIP level of each non-bottleneck work-center derived from queuing models. Equation (3.3.2) [Kuo et al. 2008] is used to determine the average WIP level of non-bottleneck work-centers which is proportional to the total WIP level of the fab.

$$WIPL_i^{BPNN} = \frac{WIPL_i^{QM}}{\sum_{i=1}^{104} WIPL_i^{QM}} \times Total\, WIP\ Level\, of\ \ Fab \qquad (3.3.2)$$

Where $WIPL_i^{QM}$ is the average WIP level of work-center derived from queuing model (Equation (3.3.1)). *Total WIP level of Fab* is the total WIP level of the fab derived from BPNN.

The average WIP level of each work-center, which is based on queuing model and BPNN under 75%, 85% and 95% fab loadings, is listed in Table 3.3.6 in the Appendix. Those average WIP levels for work-centers are used for the simulation study in next section.

3.3.5 Simulation Results and Performance Analysis

First of all we will introduce the FIFO-based-simulation to obtain the target WIP that is applied on the MWVS_1 approach under 95% fab loading. (1):18 months simulation runs of MIMAC6 with FIFO dispatching and 95% fab loading were carried out and the average WIP level for each work-center was acquired; (2): The top 10 work-centers that have the highest average WIP were determined (listed in Table 3.3.5); (3): The average WIP of these 10 work-centers were self-decreased from 2% to 20% in steps of 2% like

‘Avg.WIP’-2%×‘Avg.WIP’, -4%×‘Avg.WIP’, -6%×‘Avg.WIP’ … -20%×‘Avg.WIP’, and used as the target average WIP level for themselves. (4): The average WIP of other work-centers were used as the target WIP level for themselves. By simulation experiments, ‘Avg.WIP’-18%×‘Avg.WIP’ as the target WIP level for the these 10 work-centers could achieve the best performance for MWVS_1 approach listed in Table 3.1.8 (P.58). Actually, these 10 work-centers with high average WIP are high-utilized as well, and the sum of their average WIP accounts for a large part of the total WIP of the whole fab. This is the reason why these 10 work-centers contribute considerable cycle time since the high WIP causes long queue times. Thus, we have the reason to believe the target WIP levels for these 10 work-centers have to decrease, as a result of the importance of preventing long queues instead of preventing starvation. With respect to 75% and 85% fab loading cases, 18 months simulation runs of MIMAC6 with FIFO dispatching and under 75% and 85% fab loading were carried out and the average WIP level for each work-center was acquired and applied as target WIP level for the MWVS_1 approach.

Work-center	***Capacity Loading (%)***	***Rank***	***Work-center***	***Capacity Loading (%)***	***Rank***
20540_CAN_0.43_MII	95.0	1	***13621_IPC_3200***	79.1	7
12553_POSI_GP	93.5	2	***13024_AME_4+5+7+8***	59.8	17
11026_ASM_B2	90.6	3	***11029_ASM_C1_D1***	52.9	26
15121_LTS_3	90.1	4	***11027_ASM_B3_B4_D4***	41.9	38
12331_RST100_1+2	85.6	5	***11024_ASM_A4_G3_G4***	41.4	39

Table 3.3.5: The top 10 work-centers with high average WIP, from FIFO dispatching

We applied the target WIP level derived from these three target WIP determination scenarios described above on the MWVS_1 approach under three fab loading levels. We also consider average cycle times, cycle time variances and cycle time upper 95% percentiles as performance measures. The results are illustrated in Table 3.3.6. For the low fab loading like 75% and 85%, all these three performance measures of three target WIP determination scenarios have no significant difference. It is consistent with the theory that the effect of lot dispatching rule highly relies on the number of bottlenecks in the fab (fab loading) [Waikar et al. 1995, Wein 1988]. As a matter of fact, we can consider these three target WIP determination scenarios as three dispatching rules using target WIP as a parameter. When the fab is running under low loading, most of the work-centers have no congestion, which indicates the lots do not experience long queue time. In such circumstances, the major task of applying target WIP, which is to avoid long queue time, is quite modest. On the contrary, when the fab runs under high loading like 95%, the target WIP has enormous impact on the dispatching rule. The performances of ‘Based on FIFO’ and ‘BPNN’ are superior over the case of ‘Queuing Model’. If we look at the target WIP levels of the top 10 work-centers (listed in Table 3.3.5) in Table 3.3.7 (in Appendix), we can find out that the target WIP levels from ‘Based on FIFO’ and ‘BPNN’ are smaller than ‘Queuing Model’. It tells us that because the target WIP levels of ‘Based on FIFO’ and ‘BPNN’ scenarios are determined more precisely than ‘Queuing Model’, ‘Based on FIFO’ and ‘BPNN’ scenarios can prevent lots from piling up in front of high-utilized work-centers more accurately than ‘Queuing Model’ scenarios, which directly results in queue time reduction. However, we cannot conclude that the queuing model scenario is worse than BPNN scenario in general, because in this study we only apply them on the MIMAC6 model, and we do not have simulation experiments to support when both scenarios are

			Fab Loading (%)		
Performance Measure	***Dispatching Rule***	***Target WIP for Work-center***	***75***	***85***	***95***
Average Cycle Time (days)	***MWVS_1***	***Based on FIFO***	20.4	23.0	28.2
		Queuing Model	20.5	23.0	28.8
		BPNN	20.4	22.9	28.0
	FIFO		20.5	23.3	29.6
Cycle Time Variance (days^2)	***MWVS_1***	***Based on FIFO***	1.4	2.9	6.7
		Queuing Model	1.2	3.1	8.4
		BPNN	1.0	2.6	7.2
	FIFO		0.9	1.2	1.7
Cycle Time Upper 95% Percentile (days)	***MWVS_1***	***Based on FIFO***	25.7	29.5	37.8
		Queuing Model	25.7	29.8	38.5
		BPNN	25.5	29.0	37.0
	FIFO		26.2	29.8	39.1
MWVS_1: Minimum Workload Variability Scheduling with target WIP level					

Table 3.3.6: Three performance measures comparison of MWVS_1 approach setting target WIP level with three different ways

applied on other wafer fab models. In principle, applying historical data to determine target WIP in a statistical manner like BPNN should approach the same value as predicted by queuing model. Nevertheless, the BPNN scenario can overcome the overestimation or underestimation effect of queuing models due to historical data can reflect the internal relations among work-centers which are overlooked by queuing models [Burman et al. 1986, Connors et al. 1996, Kuo et al. 2008, Lin and Lee 2001]. Another conclusion that is attention must be paid to the work-centers both with high historical WIP and high utilization. The ‘Based on FIFO’ scenario tells us that the target WIP levels of work-centers both with high historical WIP and high utilization are more important than the other work-centers, since an overestimated target WIP level leads to longer queue times.

3.3.6 Conclusions

In this section, we investigated three target WIP estimation scenarios which are FIFO-based-simulation, queuing models and back-propagation neural networks, to determine the target WIP for MWVS_1 rule. These three scenarios have their own strengths and weaknesses. The first scenario FIFO-based-simulation is fast, easily adapted and without much complex computation effort. Because it uses the historical data of FIFO dispatching, and adjusts them by simulation experiments based on the fact that the high-historical-WIP and high-utilization work-centers are easier to suffer from long queue and have negative influence in the fab. However, it requires a profound understanding that appropriate target WIP of high-utilized work-centers could lead to queue time reductions under high fab loading. The queuing model is a standard procedure and easily applied because it ignores certain details and responds the questions very quickly under certain conditions. The drawback of queuing models is that they may overestimate or underestimate the target WIP level. In comparison with queuing models, back-propagation neural networks have the advantage that it is not necessary to develop complex algorithms and statistical models since they can be trained with observed data to figure out the hidden relationship between the input data and expected output data. However, it requires huge data training and computation effort to acquire a proper target WIP level.

In this study we cannot judge the merits of these three scenarios because of a lack of comprehensive simulation experiments on different fab models. The objective of this study is not to point out which approach will have the best performance. In contrast, the main perspective is to find out an insight into determining target WIP levels. We can come to the following two conclusions from this study:

- The target WIP plays two major roles which are starvation avoidance and congestion prevention. When the fab runs under a low loading, most of the work-centers have a low WIP. The major effect of target WIP is reduced to only starvation avoidance. Actually, the effect of starvation avoidance is rather limited due to low release rates of products. Therefore, saving capacity cannot directly convert into cycle time reduction. In this case, WIP balance would not be able to achieve cycle time reduction significantly compared to other dispatching rules. These are the reasons why the performance measures of these three target WIP determination scenarios have no prominent differences. Conversely, when the fab runs under a high loading, the situation is different and explained in the following paragraphs.
- The historical average WIP from FIFO dispatching used as the target WIP was already mentioned in the literatures. In fact, the historical average WIP of critical and high-utilized work-centers from FIFO are too high under high fab loading for MIMAC6 model, because FIFO is short of WIP balance mechanisms and dispatches lots inefficiently, which leads to excessive average WIP for high-utilized work-center. We realize that an overestimated target WIP for high-utilized work-center causes the serious problem of increased queue times. Therefore, an appropriate adjustment for the target WIP of high-utilized work-center is very necessary. It tells us that under high fab loading the target WIP of high-utilized work-center has to be taken good care of. Avoiding long

queue times plays a more important role than avoiding capacity loss for high-utilized work-center to achieve cycle time reduction.

This study demonstrated that there are different ways to determine target WIP level for work-centers. However, in general it is hard to say that which one can provide ‘accurate’ target WIP levels. Besides, the performances of WIP balance approaches lie in ‘appropriate’ target WIP level. Moreover, for a customer oriented wafer fab the product release rates change all the time due to frequent changes of customer orders. In this case, the fab runs under different capacity loadings daily, weekly and monthly. The target WIP has to be adjusted and updated accordingly. Thus, in practice it is extremely difficult to determine and apply ‘appropriate’ target WIP levels. Since there are hundreds of work-centers and thousands of operations in wafer fabs, if we consider target WIP as simulation parameter for each work-center or operation, it will lead to huge parameter explosion. Hence, it gives rise to the question, whether WIP balance can be achieved without the requirement of target WIP. Based on the theory that an optimal schedule can be achieved if a large enough information set can be utilized based on which decision is made [Li et al. 1996]. To replace the role of target WIP, we have to take, not only local WIP information of work-centers, but also global WIP information of so called *K*-step ahead and *J*-step back (like upstream and downstream) of work-centers and the whole fab into account. In Section 4.1 of Chapter 4, a global dispatching scheme for work-center extended from MIVS and BMW policies [Ham and Fowler 2007] will be explained in detail with the objective to illustrate how to achieve WIP balance without the need of target WIP.

3.4 Due Date Control - Due Date Oriented Rules

3.4.1 The Characteristics of Due Date Control

To begin with, first we would like to explain the reasons why we include ‘Due Date Control’ to ‘WIP Balance’ chapter. As we discussed above, it seems that WIP balance and due date control are two different goals in the fab, they might even conflict with each other under some circumstances. However, the inherent characteristic of due date control, which is the capability of minimizing lateness variance that leads to cycle time variance minimization, exactly makes up for the deficiency of WIP balance. In addition, the due date control rules present interesting facts that they can achieve WIP balance as well with proper due date setting, even though utilizing no WIP balance information. It drives us to acquire some insight for due date control rules.

When the performance measures involve on-time delivery and lateness, due date information is spontaneously employed for dispatching decision, which is well known as due date oriented rule. In general, the lateness is defined as the difference between the due date and the completion time. Suppose each 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:

$$L_i = F_i - D_i \begin{cases} < 0, \textit{ early} \\ \\ > 0, \textit{ tardy} \end{cases} \qquad (3.4.1)$$

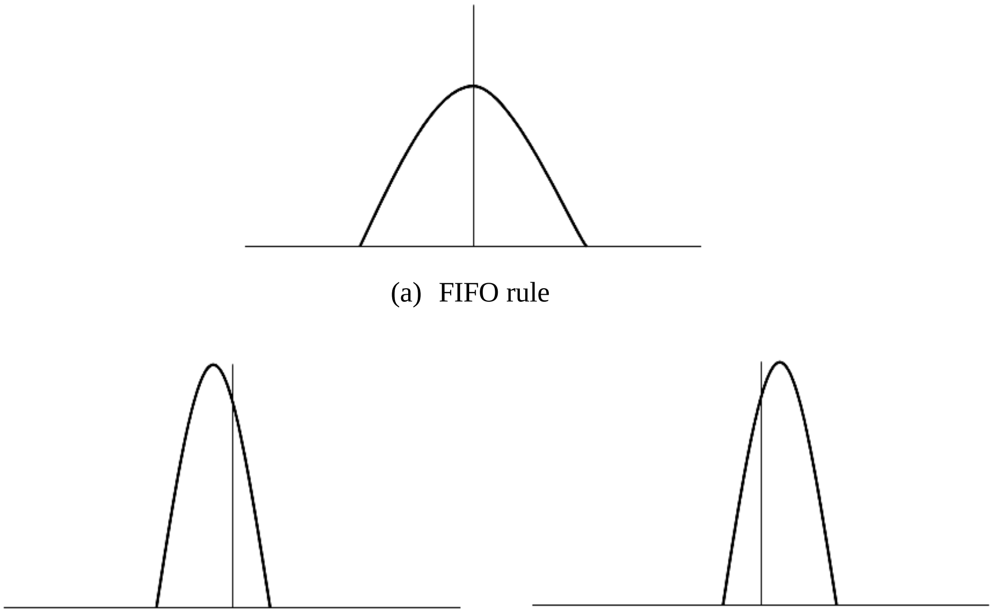
Regardless of being early or tardy, the due date oriented rules are designed to minimize the average lateness and lateness variance [Baker and Trietsch 2009]. The due date plays the role in setting a target which forces lots to catch up with. As soon as a due date is established, the due date rule keeps lots going through the fab as close to their due date as possible. Figure 3.4.1 presents us three different cases regarding to lateness distributions. The due date is represented by vertical axis, which means on the right side of vertical axis is tardy while left side is early. Figure 3.4.1 (a) represents lateness generated by the rules ignoring due date information like FIFO. When due date rules are employed, Figure 3.4.1 (b) represents a low variance of lateness while a low average lateness is achieved as well. In such circumstance, the due date rules manage to finish the lots before the due date as much as possible. However, for some reasons like when a tight target due date is required, the due date rules fail to finish lots before their due dates but as close to the due dates as possible, which still leads to a low variance but increased average as depicted in Figure 3.4.1 (c).

- In a word, disregarding low or high average lateness, the due date rules tend to minimize the variance of lateness, which is the most powerful strength of due date rules. This advantage can directly give rise to good cycle time variance performance.

Suppose at time *t*, lot *i* is released into the fab and has a release time R_i. We simply define the due date D_i of lot *i* as its release time R_i and assume lot *i* is tardy. The following comes true from Equation (3.4.2):

$$L_i = F_i - D_i = F_i - R_i = CT_i \qquad (3.4.2)$$

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



(a) FIFO rule

(b) Minimized variance with lower average

(c) Minimized variance with higher average

Figure 3.4.1: Hypothetical lateness distribution created by due date rules

3.4.2 Due Date Oriented Rules

There are different kinds of due date rules which can be classified into four main types according to the due date information.

- Allowance-based due date rules;
- Slack-based due date rules;
- Ratio-based due date rules;
- Composite due date rules.

(1). Allowance-based rules

The lot's allowance is the time difference between the due date and release time. It is the time that a lot is expected to stay in the fab, and it cannot be tolerated if the allowance is past and the lot still remains in the fab. The remaining allowance is used to measure the urgency of lots. If we need to make a dispatching decision at time *t*, then the remaining allowance of lot *i* $A_{i,t}$ is represented as: $A_{i,t} = D_i$ - *t*. The smaller the allowance, the more urgent the lot is. Since *t* is the same for all lots in the same queue, the urgency can be expressed as due date, which makes the classical allowance-based rule - Earliest Due Date (EDD).

- **Earliest Due Date (EDD)**: The lot with the earliest due date has the highest priority.

The due date used in the EDD rule, which is lot-based due date, sets a milestone to a lot to catch up with. From operation-based viewpoint, each operation could have a due date as well which is called operation due date. To better distinguish operation due date, the due date in EDD rule is also called final due date. Once the final due date is established, the operation due date breaks up the lot's allowance into as many segments as the number of the operations of a lot, which makes the Operation Due Date (ODD) rule.

- **Operation Due Date (ODD)**: The lot with the earliest operation due date has the highest priority. The operation due date is determined by dividing the allowance into as many sub-allowances as the number of operations.

The ODD for operation *j* of lot *i* is defined as follows:

$$ODD(i,j) = ReleaseTime(i) + RPT(i,j)*DDFF \qquad (3.4.3)$$

Where *ReleaseTime(i)* is the release time of lot *i*, *RPT(i,j)* denotes the raw processing time for a sequence of operations from operation 1 to operation *j* (including operation *j*) of lot *i*. *DDFF* is defined as the target cycle times divided by the raw processing time. For instance, a *DDFF* of 2 says that a lot spends half of its cycle time in processing state and the other half in non-processing states like waiting. Thus, the due date of a lot is the time when it enters the fab plus *DDFF**RPT*. For the final operation of a lot the ODD is equal to the classical due date as used in EDD. Equation (3.4.3) is only one way to determine ODD. Interested readers can find other ODD expressions in literature [Baker 1984]. In the following simulation experiment, the *DDFF* values for all products at every operation are equal. In reality, this is usually not the case because some products are more important than others, or even some operations are more important than other operations.

The ODD rule tends to keep lots going through the fab strictly at the right pace toward on-time completion. Therefore, the ODD rule is expected to outperform EDD rule regarding the cycle time variance performance.

Except allowance, another factor to measure the urgency of lot is the number of operations remaining. When two lots have the same allowance, the lot with the larger number of operations is more urgent since it will encounter more chances to delay in the queue. Hence, one more allowance-based rule called Allowance per Operation (A/OPN) appears.

- **Allowance per Operation (A/OPN)**: The lot with the least remaining allowance per operation obtains the highest priority.

$$A/OPN(i) = (Due(i) – Now) / RemainingOPN \qquad (3.4.4)$$

where *Due(i)* is the final due date of lot *i*, *Now* is the current time, and *RemainingOPN* denotes the number of remaining operations.

(2). Slack-based rules

Besides allowance information, slack-based rules consider the raw processing time information of lots as well. The slack is the remaining allowance re-adjusting the remaining raw processing time. One representative of slack-based rules is called Least Slack Time (LST) defined as follows:

- **Least Slack Time (LST)**: The lot with the least slack time has the highest priority. The slack time is the difference between the due date and remaining raw processing time.

$$LST(i) = Due(i) - Now - RemainingRPT(i) \qquad (3.4.5)$$

where *Due(i)* is the final due date of lot *i*, *Now* is the current time, and *RemainingRPT(i)* denotes the remaining raw processing time of lot *i*.

The LST rule is an extension to the EDD rule for the reason that it tells us if two lots have the same due date, the lot with longer remaining raw processing time is more urgent because its due date allows less delay.

Since the allowance can be divided into as many sub-allowances as the number of operations that leads to ODD, similarly, the slack time can also be divided for operation which leads to Least Operation Slack Time (LOST).

- **Least Operation Slack Time (LOST)**: The lot with the least slack time for operation has the highest priority.

$$LOST(i,j) = ODD(i,j) - Now - RemainingRPT(i) \qquad (3.4.6)$$

Where *ODD(i,j)* is the operation due date of lot *i* at operation *j*, *Now* is the current time, and *RemainingRPT(i)* denotes the remaining raw processing time of lot *i*.

When the number of remaining operations are considered to the slack, similar to A/OPN, a slack-based rule called Slack per Operation (S/OPN) is created.

- **Slack per Operation (S/OPN)**: The lot with the least slack per operation obtains the highest priority.

$$S/OPN = (Due(i) - Now - RemainingRPT(i)) / RemainingOPN \qquad (3.4.7)$$

where *Due(i)* is the final due date of lot *i*, *Now* is the current time, and *RemainingRPT(i)* denotes the remaining raw processing time of lot *I*, *RemainingOPN* denotes the number of remaining operations.

(3). Ratio-based rules

The ratio-based rules are variants of slack-based rules. The ratio-based rules use ratio between the remaining time to due date and remaining raw processing time instead of difference like slack-based rules to measure the urgency of the lot. One classical rule is called Critical Ratio (CR).

- **Critical Ratio (CR)**: The lot with the smallest CR value gets the highest priority. A CR value of less than 1 denotes a lot which falls behind schedule, a CR value equivalent to 1 means that a lot is on schedule, a

CR value of greater than 1 represents a lot which is ahead of schedule and has slack time left.

$$CR(i) = (Due(i) - Now) / RemainingRPT(i) \tag{3.4.8}$$

where *Due(i)* is the final due date of lot *i*, *Now* is the current time, and *RemainingRPT(i)* denotes the remaining raw processing time of lot *i*.

Likewise, we can also determine a ratio for each operation like LOST rule, which leads to Operation Critical Ratio (OCR).

- **Operation Critical Ratio (OCR)**: The lot with the smallest operation critical ratio gets the highest priority.

$$OCR(i,j) = (ODD(i,j) - Now) / RemainingRPT(i) \tag{3.4.9}$$

where *ODD(i,j)* is the operation due date of lot *i* at operation *j*, *Now* is the current time, and *RemainingRPT(i)* denotes the remaining raw processing time of lot *i*.

(4). Composite rules

The performances of those due date rules above are mainly affected by how tight or loose the due date is set [Elvers 1973]. Some rules perform better with tight target due dates like SPT, although SPT does not use any due date information, while some rules perform better with loose target due dates like EDD and ODD. By noticing the complementary strengths of different rules working with different target due dates, Baker and Bertrand [1981] presented the composite Modified Due Date (MDD) rule which is a combination of Least

Work Remaining (LWR) and EDD. It performs like LWR if the target due date is tight and like EDD if the target due date is loose. Baker and Kanet [1983], proposed to use ODD to replace EDD, and SPT to replace LWR, thus resulting in a new composite rule called Modified Operation Due Date (MOD). They believed that the MOD rule is more effective than MDD rule with respect to tardiness performance in the job shop.

- **Modified Due Date (MDD)**: The lot with the earliest modified due date has the highest priority. The modified due date is either the original final due date or the earliest finish time, whichever is larger.

$$MDD(i) = Max\{Due(i), Now + RemainingRPT(i)\} \qquad (3.4.10)$$

where *Due(i)* is the final due date of lot *i*, *Now* is the current time, and *RemainingRPT(i)* denotes the remaining raw processing time of lot *i*.

- **Modified Operation Due Date (MOD)**: The lot with the earliest modified operation due date has the highest priority. The modified operation due date is either the original operation due date or the earliest operation finish time, whichever is larger.

$$MOD(i,j) = Max\{ODD(i,j), Now + PT(i,j)\} \qquad (3.4.11)$$

Where *ODD(i,j)* is the operation due date of lot *i* at operation *j*, *Now* is the current time, and *PT(i,j)* is the processing time of lot *i* at operation *j*.

The MDD and MOD rules attempt first to complete the lots early or on time, second to complete the lots as soon as possible when the requested due date is unattainable. Assuming that with a loose target due date, all lots have positive slack, MDD rule performs as EDD rule. While with a tight target due date, all lots have negative slack, MDD rule performs as LWR rule. The MOD rule performs in a similar way like MDD rule. When all lots have positive slack for operation, MOD dispatches on the basis of ODD, but when all lots have negative slack for operation, MOD dispatches on the basis of SPT.

3.4.3 Due Date Tightness Setting

The most critical and difficult task when applying due date rules is about the due date tightness, since the performance of due date rules is highly sensitive and dependent on the due date tightness. In a real wafer fab, the due dates are usually provided by the planning department through negotiating with customers. The due dates can be changed during the manufacturing process due to the dynamic market and manufacturing fashion, e.g., the earlier delivery requirements from customers, which imposes additional challenges to apply due date rules. In general, there are two types of due date information which are lot-based and operation-based. For instance, classical rules like EDD, LST and CR only specify one due date (also called final due date for the last operation), which leads to lot-based due date, in contrary, the variants of classical rules specifying due date for each imminent operation give rise to operation-based due date versions like ODD, LOST and OCR. In this study, we are interested in finding out the general performance of due date rules corresponding to the change of due date tightness. Additionally, two main issues that are (1): which due date setting is more effective for due date rules; (2): how the average cycle time and cycle time variance are affected by the due date tightness, are expected to be figured out.

Normally, the due date of a lot of a specific product is given in terms of a target due date flow factor (DDFF) discussed above. It is difficult to decide exactly which DDFF values should be given to which products from an academic viewpoint without knowing any requirements and constrains to set due dates, especially without any support from a real planning department. As a consequence, in this study we apply the same DDFF for all products to avoid the explosion of the parameter space for the simulation experiments. Typically, as we divide the fab loading into three levels, we determine that for 95% and 85% fab loading the DDFF ranges from 1.5 to 2.9 in steps of 0.2, for 75% fab loading the DDFF ranges from 1.3 to 2.7 in steps of 0.2. The range of DDFF can be split into three parts, e.g., for the 95% fab loading:

- Low DDFF (1.5 to 1.9): the target due date is tight in terms of these DDFFs. When FIFO is applied 100% of lots are tardy. Although in practical manufacturing it is very hard to utilize these DDFFs, one interesting phenomenon arising from them is that they cause WIP imbalance for the fab. Since we spend much effort to investigate the cause and solution for WIP imbalance, it is well worth studying these low DDFFs. In particular more attention has to be paid if the fab is running with tight target due dates.

- Medium DDFF (2.0 to 2.4): These DDFFs are reasonable because they can be achieved. Normally, the best performance of due date rules is expected from them. They provide insight into due date rules with regard to tardiness control, for the reason that tight DDFFs are difficult to achieve while loose DDFFs are easy to meet. It is also interesting to notice that the WIP achieved by the medium DDFFs is more balanced than the low and high DDFFs.

- High DDFF (2.5 to 2.9): These DDFFs are considered as loose because all lots achieve zero tardiness when applying FIFO, thus, they are of less interest. In practical application, high DDFFs are hardly considered since they cannot accelerate lot movement. Moreover, high DDFFs bring less competitiveness to customer oriented companies in today's fierce market competition.

3.4.4 Simulation Results and Performance Analysis

Average cycle time, cycle time variance, cycle time upper 95% percentile, percent tardy lots and average tardiness of tardy lots are considered as performance measures. We focus on 95% fab loading case since we can obtain more insight into the relation that is the interaction between the behavior of due date rules and WIP imbalance under high fab capacity loading. We list the simulation results of 85% and 75% fab loading in the appendix as well.

3.4.4.1 Allowance-based Rules vs. Composite Rules

Average cycle time: In Figure 3.4.2, firstly for the allowance-based rules, the lot-based due date version EDD rule produces considerable average cycle time regardless of due date tightness. Whereas, the operation-based due date version ODD and A/OPN rules perform quite well except for the tight due dates, the average cycle times under tight due dates are relatively higher than in the case of medium and loose due dates, and both rules achieved the best performance at a medium due date (2.3 DDFF). The ODD and A/OPN rules have a similar performance since they progress lots based on operation milestones. These results suggest that the pacing introduced by operation milestones (operation-based due dates) is more stable and robust than final due date (lot-based due dates), thus the WIP produced by ODD and A/OPN rules is more balanced than EDD rule does. The composite rules MDD and MOD which are variants of EDD and ODD, respectively, show an interesting performance. When the target due date is tight, let us say DDFF 1.5, 1.7 and 1.9, EDD and ODD rules perform poorly because overemphasizing due date control causes WIP imbalance for critical work-centers, thus resulting in high cycle time. In contrast, for the composite rules MDD and MOD, the introduction of LWR rule to EDD rule and SPT rule to ODD rule show remarkable improvements. The LWR and SPT rules breaks the dominance of EDD and ODD rules under tight due dates. Therefore, the MDD and MOD rules achieve lower average cycle times compared to EDD and ODD, respectively. As due dates become larger and larger, since MDD rule performs like EDD and MOD performs like ODD, the average cycle time of MDD rule has a trend to come close to the average cycle time of EDD rule, and MOD rule differs slightly from the ODD rule. This behavior of the composite rules MDD and MOD provide us a valuable insight into dealing with WIP imbalance which is caused by tight due dates. In Section 6.2, a new composite rule combining MOD and LWNQ rules is inspired by the observation of simulation results of composite rules in this section.

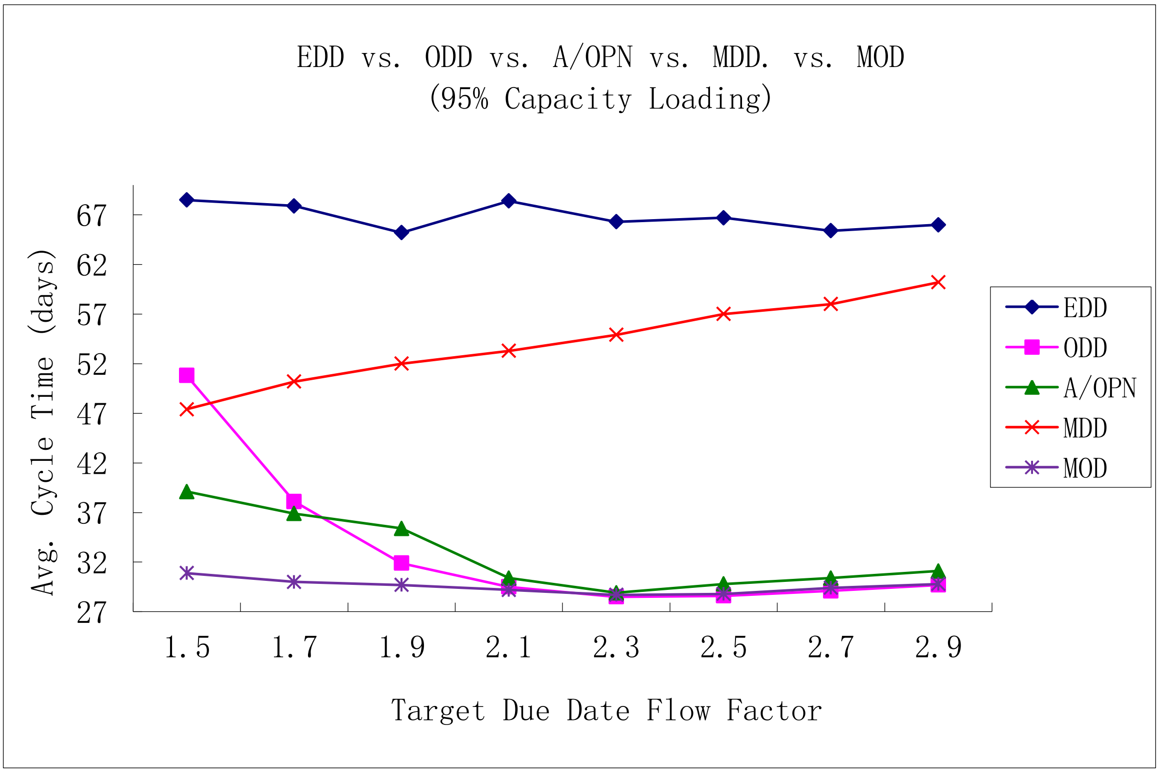


Figure 3.4.2: Average cycle time comparison of allowance-based rules vs. composite rules

Cycle time variance: Figure 3.4.3 presents the cycle time variance performance and Figure 3.4.4 shows the cycle time upper 95% percentile performance. The foregoing discussion suggests that the operation-based due date is more effective than lot-based due date with regard to lot pace movement, because the operation milestones strictly keep lots progressing at the right pace to meet the operation due dates, which is certainly confirmed by the simulation results. Combined with the cycle time upper 95% percentile performance, ODD and A/OPN rules definitely outperform EDD rule, and MOD is superior to the MDD rule. In fact, due to the considerable cycle time performance of EDD and MDD, their cycle time variance performances become less interesting. The ODD and MOD rules dominate the A/OPN rule, as ODD is a part of MOD, we can say that ODD rule has a remarkable performance of low cycle time variance.

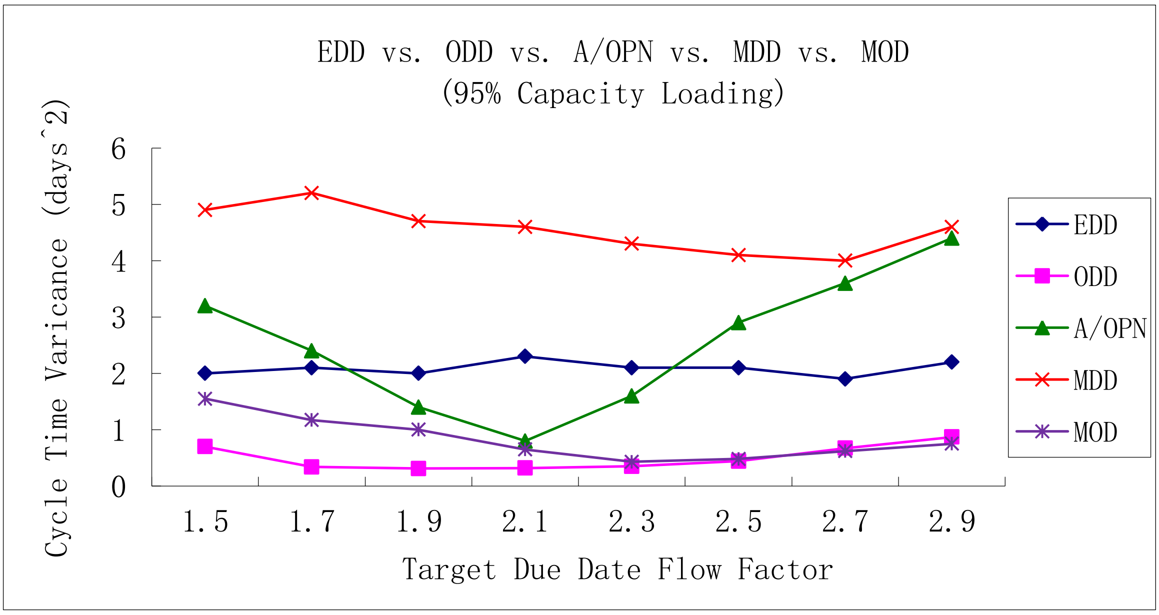


Figure 3.4.3: Cycle time variance comparison of allowance-based rules vs. composite rules

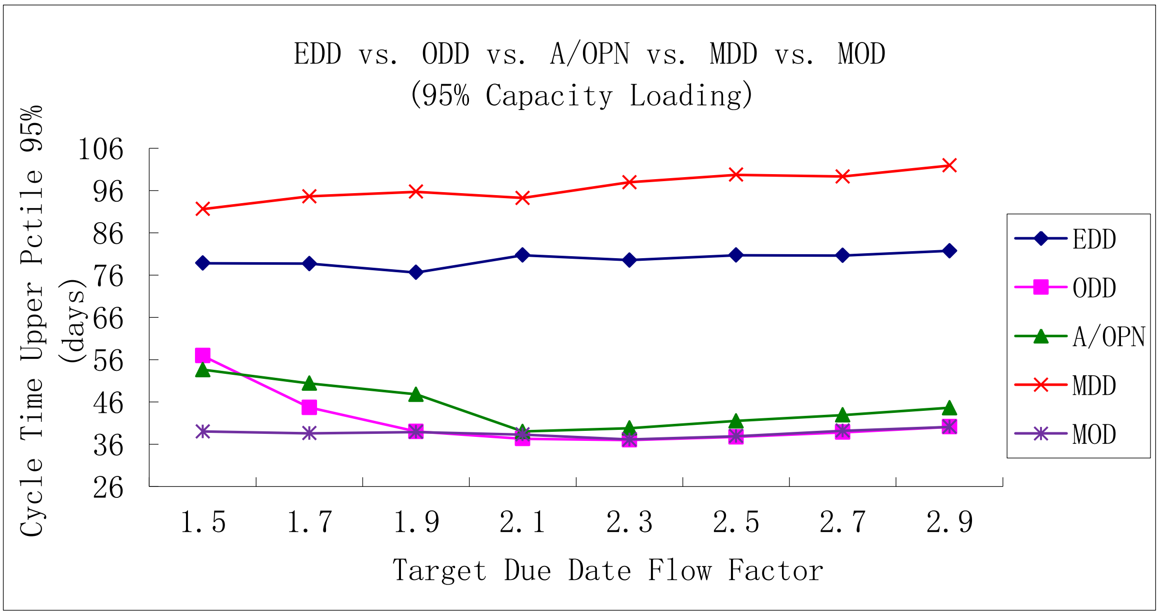


Figure 3.4.4: Cycle time upper 95% percentile comparison of allowance-based rules vs. composite rules

On-time delivery: EDD and MDD rules achieve almost 100% tardy lots and produce high tardiness for tardy lots under all target due dates in Figure 3.4.5 and 3.4.6. The MOD rule dominates ODD and A/OPN when target due dates are tight. It can be explained by the fact that MOD performs like SPT which intends to process lots quickly without consideration of due date control. As due dates become loose, MOD is slightly outperformed by ODD and A/OPN.

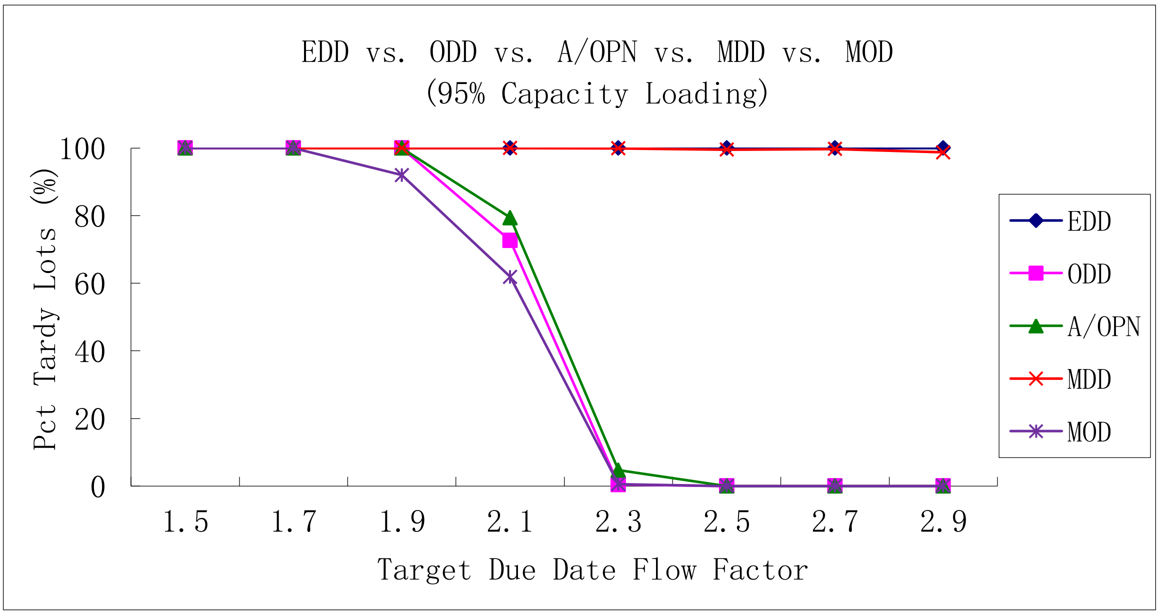


Figure 3.4.5: Percent tardy lots comparison of allowance-based rules vs. composite rules

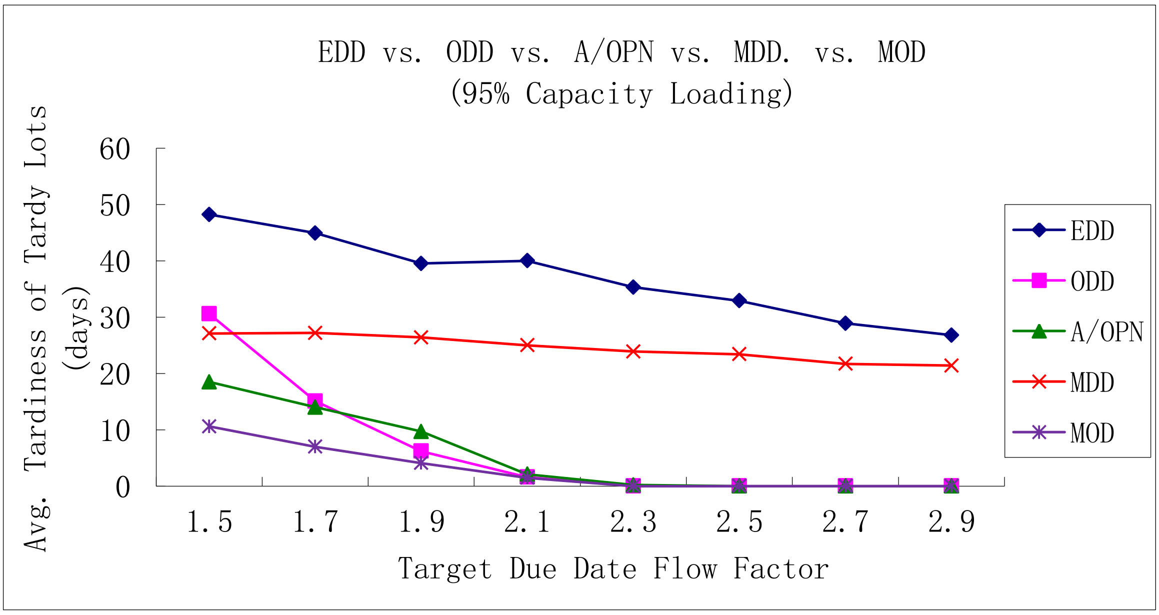


Figure 3.4.6: Average tardiness of tardy lots comparison of allowance-based

rules vs. composite rules

3.4.4.2 Slack-based Rules vs. Ratio-based Rules

Average cycle time: In Figure 3.4.7, the LST rule has a similar performance as EDD rule, as lot-based due date plays a major role in LST. On the contrary, when the lot-based due date is replaced by operation-based LOST rule achieves promising improvements. It is surprising that LOST still produces low average cycle times under tight due dates. Apparently, there is a crossover for the two operation-based rules: the LOST rule is better than S/OPN under tight due dates, while S/OPN outperforms LOST under medium and loose due dates. CR rule utilizes lot-based due date as well, whereas, it shows different behavior in comparison to the LST rule. The CR rule achieves considerable cycle time as LST does when due dates are tight (DDFF 1.5, 1.7 and 1.9). Then CR has sudden performance improvements under medium and loose due dates. Different from LST, CR is a dynamic rule and assigns priority to lot dynamically over time, which turns out to be effective to overcome the drawback of lot-based due dates. Once again, the operation-based version OCR successfully reduces average cycle times under tight due dates compared to CR. These comparisons suggest again that the rules utilizing operation-based due dates are more effective than the ones utilizing lot-based due dates.

Cycle time variance: From Figure 3.4.8 we cannot draw clear conclusions about whether slack-based or ratio-based rule can achieve the best cycle time variance, since there are some crossover points of all rules. It seems that under tight due dates LOST rule is preferable, which is confirmed by the fact that LOST rule achieves the lowest cycle time compared to other rules in Figure 3.4.7. Under medium and loose due dates, except for LST rule, the remaining four rules differ slightly.

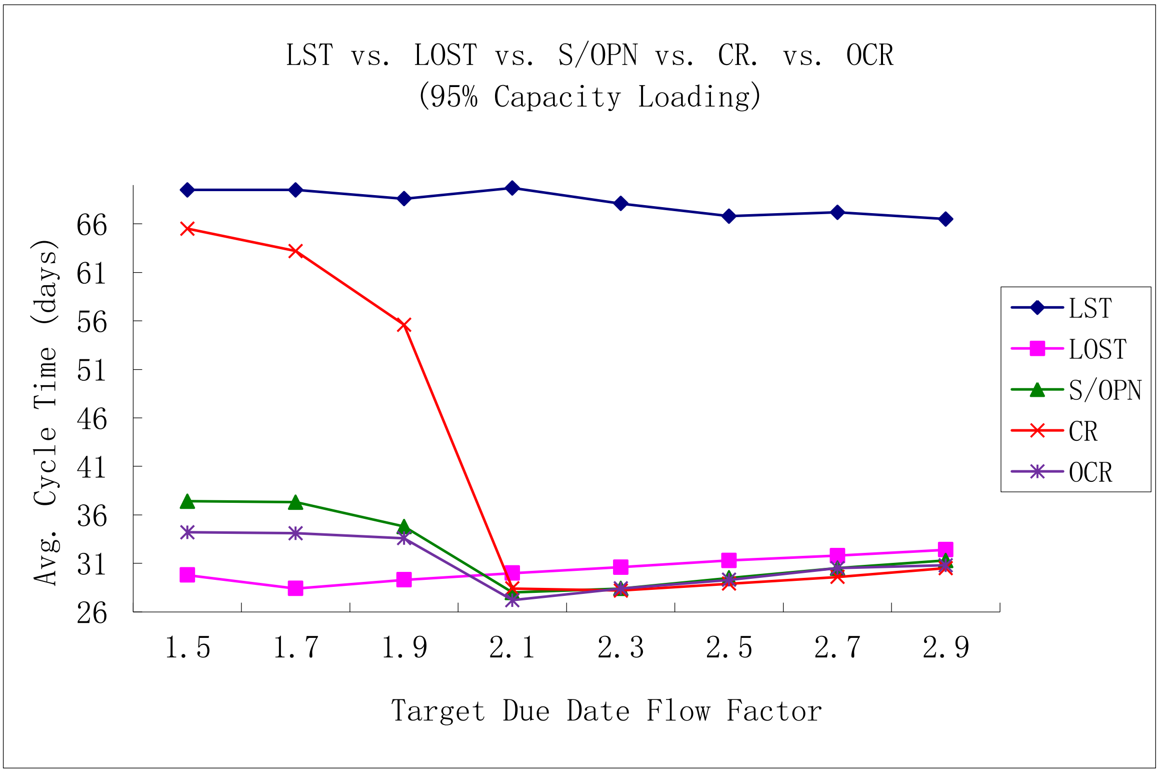


Figure 3.4.7: Average cycle time comparison of slack-based rules vs. ratio-based rules

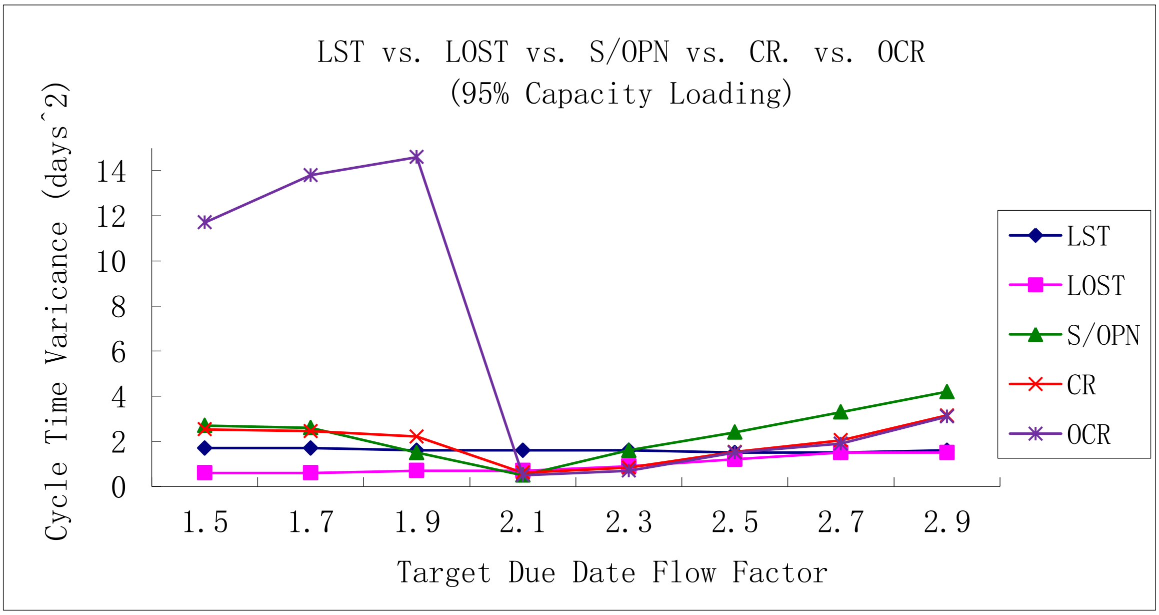


Figure 3.4.8: Cycle time variance comparison of slack-based rules vs. ratio-based rules

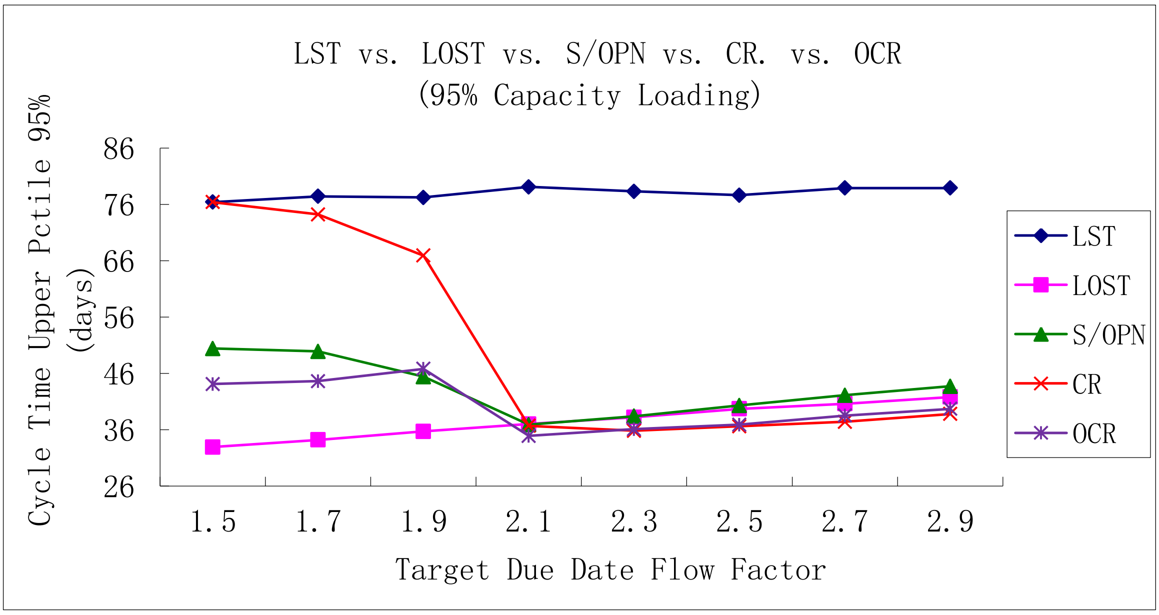


Figure 3.4.9: Cycle time upper 95% percentile comparison of slack-based rules vs. ratio-based rules

On-time delivery: In Figures 3.4.10 and 3.4.11, the LST rule always yields 100% tardy lots and high tardiness for tardy lots regardless of the tightness of due dates. Under tight due dates (DDFF 1.5, 1.7 and 1.9), the remaining four rules have similar performance of percent tardy lots, but LOST rule dominates other three rules with regard to tardiness. The picture is different for the medium due date (DDFF 2.1), OCR rule produces the lowest percent tardy lots and tardiness. For the loose due dates, S/OPN, CR and OCR rules achieve zero tardiness, but LOST rule performs differently and produces still significant tardiness until DDFF 2.7.

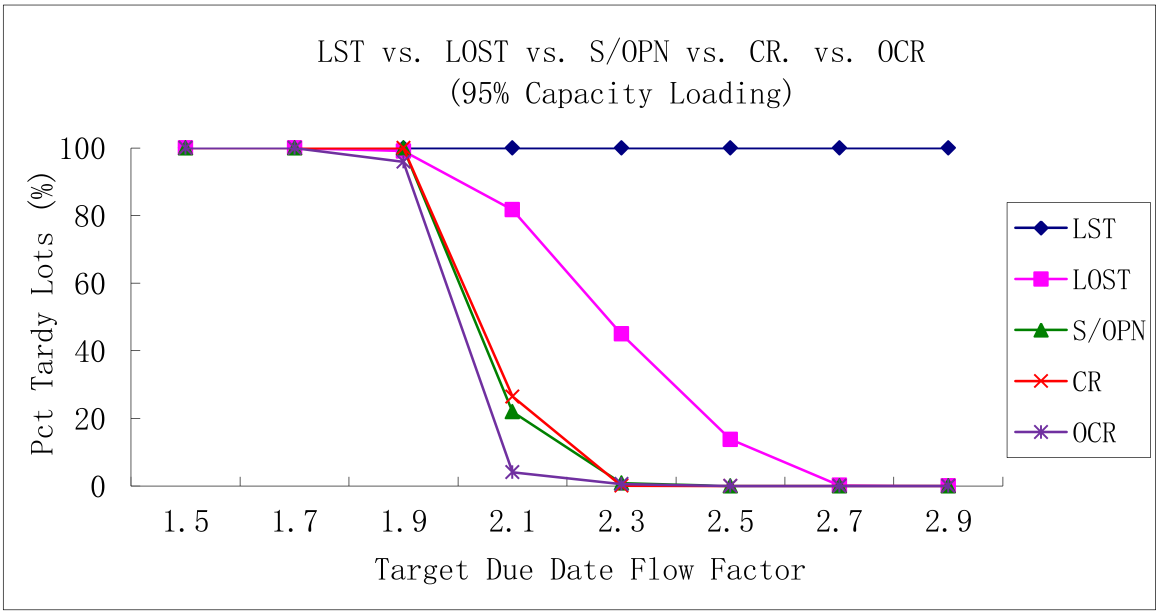


Figure 3.4.10: Percent tardy lots comparison of slack-based rules vs. ratio-based rules

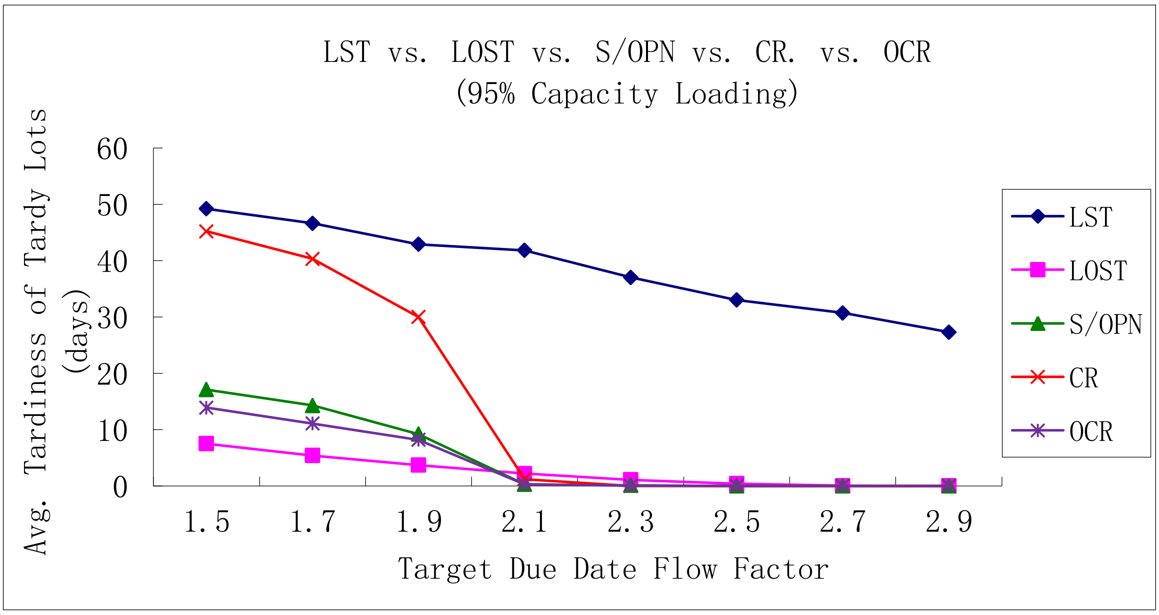


Figure 3.4.11: Average tardiness of tardy lots comparison of slack-based rules vs. ratio-based rules

3.4.5 Conclusions

A comprehensive study about the performance of various due date oriented rules was investigated in this section. These due date rules can be classified into allowance-based, slack-based, ratio-based and composite rules. They utilize two kinds of due date information which are lot-based and operation-based. Because there are some crossover points for the due date rules from the simulation results, we cannot draw a clear conclusions which kind of due date rule performs better. Whereas, if we look at the due date rules from the viewpoint of due date information, we can obtain the following facts:

In MIMAC6 model, we found out that it was difficult to apply lot-based due date rules that are EDD, MDD, LST and CR under high capacity loading case. EDD, MOD and LST always produced considerable cycle time in spite of due date tightness, and CR had a sudden performance degradation when due dates were tightened. In contrast, operation-based due date version rules that are ODD, A/OPN, MOD, LOST, S/OPN and OCR performed differently. Under tight due dates they achieved relatively high average cycle times but performed well under medium and loose due dates.

- As a general guideline, the rules utilizing operation-based due date are more effective than the ones utilizing lot-based due date.

The average cycle time performance in Figures 3.4.2 and 3.4.7 demonstrated that under tight due dates most of the rules, except for LOST rule, achieved relatively higher cycle time than in the cases of medium and loose due dates. In other words, tight due dates lead to WIP imbalance for those rules in MIMAC6 model, which indeed may not be true in other models. However, since the WIP imbalance occurred because of the tight due dates, we concentrated more on how to solve this problem instead of investigating the performance of due date rules under tight due dates in other models. If we made comparisons between, for example, EDD and ODD (or A/OPN), LST and LOST (or S/OPN), CR and OCR, we observed that the WIP of operation-based due date rules was more balanced than lot-based due date rules. Furthermore, the comparisons between EDD and MDD, ODD and MOD told us that the composite rules achieved relatively balanced WIP because they broke the dominance of due date control under tight due dates.

- When the fab is running with tight due dates products under high capacity loading, in order to prevent from high WIP, either using operation-based due date rules or using composite rule like MOD to break the dominance of due date control would be preferable.

In particular, the remarkable performance of the MOD rule appear to make it a desirable choice under conditions where we cannot guarantee medium or loose due dates to the products. As a matter of fact, the MOD rule provides us a new method to tackle the conflict between WIP balance and due date control. Although SPT rule does not target at WIP balance, its anti-due date control embodied in MOD rule has the positive effect to avoid WIP imbalance.

Among the five performance measures, we highlighted the cycle time variance since it is the key point to overcome the drawback arising from WIP balance approaches we studied in the previous sections.

- If we consider the cycle time variance alone in Figures 3.4.3 and 3.4.8, the ODD rule absolutely dominated other rules. Along with cycle time upper 95% percentile in Figures 3.4.4 and 3.4.9, MOD and LOST rules showed competitiveness to ODD under tight due dates. Obviously, MOD and LOST originated from ODD.

The major concern for WIP balance approaches is to reduce the cycle time variance, thereby, the operation milestone due date from ODD provides fairly comprehensive understanding for operational control.

CHAPTER 4

EXTENSION TO WORK-CENTER ORIENTED WIP BALANCE

This chapter is an extension to Chapter 3 to address three issues.

Section 4.1 intends to deal with the first issue that is to balance the WIP for work-centers without the need of target WIP. This is important because there is a strong argument from industry that without WIP targets the WIP balance approach is not feasible as there are no criteria to determine if a work-center is starved or crowded.

Section 4.2 attempts to underline the other two issues. Firstly, it will report the importance and possibility to control the WIP flow by taking workload information of work-center and lot status information into account. Since WIP imbalance causes WIP increase and fluctuation, a new WIP imbalance monitor and calibration approach will be introduced to smooth the WIP flow.

4.1 A Global WIP Oriented Dispatching Scheme: Work-center Workload Balance without Requirement of Target WIP

4.1.1 Why Abandon Target WIP

In previous chapter, Section 3.1 we proposed MWVS rule using target WIP to balance WIP for work-centers, and we discussed the related issues of the importance and ways to determine target WIP for MWVS in Section 3.3. In spite of the promising results achieved by MWVS with target WIP, we still have a strong desire to abandon target WIP. In the following we explain several major reasons to support our viewpoint.

- Much effort has to be spent to apply target WIP.

(1): Since the performance of the fab is sensitive and highly relies on the target WIP, uncertainty of product volume mix and almost daily changing lot release rate due to frequent change of customer orders cause the necessity to update the target WIP weekly, daily, or even hourly accordingly.

(2): According to the Theory of Constraints (TOC), the bottlenecks determine the throughput of the fab. The objective of target WIP for bottlenecks (critical work-centers) is to waste no capacity and avoid starvation. Here comes the problem that a misleading target WIP could result in starvation or overload to the bottlenecks, which has significant impact on throughput of the fab. Not to mention that the bottlenecks are dynamically changing in reality. Hence, it is not easy to manage the target WIP of bottlenecks. We should realize that stressing the importance of the target WIP of bottlenecks does not weaken the importance of target WIP of non-bottlenecks. On the contrary, because wafer fab includes hundreds of work-centers (including bottlenecks and non-bottlenecks), the situation becomes more complicated.

- Much effort has to be spent to acquire target WIP.

(1): There are different ways to set up target WIP for work-centers. In practice, the target WIP is determined by trial-and-error approaches or by experiences of engineers. As described in Section 3.3, we explored three different ways and tried to establish ‘appropriate’ target WIP for work-centers. However, we cannot draw a conclusion which way is the best, since they all have advantages and disadvantages.

(2): On one hand there should be a total target WIP for the whole wafer fab, which makes sure there is sufficient WIP to achieve high utilization of the bottlenecks. On the other hand, the total target WIP should be allocated to the work-centers as individual target WIP for themselves. Due to the dynamic environment of wafer fabs, the loadings of work-centers change all the time, which requires the target WIP of the whole fab and the work-center to be adjusted and updated accordingly. How to deal with the interaction between the total target WIP for the whole wafer fab and the individual target WIP for the work-centers is a challenging task.

(3): As we all know, the more parameters are taken into account, the more difficulties are imposed on simulation. Due to hundreds of work-centers in the wafer fab, the explosion of the parameter space will be hard to handle for the simulation experiment if we consider the target WIP of each work-center as simulation parameters.

- What is an appropriate or acceptable target WIP level?

Last but not least, in practice even if we have an effective way to determine target WIP and have an outstanding WIP balance approach to apply target WIP, for such an important key performance indicator, it is worrying that a lot of engineers simply do not know how much WIP they have at any point in time for the work-centers or the whole fab, let alone control or balance it. A misleading target WIP results in starvation or overload, which has a big impact on cycle time and on-time delivery. Appropriate or acceptable target WIP varies greatly according to a number of factors, e.g., product type, production system, upstream and downstream supply chain, management capability and so on. The planning department has to make a deep and full research on the capacity analysis to determine the capacity loadings and constraints, the daily output, etc. which strongly influence the appropriate or acceptable target WIP.

In the literature, seldom papers address the possibility of balancing WIP without target WIP. In Section 3.1.3 of Chapter 3 we introduced one-step-ahead and one-step-back MWVS as our first step, because the information set employed by one-step-ahead and one-step-back MWVS is rather limited, the performance is not promising. The objective, which is to develop WIP balance scenarios to replace the role of target WIP, drives us to this section.

In order to do this, firstly we need to understand what kind of role the target WIP plays. Basically, the WIP control philosophies used in wafer fab can be classified into push and pull [Fowler et al. 2002, Strum et al. 1999]. Generally speaking, push approaches are about what upstream work-center desires to produce and pull approaches are about what downstream work-center is capable of producing. For instance, if the wafer fab is controlled by the push approaches, a work-center is scheduled to work whenever it is available for processing, regardless of the downstream status. It causes the problem that some work-centers are overloaded while some are starved, and WIP becomes imbalanced in the manufacturing line. When the wafer fab is controlled by pull approaches, a work-center is scheduled to work according to the request from downstream work-centers. Therefore, it can avoid the problem arising from push approaches. It is obvious that the target WIP is used to regulate the workload of work-center.

- It is not difficult to figure out that the role of target WIP is to measure the pull request of downstream work-centers (also described as starvation avoidance and congestion prevention). If we expect to get rid of the target WIP, we need to develop dispatching scenarios replacing target WIP to measure the pull requests of downstream work-centers.

Next, we will discuss what kind of information can be employed to replace the target WIP. In Section 3.1.2 of Chapter 3 we have introduced the MIVS rule (also called one-step-ahead MIVS). Indeed, because wafer fab environments dynamically change, WIP distributes dynamically as well. Only considering one-step-ahead information may be too myopic to achieve better decision making. Thus, the one-step-ahead MIVS rule could be extended to *K*-step-ahead and *J*-step-back MIVS [Collins and Palmeri 1997, Li et al. 1996]. In practice, *K* and *J* can be large enough to include the entire line in wafer fab. This look-ahead and look-back policies use global information to dynamically analyze the WIP distribution. In fact, it is consistent with the theory that an optimal schedule can be achieved, if the information set based on which decision is made is large enough. Similar to MIVS considering WIP flow as the viewpoint of operation, we can also consider the work-center as a subject for WIP flow. Each work-center has its upstream and downstream work-centers based on the WIP flow, which means the work-centers are connected by the WIP flow. If we define the workload of work-center *i* at operation *t* as follows:

$$W_{i,t} = \frac{Lots_t \times (LoadTime_t + Process\,Time_t + UnloadTime_t)}{NumberofMachines_i} \qquad (4.1.1)$$

Where $Lots_t$ is the number of lots at operations *t* at the queue of work-center *i*. $LoadTime_t$, $Process\,Time_t$ and $UnloadTime_t$ are the load time, raw processing time and unload time of lots at operation *t*, respectively. $NumberofMachines_i$ is the number of machines in work-center *i*.

Then we can calculate the sum workload of work-center *i* as follows:

$$W_{i,sum} = \sum_{t=0}^{n} W_{i,t} \qquad (4.1.2)$$

Once we obtain the workload information of work-centers dynamically, we can draw the workload flow of work-centers similar to the process flow (WIP distribution at operations) of MIVS rule described in Figure 3.1.2 (P.49). Figure 4.1.1 and 4.1.2 show us real examples from MIMAC6 model

Figure 4.1.1 presents workload information of work-centers at the viewpoint of *k*-work-centers ahead of lot flow. Actually we can only see the real workload of work-centers instead of target workload in this figure. The question is how to make decision on which lot should be processed. If we only consider the local work-center at which the dispatching decision takes place like ‘AUTO-CL_dot’, 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 (LQF) rule. In fact, different dispatching decision can be made according to different viewpoint and information. If we consider 1-work-center ahead, we realize that

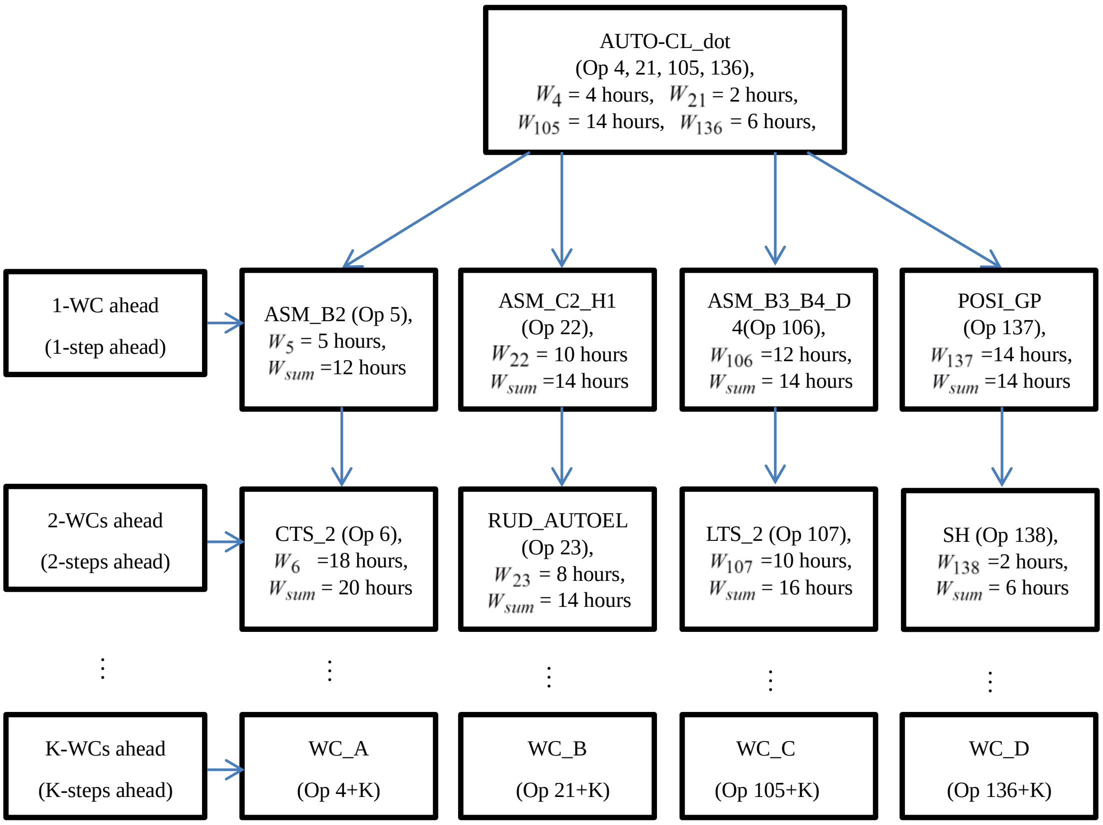


Figure 4.1.1: Workload information of work-centers at the viewpoint of *K*-work-centers ahead of lot flow

the downstream work-center ‘ASM_B2’ has the minimum workload. 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 (LWNQ). 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’ that can process the lot at operation 6 is high loaded. If we still send the lot to it, it will only cause heavier burden on downstream work-centers. In this case, the lot at operation 136 is recommended to be sent to work-center ‘POSI_GP’. ‘POSI_GP’ is already high loaded, but its downstream work-center ‘SH’ is low loaded. This example tells us that if we only focus on local information to make dispatching decisions, the results turns out to be ineffective since local information is too limited for full insight. Likewise, we can extend the 1-work-center ahead to *k*-work-centers ahead to utilize the global workload information to make decision, even without target workload.

Similarly, we can apply this idea to draw workload information of work-center from the viewpoint of *J*-work-centers back, as showed in Figure 4.1.2. From 1-work-center back viewpoint, the lot at operation 21 is recommended to be processed since the upstream work-center ‘NF-2’ has a high workload at operation 20 which is expected to arrive at ‘AUTO_CL-dot’ very soon. However, the 2-work-center back leads to the different opinion that the lot at operation 4 should be the first because the WIP at ‘AMC-EPI_1+2’ is extremely high and will arrive at operation 3 and 4 very soon as well. Similarly, this 1-work-center back can be extended to *J*-work-centers back

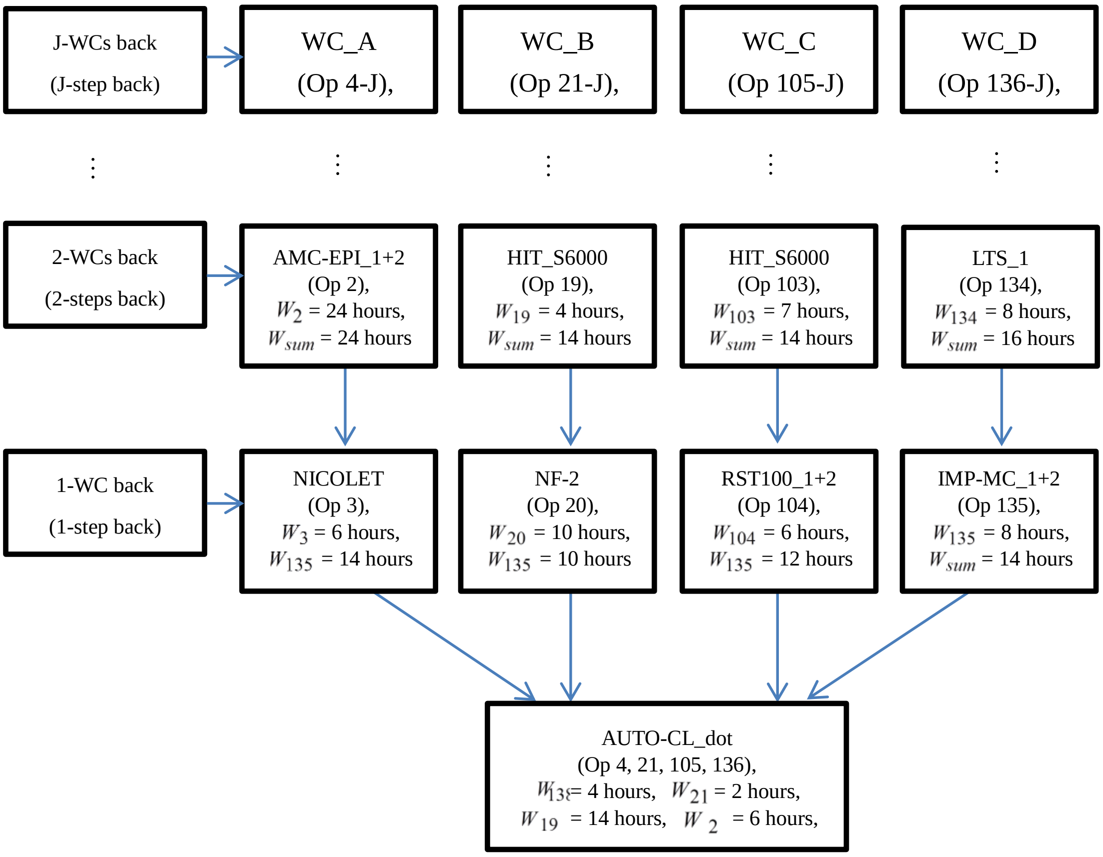


Figure 4.1.2: Workload information of work-centers at the viewpoint of *J*-work-centers back of lot flow

As a result, the above examples indicate that we can still balance the workload of work-centers even without target workload, if we can employ real workload information from *K*-work-centers ahead and *J*-work-centers back concepts. Actually, in order to make it happen, we divide the workload information into three components which are workload information of upstream work-centers (*J*-work-centers back), local work-center, downstream work-centers (*K*-work-centers ahead). The challenge is how to combine these three components into one rule that can reflect the pull request of work-centers precisely.

4.1.2 Workload Indicator (WI) to Measure the Pull Request of Work-center

4.1.2.1 Methodology

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

i: work-center identifier;

t: operation identifier;

j: candidate operation (operation in the local work-center to be assigned a priority) identifier;

J, K: number of steps look back/ahead;

J', K': the first downstream/upstream operation from the candidate operation in the bottleneck work-center;

$W_{i(up),t}$, $W_{i(down),t}$: workload of operation *t* at upstream/downstream work-centers;

$W_{i(local),j}$: workload of candidate operation *j* at local work-center;

Mod: modification factor;

b_u: batch utilization.

Local work-center: at time *t*, when a machine in a work-center is available for processing. The work-center needs to make dispatching decision (assign priority to lot).

(1). Workload indicator for local/upstream/downstream work-centers

In Figures 4.1.1 and 4.1.2 we use the amount of work hours left to represent the workload of work-centers. However, it is too simple and rough to calculate the priority of lot by only summing up the workload. Obviously, we need to transfer the workload to an indicator which can express the pull requests of work-centers and the urgency of lots. In literature [Ham and Fowler 2007] the workload is suggested to be weighted by accumulated cycle time from the local work-center to each up/downstream work-center, as described in Equation (4.1.2).

$$Upstream_{J\text{-}work\text{-}centers\ back} = \sum_{t=j-J}^{j} \frac{W_{i(up),t}}{1+\sum_{r=t+1}^{j} RPT_r}$$

$$Downstream_{K\text{-}work\text{-}centers\ ahead} = \sum_{t=j+1}^{j+K} \frac{W_{i(down),t}}{1+\sum_{r=j+1}^{t} RPT_r} \qquad (4.1.2)$$

It is easy to understand that 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-centers, 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, we propose a simple and sufficient workload indicator (where we prove its effectiveness by simulation later on) as follows:

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

$$Downstream_{K\text{-work-centers ahead}} = \sum_{t=j+1}^{K} \frac{W_{i(down),t}}{2^{t-j}} \qquad (4.1.3)$$

With respect to the local work-center, the workload of operation *j* is used as workload indicator, that is

$$Local = W_{i(local),j} \qquad (4.1.4)$$

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

$$WI_j = Upstream_{J\text{-work-centers back}} + Local + Downstream_{K\text{-work-centers ahead}}$$

$$= \sum_{t=j-J}^{j\text{-}1} \frac{W_{i(up),t}}{2^{j-t}} + W_{i(local),j} - \sum_{t=j+1}^{K} \frac{W_{i(down),t}}{2^{t-j}} \qquad (4.1.5)$$

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

(2). Incorporate batch and setup into workload indicator for the batch processing work-centers

One benefit of WIP balance from the viewpoint of work-centers is that we can incorporate batch and setup strategies into dispatching decisions. As matter of fact, batches and setups are significant factors in wafer fabs, especially how to increase batch size and utilization and reduce setup times show great appeal in literature, for the reason of a large cycle time reduction benefit [Demeester and Tang 1996, Glassey and Weng 1991, Kim et al. 2008 , Robinson et al. 1995, Robinson 2002].

The workload indicator in Equation (4.1.5) demonstrates that an operation should have high priority if its upstream is high loaded and downstream is low loaded. When batch and setup are involved, the situation becomes different and complicated. We can show that the workload indicator has a potential flaw. In Figure 4.1.3, work-center *A* performs batch processing and the maximum batch size is 4 lots. Operation 4 and 8 both have 2 lots, and operation 16 has 4 lots. According to the workload indicator, apparently work-center *A* would prefer to process lots at operation 4 because it has high workload in upstream and low workload in downstream work-centers. However, it would result in the half batch size and capacity loss. If work-center *A* turns out to be the bottleneck, this would directly impact the throughput and cause a long queue. In this case, lots at operation 16 could be a better choice since this can lead to a full batch size and reduce the queue length.

There is no doubt that batches and setups have to be considered and incorporated into the workload indicator to avoid empty batches and constantly new setups for the batch processing work-center. Therefore, a modification factor is introduced to the workload indicator to weight the influence of batches and setups, as show in Equation (4.1.6).

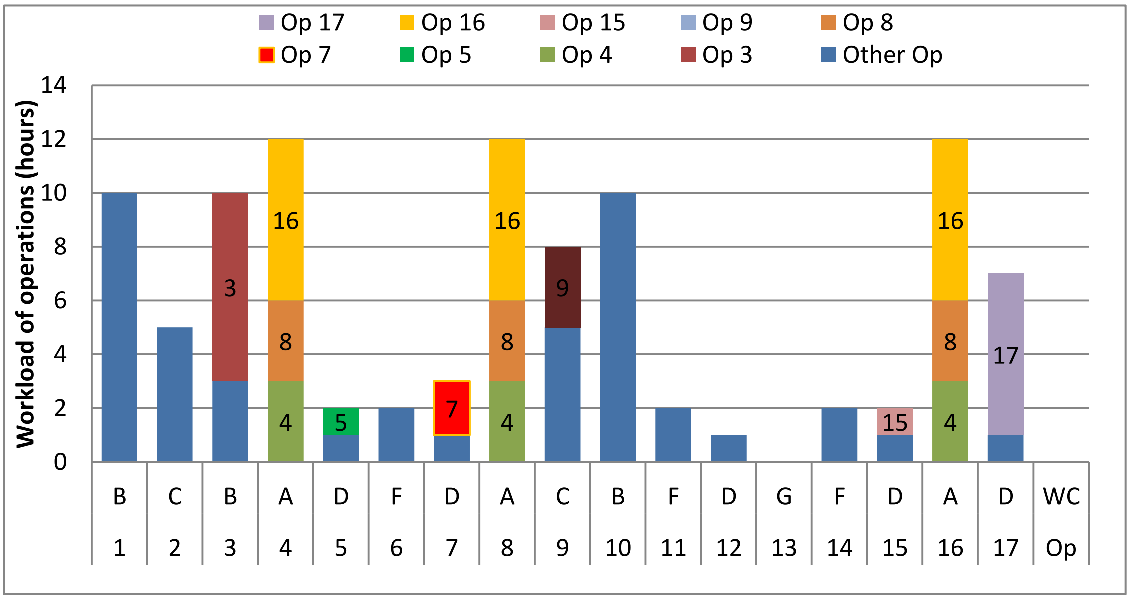


Figure 4.1.3: Batch processing to influence workload indicator

$$WI_{j,final} = mod * Upstream_{J\text{-work-centers back}} + mod * Local$$

$$+ mod^{-1} * Downstream_{K\text{-work-centers ahead}}$$

$$= \sum_{t=j-J}^{j-1} mod * \frac{W_{i(up),t}}{2^{j-t}} + mod * W_{i(local),j} - \sum_{t=j+1}^{K} mod^{-1} * \frac{W_{i(down),t}}{2^{t-j}} \qquad (4.1.6)$$

Where, for single processing work-center, *mod* = 1. For batch processing work-centers, we introduce a factor called batch utilization which is the percentage of lots that can form a batch resulted from the to be scheduled lot, as presented in Equation (4.1.7). Besides, some batch processing work-centers have setup requirement. If the candidate operation needs a new setup compared to the current setup, we use a simple approach to measure the setup requirement as depicted in Equation (4.1.8).

Three parameters '*X*', '*Y*' and '*Z*' in Equations (4.1.7) and (4.1.8) 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 (4.1.7); (2). For the batch utilization less than 50%, obviously, we need to lower its influence to the WI. Thus, we determine three levels of the parameter '*X*' that 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). The simulation results tell us that the combination (0.3, 1.4, 0.2) could achieve the best performance. Therefore, it leads to the Equations (4.1.9) and (4.1.10).

$$b_u : batch\ utilizatio\ n\ (\%)$$

$$mod = \begin{cases} X & b_u < 50\% \\ X + (b_u - 0.5)*Y & 50\% \le b_u < 100\% \\ 1.0 & b_u = 100\% \end{cases} \qquad (4.1.7)$$

$$\begin{array}{ll} If\ \ new\ setup\ is\ required: & \\ & mod = mod - Z; \\ else & \\ & mod\ remains\ unchanged. \end{array} \qquad (4.1.8)$$

$$b_u : batch\ utilizatio\ n\ (\%)$$

$$mod = \begin{cases} 0.3 & b_u < 50\% \\ 0.3 + (b_u - 0.5)*1.4 & 50\% \le b_u < 100\% \\ 1.0 & b_u = 100\% \end{cases} \qquad (4.1.9)$$

$$\begin{array}{ll} \textit{If new setup is required :} & \\ & mod = mod - 0.2; \\ \textit{else} & \\ & \textit{mod remains unchanged}. \end{array} \qquad (4.1.10)$$

Setup is also required by some single processing work-centers. However, a setup avoidance strategy (Section 2.3, P.39) is applied for the single processing work-center in MIMAC6 model. This strategy modifies the dispatch rule in force at the work-center to include setup minimization as a primary goal behind priorities. Hence, setup for the workload indicator of single processing work-center is not discussed here.

(3). Determine the look ahead and look back distance (the 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 TOC, 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.

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 for the bottleneck, the operations called bottleneck operations are sorted by workload in descending order and marked as the upstream or downstream of the candidate operation *t* (the lot needs to be assigned priority in the local work-center). 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 *i* are determined as follows [Ham and Fowler 2007]:

$$K = MAX\{MIN\{\text{number of downstream operations}, 5\}, K'\} \text{ and}$$

$$J = MAX\{MIN\{\text{number of upstream operations}, 5\}, J'\} \qquad (4.1.11)$$

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 (4.1.11) still guarantees that the sizes of *K* and *J* are at least 5 that is recommended by [Ham and Fowler 2007] .

(4). Strong pull at the end/weak pull at the beginning of the process flow and unidirectional preference of work-center

There are two unusual cases caused by the workload indicator. The WI in Equation (4.1.6) has two parts concerning upstream and downstream work-centers. When WI is used to calculate the priority of candidate operations, which in other words is to determine the pull request, 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 a high relativity to obtain a higher priority from WI compared to the operations at the beginning or middle of process flow, if a work-center can perform the operations both at the beginning and end of process flow.

The other problem is that some work-centers are designed to be low utilized, which means the WI may be particularly low to force the upstream work-center to send lots to it. This unidirectional preference causes the upstream work-center to always send a lot to the low utilized downstream work-center as long as a lot for this particular low utilized downstream work-center arrives at the upstream work-center.

Although the WI is expected to measure the pull request of work-centers accurately and speed up the process flow, the WI is inherently affected by the strong/weak pull request at the end/beginning of process flow and unidirectional preference, which might result in a poor pace of lot movements. We should notice that some lots spend long queue times inevitably when unidirectional preference takes place. In some extreme cases, those lots are delayed significantly because they cannot be processed in time. The operation due date (ODD) dispatching rule can be applied under this circumstance. We use ODD to override the WI if lots are delayed because of the unusual events mentioned above. This special strategy can overcome the weakness of WI and force the delayed lots to catch up with their due dates. In addition, the cycle time variance is improved as well.

4.1.2.2 Detailed algorithm

If a machine of a work-center is available to process lots, the following algorithm is carried out to calculate the WI for each candidate operation.

Step 1. Calculate the workload of all operations according to Equation (4.1.1);

$$W_{i,t} = \frac{Lots_t \times (LoadTime_t + Process\ Time_t + UnloadTime_t)}{NumberofMachines_i}$$

Step 2. Calculate the workload of all work-centers according to Equation (4.1.2);

$$W_{i,sum} = \sum_{t=0}^{n} W_{i,t}$$

Step 3. Find out the bottleneck work-center with the highest workload, and determine the look ahead and look back distance *K* and *J* according to Equation (4.1.11);

$$K = MAX\{MIN\{number\ of\ downstream\ operations, 5\}, K'\}$$

$$J = MAX\{MIN\{number\ of\ upstream\ operations, 5\}, J'\}$$

Step 4. Check whether a candidate operation misses its operation due date. If this is true and the local work-center is not the bottleneck, the ODD rule is used to override the WI;

If (Now>ODD), ODD rule is used for dispatching;

Else continue to Step 5

Step 5. Calculate the modification factor *mod* according to Equation (4.1.9) and (4.1.10);

For single processing work-center, *mod* = 1;

For batch processing work-center,

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

$$If\ \ new\ setup\ required\ : mod = mod - 0.2$$

Step 6. calculate the WI for each candidate operation according to Equation (4.1.6);

$$\begin{aligned} WI_{j,final} &= mod * Upstream_{J\text{-}work\text{-}centers\ \ back} + mod * Local \\ &\quad + mod^{-1} * Downstream_{K\text{-}work\text{-}centers\ \ ahead} \\ &= \sum_{t=j-J}^{j\text{-}1} mod * \frac{W_{i(up),t}}{2^{j-t}} + mod * W_{i(local),j} - \sum_{t=j+1}^{K} mod^{-1} * \frac{W_{i(down),t}}{2^{t-j}} \end{aligned}$$

Step 7. Set the WI as priority to the candidate operation (Factory eXplorer simulation tool considers the smaller value as higher priority).

$$Priority_j = -WI_{j,\,final}$$

4.1.3 Simulation results and performance analysis

We addressed above that dispatching decisions are different based on different viewpoints. Thus, we proposed several factors which are influencing the pull request of work-centers and incorporated them into WI to overcome some of its inherent weaknesses. In order to understand the positive or negative effect, and the interaction among those factors, we carried out experiments using four approaches from a simple one that only considers the workload of local work-center to a sophisticated one that combines all factors discussed above.

The first approach represented by '*WI(1):L*' is to only consider the workload at the local work-center as described in Equation (4.1.4). 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 fab. 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 Equation (4.1.5). The sizes of *K* and *J* are dynamically determined according to the bottleneck detection and described in Equation (4.1.11). 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 Equation (4.1.6), (4.1.9) and (4.1.10). In the fourth approach represented by '*WI(4):Mod(U+L+D)+ODD*', we include ODD rule to override WI to overcome the tardiness 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 4.1.2.2.

As usual we consider average cycle time, cycle time variance and cycle time upper 95% percentile as performance measures under 95% fab loading. We compare the results to FIFO, MIVS and MWVS_1. The results are shown in Table 4.1.1.

Approach	***Average 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
MWVS_1	28.2	6.7	37.8
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 (2.0 DDFF)	26.2	2.0	34.0

Where MIVS is Minimum Inventory Variability Scheduling;

MWVS_1 is Minimum Workload Variability Scheduling 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 override the WI.

Table 4.1.1: Four performance measures comparison among four variations of WI, FIFO, MIVS and MWVS_1 under 95% fab loading

We focus more on the results of four variation approaches of WI because we discussed the results of MIVS and MWVS_1 before. Only considering the workload of local work-center '*WI(1):L*' 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 quite limited and cannot achieve a global balance for the whole wafer fab. When the upstream and downstream workload information are included and the sizes of *J* and *K* are determined dynamically represented as ‘*WI(2):U+L+D*’, it turns out to be major improvement. The average cycle time is reduced from 42.4 to 28.5 days, which becomes comparable with MIVS and MWVS_1. 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. 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, the batch and setup requirement are important factors in wafer fab when workload balance is desired. Since increasing batch utilization and avoiding constant new setup lead to faster lot movement, thus achieving cycle time reduction. Hence, incorporating batch and setup requirement to WI is necessary, certainly, it is worthy and affordable. Although the modification factor has positive effect on average cycle times, with respect to the cycle time variance it shows limited improvement. The intention to include ODD, which is ‘*WI(4):Mod(U+L+D)+ODD*’, to overwrite WI is to avoid some lots becoming tardy and indirectly improve cycle time variance. It is true that the cycle time variance is improved significantly after ODD is applied, meanwhile, the average cycle time maintains as good as ‘*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 MWVS_1 which is the one achieving workload balance with the requirement of target WIP. Obviously, all factors concerned show great advantages to replace target WIP to achieve balanced WIP, which makes them strongly competitive and promising.

Table 4.1.2 shows the average batch size comparison of 17 batch processing work-centers among MIVS, MWVS_1 and WI(4). We can see that the complete WI approach has the advantage in forming large batches in comparison to MIVS and MWVS_1, and it also explains the reason why the cycle time is reduced greatly from microscopic viewpoint of work-centers.

Work-center	***Average Batch Size (wafers per batch)***		
	MIVS	***MWVS_1***	***WI(4): Mod(U+L+D)+ODD***
11022_ASM_A2	24.1	24.5	26.6
11021_ASM_A1_A3_G1	38.2	39.6	41.5
11026_ASM_B2	89.4	91.3	92.2
11027_ASM_B3_B4_D4	40.7	41.6	43.1
11029_ASM_C1_D1	50.8	51.7	53.1
11030_ASM_C2_H1	42.2	43.4	45.5
11032_ASM_C4	26.4	26.8	29.3
11122_ASM_D2	26.9	27.7	29.2
11128_AMS_E4	48.5	49.2	50.6
11524_MAX1+2_AL-TEMP	50.7	51.5	53.1
12021_AUTO-CL_undot	54.4	54.6	55.4
12022_AUTO-CL_dot	53.6	53.7	55.2
14131_AMT-PREC_1+3	52.2	53.9	54.6
14137_AMT-PREC_7	56.6	58.4	59.9
14821_DNS-SOG_1	62.4	66.3	68.9
12221_HF-DIP-5_B	91.8	91.9	92.2
11132_ASM_F4_D3	33.0	34.5	35.8

Table 4.1.2: Average batch size comparison of 17 batch processing work-centers among MIVS, MWVS_1 and WI(4) under 95% fab loading

Figure 4.1.4 presents the average cycle time comparison among FIFO, MIVS, MWVS_1 and WI(4) from low to high loadings. When the fab loading is low, basically, MIVS, MWVS_1 and WI(4) show no significant difference in comparison to FIFO. As the loading becomes high, the improvement becomes obvious. MIVS, MWVS_1 and WI(4) balance WIP via different mechanisms, but in MIMAC6 model, they take effect tremendously compared to FIFO only under high fab loading cases. Based on the fact that the effect of lot dispatching rule highly relies on the fab loading [Waikar et al. 1995, Wein 1988], so far, we can conclude that the higher the fab loading with complex variability is, the more obvious the cycle time improvement can be achieved by WIP balance (with/without target WIP).

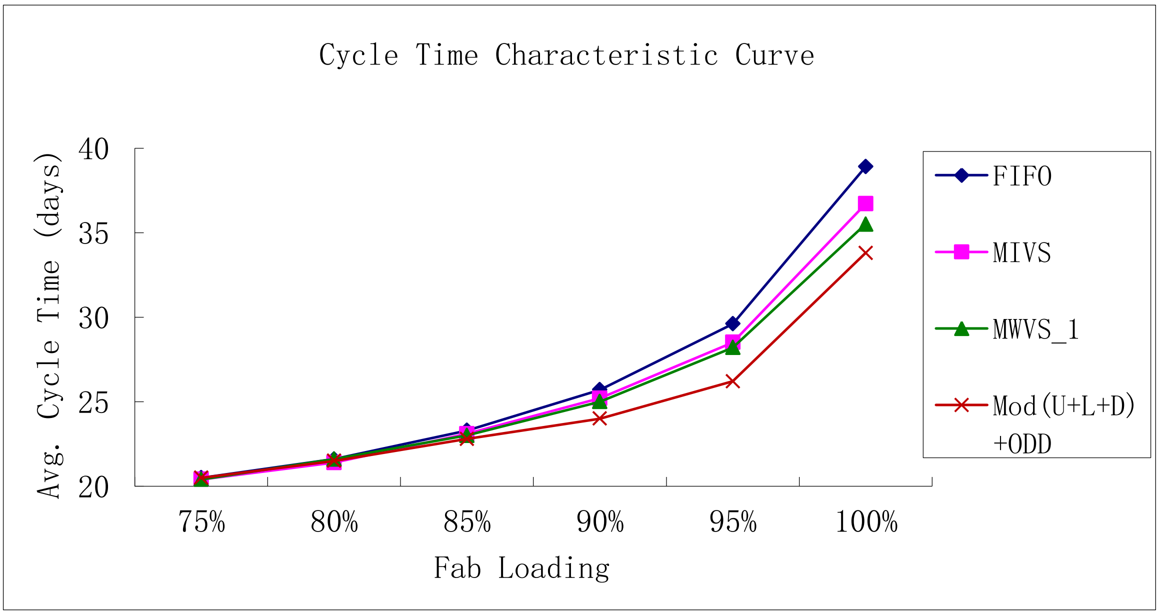


Figure 4.1.4: Average cycle time comparison among FIFO, MIVS, MWVS_1 and WI(4) under different loadings

4.1.4 Conclusions

Section 3.3 addressed the difficulties and challenges to determine and apply target WIP to achieve WIP balance. Thus, in this section we sought to develop a dispatching scheme to achieve workload balance for work-centers, in particular, without the requirement of target WIP. Inspired by MIVS and BMW rules, we developed a workload flow of work-centers in which we can observe the dynamic workload information of the line. Based on that, we suggested the Workload Indicator (WI) to measure the pull request of work-centers. To replace the target workload, 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 by adding a new factor the WI each time to see how those factors interact with each other. The simulation results demonstrated that the proposed WI approach achieved workload balance for work-centers in comparison with MWVS with target WIP and MIVS. The success of WI, which is to balance WIP but abandoning target WIP, can be summed up in the following three aspects:

(1). Different dispatching decisions can be made according to different standpoints, it is not sufficient to only take the workload of local work-centers into account when workload balance for the whole wafer fab is desired.

- Considering both the upstream and downstream workload situation can overcome the weakness of a local constraint.

With respect to the look ahead and look back distance, because there is no solid simple rule that which sizes of *J* and *K* are optimal, it would be beneficial to detect the workload of upstream and downstream dynamically on the basis of the line situation, like the Equation (4.1.11) (P.137) told us the way to determine the size of *J* and *K* based on the bottleneck. The minimum size 5 of *J* and *K* is recommended by the literature [Ham and Fowler 2007].

(2). Small batch sizes and constantly new setup requirements not only cause capacity loss, but also long queues are piling up.

- Batches and setups are crucial factors affecting the workload balance.

In fact, one advantage of balancing workload for work-centers is to incorporate batching and setup strategies into decision making. In our study, we used a Modification Factor to express the needs of batching and setup, and combined them with the WI, which brings further workload balance.

(3). There are some inherent weaknesses of WI which can be explained by unusual events like unidirectional preference which causes long queue times and tardiness of lots. Special strategy should be applied to avoid the drawbacks of WI. We used the ODD rule to override WI if lots become tardy for their operation due dates. The cycle time variance is improved remarkably.

- It tells us that considering lot status for dispatching decision is also a complement to balance workload of work-centers.

In Section 5.2, we will discuss further about considering lot status. This is as important as the workload of work-centers to achieve the whole fab balance.

4.2 WIP Control and Calibration

4.2.1 The Necessity of WIP Control and Calibration

On one hand, the MWVS rule in Section 3.1 and the WI approach (WI(3): without using ODD rule to overwrite WI) in Section 4.1 achieve both faster and poorer paced lot movement. The reason is that they dispatch lots only on account of WIP (workload) information of work-centers. The due date rules in Section 3.4 have the advantage of progressing lots with a good pace based on lot status (due date information), but have the disadvantage of losing WIP balance under some circumstances because of ignoring the WIP situation in the fab. These facts indicate that overemphasizing on one side definitely impairs the other side.

- In order to make sure that the wafer fab runs in a smooth way and lots go through the fab at the right pace to avoid WIP fluctuation and achieve shorter cycle times, higher throughput and better on-time delivery, we have to take both workload information of work-centers and lot status information into consideration.

On the other hand, although huge effort has been spent to balance WIP, as a matter of fact, WIP imbalance (also called WIP exception or WIP abnormity) still occurs constantly [Guo et al. 2007, Hopp and Spearman 2011, Kuo et al. 2008, Tu et al. 2005, Uzsoy et al. 1993, Yeh et al. 2008], due to the production variations of wafer fabs, e.g., machine breakdowns, batch processing, setup requirements, hot lots and so on. WIP imbalance can be expressed in three ways: (1) From operation viewpoint, WIP imbalance means that WIP piles up in one or some operations. It is dangerous if the high WIP operations are only performed by one work-center and when the work-center has a breakdown, the process flow is suspended. (2) From work-center viewpoint, WIP imbalance represents some work-centers are overloaded, while others are starved. Lots experience long queue times at the high loaded work-centers, while the capacity is lost for the starved work-centers. (3) From the whole fab viewpoint, one obvious symptom of WIP imbalance is the degradation of throughput. When the wafer fab runs in a steady state with continuous lot arrival, more and more lots remain in the fab because WIP imbalance blocks the product flow, which leads to WIP accumulation to decrease throughput.

- These WIP imbalance phenomena gives rise to WIP fluctuation causing unpredictable cycle time and excessive tardiness, which has a large impact on both cycle time and on-time delivery. Therefore, it is essential that an effective WIP imbalance detection and calibration policy can be applied to smooth the WIP flow when WIP imbalance occurs.

To deal with WIP imbalance, first of all we need to know how to distinguish between WIP balance and WIP imbalance. In other words, what kind of criteria can be used to determine WIP imbalance? The most popular and understandable way is to predefine a target WIP to an operation or a work-center [Guo et al. 2007, Kuo et al. 2008, Yeh et al. 2008]. We have mentioned that target WIP plays the role of measuring the pull requests of operations or work-centers in WIP balance dispatching rules. Indeed, target WIP has another crucial role which is considered as a trigger event to determine WIP imbalance. For instance, in [Guo et al. 2007] a target WIP called Acceptable WIP Deviation Levels (AWDL) is set for the work-centers. The AWDL includes upper-limit AWDL (UAWDL) and lower-limit AWDL (LAWDL). The actual WIP between the UAWDL and LAWDL is considered as normal situation. While the actual WIP out of the boundary of UAWDL and LAWDL is considered as WIP imbalance. The actual WIP higher than the UAWDL means the work-center is congested, in contrast, the actual WIP lower than the LAWDL means the work-center is starved. After detecting the WIP imbalance, a WIP correction approach that is MIVS is employed to correct for the WIP imbalance to avoid uncertainties getting worse.

This WIP calibration approach is also based on the appropriate target WIP as the WIP balance approaches like MWVS and MIVS. The difficulties and challenges to determine the appropriate target WIP drives us to seek a different way to achieve WIP balance described in Section 5.1. Here it turns out to be the same problem as above, whether we can correct for WIP imbalance without target WIP. In this case, the foremost point to replace target WIP is to find out an alternative to monitor and detect the WIP imbalance occurrence. If we abandon target WIP, we will lose the critical information where WIP imbalance occurs. Whereas, by noticing that throughput decrease is a symptom of WIP imbalance, we can utilize throughput decrease as a trigger event to determine WIP imbalance. If throughput has a sudden degradation, WIP imbalance definitely happens ‘somewhere’ in the fab to interfere the process flow. Therefore, more and more lots stay in the fab and the WIP keeps building up. In this case we have less concern about the accurate location where WIP imbalance takes place, as we decide to give up target WIP. On the contrary, the objective we are facing is how to make sure that the throughput goes back to the right track. Simply speaking, to correct for WIP imbalance, we have to increase the throughput again. The problem would become clear if we consider WIP flow from the viewpoint of operation as described in Figure 3.1.2 (P.49). It is not difficult to see that due to the intensive WIP distribution at some operations, in particular WIP accumulates in the early and middle operations. The blocked

WIP flow results in a throughput decrease. To make sure that throughput goes back to the normal state, the WIP calibration has to smooth the WIP flow and push lots from the congested operations to downstream. It follows that to speed up the WIP flow, WIP position analysis at operations is very necessary.

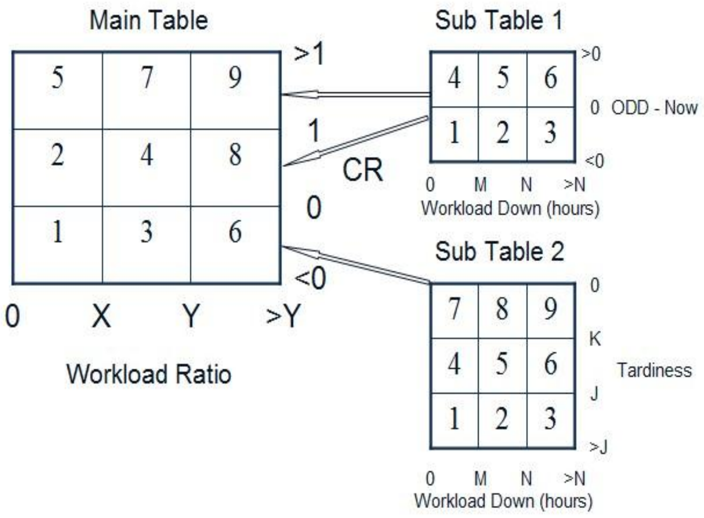
In this section, firstly we introduce a priority matrix table which prioritizes lots according to the workload information of work-centers and due date information of lots. Our goal is to keep the lots going through the fab smoothly and at the right pace to achieve WIP balance. However, as WIP balance is relative and time dependent, the simulation results show that the WIP evolution curve of matrix table still appears to be imbalanced from a local viewpoint. Therefore, secondly we propose to use throughput decrease as trigger event to detect WIP imbalance, which differentiates our approach from the one in literature applying target WIP for work-centers. Besides that, a WIP calibration approach utilizing a WIP position analysis to speed up the process flow to increase throughput is developed as well.

4.2.2 Priority Matrix Table for WIP Control

Previous sections addressed the benefit of WIP balance for work-centers. Nevertheless, due to a lack of consideration of the lot status like due date information, lots progress in poor pace causing a random WIP distribution which leads to poor cycle time variance. We also noticed that the due date rules have a mechanism to minimize lateness variance to minimizes cycle time variance. It tells us that WIP balance means not only cycle time reduction but also disciplined lot movement. Thereby, this study proposes a priority matrix table which employs workloads of work-centers, critical ratio (CR) and operation due date (ODD) of lots with the purpose to keep lots progressing smoothly toward on-time completion without causing serious WIP fluctuation.

The priority matrix table includes a main table and a sub table. The main table ranks the dispatching sequence according to the CR value of lot and workload ratios between downstream and upstream work-centers. As shown in Figure 4.2.1, the CR and workload ratio are divided into 3 levels which leads to a combination of nine priorities. The CR value larger than 1 means the lot is ahead of schedule, between 0 and 1 represents the lot falls behind schedule and comes close to its final due date, less than 0 means that the lot is tardy. The workload ratio is the ratio between the remaining production hours of downstream and upstream work-centers. The larger the workload ratio is, the lower priority the lot has. As a result, these nine priorities can lead to 362,880 (factorial function 9!) combinations. It is extremely time consuming if we tested all these combinations to find out the best one. However, we notice that the tardy lot has higher priority than the non-tardy lot, and the lot which is heading towards a work-center with low workload has higher priority than the lot which is heading towards a high loaded work-center. Based on this observation, we can pre-determine the places for priority 1, 7, 8 and 9 in the main table of Figure 4.2.1. Consequently, the remaining priorities 2, 3, 4, 5 and 6 lead to 120 combinations (factorial function 5!). Thus, 120 simulation runs are carried out to find out the best combination that is shown in the main table of Figure 4.2.1.

In some cases, several lots have the same priority in the main table. Therefore, two sub tables are used to subdivide priorities on a detailed level. Sub table 1 is used for the case of CR larger than 0, which denotes that the lot is still on schedule. Sub table 1 for delayed lots is used to divide priorities into two levels, and the remaining production hours of downstream work-centers is divided into three levels. As a result, there are six priorities in sub table 1. Each operation has its own operation due date, the lots late for their operations due dates have a higher priority than the lots ahead of their operation due dates. Sub table 2 focuses on the tardy lot case (CR less than 0). Three tardiness levels and three levels of remaining production hours of downstream work-centers lead to nine priorities in sub table 2. The more tardiness the lot has, the more urgent the lot is. Different from main table, the places of priorities in sub table 1 and 2 are pre-determined without simulation experiments, because the sub tables are designed to focus more on due date control given the same workload condition.



CR = (Due - Now) / RemainingRPT;
ODD = Release Time + RPT(i) * FF;
Workload Ratio =RPH(down)/RPH(up);
Tardiness = Now - Due;
Where：
Due: due date of a lot;
Now: current time;
RemainingRPT: remaining raw processing time;
RPT(i): the RPT for a sequence of processing steps or operations from operation 1 to operation i (including operation i)；
FF: the ratio of target cycle time and raw processing time of a lot;
RPH(up): remaining production hour of upstream work-center;
RPH(down): Workload Down: remaining production hour of downstream work-center;

Figure 4.2.1: Priority main table and sub table

In the main table and sub tables, smaller values represent higher priorities, i.e., priority 1 is higher than priority 2. The lots in the queue are sequenced step by step as follows:

Priority(main table) -> Priority(sub table) -> ODD

If lots have the same priority from the current rule, then the next rule is used to distinguish until ODD is applied as the final rule, e.g., if lots have the same priority due to the main table, then the priority from sub table is used.

Besides how to determine the best combination of priorities, another challenge is how to specify the parameters *X*, *Y*, *M*, *N*, *K* and *J*.

(1). One year simulation of MIMAC6 model with FIFO dispatching was carried out. We obtained approximate 180000 workload ratio values of downstream and upstream work-centers. We summarized and divided these 180000 workload ratio values into three categories evenly, which results in (0<=ratio<=2), (2<ratio<=5) and (ratio>5). Thus, the workload ratio levels *X* and *Y* from the main table are defined as 2 and 5, respectively.

(2). In sub table 1 and 2, the workload of downstream levels *M* and *N* are defined as one shift 8 hours and two shifts 16 hours, respectively. The tardy levels *K* and *J* are defined as 12 hours and 24 hours, respectively. Actually, the values of *M*, *N*, *K* and *J* can vary differently for the simulation experiments, which means huge effort is required to find out the optimal combination. In this study the *M*, *N*, *K* and *J* are specified by engineers at Infineon Technology Dresden Germany, to avoid complicated parameter setting. In the future research, we intend to carry out a full simulation study on the these four parameters setting.

4.2.3 WIP Imbalance Monitor and Calibration Approach

When the fab is running in a steady state, the ideal WIP curve should evolve smoothly without increasing dramatically or fluctuating seriously over time. In reality, due to the characteristics of wafer fab, WIP imbalance occurs inevitably anywhere in the fab. In order to detect and correct the WIP imbalance, a WIP abnormity monitor and calibration approach are proposed. In contrast to the traditional WIP abnormity monitor approach such as predefined target WIP level for operations or work-centers, throughput decrease is applied as the trigger event to monitor WIP imbalance. If the throughput decreases suddenly and is less than the release, which means that more and more lots stay in the fab and the WIP will build up. WIP imbalance, represented by WIP piling up in some operations, definitely takes place somewhere in the fab. Hence, using WIP position analysis for operations to determine which lots at high WIP operations should be pushed to balance the low WIP operations is essential.

WIP position analysis is not a novel technique, and WIP distribution histograms used by MIVS in Figure 3.1.2 (P.49) were already presented as an example. WIP position analysis means, simply speaking, via analyzing WIP distributions in operations, we can identify which manufacturing area has too much WIP and which has too little WIP. It consists of two parts which are WIP in blocks and WIP in operations.

4.2.3.1 WIP in Block

Each lot’s process flow can be divided into *B* blocks. Blocks correspond to a logical separation that allows having intermediate controls on lot manufacturing. In the MIMAC6 model, average 30 process operations (one mask layer) form a block, which is provided by the experienced industrial engineers. In practice, each block has an output goal for the next block to balance the WIP. Due to computational and algorithmic limitations, the accurate output goal of each block is not defined in this study, which is also consistent as that no target WIP level is specified for operations or work-centers. As presented in Figure 4.2.2, 'total WIP[this]' means the total WIP level of the current block *i*, and 'total WIP[next]' means the total WIP level of next block *i+1*. If the total WIP of the current block is higher than the total WIP of the next block, there is the need of pushing WIP from the current block to the next block to maintain the throughput. Otherwise, the WIP builds up in the current block. There are three priorities of lots in Figure 4.2.2. For instance, the lots in Block 5 get higher priority than the lots in Block 4.

- Priority 1: If Total WIP[this] > Total WIP[next]
- Priority 2: If Total WIP[this] == Total WIP[next]
- Priority 3: If Total WIP[this] < Total WIP[next]

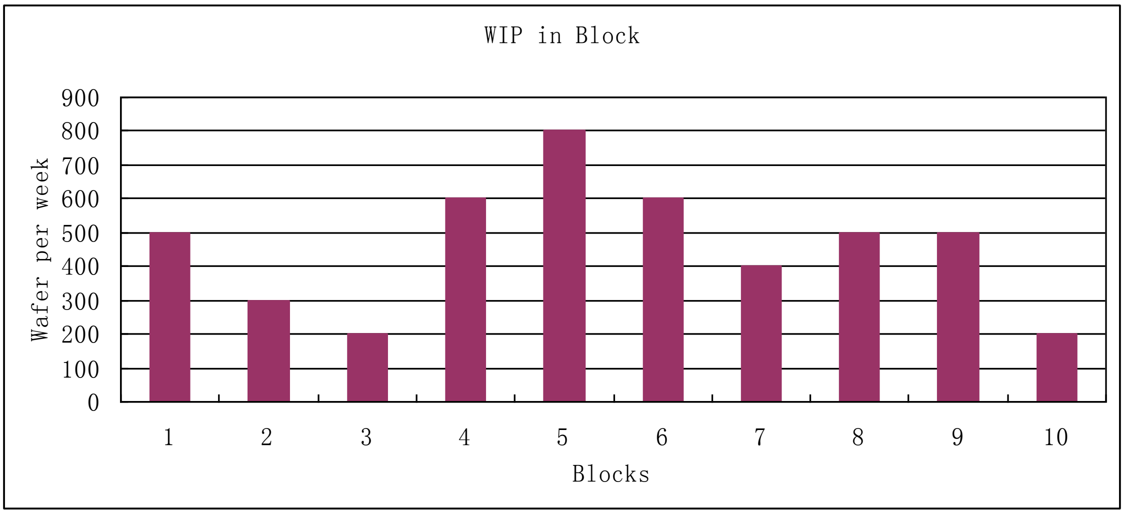


Figure 4.2.2: WIP in block (if we consider Block 4 is the current block, then Block 5 is the next block, and 'total WIP[this]' ('total WIP[4]') means the total WIP level of the Block 4, and 'total WIP[next]' ('total WIP[5]') means the total WIP level of Block 5.)

4.2.3.2 WIP Position Analysis at Operation

Each block includes dozens of operations. After determining which block should push the WIP to balance the downstream block, the next step is to balance the WIP inside the block to ensure the throughput. Figure 4.2.3 demonstrates an example of the WIP position analysis at operations. This WIP quantity histogram presents a visual picture of WIP balance. The 'WIP[this]' is the current WIP level of operation *i*, the 'WIP[next]' is the current WIP level of next downstream operation *i+1*, the 'WIP[last]' is the average historical WIP level of operation *i* during the last '*X*' hours. Instead of predefining a target WIP level for the operation like MIVS, we compare the following two parts of measurement 'WIP[this]' and 'WIP[next]', 'WIP[this]' and 'WIP[last]', e.g., operation 2 has more WIP than operation 3, and the WIP is increasing compared to its average historical WIP level. Hence, the WIP of operation 2 should be pushed to operation 3 immediately since operation 2 has a trend to accumulate to starve operation 3. There are the following four priorities of lots in Figure 4.2.3.

- Priority 1: If WIP[this] > WIP[next] and WIP[this] > WIP[last]
- Priority 2: If WIP[this] > WIP[next] and WIP[this] <= WIP[last]
- Priority 3: If WIP[this] <= WIP[next] and WIP[this] > WIP[last]
- Priority 4: If WIP[this] <= WIP[next] and WIP[this] <= WIP[last]

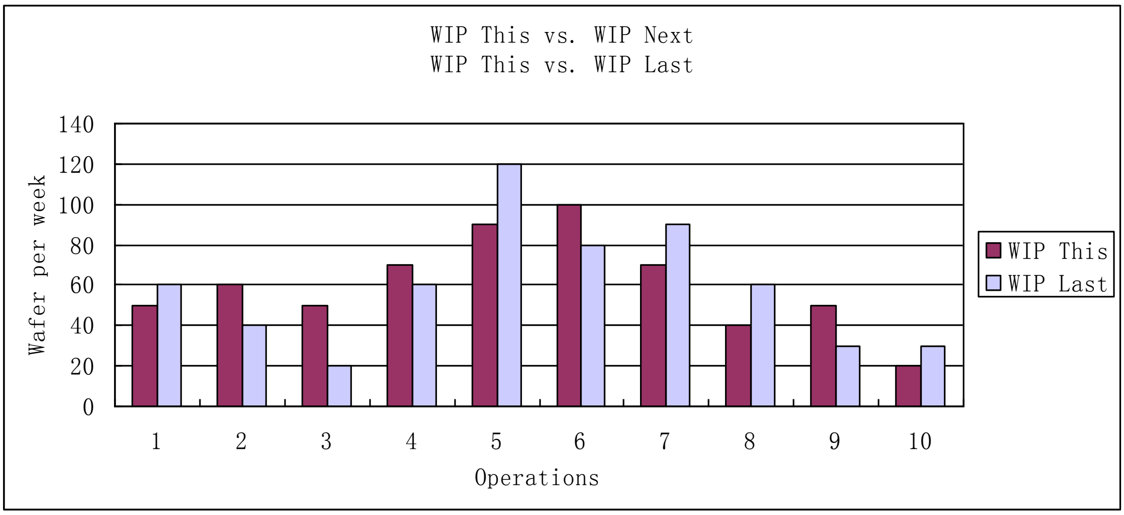


Figure 4.2.3: WIP position analysis at operation (If we consider Operation 4 is the current operation, then Operation 5 is the next operation. WIP[this] (WIP[4]) means the current WIP level of Operation 4, WIP[next] (WIP[5]) means the current WIP level of Operation 5, WIP[last] (WIP[4]) means the average historical WIP level of operation 4 during the last '*X*' hours)

4.2.3.3 Balance Work-center

Balancing WIP for operations ensures the throughput for the blocks, thus achieving balance for block. However, due to the reentrant nature of wafer fabs, the same work-center can perform different operations. Achieving WIP balance for operations does not mean WIP balance for work-centers. The lots acquiring priorities from block and operation analysis above may have the same priority. Here Least Work at Next Queue (LWNQ) is applied to balance the WIP of work-centers. The lot heading towards a work-center with less workload has a higher priority than the lot heading towards a high loaded work-center.

4.2.3.4 Reschedule Lots

If the WIP calibration approach is applied to correct WIP imbalance, the lots queuing in the work-centers are re-sequenced according to the following rules:

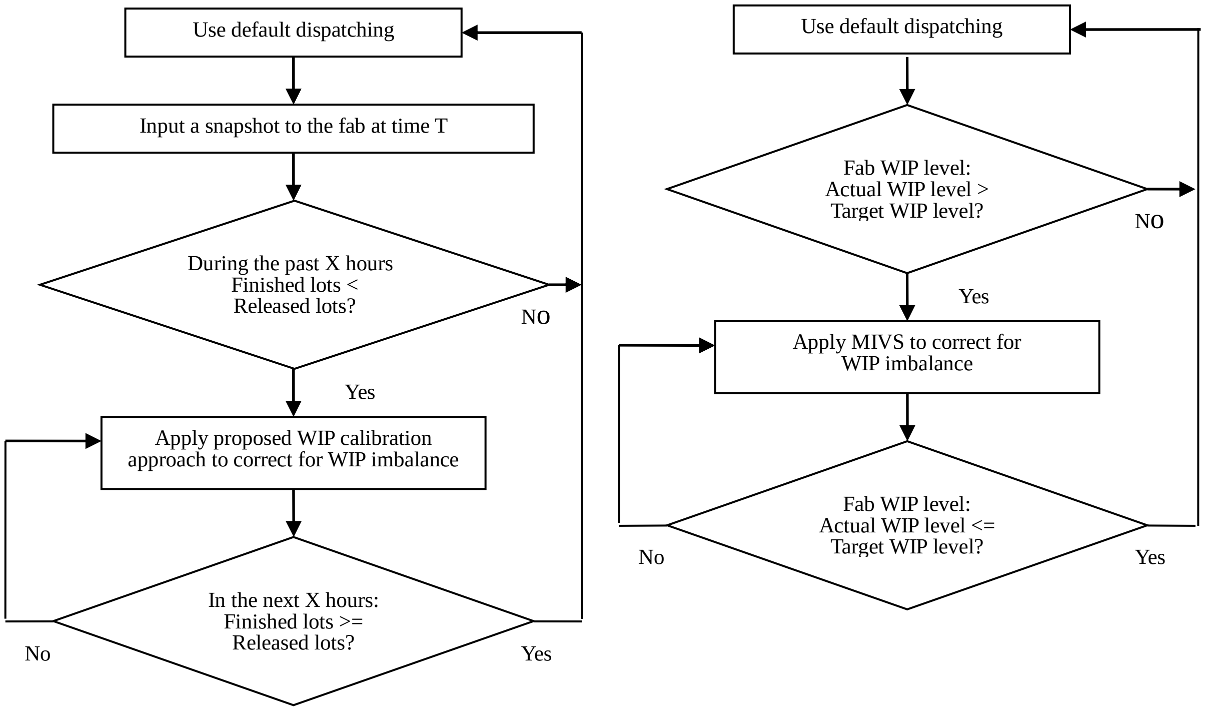
- Priority(block) -> Priority(operation) -> LWNQ -> ODD

If lots have the same priority from the current rule, then the next rule is used to distinguish until ODD is applied as the final rule, i.e., if lots have the same priority of the block (Figure 4.2.2), then the priority of the operation (Figure 4.2.3) is used to differentiate.

4.2.3.5 Monitor and Calibrate WIP Imbalance

Different from the traditional approach such as setting a target WIP level for the operation or work-center to monitor WIP imbalance, throughput degradation is used as a trigger event in this study. In case the throughput decreases, the manufacturing process is blocked somewhere in the fab. For this reason, the lots are rescheduled according to the priorities described above. The purpose is to make sure (1): WIP balance is inside the congestion block; (2): The congestion block pushes WIP to downstream block to ensure throughput. Figure 4.2.4 (a) presents a procedure to correct WIP imbalance. We will explain why the proposed WIP calibration approach works in Section 4.2.4.3.

We also develop another approach, using total WIP level of the fab as trigger event and MIVS rule to correct WIP imbalance as shown in Figure 4.2.4 (b), as benchmark. The main perspective is to figure out whether it is feasible to correct WIP imbalance without the need of target WIP.


Use default dispatching
Input a snapshot to the fab at time T
During the past X hours Finished lots < Released lots?
NO
Yes
Apply proposed WIP calibration approach to correct for WIP imbalance
In the next X hours: Finished lots >= Released lots?
No
Yes
Use default dispatching
Fab WIP level: Actual WIP level > Target WIP level?
NO
Yes
Apply MIVS to correct for WIP imbalance
Fab WIP level: Actual WIP level <= Target WIP level?
No
Yes

(a) Using throughput decrease to monitor WIP imbalance and the proposed WIP calibration approach to correct for WIP imbalance

(b) Using fab WIP level to monitor WIP imbalance and the MIVS rule to correct for WIP imbalance

Figure 4.2.4: Two different methods to monitor and calibrate WIP imbalance

4.2.4 Simulation Results and Performance Analysis

The simulation of MIMAC6 was carried out for 72 weeks. The first 24 weeks were considered as warm-up periods, and not taken into account for statistics. The fab capacity loading was set to 95%, which means the bottleneck work-center of the fab is driven to 95% utilization. The target due date flow factor was set to 2.0, this is a reasonable value because according to the fab running in the reign of FIFO, the percent tardy lots is 78% under this condition. If the target due date flow factor is too tight like 1.8, 100% of the lots will be tardy. If too loose like 2.2, only 23% of the lots will be tardy.

4.2.4.1 Priority Matrix Table for WIP Control

First of all, we compare the WIP evolution curve of our proposed priority matrix table with FIFO and ODD. Figure 4.2.5 shows the three different WIP curves over 48 weeks. Apparently, the WIP performance of the matrix table outperforms FIFO and ODD cases since it is lower and flatter than FIFO and ODD, which demonstrates the matrix table succeeds in scheduling lots in a more balanced way instead of causing serious WIP fluctuation like FIFO. It also tells us that the matrix table helps to reduce WIP variation since it does not jump oftentimes. The WIP curve of the matrix table has a similar trend as the ODD case. However, it is not as smooth as the ODD case. Beginning from the 30th weeks, the WIP of matrix table keeps climbing, then it stays around the 5000 wafers level. The reasons are the tardy lots and the continuous arrival of fresh lots. According to the matrix table, the tardy lots obtain the higher priorities than the new lots which are still in their early operations. In addition, the work-centers have breakdowns. The consequence is that fresh lots cannot be processed until the tardy lots leave the queue. Hence, the WIP builds up due to the continuous arrival of fresh lots. After the 37th week, the WIP goes down. This tells us that if the fab is running with considerable number of tardy lots, only focusing on due date control could lead to excessive WIP, because speeding up the tardy lots leads to longer waiting times of fresh lots. If a critical work-center has a failure, the fab becomes unstable and difficult to control.

Secondly, the average cycle time, cycle time variance, percent tardy lots and average tardiness of tardy lots are considered as major performance measures. The target due date flow factor is ranging from 1.8 to 2.6 in steps of 0.2. We take a close look at how these four performance measures change corresponding to the target due date flow factor change. Figure 4.2.6 shows the results. We

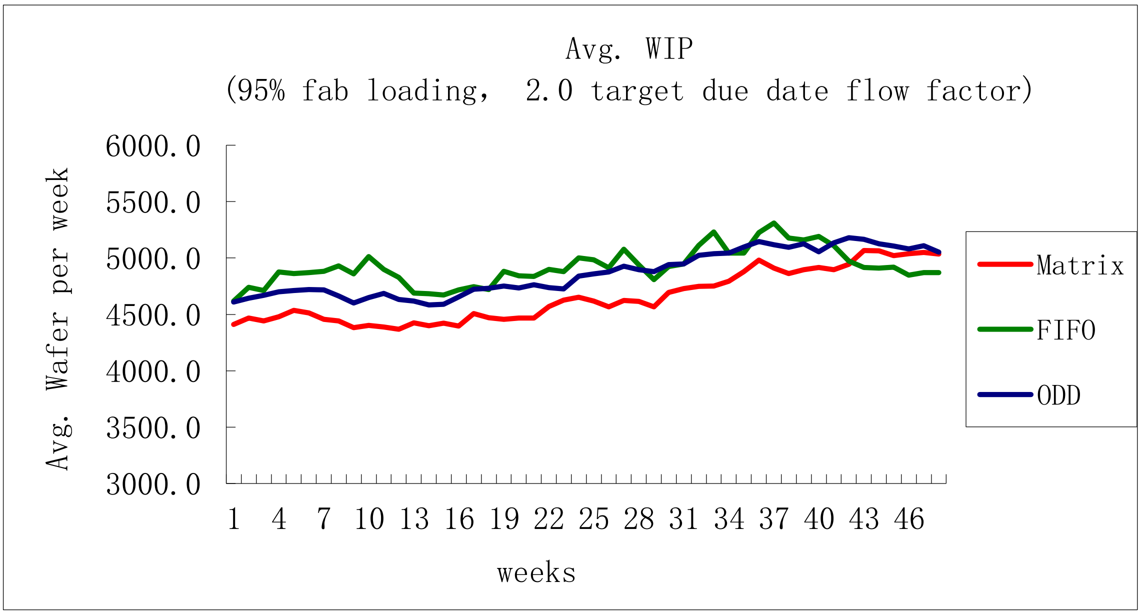
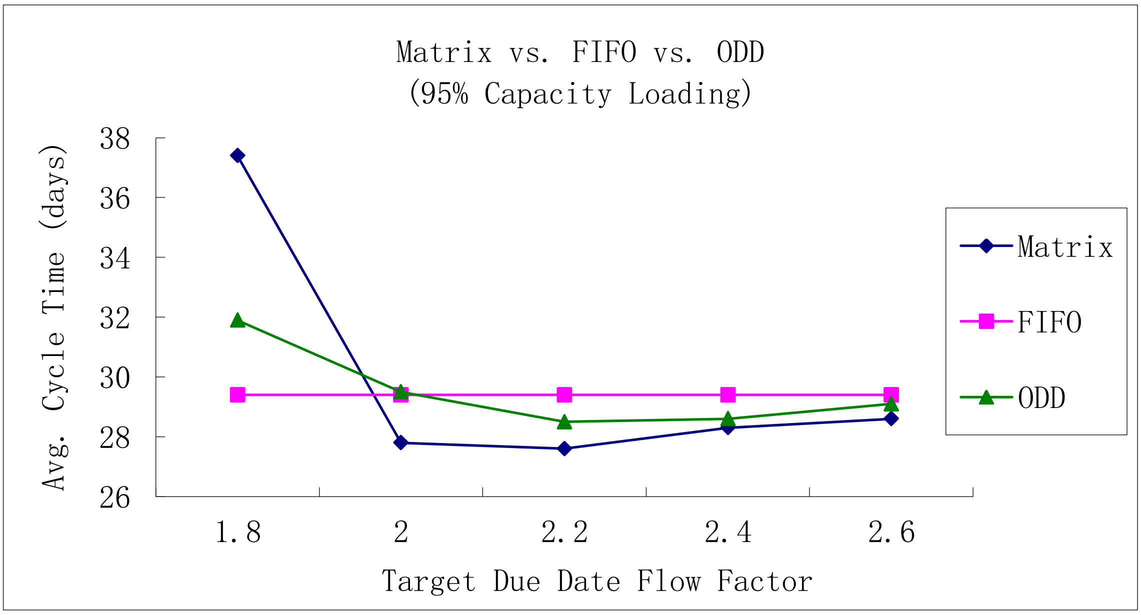
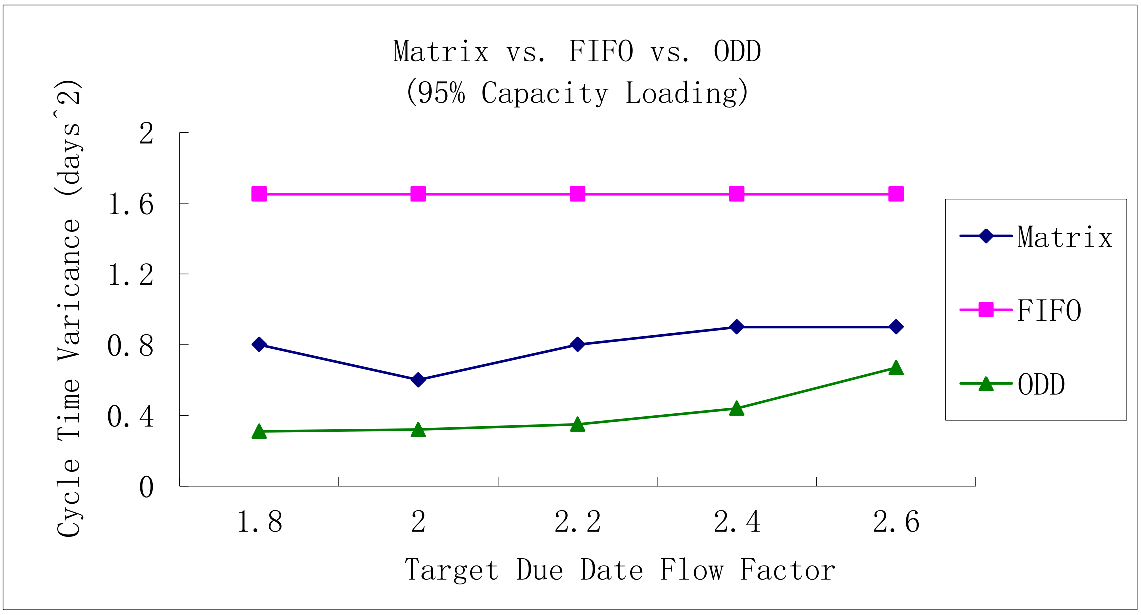


Figure 4.2.5: WIP evolution curve comparison among three different rules

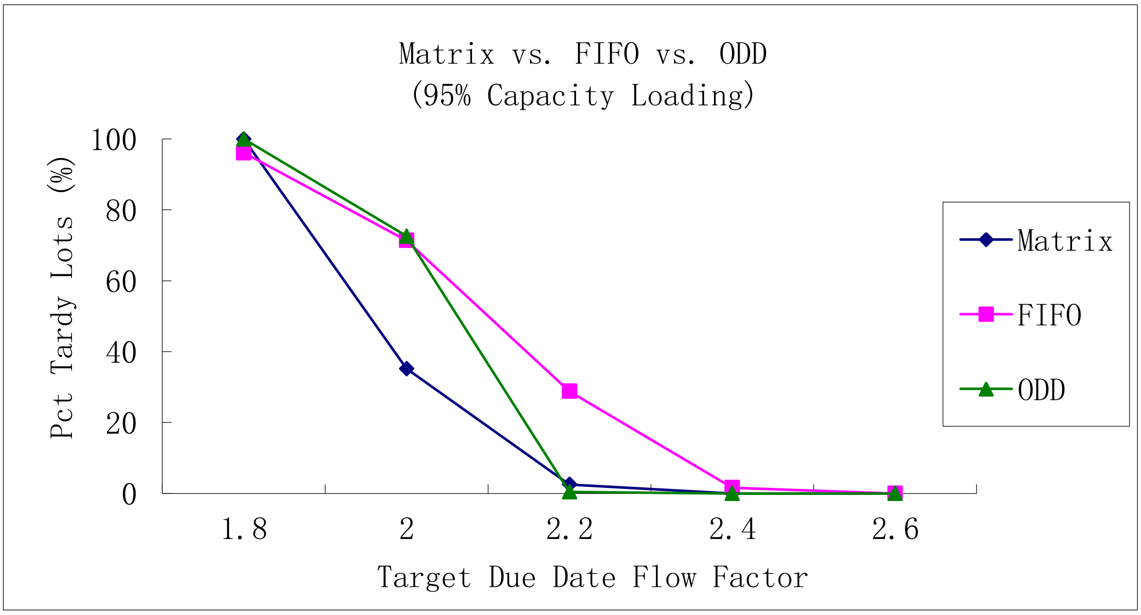
observe that the matrix table has a sudden performance degradation if the target due date is tight. Figure 4.2.6 (a) shows that the average cycle time of the matrix table is considerably higher compared to FIFO and ODD at due date flow factor 1.8. Under this tight due date, 100% of the lots are tardy not only for the matrix table but also for the FIFO and ODD cases. In this case, during warm-up periods, the lots in the middle operations are already tardy. The matrix table assigns high priority to those tardy lots to speed them up. Hence, the new arrival lots or the lots in the early operations are blocked. As time goes by, more and more lots become tardy. The lots in the early operations cannot be processed and become tardy too. The matrix table only focuses on the tardy lots instead of new arrival lots. The consequence is that the fab is running with a large number of tardy lots. The throughput decreases due to overemphasizing due date control. The WIP keeps building up because of decreased throughput and continuous arrival of new lots, which causes considerable cycle time. It indicates again that overemphasized due date control could lead to high WIP levels. The matrix table is not suitable for the case that 100% of lots are tardy under high fab capacity loading because of sudden performance degradation. With regard to the medium and loose target due date, the matrix table is superior to FIFO and ODD considering average cycle times and on time delivery. With respect to the cycle time variance, the matrix table provides a mechanism to ensure that lots go through the fab at the right pace and acquires better performance than FIFO. However, it is outperformed by ODD.



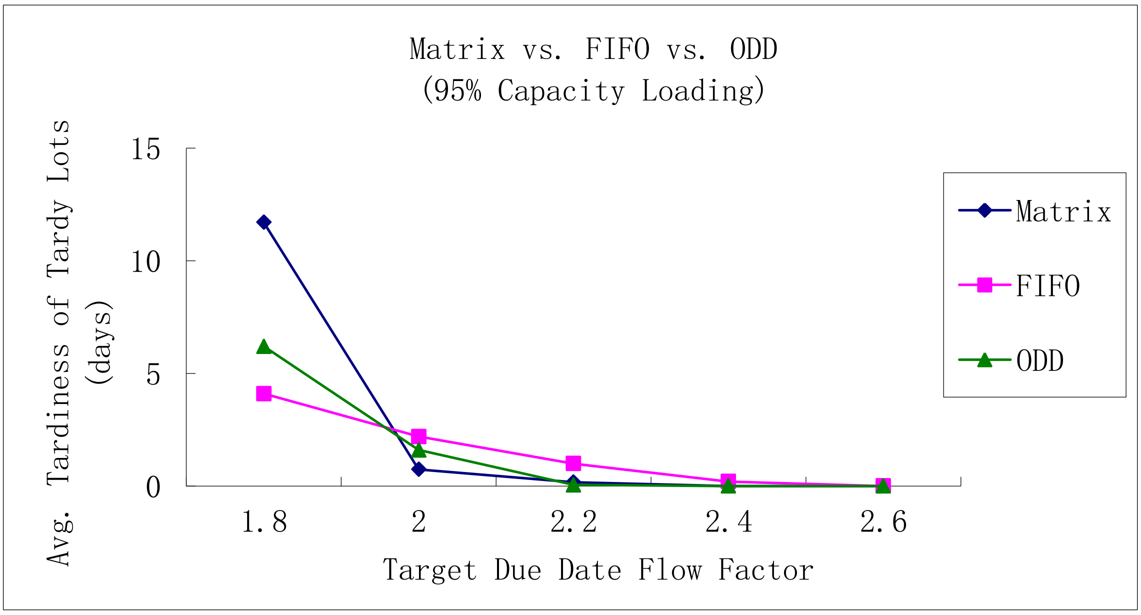
(a) Cycle time comparison



(b) Cycle time variance comparison



(c) Percent tardy lots comparison



(d) Average time tardy for tardy lots comparison

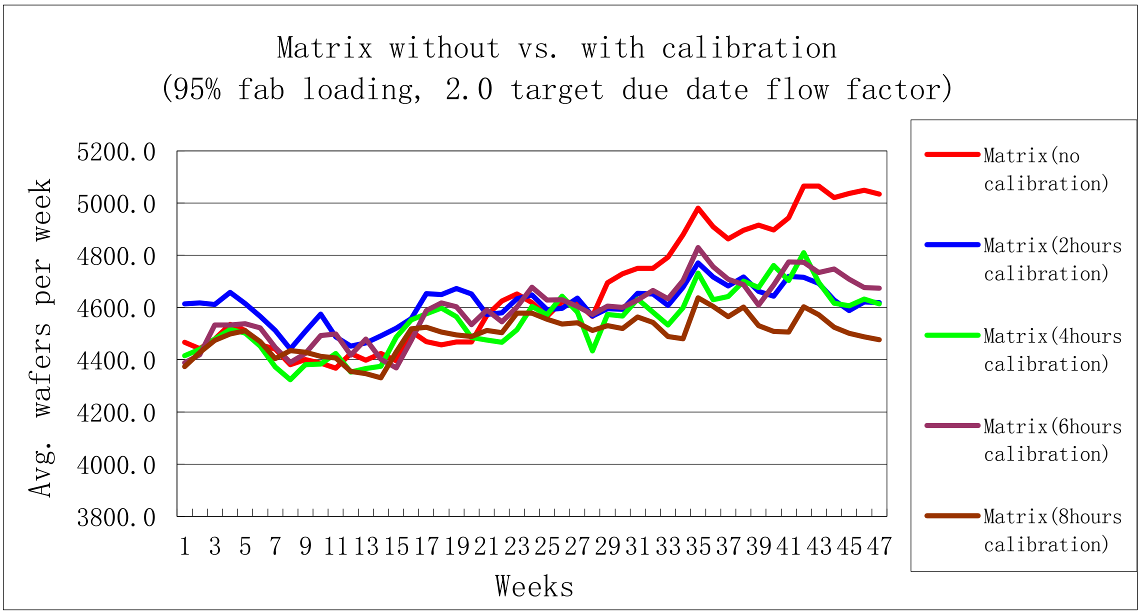
Figure 4.2.6: Four performance measures comparison among three rules with different target due date flow factors

4.2.4.2 Proposed WIP Imbalance Monitor and Calibration Approach

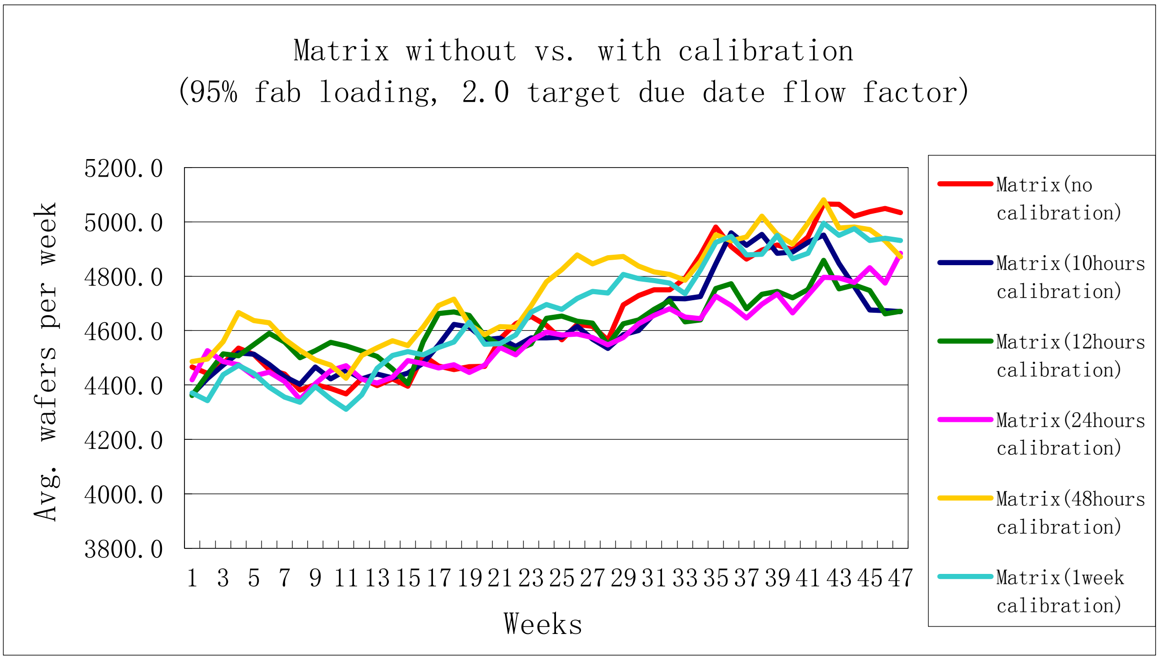
From Figure 4.2.5, we observe that from a global viewpoint the matrix table achieves more balanced WIP than FIFO and ODD for medium target due dates. From a local viewpoint, the WIP curve of matrix table still has WIP imbalance, such as the WIP keeps climbing from 30th weeks. One of the objectives of this study is to reduce the occurrence of WIP fluctuation and prevent WIP from accumulating. In case the throughput decreases, the WIP calibration approach is applied to calibrate the WIP abnormality and avoid WIP building up gradually.

Figure 4.2.7 shows the WIP curve differences between using matrix table with/without calibration. Figure 4.2.7 (a) represents the calibration is applied every 2, 4, 6 and 8 hours, and the WIP calibration approach succeeds in preventing WIP from climbing starting from the 30th weeks. It seems that the WIP calibration with an eight hour interval could achieve the best WIP curve. The WIP curve with calibration behaves smoothly, which implies the WIP calibration approach captures the WIP imbalance phenomenon and balance the block, operation and work-center simultaneously. Additionally, the smooth WIP curve also represents a better cycle time and cycle time variance performance. In Figure 4.2.7 (b), the calibration intervals are 10, 12, 24, 48 and 168 hours and the calibration effect is not as evident as in Figure 4.2.7 (a), even worse for 48 and 168 hours cases. Table 4.2.1 shows four performance measures of the matrix table without calibration and with calibration during 2, 4, 6, 8, 10, 12, 24, 48, 168 hours interval. It is obvious that the performances are improved when the calibration interval is small such as from 2 hours to 24 hours. When the calibration interval becomes large, the calibrated WIP curve has a trend to come close to the un-calibrated WIP curve, which means the performance improvement becomes limited. For the 48 hours and 168 hours cases, the performance is even worse than the one without calibration. The reason is that the earlier we monitor WIP abnormity, the more accurate we capture the WIP imbalance phenomenon. If the interval is large, we might miss the opportunity to calibrate when WIP imbalance occurs. The consequence is the small WIP imbalance develops into serious problem over time. It is a huge challenge to correct the enlarged WIP imbalance. It demonstrates that the WIP abnormity phenomenon has to be monitored any time and calibrated as soon as possible.

Figure 4.2.8 also shows us that the proposed WIP monitor and calibration approach can be adapted to other cases that different dispatching rules like FIFOand ODD are applied as default rules. The robustness of the proposed WIP calibration approach is tested in the following way. First, the default dispatching rule is changed from Matrix table to FIFO and ODD. Then the fab loading is changed from 95% to 85% and 75% cases. The simulation results indicate that the proposed WIP calibration approach succeeds in correcting WIP imbalance as long as it occurs, no matter which default rules are used.



(a) Every 2, 4, 6, 8 hours



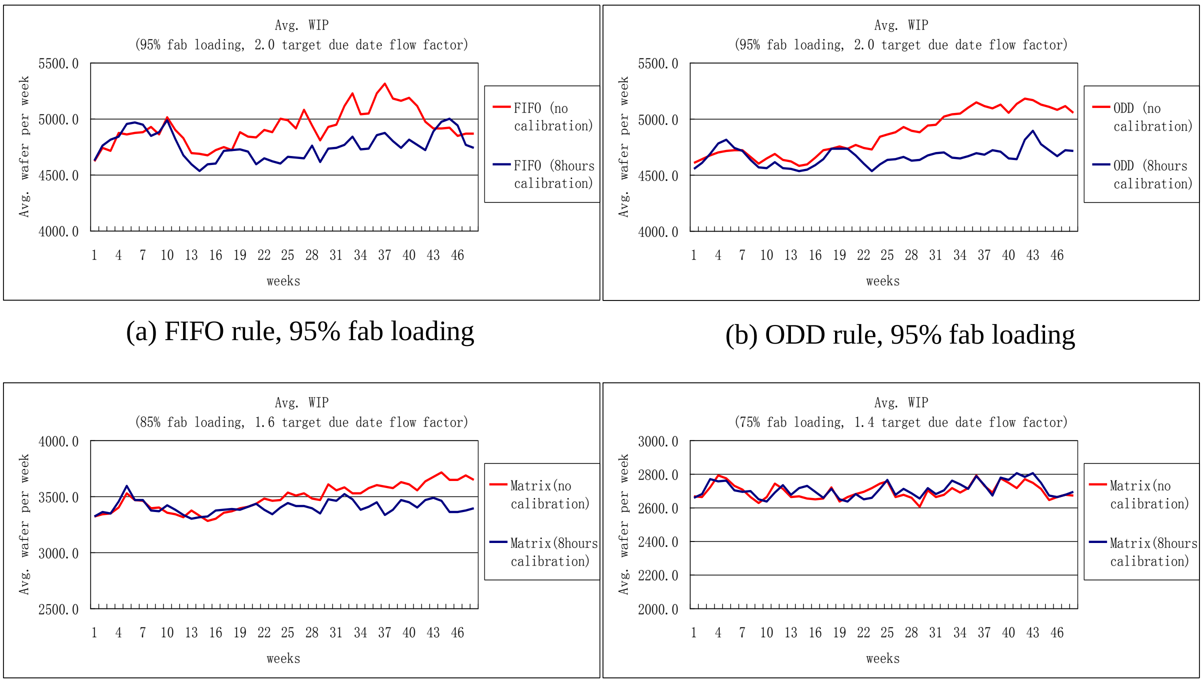
(b) Every 10, 12, 24, 48, 168 hours

Figure 4.2.7: WIP evolution curve of priority matrix table applied with WIP calibration approach every 2, 4, 6, 8, 10, 12, 24, 48, 168 hours

Matrix	***Average Cycle Time (days)***	***Cycle Time Variance (days^2)***	***Percent Tardy Lots (%)***	***Average Tardiness for Tardy Lots (days)***
No calibration	27.8	0.64	35.3	0.74
2 hours	27.6	0.62	21.2	0.59
4 hours	27.2	0.65	15.3	0.61
6 hours	27.4	0.62	19.4	0.64
8 hours	26.7	0.81	7.3	0.95
10 hours	27.6	0.62	27.9	0.62
12 hours	27.6	0.60	23.9	0.60

24 hours	27.4	0.61	20.5	0.56
48 hours	28.3	0.80	49.5	0.73
1 week	27.8	0.61	39.5	0.66

Table 4.2.1: Four performance measures of priority matrix table with and without WIP calibration



(a) FIFO rule, 95% fab loading

(b) ODD rule, 95% fab loading

(c) Matrix table, 85% fab loading

(d) Matrix table, 75% fab loading

Figure 4.2.8: WIP evolution curve of different rules with different fab loading, with calibration vs. without calibration

4.2.4.3 WIP Calibration with vs. without Target WIP

The previous section demonstrates that the proposed WIP calibration approach is able to drag the WIP curve back to the right track without applying target WIP levels if WIP imbalance occurs. However, we are curious about whether the performance of the proposed WIP calibration approach is as good as the one applying target WIP levels. Therefore, another WIP correction experiment with the MIVS rule depicted in Figure 4.2.4 (b) is employed. We only select the WIP curve of the matrix table with the proposed calibration approach of an 8 hour interval as a bench mark. In Figure 4.2.9, when the target WIP levels of fab are 180 and 190 lots, the WIP curves of the MIVS correction are as flat as the case of proposed WIP calibration approach. As the fab target WIP level raises, and because the frequency of WIP monitor and correction decreases, some WIP imbalance phenomenon cannot be captured and calibrated, which leads to a similar climbing WIP curve as the one without any correction. The detailed performance measures are showed in Table 4.2.2. The cycle time difference is almost 1 day if the fab target WIP level rises from 180 to 200 lots. This shows again that the target WIP level correction oriented approach is difficult to apply since a misleading target WIP level could result in a huge performance degradation.

Our proposed WIP calibration approach is inspired by MIVS. These are the two reasons why MIVS is successful (1): When operation *i* has high WIP and its next downstream operation *i*+1 has low WIP, MIVS gives higher priority to operation *i* in order to avoid the starvation at operation *i*+1; (2): MIVS uses target WIP levels to minimize the deviation between the actual WIP and target WIP level. Firstly, our proposed WIP calibration approach has a mechanism to push WIP from high WIP upstream to low WIP downstream. On one hand, it can ensure the high WIP block has output for the low WIP block. On the other hand, it can make sure that WIP is balanced inside the block. Secondly, our approach uses historical average WIP level to replace target WIP level in MIVS. This historical average WIP level plays the role for target WIP level. If the WIP imbalance occurs, the historical average WIP level represents the average WIP during the last ‘*X*’ hours, and during the last ‘*X*’ hours the WIP is assumed to be in a balanced state. We can see that when the WIP degrades from balanced state to imbalanced state, the historical average WIP during the last ‘*X*’ hours becomes the theoretical target WIP level which we want to achieve. We want to minimize the deviation between the current WIP and the historical average WIP, to make sure that the WIP turns into balanced state again. In comparison to the target WIP level the historical average WIP level changes all the time and depends on the ‘*X*’ hour time window. In addition, with the assistance of the ODD rule we know exactly which lots need to be pushed to downstream to balance the WIP.

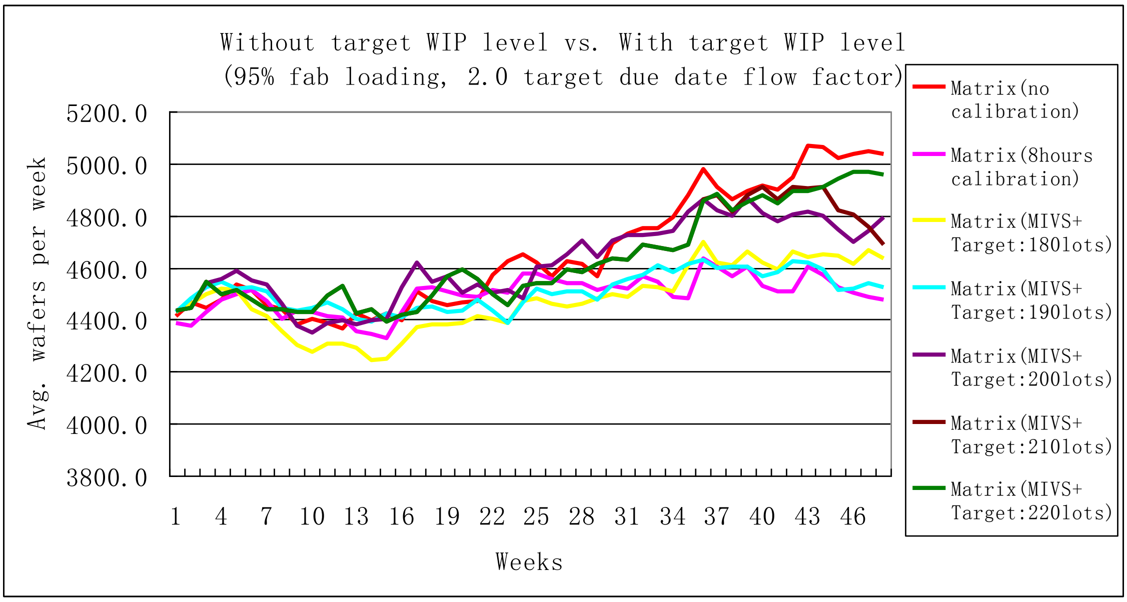


Figure 4.2.9: WIP curve evolution of matrix table applied calibration with vs. without target WIP level

Matrix	***Average Cycle Time (days)***	***Cycle Time Variance (days^2)***	***Percent Tardy Lots (%)***	***Average Tardiness for Tardy Lots (days)***
No calibration	27.8	0.64	35.3	0.74
8 hours	26.7	0.81	7.3	0.95

calibration				
MIVS + 180Lots	26.8	0.49	8.0	0.18
MIVS + 190Lots	27.0	0.50	5.2	0.25
MIVS + 200Lots	27.6	0.66	30.9	0.72
MIVS + 210Lots	27.5	0.67	27.0	0.61
MIVS + 220Lots	27.6	0.63	30.5	0.65

Table 4.2.2: Four performance measures comparison of matrix table applied calibration with vs. without target WIP level

4.2.5 Conclusions

In this section, we proposed a matrix table combining workload information of work-centers and due date information of lots together to ensure disciplined lot movement while achieving WIP balance. Although the WIP curve (Figure 4.2.5, P.161) showed superiority to FIFO and ODD in a macro viewpoint, actually, it still had a relative WIP imbalance from a micro viewpoint, since WIP balance is time dependent. Accordingly, we proposed a WIP imbalance monitor and calibration approach which considers throughput decrease as trigger event to detect WIP abnormity and WIP position analysis to correct the WIP imbalance. The simulation results demonstrated that the proposed WIP imbalance monitor and calibration approach is able to detect and calibrate the WIP imbalance arising from the matrix table and yield a smoother and flatter WIP curve. Furthermore, the WIP calibration approach is robust since it can calibrate different rules used as default dispatching rules in the fab under different loading cases. Most importantly, the proposed WIP calibration approach does not need any assistance from target WIP and still achieves promising performance compared to the traditional calibration approach applying target

WI. Based on these arguments, we can draw conclusion as follows:

(1). Workload balance for work-centers can achieve shorter cycle times but at the cost of lot pace. A higher requirement in WIP balance is not just cycle time reduction, but disciplined lot movement is also desired.

- Apparently, it demonstrates that not only workload information of work-centers but also the lot status like due date information should be taken into account simultaneously, if both WIP balance and on-time delivery are desired. The proposed priority matrix table is an approach based on these arguments.

(2). We have to notice that WIP balance is a relative term and time dependent. Due to the characteristics of wafer fabs, WIP imbalance occurs inevitably from time to time. To prevent minor imbalances from accumulating and becoming a serious problem, an effective WIP monitor and calibration approach is essential. Differentiating from traditional approaches, throughput decrease is used to detect the WIP imbalance and WIP position analysis is employed to correct it, with one precondition that is continuous lot arrival. Although the proposed approach is able to stop WIP from climbing, there is one thing we should pay attention to, we monitor the WIP imbalance every *X* hour that can range from short to long interval. On one hand, if the interval is too short, of course we have more chances to capture the imbalance phenomenon. However, the calibration may take effect for some fake imbalance cases that can be self-calibrated. On the other hand, if the interval is too long, the opportunities to capture and calibrate are missed, which leads to enlarged WIP imbalance and it may be too late to correct it. That is the reason why the simulation results indicate that the medium interval which is 8 hours in this study could achieve the best calibration performance.

- In reality, WIP imbalance has to be monitored at any time, but there is no clear conclusion that WIP calibration always brings positive effect all the time because it depends on how often the detection and calibration are carried out, and on the real situation in the fab as well.

(3). The reasons, why the proposed approach can achieve performance as good as the traditional one even without target WIP, are the following: (1). The WIP position analysis aims at balancing WIP inside the block. Particularly, the 'historical WIP' plays the role that target WIP does when WIP imbalance occurs, which assures that WIP turns into balanced state again; (2). We make sure the high WIP block has output to the low WIP block to ensure that the throughput goes back to the right track; (3) ODD and LWNQ rules help to identify which WIP should be pushed to downstream as well.

- Utilizing more information sets to replace target WIP is the key success point of our proposed WIP monitor and imbalance calibration approach, which is also a major contribution of this study.
- Another advantage of our proposed approach is that there is only one simulation parameter that is the X hour interval to monitor WIP imbalance.

Although we only consider the target WIP for the whole fab as simulation parameter. In reality, we should consider the target WIP of every operation or work-center as simulation parameter, which brings the problem that it is extremely difficult to figure out the interactions among those target WIP levels, not to mention the explosion of the huge parameters setting. This is one of the reasons that makes our proposed approach competitive.

CHAPTER 5

COMBINING WIP BALANCE AND DUE DATE CONTROL

In Chapter 3, we described and analyzed the performances of WIP balance and due date control in detail individually. We observed that WIP balance has the drawback of poor cycle time variance which might be a potential problem if due date performance is concerned (e.g. Table 3.1.8, P.58), while due date control potentially produces excessive WIP if due dates are too tight (e.g. Figure 3.4.2, P.112). Under some circumstances WIP balance and due date control conflict with each other. It is no doubt that WIP balance is critical as it brings average cycle time reduction. However, as many companies move from make-to-stock to make-to-order to satisfy their customers, on-time delivery is of importance as well. One issue arising from here is how to deal with the conflict between WIP balance and due date control. Indeed, from operational control viewpoint sometimes it is hard to take both targets into account, unless the WIP balance can achieve significant cycle time reduction which causes considerable tardiness minimization given the same due date. Otherwise we need to make a trade-off between WIP balance and due date control. This practical issue brings forward a challenging task that short cycle time and good on-time delivery are targeted simultaneously, which motivates us to carry out a preliminary study which investigates the interaction between them.

In Section 5.1, by noticing the excellent cycle time minimization effect of due date oriented rules, we develop a priority-based two-layer hierarchical dispatching scheme with the purpose to investigate the complementary effect of due date control on WIP balance. This two-layer hierarchical dispatching scheme turns out to be effective if low cycle time as well as good due date performance are desired concurrently.

Section 5.2 makes use of different ways to achieve both WIP balance and due date control targets in comparison to Section 6.1. We are aware that due date oriented rules like ODD cause WIP imbalance when the target due date is tight. However, the composite rule MOD that is a variant of ODD performs well under tight due dates (Figure 3.4.2, P.112), for the reason that MOD breaks the dominance of ODD by the introduction of SPT. Inspired by this observation, we integrate an additional WIP balance rule called Least Work at Next Queue (LWNQ) into the MOD rule to form a new composite rule. In contrast to the two-layer hierarchical dispatching, each single rule takes effect in parallel inside the composite rules. The contribution of each single rule to the composite rule under different due dates is determined by a scaling parameter that is estimated by means of design of experiment.

In Section 5.3, we apply the theory of WIP balance and due date control on a practical problem in a customer oriented wafer fab. In such a wafer fab, products are classified as low volume and high volume products, and low volume products are more critical than high volume products regarding cycle time and due date commitment. In general, due date oriented rules are applied to this kind of wafer fab. By noticing the benefits achieved by WIP balance, we intend to apply WIP balance to this fab as well. Nevertheless, two main questions for low volume products arise (1): Whether due date performance is sacrificed by achieving WIP balance for high volume products; (2): How to make the trade-off if due date is desired more than WIP balance.

5.1 Incorporating Due Date Oriented Rule into WIP Balance Approach to Achieve Cycle Time Reduction and On-time Delivery Improvements

5.1.1 What Happens When Due Date Meets WIP Balance

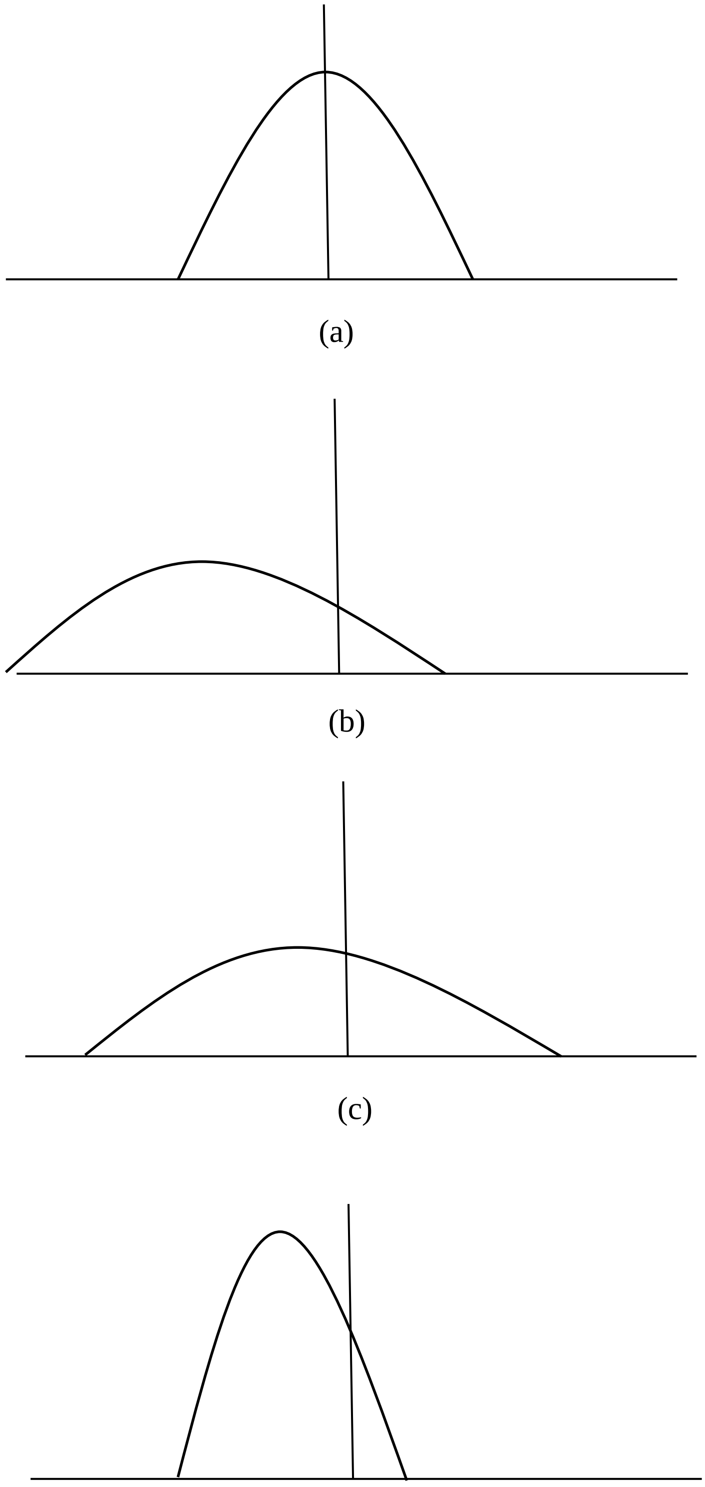
In the literature [Chung and Jang 2009, Collins and Palmeri 1997, Dabbas and Fowler 2003, Ham and Fowler 2007, Leachman et al. 2002, Lee et al. 2001, Li et al. 1996, Toba 2000, Vargas-Vilamil et al. 2003] WIP balance approaches, from the local (work-center) viewpoint only focusing on either avoiding bottleneck starvation or preventing non-bottleneck congestion, or from the global (wafer fab) viewpoint to reduce the WIP variability and smooth the process flow to achieve average cycle time reduction. They seldom address the cycle time variance performance. Based on the observations from simulation results, WIP balance progresses lots fast but with poor pace, which means that some lots are accelerated while other lots are delayed. This is inherently driven by the characteristics of wafer fabs, e.g., re-entrant flows, batch processing and setup time requirements. For instance, an early-arrival lot can be bypassed by a late-arrival lot because the late-arrival lot fulfills the batch or setup requirements. This can be illustrated by a simple example. Suppose machine *M0* processes three different products *P1*, *P2* and *P3*, and has two downstream machines *M1* and *M2*. *P1* and *P3* are processed by *M0* and *M1*, *P2* is processed by *M0* and *M2*. Lot *L1*, *L2* and *L3* belonging to *P1*, *P2* and *P3* respectively are available to be processed by *M0* at a given time. *L1* falls behind schedule but *L2* and *L3* are ahead of schedule. *L3* is at the top of the queue, *L1* is in the middle and *L2* is the last. In the meantime, *M1* has 10 lots in the queue, which is beyond the target WIP level. While *M2* only has 2 lots, which is lower than the target WIP level. Therefore, it makes sense that *M0* chooses to process *L2* although *L1* and *L3* arrive first. The consequence of such a WIP balancing action is to avoid capacity loss of *M2* and to prevent a long queue in front of *M1*, whereas, at the cost of bringing longer queue time to *L1* that has fallen behind schedule. Actually, there is still one more issue that WIP balance does not take into consideration. Although *L3* is ahead of *L1*, *L1* is more urgent than *L3*. How to distinguish between *L1* and *L3* requires more elaborated work for WIP balance. Without an effective mechanism of ensuring lot movement at the right pace, WIP balance sacrifices some lots by excessive tardiness to achieve average cycle time reduction, which is the root cause that is in conflict with due date control.

- The broadened cycle time distributions arising from WIP balance becomes a potential problem as long as due date performance is involved.

Figure 5.1.1 presents four hypothetical cycle time distributions, with the due dates represented by the vertical axis. Figure 5.1.1 (a) shows the distribution of a dispatching methodology ignoring WIP balance like FIFO. When WIP balance is applied, Figure 5.1.1 (b) shows that the tardiness is minimized because of a significant average cycle time reduction, although the cycle time distribution is broadened. However, when the average cycle time reduction is not sufficiently achieved, the tardiness performance becomes worse since the degraded variance causes that a proportion of lots have excessive tardiness, as represented by Figure 5.1.1 (c). Actually, the difference between Figure 5.1.1 (b) and (c) indicates that a high cycle time variance may cause missed due dates if the average cycle time cannot improve sufficiently. Figure 5.1.1 (d) shows the way to solve the problem in Figure 5.1.1 (c). The average cycle time is still maintained at the same level as in Figure 5.1.1 (c), but the tardiness performance is improved due to a low cycle time variance.

From the above observations, the key point to deal with the conflict between WIP balance and due date control is to minimize the degraded cycle time variance of WIP balance. In Section 4.2 we proposed three rules, one of which is the due date oriented dispatching rule ODD to minimize cycle time variance for MWVS and MIVS. The simulation results suggested that ODD rule has a significant complementary effect on MWVS and MIVS, which leads to not only cycle time variance minimization but also average cycle time reduction. We also conducted a comprehensive study about the due date oriented rules in Section 4.4. The excellent performance of ODD is the inherent characteristic of due date control. Thus, the problem that WIP balance shows weaknesses when good due date performance is required, can be resolved by due date control itself.

Once we obtain insight into the internal relationship between WIP balance and due date control, the next step is how to utilize due date oriented rules to solve the conflict. Section 4.2 showed that a hierarchical dispatching scheme may accomplish our goal. If we consider two-layer priorities for hierarchical dispatching, the first layer priority tells us which lots fulfill the WIP balance requirements, the second layer priority tells us the urgency (which lot is the optimal one to optimize a certain target) among the lots selected from the first layer, when other performance indicators such as cycle time variance have to be optimized. Consequently, both targets are taken into consideration, such that the



(a)

(b)

(c)

(d)

Figure 5.1.1: Hypothetical cycle time distribution when WIP balance is applied and due date is concerned

lots progress smoothly without serious fluctuation to achieve WIP balance. Furthermore, the narrowed cycle time distributions ensure that as many lots as possible complete before reaching their due dates.

- Therefore, it is our hypothesis that this two-layer priority hierarchical dispatching scheme can achieve both cycle time reduction and on-time delivery improvement simultaneously.

In this section, we will introduce another work-center oriented WIP balance approach named WIP Control Table (WIPCT) [Zhou and Rose 2010] to replace MWVS, as we highlight two-layer hierarchical dispatching. We apply WIPCT and MIVS as the top layer for WIP balance, and ODD as the bottom later for due date control.

5.1.2 Introduction to WIPCT

In the literature [Zhou and Rose 2010], a work-center oriented WIP balance approach called WIP Control Table (WIPCT) is proposed. This idea is based on a so-called pull-push concept which includes both pull and push philosophies. Pull means that a target WIP is specified to the work-center to determine which lots fulfill the pull request of downstream work-centers. While push means that among the lots fulfilling pull requests the upstream work-center chooses an optimal lot to push downstream, with the result that on one hand the pull request is satisfied, and on the other hand the desired performance target is achieved. Indeed, WIPCT expresses the same idea as the priority-based two-layer hierarchical dispatching methodology mentioned above to deal with WIP balance and due date control.

The objectives of WIPCT are (1): Evaluating the pull requests of downstream work-centers; (2): 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., target WIP levels, actual WIP levels, WIP differences and utilizations. Suppose work-center 1 has three downstream work-centers 2, 3 and 4. Table 5.1.1 describes a sample WIPCT of work-center 1.

Downstream Work-center	***Target WIP (lot)***	***Actual WIP (lot)***	***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 that needs to be maintained. The target WIP used in the MIVS and WIP Control Table 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 the work-center is running out of WIP. The smaller the difference is, the stronger pull request the work-center has;

The Actual WIP, WIP Difference and Utilization will be updated in case of lot move in/out and machine status change in work-center 1.

Table 5.1.1: An example of WIPCT

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 differences, the smaller the WIP difference, the higher the rank is. If work-centers have the same WIP difference, the one with higher utilization has a higher rank (From the opinion of engineers in Infineon Dresden, the high utilization work-center means the work-center is utilized quite often. Because the configurations like temperature, mask and so on are ready, this kind of work-center has a preference to get a lot to process). Based on this algorithm, in Table 5.1.1 work-center 3 has the strongest pull request, the next is work-center 2 and work-center 4 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.

5.1.3 Simulation Results and Performance Analysis

5.1.3.1 Benefit of Cycle Time Variance Minimization

We already addressed in Section 3.2 of Chapter 3 that the significant cycle time variance minimization performance is achieved when ODD is incorporated into MIVS. This section is an extension of Section 3.2 when on-time delivery performance is concerned, which intends to point out the benefit of cycle time variance minimization. To do this, we apply ODD to better distinguish the urgency of lots for WIPCT as we do to MIVS, and take two more performance measures which are percent tardy lots (pct. tardy lots) and average tardiness for tardy lots (avg. tardiness), along with average cycle time (avg. CT), cycle time variance (CT variance) and cycle time upper 95% percentile (CT upper pctile 95%).

In order to figure out the significant complementary effect of the ODD rule on WIPCT, we modify the second layer priority slightly 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 5.1.2. The cycle time variance is assumed to degrade considerably by this rule since the batch rule results in serious overtaking movements 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 5.1.2. 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).

	95% Fab Loading				
	Avg. Cycle Time (days)	***CT Variance (days^2)***	***CT Upper Pctile 95%(days)***	***Pct. Tardy Lots (%)***	***Avg. Tardiness (days)***
FIFO (DDFF 2.2)	29.6	1.8	39.1	62.4	2.2
MIVS+FIFO (DDFF 2.2)	28.7	1.8	37.0	44.4	1.0
WIPCT+FIFO (DDFF 2.2)	28.9	3.2	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:

DDFF: Due date flow factor; FIFO: First-in-first-out;

Batch+FIFO: For the batch processing work-center, if more than one lot has the same priority, the one that can achieve the largest batch size gets the highest priority; For the single processing work-center, if more than one lot has the same priority, the one that enters the queue first gets the highest priority.

Batch+ODD: For the batch processing work-center, if more than one lot has

	the same priority, the one that can achieve the largest batch size gets the highest priority; For the single processing work-center, if more than one lot has the same priority, the one that has the smallest ODD gets the highest priority.

Table 5.1.2: The benefit of cycle time variance minimization when due date is involved

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 losses and speed up the lot movements. 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 results in some lots becoming 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 95% percentile is not improved sufficiently. The percent tardy lots increases from 25.5% to 78.7%, and the average tardiness for 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.

5.1.3.2 General Performance of MIVS and WIPCT Combined with ODD

Because the performance of the ODD rule is influenced by the tightness of the target due dates, this section intends to examine the general performance of MIVS and WIPCT combined with ODD when target due dates are involved. We still take those five performance measures from Table 5.1.3 into account and compare them to FIFO, MIVS+FIFO, WIPCT+FIFO, ODD, MIVS+ODD and WIPCT+ODD under 95% fab loading. Furthermore, the target due date flow factor is taken into consideration, and ranges from 1.5 to 2.9 in steps of 0.2. The results are illustrated in Figures 5.1.2, 5.1.3, 5.1.4, 5.1.5 and 5.1.6.

From Figure 5.1.2, we observe that the average cycle times of the ODD rule are considerably larger when the target due date is set too tight such as 1.5 and 1.7, while ODD can perform as good as MIVS+FIFO and WIPCT+FIFO when the target due date is set appropriately such as 2.3 and 2.5. Because due date oriented rules are sensitive to due date tightness, it is not a trivial task to assign an appropriate target due date to each product. We can see the complementary strength of MIVS and WIPCT combining with ODD. On one hand, obviously, MIVS+ODD and WIPCT+ODD achieve shorter average cycle times than MIVS+FIFO, WIPCT+FIFO and ODD. On the other hand, MIVS+ODD and WIPCT+ODD are rather robust because they are less sensitive than ODD when target due dates change from tight to loose. This is a major improvement of WIP balance combining with due date control in comparison with applying WIP balance or due date control individually.

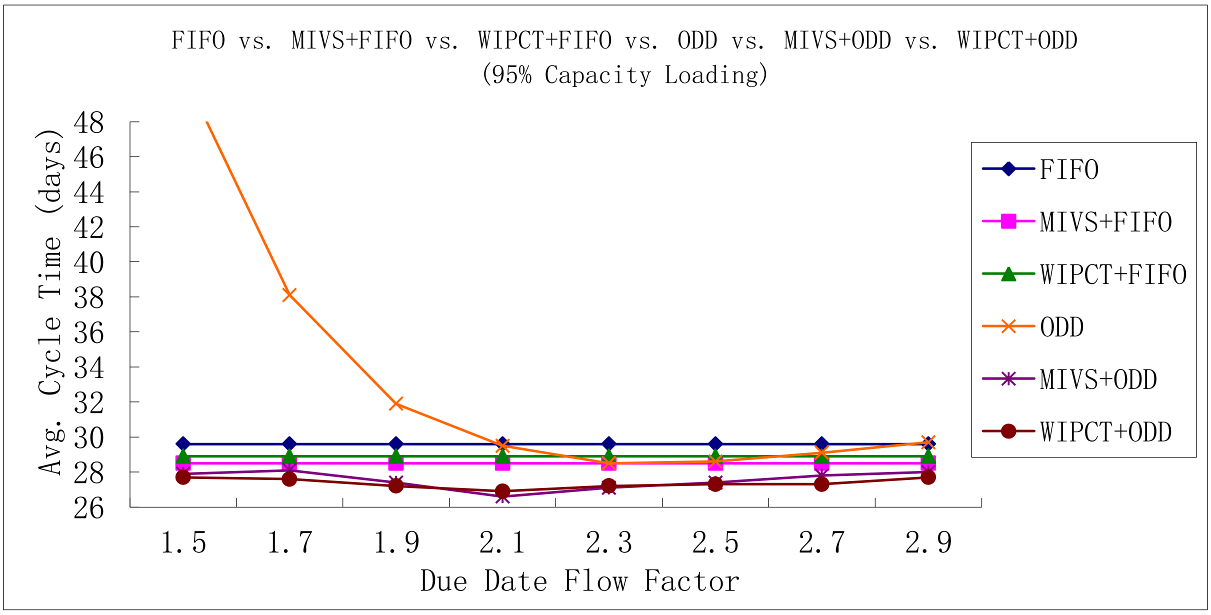


Figure 5.1.2: Average cycle time comparison

Figures 5.1.3 and 5.1.4 show characteristics of the cycle time distributions. As we can see, the ODD rule has a smooth and excellent cycle time variance curve. When ODD is introduced to MIVS and WIPCT, apparently both cycle time variance curves have a similar trend as ODD. More importantly, the variance performances of MIVS+ODD and WIPCT+ODD are improved compared to MIVS+FIFO and WIPCT+FIFO, respectively. The cycle time upper 95% percentile of MIVS+ODD and WIPCT+ODD are superior to the other four rules, except for the case of loose target due date flow factor 2.9.

The tardiness performance is presented in Figures 5.1.5 and 5.1.6. Due to the shorter average cycle times and narrower cycle time distributions achieved by MIVS+ODD and WIPCT+ODD, the percent tardy lots performance is improved considerably. For the due date flow factor 2.1, approximately 60% lots become tardy for both MIVS+FIFO and WIPCT+FIFO. MIVS+ODD and WIPCT+ODD manage to reduce the tardy lots from 60% to 10%. For the flow factor 2.3, MIVS+ODD and WIPCT+ODD achieve zero tardiness, while MIVS+FIFO and WIPCT+FIFO still have a portion of tardy lots. With regard to the average tardiness for tardy lots, MIVS+ODD and WIPCT+ODD outperform the other four rules.

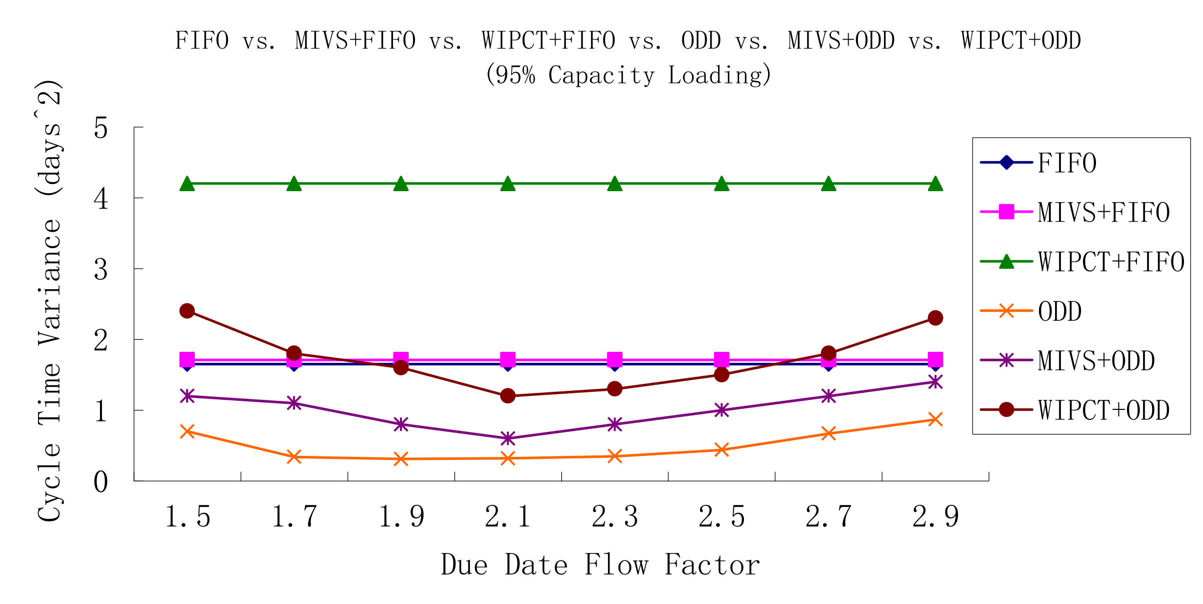


Figure 5.1.3: Cycle time variance comparison

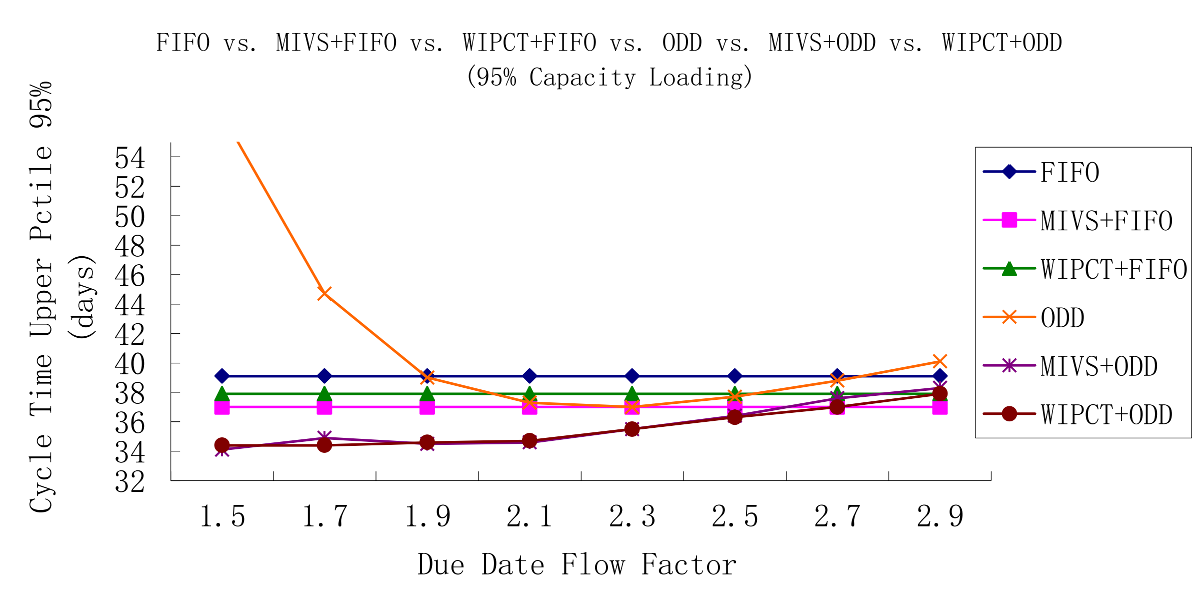


Figure 5.1.4: Cycle time upper 95% percentile comparison

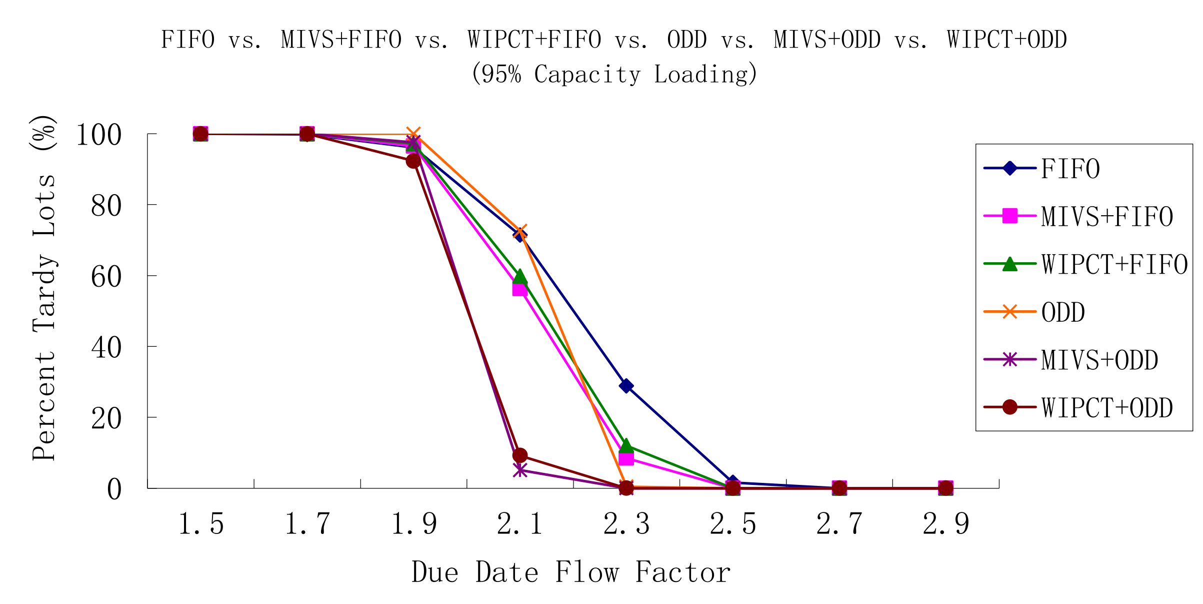


Figure 5.1.5: Percent tardy lots comparison

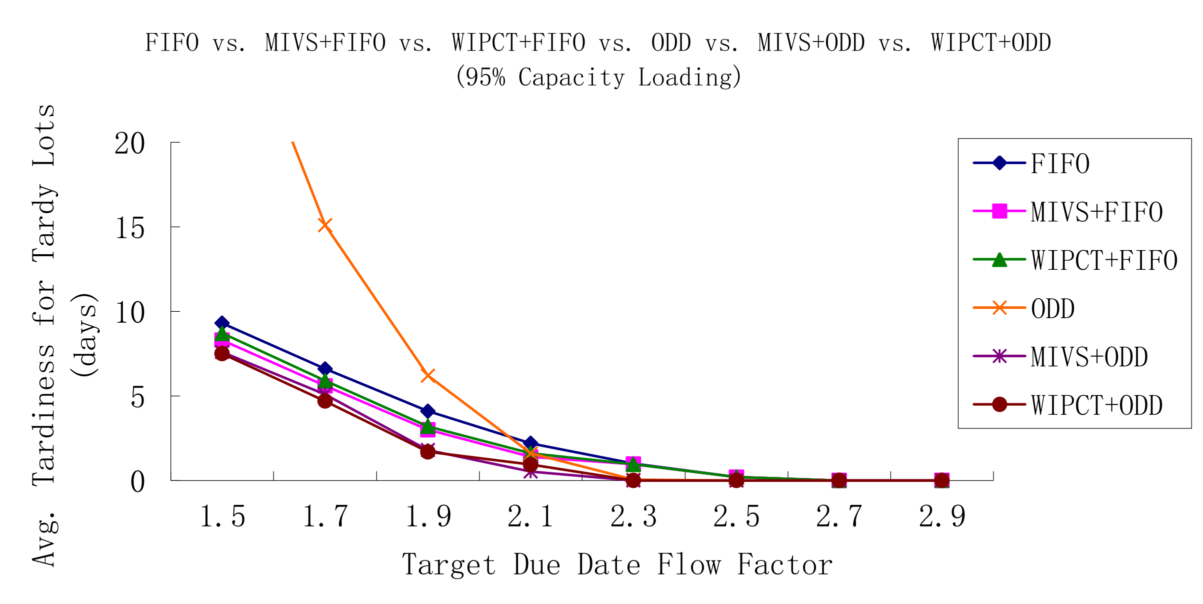


Figure 5.1.6: Average tardiness for tardy lots comparison

5.1.4 Conclusions

Based on the previous observations from using WIP balance or due date control individually, the key point to solve the confliction between WIP balance and due date control, when on-time delivery is concerned, is to minimize the cycle time variance for WIP balance. Inspired by the excellent cycle time variance minimization performance of due date oriented rules, this section intended to apply the ODD rule to deal with the problem arising from WIP balance. We took two WIP balance approaches that are MIVS and WIPCT into consideration, and integrated the ODD rule to them to form a priority-based two-layer hierarchical dispatching scheme which 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 account, such that the lots progress smoothly without serious fluctuations to achieve WIP balance. Furthermore, the narrowed cycle time distributions ensures that as many lots as possible complete before reaching their due dates.

Our assumption was confirmed by the simulation results. The performance of MIVS and WIPCT combining with ODD are promising.

- On one hand, MIVS+ODD and WIPCT+ODD are superior to MIVS+FIFO and WIPCT+FIFO respectively with regard to the cycle time reduction improvement. On the other hand, MIVS+ODD and

WIPCT+ODD are less sensitive than individually applying ODD with respect to the due date tightness change.

These two advantages make MIVS+ODD and WIPCT+ODD effective, robust and competitive. In other words, the introduction of ODD to MIVS and WIPCT not only improves the cycle time variance, but also achieves average cycle time reductions, which directly improves on-time delivery.

- Therefore, we can make a safe conclusion that due date control has significant compensation effect on WIP balance if short cycle times as well as good on-time delivery are desired, simultaneously.

5.2 A Composite Rule Combining WIP Balance and Due Date Control

5.2.1 Introduction

In Section 5.1 we saw that due date oriented rules have significant compensation effect on WIP balance. Conversely, we are curious about whether WIP balance can overcome the drawback arising from due date control as well. The problem of due date control rules e.g., EDD, LST, CR and ODD lead to considerable cycle times when target due date is set tight under high fab loading (Figure 3.4.2, P.112).

- This WIP imbalance is caused by the fact that overemphasizing due date control (1): ignores the global WIP status of the fab, thus the upstream work-centers mistakenly send lots to the congested downstream work-centers; (2): does not provide a mechanism to speed up lots before they are very close to their due dates in front of the congested work-centers, which leads to poor batch size for critical batch processing work-centers like '11026_ASM_B2' in the MIMAC6 model.

In contrast, the MDD and MOD rules which are extensions of the EDD and ODD rules solve the problem addressed above favorably as shown in Section 3.4. Let us take the MOD rule as an example. The original intention that SPT rule is introduced to ODD rule is to form a good combination to achieve short and predictable cycle times simultaneously. Although both SPT and ODD rules seem to include no information about WIP, the combination of them shows that the WIP is more balanced than for applying SPT or ODD individually. If we want to know the cause why WIP balance can be accomplished by the MOD rule, we need to understand how MOD works. The MOD rule attempts firstly to complete the lots early or on time, and secondly to complete the lots as soon as possible when the requested due date is unattainable. Assuming a loose target due date, and all lots having positive slack, the MOD rule performs as the ODD rule. With a tight target due date, all lots have negative slack, and the MDD rule performs as SPT rule, which exactly provides the mechanism to accelerate lots while the ODD rule shows poor performance under tight due dates.

The MOD rule shows us a way to deal with WIP balance and due date control. The composite rule shows the strength that is missing in the conventional single rules. Thus, firstly it is necessary to differentiate composite rules from conventional single rules. Single dispatching rules only focus on one objective, for instance, SPT rule is good at minimizing cycle time and ODD rule intends to minimize lateness variance to achieve good on-time delivery. In most cases, single rules are local because they only utilize information about the lots represented at the individual queue. Hence, single rules have their limitations and show restricted use in practice. In contrast, composite rules combine the characteristics of several basic single rules into one composite dispatching rule. Composite rule is a ranking expression which considers a basic rule as a function of attributes of lots or work-centers. In a composite rule, each basic single rule has its own scaling parameter which is chosen appropriately to determine the contribution of the basic rule to the composite rule. That is the difficult part in using composite rules.

- Inspired by the MOD rule, we intend to introduce a single WIP balance rule called Least Work at Next Queue (LWNQ) to MOD rule to achieve further WIP balance, for the reason that the average cycle time of MOD is still outperformed by FIFO when the target due date is tight.

The introduction of the SPT rule successfully handles the problem that the ODD rule does not accelerate lots before they are close to their due dates. We have the reason to believe that the LWNQ rule can deal with the problem that the ODD rule mistakenly send lots to the congested downstream work-centers. The LWNQ rule is a simple workload control rule which looks at the WIP flow from the viewpoint of work-centers. The lot that is to be processed by the next work-center with the least production hours remaining obtains the highest priority among the waiting lots. The scope of this study is to examine in detail the behavior of the proposed composite rule when the target due dates change from tight to loose, to confirm our assumption that WIP balance can be a compensation to due date control as well. As we mentioned above, how to determine the proper scaling parameters is the key to apply a composite rule. In this study, three scaling parameters with three levels are pre-determined and a design of experiment is used to acquire suitable levels for the parameters.

5.2.2 Proposed Composite Rule

5.2.2.1 Ranking Expression

The proposed composite rule is a ranking expression combining ODD, SPT and LWNQ (see Equation (5.2.1)). Each single rule has its own scaling parameter determining the contribution of itself to the total ranking expression. In this composite rule an index value is calculated for each lot and the lot with lower index value is favored.

$$I(i,t) = \frac{ODD}{P1} + \frac{PT + Now}{P2} + \frac{LWNQ}{P3} \qquad (5.2.1)$$

where *I(i,t)* represents the index value of lot *i* at time *t*, *ODD* is the operation due date value of lot *i*, *PT* is the processing time of lot *i*, *LWNQ* is the remaining production hours of the work-center at which lot *i* will be processed next, *Now* is the current time. *P1*, *P2* and *P3* are the scaling parameters.

These three scaling parameters should be related to the due date and tardiness of lots, workload of upstream and downstream work-centers, so as to determine the contribution of each basic rule. The following are the factors designed to determine *P1*, *P2* and *P3*.

- MOD factor: *M = Due(i, op) / (PT(i) + Now)*;
- Due date tightness factor: *T1 = 1 – Due(avg, final) / (Workload + Now)*;
- Due date tightness factor: *T2 = 1 – Due(avg, op) / (Workload + Now)*;
- Tardiness factor: *Tar1 = Tardiness(i) / Tardiness(avg)*;
- Tardiness factor: *Tar2 = Tardiness(i) / MaxTardiness(down)*;
- Slack time ratio factor: *S = (Due(i, op) - Now) / (Due(i, final) - Now).*

Where, in the queue of a work-center, *Due(avg, final)* is the average final due date of lots, *Due(avg, op)* is the average operation due date of lots, *Tardiness(i)* is the tardiness of lot *i*, *Tardiness(avg)* is the average tardiness of all lots in the queue, *MaxTardiness(down)* is the maximum tardiness in the downstream work-centers where lot *i* is heading, *Due(i, op)* is the operation due date of lot *i*, *Due(i, final)* is the final due date of lot *i*, *PT(i)* is the processing time of lot *i*, *Workload* is the remaining production hours of the work-center in which lot *i* is queuing, *Now* is the current time.

The factor *M* originates from the MOD rule. It decides whether the ODD rule dominates over the SPT rule or vice versa, working with different target due dates. *T1* represents the final due date tightness of lots. If *T1* is large, the average final due date is small, and most of the lots seem to be tardy with respect to their final due dates. Conversely, if *T1* is small, the average final due date is large, which means most of the lots likely complete on time. *T2* has the same meaning as *T1*, the difference lies in that *T2* considers the average operation due date for lots. If *T2* is large, which demonstrates that most of lots seem to be tardy for the due date of operation and vice versa. *T2* is more sensitive than *T1*, since operation due date considers due date for all intermediate operations, it reflects the tardiness problem more precisely than the final due date. *Tar1* is the measure of tardiness emergency in the queue. The larger the *Tar1* is, the more tardy the lot is. In contrast to *Tar1*, *Tar2* calculates whether the tardy lot has opportunity to be speeded up to next operation to catch up with the due date. If *Tar2* is larger than 1, which means the tardiness in the downstream work-centers for the lot is less serious than in the current work-center. The lot probably needs to be accelerated to the next operation. The factor *S* measures the slack time ratio between operation due date and final due date.

5.2.2.2 Design of Experiments

There are three scaling parameters *P1*, *P2* and *P3* which are considered as factors. In this study, each factor has three different levels as shown in Table 5.2.1. Therefore, a full factorial design with 27 possible combination is applied to figure out which level combination can achieve the best performance.

These three levels of each factor are designed for purpose in Table 5.2.1. We are aware that in the MOD rule, SPT rule dominates over ODD under tight due dates and ODD dominates over SPT under loose due dates. The main idea is that the LWNQ rule is designed to play the second role in the proposed composite rule. When target due dates are tight, "*P1>P3>P2*" causes that SPT plays the primary role, LWNQ plays the secondary role and ODD contributes the least. When the target due dates are not tight (medium or loose), "*P2>P3>P1*" brings that ODD contributes the most, next is the LWNQ and SPT has the least effect.

For *P1* and *P2*, Level 1, 2 and 3 show similarities. In principle, the MOD factor *M* is used to decide whether ODD or SPT rule contributes the most in the composite rule. This is the reason why *P1* is smaller than *P2* when target due dates are medium or loose (*M*>=1), and *P1* is larger than *P2* when target due dates are tight (*M*<1). Regarding the value of *P1* and *P2*, Level 1, 2 and 3 have different expressions, Level 1, 2 and 3 are determined by *M*, *T1* and *T2*, respectively. The basic idea is the ranges of *P1* and *P2* become smaller from Level 1 to Level 3. We use 0.3 and 0.5 as dividing points for *T1* and *T2* respectively. The reason is we ran one year simulation experiment with FIFO dispatching. We obtained approximately 220000 different values of *T1* and *T2*. Then we summarized and divided these 220000 different values of *T1* and *T2* into two levels evenly, and found out that 0.3 and 0.5 are the dividing points for the *T1* and *T2*, respectively.

There are two different cases for *T3*. The first case is to utilize *Tar1* in Level 1 and *Tar2* in Level 2. It is interesting to see how this composite rule behaves as we cannot guarantee that the value of *P3* is always between *P1* and *P2,* for the reason that *Tar1* and *Tar2* can be very large due to tight due dates. The second case is to apply slack time ratio factor *S* in Level 3. *P3* of Level 3 is designed to be between $P1$ and $P2$. However, with one special situation that S is larger than 1, which means the lot is extremely tardy. Thus, a negative value is applied to S to accelerate the lot.

Factors	***P1***	***P2***	***P3***
Level 1	If (M>=1) $P1$=1 Else $P1$=8*M	If (M>=1) $P2$=8*M Else $P2$=1	If ($Tar1$>=1) $P3$=$Tar1$ Else $P3$=1/$Tar1$
Level 2	If (M>=1) $P1$=1+$T1$ for $T1$<=0.3 $P1$=2-$T1$ for $T1$>0.3 Else $P1$=4.5+$T1$ for $T1$<=0.3 $P1$=6-2*$T1$ for $T1$>0.3	If (M>=1) $P2$=4.5+$T1$ for $T1$<=0.3 $P2$=6-2*$T1$ for $T1$>0.3 Else $P2$=1+$T1$ for $T1$<=0.3 $P2$=2-$T1$ for $T1$>0.3	If ($Tar2$>=1) $P3$=$Tar2$ Else $P3$=1/$Tar2$
Level 3	If (M>=1) $P1$=1.5+$T2$ for $T2$<=0.5 $P1$=3-$T2$ for $T2$>0.5 Else $P1$=5.5+$T2$ for $T2$<=0.5 $P1$=7-2*$T2$ for $T2$>0.5	If (M>=1) $P2$=5.5+$T2$ for $T2$<=0.5 $P2$=7-2*$T2$ for $T2$>0.5 Else $P2$=1.5+$T2$ for $T2$<=0.5 $P2$=3-$T2$ for $T2$>0.5	If (0<=S<1) $P3$=4+S Else If (S>=1) $P3$=- (S+2) Else $P3$= 4+|S|

Table 5.2.1: Design of experiment to determine the scaling parameters

5.2.3 Simulation Results and Performance Analysis

Firstly we consider that the fab is running with a tight target due date and under 95% capacity loading. The target due date flow factor was set to 1.5 to the all products, which means all products tend to be tardy. In the MOD rule, SPT plays a more important role than ODD under this tight target due date. By noticing this, the LWNQ rule was introduced to play second role in this composite rule. This is the reason why the scaling parameters were set in Table 5.2.1. Table 5.2.2 shows 27 possible average cycle times of all products corresponding to the different levels of scaling parameters in Table 5.2.1.

From Table 5.2.2, we can see that among all the combinations, the best average cycle time performance is achieved by *P1(L2)P2(L2)P3(L3)*, which is Level 2 for *P1*, Level 2 for *P2* and Level 3 for *P3* in Table 5.2.1. After that, we continued the simulation experiment with due date flow factors ranging from 1.7 to 2.9 in steps of 0.2. To each due date flow factor, we used the same design of experiment with three different levels of scaling parameters like Table 5.2.1. We found out that different levels should be set corresponding to different due date flow factors to acquire good average cycle time performance. In Table 5.2.3, we list the best levels of *P1, P2* and *P3* corresponding to different due date flow factors under 95% fab loading.

Avg. Cycle Time (days)					
P1(L1)P2(L1)P3(L1)	31.0	***P1(L2)P2(L1)P3(L1)***	30.5	***P1(L3)P2(L1)P3(L1)***	30.2
P1(L1)P2(L1)P3(L2)	31.2	***P1(L2)P2(L1)P3(L2)***	30.4	***P1(L3)P2(L1)P3(L2)***	31.0
P1(L1)P2(L1)P3(L3)	30.2	***P1(L2)P2(L1)P3(L3)***	29.9	***P1(L3)P2(L1)P3(L3)***	30.8
P1(L1)P2(L2)P3(L1)	30.8	***P1(L2)P2(L2)P3(L1)***	29.8	***P1(L3)P2(L2)P3(L1)***	30.1
P1(L1)P2(L2)P3(L2)	30.5	***P1(L2)P2(L2)P3(L2)***	29.9	***P1(L3)P2(L2)P3(L2)***	30.3
P1(L1)P2(L2)P3(L3)	29.9	***P1(L2)P2(L2)P3(L3)***	29.2	***P1(L3)P2(L2)P3(L3)***	30.2
P1(L1)P2(L3)P3(L1)	30.4	***P1(L2)P2(L3)P3(L1)***	30.0	***P1(L3)P2(L3)P3(L1)***	31.5
P1(L1)P2(L3)P3(L2)	31.5	***P1(L2)P2(L3)P3(L2)***	30.3	***P1(L3)P2(L3)P3(L2)***	31.2
P1(L1)P2(L3)P3(L3)	29.7	***P1(L2)P2(L3)P3(L3)***	30.1	***P1(L3)P2(L3)P3(L3)***	30.1

Table 5.2.2: Average cycle time of MIMAC6 with 1.5 target due date flow factor and under 95% fab loading, corresponding to different combinations of levels of scaling parameters *P1*, *P2* and *P3*

Secondly, we considered average cycle time, cycle time variance, cycle time upper 95% percentile, percent tardy lots and average tardiness for tardy lots as major performance measures which are from the best levels of *P1, P2* and *P3* listed in Table 5.2.3, and compared the proposed composite rule with MOD and FIFO. Figures 5.2.1, 5.2.2, 5.2.3 and 5.2.4 show these four performance measures. From Figure 5.2.1, the composite rule's average cycle time curve has a similar trend as the MOD rule. The maximum average cycle time can be found

Level ***Due Date Flow Factor***	***P1***	***P2***	***P3***
1.5, 1.7	If (M>=1) $P1$=1+$T1$ for $T1$<=0.3 $P1$=2-$T1$ for $T1$>0.3 Else $P1$=4.5+$T1$ for $T1$<=0.3 $P1$=6-2*$T1$ for $T1$>0.3	If (M>=1) $P2$=4.5+$T1$ for $T1$<=0.3 $P2$=6-2*$T1$ for $T1$>0.3 Else $P2$=1+$T1$ for $T1$<=0.3 $P2$=2-$T1$ for $T1$>0.3	If (0<=S<1) $P3$=4+S Else If (S>=1) $P3$=- (S+2) Else $P3$= 4+|S|
1.9, 2.1, 2.3	If (M>=1) $P1$=2+$T1$ for $T1$<=0.5 $P1$=3-$T1$ for $T1$>0.5 Else $P1$=4.5+$T1$ for $T1$<=0.5 $P1$=6-2*$T1$ for $T1$>0.5	If (M>=1) $P2$=4.5+$T1$ for $T1$<=0.5 $P2$=6-2*$T1$ for $T1$>0.5 Else $P2$=2+$T1$ for $T1$<=0.5 $P2$=3-$T1$ for $T1$>0.5	If (0<=S<1) $P3$=3+S Else If (S>=1) $P3$=- S Else $P3$= 3+|S|
2.5, 2.7, 2.9	If (M>=1) $P1$=1+$T1$ for $T1$<=0.7 $P1$=2-$T1$ for $T1$>0.7 Else $P1$=5.5+$T1$ for $T1$<=0.7 $P1$=7-2*$T1$ for $T1$>0.7	If (M>=1) $P2$=5.5+$T1$ for $T1$<=0.7 $P2$=7-2*$T1$ for $T1$>0.7 Else $P2$=1+$T1$ for $T1$<=0.7 $P2$=2-$T1$ for $T1$>0.7	If (0<=S<1) $P3$=5+S Else If (S>=1) $P3$=- (S+2) Else $P3$= 5+|S|

Table 5.2.3: Determination of levels of scaling parameters for different due date flow factors ranging from 1.7 to 2.9 with 95% fab capacity loading

for a tight due date flow factor 1.5, however, there is an almost 2 days improvement compared to the MOD rule. The introduction of LWNQ rule takes effect and brings further WIP balance for the fab. The average cycle time becomes smaller as due date flow factors change from tight to loose and reaches its minimum at a due date flow factor of 2.5 which differentiates from MOD for minimum average cycle time at due date flow factor 2.3. Besides that, the difference between the proposed composite rule and MOD rule starts at a larger magnitude, then becomes smaller when tight due date is changed to medium due date. The minimum difference is at medium due date flow factor of 2.1.

After that, it becomes larger again under loose due dates. This tells us that the LWNQ rule has more influence under tight and loose due dates than medium due dates. For the MOD rule, SPT dominates ODD under tight due dates and ODD dominates SPT under loose due dates. The LWNQ can overcome the WIP imbalance that happens due to only SPT or ODD dominance. No matter how the due date changes, the composite rule always outperforms FIFO rule.

With respect to cycle time variance, it seems that the due date flow factor of 2.3 is a watershed. Figure 5.2.2 shows that before flow factor 2.3 the composite rule is superior to the MOD rule, and the MOD rule outperforms the composite rule after 2.3. As we mentioned above, SPT plays a more important role than ODD under tight due dates in the MOD rule. However, SPT does not have a mechanism to reduce cycle time variance. The introduction of LWNQ helps SPT to achieve a better cycle time variance performance. In contrast, ODD has a major influence under loose due dates, and ODD can reduce the lateness relative to due date, thus reducing cycle time variance. The LWNQ can help to achieve WIP balance, however, at the cost of reducing the ODD effect. Therefore, the composite rule is outperformed by the MOD rule under loose due dates. With respect to the cycle time upper 95% percentile performance we can see that the composite rule is superior to MOD and FIFO rules in Figure 5.2.3.

Concerning the on time delivery performance, if the target due date is too tight, 100% of lots are delayed, while 0% are delayed for the loose target due date. Therefore, we only focus on the difference of selected rules with due date flow factor 1.9, 2.1 and 2.3. For other flow factors, the on time delivery percentage is either 100% or 0%. Figure 5.2.4 indicates that the composite rule is superior to MOD and FIFO. For flow factor 1.9 and 2.1, the composite rule has less percentage of tardy lots than MOD and FIFO. For flow factor 2.3, the composite rule achieves zero tardy lots while FIFO still produces around 30% tardy lots. Regarding to the average tardiness for tardy lots, Figure 5.2.5 illustrates that the composite rule also achieves a better performance than MOD and FIFO, since the tardiness curve of the composite rule is lower and flatter than MOD and FIFO cases.

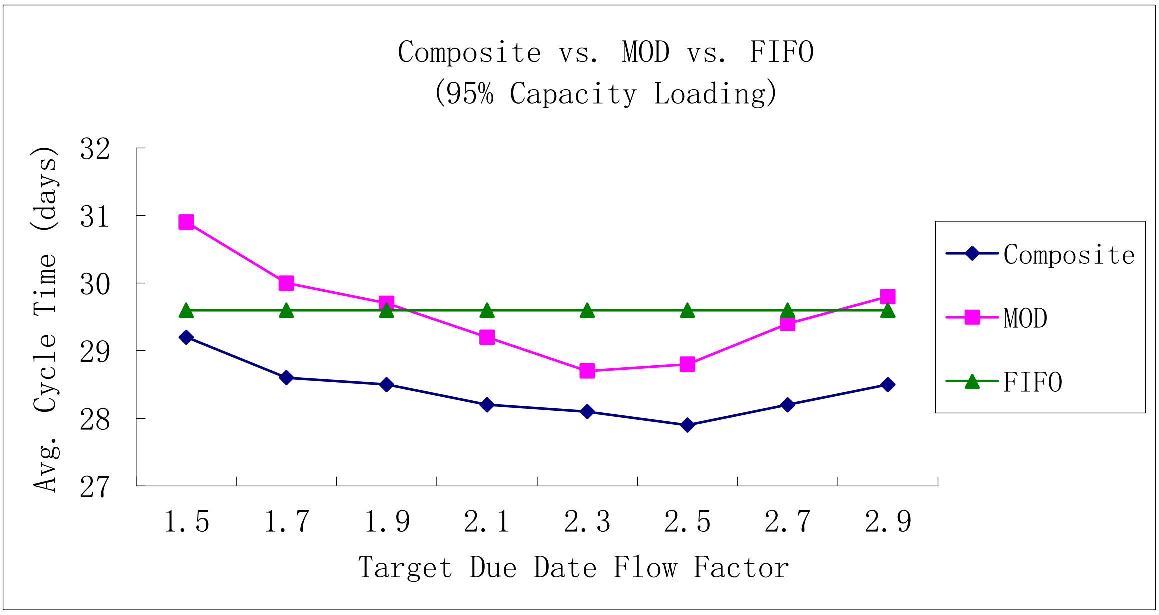


Figure 5.2.1: Average cycle time comparison

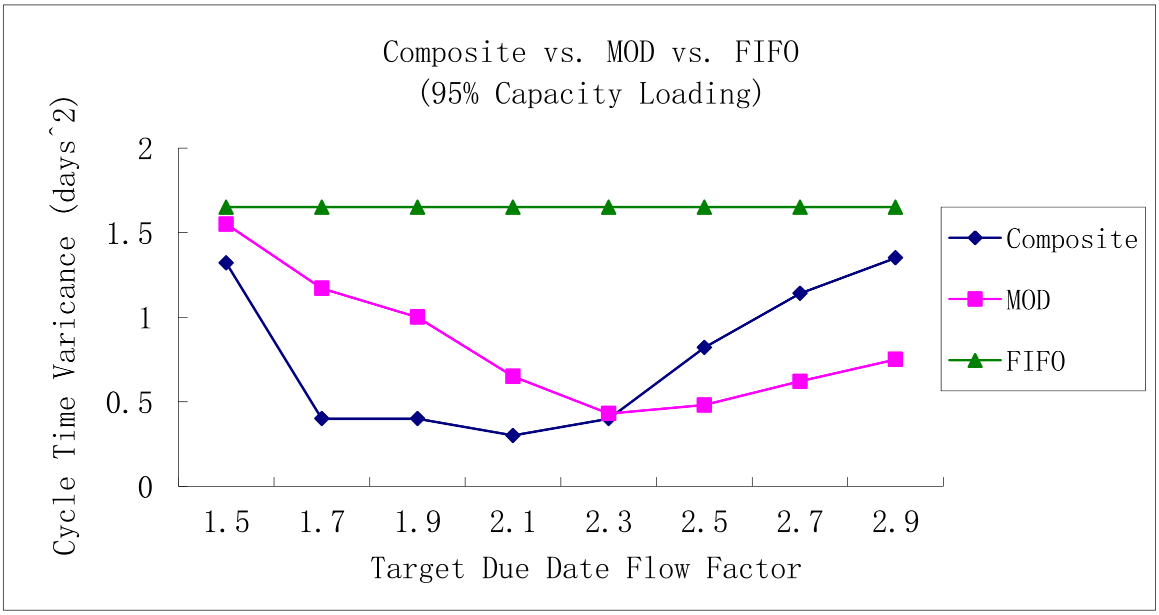


Figure 5.2.2: Cycle time variance comparison

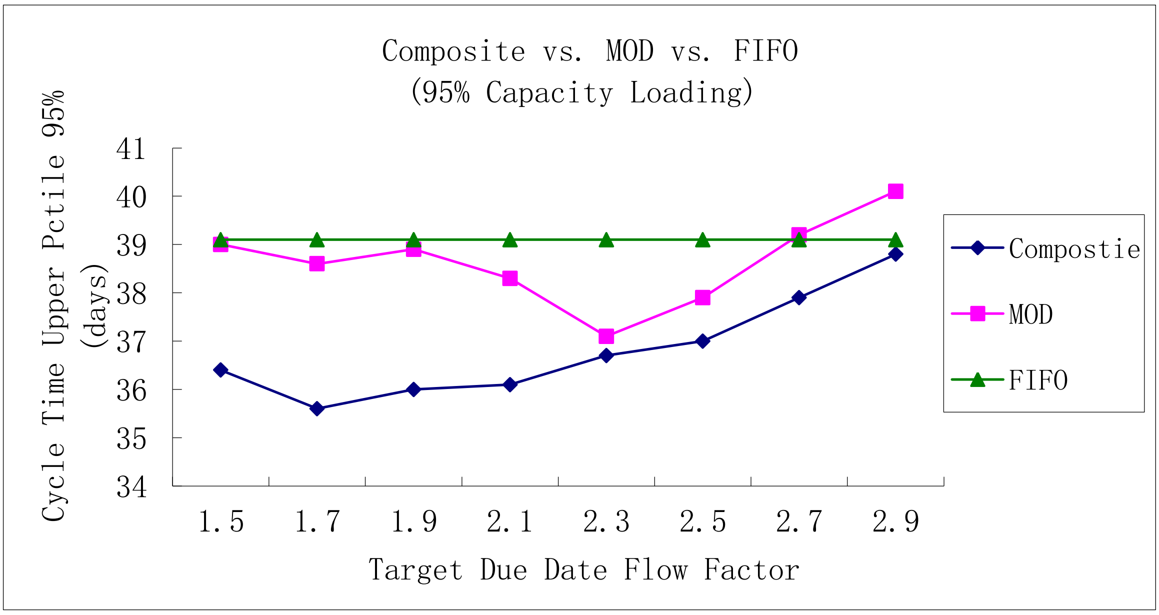


Figure 5.2.3: Cycle time upper 95% percentile comparison

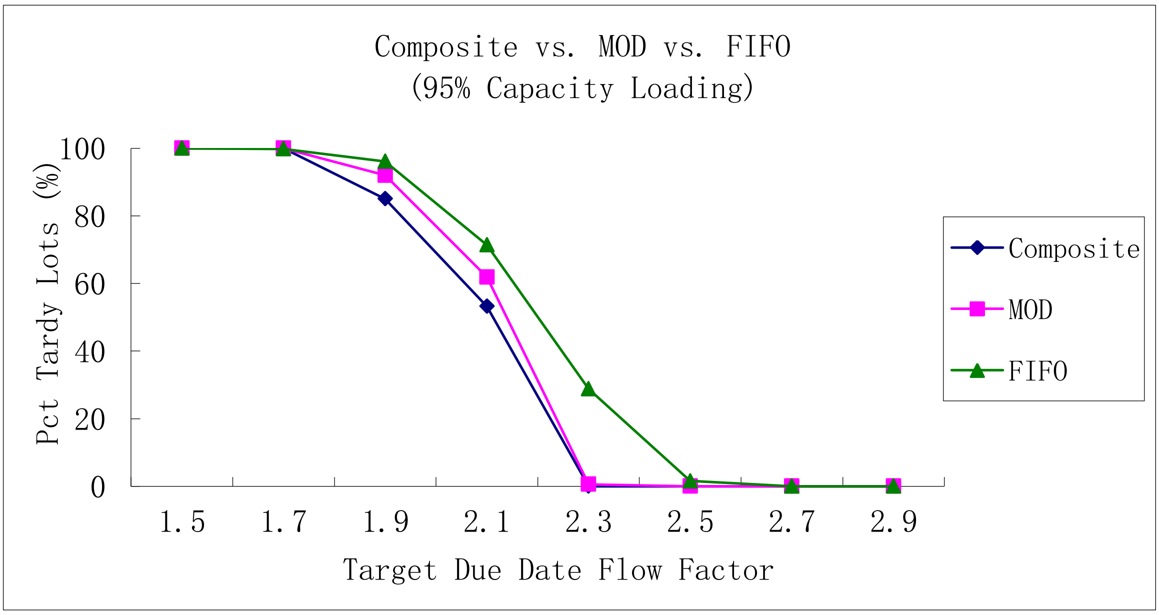


Figure 5.2.4: Percent tardy lots comparison

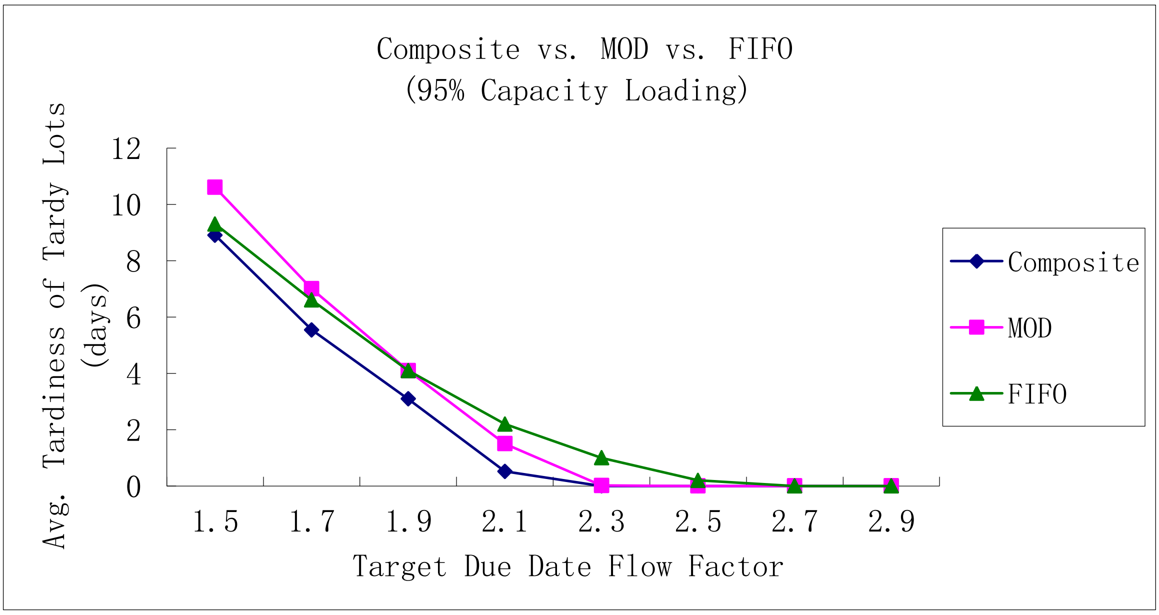


Figure 5.2.5: Average tardiness for tardy lots comparison

5.2.4 Conclusions

This section introduced a new composite rule that was an extension to MOD rule. The objective of this study was to integrate a single local WIP balance rule called LWNQ into MOD rule to achieve further WIP balance since MOD was still outperformed by FIFO under tight due dates. The proposed composite rule was a ranking expression including SPT, ODD and LWNQ. For each single rule, there was a scaling parameter to determine the contribution of single rule to the proposed composite rule. The challenge and difficulty in using the composite rule was to acquire appropriate scaling parameters. Thus, a design of experiment was utilized to determine the best level of each parameter for each due date flow factor. The simulation results demonstrated that the composite rule with appropriate scaling parameters achieved promising results regarding average cycle time, cycle time variance, cycle time upper 95% percentile, percent tardy lots and average tardiness for tardy lots performance versus the MOD rule.

The success of the composite rule originated from the proper scaling parameters. Actually, the *P1*, *P2* and *P3* in Table 5.2.3 were designed on purpose. We were aware that in the MOD rule, SPT rule dominates over ODD under tight due dates and ODD dominates over SPT under loose due dates. The LWNQ rule was designed to play the second role in the proposed composite rule. When target due dates were tight, "*P1>P3>P2*" caused that SPT played the primary role, LWNQ played the secondary role and ODD contributed the least. When the target due dates were not tight (medium or loose), "*P2>P3>P1*" brought that ODD contributed the most, next was the LWNQ and SPT had the least effect.

The contributions of this study are:

- We propose the viewpoint that overemphasizing due date control might cause WIP imbalance. Therefore, introduction of a WIP balance mechanism is very necessary when the fab is running under tight due date products and high fab loading.
- The first time to propose a composite rule including WIP balance and due date control. WIP balance and due date control take effect in parallel inside the composite rule, which differentiates it from the two-layer hierarchical dispatching scheme described in Section 5.1.

The proposed composite rule can achieve shorter average cycle times and better on-time delivery performance simultaneously in comparison with applying the single rule individually, which is a characteristic of multiple objectives accomplishment of the composite rule.

5.3 A Global Dispatching Rule To Manage Low And High Volume Products

5.3.1 Low and High Volume Products

So far we discussed that under certain circumstances, on one hand WIP balance and due date control are in conflict with one another, and on the other hand they can mutually compensate their drawbacks. As the semiconductor manufacturing is full of uncertainty and variability, in some cases trade-off or even sacrifice for one side is necessary if both targets cannot not be obtained concurrently. This section intends to solve a practical issue based on the observed interaction between WIP balance and due date control

There are hundreds of wafer products in a customer oriented wafer fab. Some products are referred to low volume products such as tests, samples, small orders and new products which have low release rates, i.e., dozens of wafers are released per week, while some products are referred to high volume products like common commodity type semiconductors which have a higher release rates than low volume products. Low volume products often have very tight target due dates and are more critical than high volume products with respect to cycle time and delivery reliability because of due date commitment to the customers. Low volume products are expected to go through the fab as fast as possible, at least meeting the target cycle time or due date. However, there is a basic observation that low volume products suffer more from specific machine constraints like higher batch formation times, longer setup waiting times and less qualified machines available, etc. In addition, local rules change the target function of global rules in order to make a compromise between due date and local constraints, for instance, a WIP balance target between the work-centers seems to reduce the weight of due date control, because a WIP balance approach would rather push an early lot to an empty work-center instead of push a tardy lot to a crowded work-center.

In general, this kind of wafer fab is controlled by due date oriented rules. By noticing the benefits achieved by WIP balance, we expect to apply WIP balance approach to it as well.

- There are two main concerns arising for low volume products (1): Whether due dates are sacrificed by achieving WIP balance for high volume products; (2): How to make a trade-off if due date is desired more than WIP balance.

To validate our assumptions, two customized WIP balance approaches are developed, but the intention is not to take the place of the existing rules in the fab. In contrast, we intend to incorporate these two customized WIP balance approaches into the existing rules, to find out the potential problem for low volume products. If our assumptions are true, similarly, two customized due date control approaches are proposed and integrated into the existing rules as well, to support low volume products.

5.3.2 A Global Rule Integrating WIP Balance and Due Date Control

5.3.2.1 Bottleneck Workload Control

According to the Theory of Constraints (TOC), the performance of the whole fab, e.g., its throughput is mainly determined by the bottleneck performance. It is necessary to determine an adequate WIP level for the bottleneck to avoid starvation and to support the whole fab to achieve its maximum throughput while running at the minimum WIP level. However, if the WIP level of the bottleneck exceeds the desired WIP level while achieving the maximum throughput of the whole fab, the cycle time is degraded. Lots will spend a significant queue time in front of the bottleneck work-center, which will also cause a WIP imbalance to the line. Therefore, a minimum workload is defined to the bottleneck work-center. If the actual workload of the bottleneck drops to the minimum workload, the bottleneck is fed with lots to prevent starvation. A maximum workload is also taken into account. If the actual workload of the bottleneck is higher than the maximum workload, bottleneck feeding is stopped to avoid extraordinary queue time, especially, when the bottleneck is broken down. In this study, we only consider a single dynamic bottleneck in the fab where the bottleneck is the work-center with the highest utilization. The minimum and maximum workload for the bottleneck is defined as 12 hours and 24 hours respectively which are defined by the engineers at Infineon Technology, Dresden Germany.

5.3.2.2 Feeding Empty Non-bottleneck Work-centers

Although the bottleneck is the most critical work-center which determines the performance of the whole fab, feeding empty non-bottleneck work-centers can also smooth the material flow, avoid capacity losses of machines, and improve product cycle times. Therefore, a minimum workload of 1.5 hours is also defined for the non-bottleneck work-centers. If the workloads of non-bottlenecks drop to this minimum workload level, lots are scheduled to feed them to avoid starvation. These 1.5 hours are also specified by the engineers in Infineon Technology, Dresden Germany.

5.3.2.3 Acceleration of Maximum Tardiness Lot

In general, WIP balance algorithms tend to push lots to work-centers that are running out of WIP without taking due dates into consideration. In this case, overemphasizing WIP balance has a negative impact on on-time delivery. In fact, sometimes it would be better to push a delayed lot to a high WIP work-center instead of pushing an early lot to a low WIP work-center. Because of customer commitments, keeping the due date is the first priority for customer oriented companies. Therefore, a compromise is necessary in order to meet due dates and reduce tardiness. Pushing a delayed lot despite WIP balance requirements to downstream work-centers can give the delayed lot a chance to speed up, to save cycle time, and reduce tardiness, although work-center capacity might be lost. The acceleration algorithm works as follows:

Step 1: In the queue of the upstream work-center, if lots are delayed for the operation, we determine the lot which has the maximum tardiness ‘MaxTardinessUp’ for the operation.

Step 2: Then, we identify the target downstream work-center where the ‘MaxTardinessUp’ lot will be processed. Next, we find the lot which has the maximum tardiness ‘MaxTardinessDown’ for operation in the queue of the target downstream work-center (like in Step 1).

Step 3: If ‘MaxTardinessUP’ is greater than ‘MaxTardinessDown’, the lot which has ‘MaxTardinessUp’ is assigned a high priority in the upstream work-center.

5.3.2.4 Acceleration of Lots Close to Due Date

Acceleration of delayed lots can only reduce tardiness instead of improving on-time delivery performance. Thus, we also propose to speed up the lots which are close to their due dates. This provides a mechanism for those lots to catch up with their due date. If there is still 1 week (this parameter is specified by the engineers in Infineon Technology, Dresden Germany) left for the lot to chase after the due date and the lot’s CR value is less than 1 - which means the lot is close to due date and possibly falls behind schedule - this lot will obtain a higher priority since there is a high probability that it will be delayed in the future.

5.3.2.5 Integration of the Proposed WIP Balance and Due Date Control Approaches into a Global Dispatching Rule

In order to test our approaches we extended a simplified version of the global dispatching rule ‘IFD’ which is in use at Infineon Technologies Technology Dresden, a German semiconductor manufacturer, with our ideas. As we can see from Figure 5.3.1 (a), there are 3 hierarchies of lot priority for the IFD rule. In each queue of a work-center, lots are categorized into 3 classes in descending priorities according to their states.

When WIP balance approaches as described in Section 5.3.2.1 and 5.3.2.2 are incorporated into IFD rule, it becomes the one in Figure 5.3.1 (b). From priority classes 2 to 4, the priorities are divided into 2 sub-classes which are delayed lot and non-delayed lot. The goal is to avoid bottleneck starvation and capacity loss for the empty non-bottlenecks. Nevertheless, the first problem for the low volume products arises here. The WIP balance approaches intend to balance the workload of work-centers, without taking the lot’s due date information into account, i.e., it would prefer to feed a lot with a loose due date to a low WIP work-center rather to push a lot with tight due date to a high WIP work-center. This may lead to cycle time reduction at the cost of on-time delivery of products with tight target due dates.

Therefore, in order to solve this problem, due date control approaches as described in Section 5.3.2.3 and 5.3.2.4 are included into the IFD rule, which is presented in Figure 5.3.1 (c). The delayed lots which fulfill the criterion for accelerating maximum tardiness lots belong to the second priority class. This priority class is more critical than the priority class of the bottleneck workload control method and of the feeding empty non-bottleneck method because customer commitment is more important than WIP balance in this study. Accelerating maximum tardiness lots is considered as a compromise to WIP balance. The upstream work-centers would rather push the maximum tardiness lot to downstream work-centers which may be highly loaded instead of pushing an early lot to downstream work-centers which may be starved to maintain WIP balance. The maximum tardiness lot has to be moved to the next operation to minimize delay. Furthermore, the non-delayed lot class is also split into two sub-classes which separate lots close to their due dates from lots on schedule. According to the acceleration of lots close to due date method, lots which are close to due date are more preferential than lots on schedule. In Figure 5.3.1, if lots belong to the same priority class, the ODD rule is applied as the dispatching rule.

(a) IFD Rule

1. Waiting time > 48 hours
2. Delayed lot
3. Non-delayed lot

(b) IFD Rule + WIP balance

1. Waiting time > 48 hours
2. Feeding empty bottleneck
    - 2.1. Delayed lot
    - 2.2. Non-delayed lot
3. Feeding empty non-bottleneck
    - 3.1. Delayed lot
    - 3.2. Non-delayed lot
4. Lot for non-empty work center including normal bottleneck
    - 4.1. Delayed lot
    - 4.2. Non-delayed lot
5. Lot for over-loaded bottleneck
    - 5.1. Delayed lot
    - 5.2. Non-delayed lot

(c) IFD Rule + WIP balance
+ Due date control

1. Waiting time > 48 hours
2. Acceleration of maximum tardiness lot (only for low volume products)
3. Feeding empty bottleneck
    - 3.1. Delayed lot
    - 3.2. Non-delayed lot
        - 3.2.1. Close to due date
        - 3.2.2. On schedule
4. Feeding empty non-bottleneck
    - 4.1. Delayed lot
    - 4.2. Non-delayed lot
        - 4.2.1. Close to due date
        - 4.2.2. On schedule
5. Lot for non-empty work center including normal bottleneck.
    - 5.1. Delayed lot
    - 5.2. Non-delayed lot
        - 5.2.1. Close to due date
        - 5.2.2. On schedule
6. Lot for over-loaded bottleneck
    - 6.1. Delayed lot
    - 6.2. Non-delayed lot
        - 6.2.1. Close to due date
        - 6.2.2. On schedule

Figure 6.3.1: Integration of WIP balance and due date control approaches into IFD rule

5.3.3 Simulation Results and Performance Analysis

5.3.3.1 Simulation Experiment

The products like ‘B6HF’, ‘C4PH’ and ‘C6N3’ have a low release rate, while the products like ‘B5C’, ‘C5P’ and ‘C5PA’ have a high release rate in the original MIMAC6 model. In order to test our idea, we modified the release rate to make sure the low volume products are separated from high volume products, which is demonstrated in Table 5.3.1. In Case 1, products ‘B6HF’, ‘C4PH’ and ‘C6N3’ are considered as low volume products. They only release 1-2 lots per week and have tight target due dates. Products ‘B5C’, ‘C5P’ and ‘C5PA’ are considered as high volume products. The release rates in Case 1 result in a fab loading of 99.5%. In Case 2, the low volume products in Case 1 are changed to high volume products. While the high volume products in Case 1 become low volume. (The reason for these changes is we want to find out whether the general behaviors of these wafer fabs are dependent/independent upon specified low/high volume products.) The release rates in Case 2 lead to a fab loading of 99.4%, which is quite close to Case 1.

Case 1: 99.5% Fab Loading			***Case 2: 99.4% Fab Loading***		
Product	***Release Rate (wafers per week)***	***Target Due Date Flow Factor***	***Product***	***Release Rate (wafers per week)***	***Target Due Date Flow Factor***
C6N3	48	1.8	C6N3	150	2.4
B6HF	24	1.8	B6HF	165	2.4
C4PH	48	1.8	C4PH	300	2.4
C6N2	100	2.4	C6N2	100	2.2
OX2	100	2.4	OX2	100	2.2
C5F	100	2.4	C5F	100	2.2
B5C	150	2.6	B5C	48	1.3
C5PA	300	2.6	C5PA	48	1.3
C5P	350	2.6	C5P	48	1.3

Table 5.3.1: Release rate and target due date flow factor for each product in MIMAC6

5.3.3.2 Average Cycle Time and Tardiness Performances Comparison

Firstly, the MIMAC6 model is tested by the IFD rule with the setting in Case 1. Then the WIP balance approaches are incorporated into the IFD rule, to find out whether the target due date of low volume products are sacrificed by the WIP balance of high volume products. If it is true, the due date control approaches are integrated into the IFD rule, too, to see whether the tardiness of low volume products can be minimized as much as possible without losing the cycle time achieved by WIP balance. The simulation of MIMAC6 was carried out for 18 months. The first 6 months were considered as warm-up periods and not taken into account for statistics. The average cycle time, percent tardy lot and average tardiness of tardy lots are considered as major performance measures, and the results are presented in Table 5.3.2.

When the WIP balance approaches are incorporated into the IFD rule represented as IFD+W, from the fab viewpoint, the average cycle time of all products are improved compared with the case of only the IFD rule, whereas, from the product viewpoint, not every product’s cycle time is reduced. According to the IFD rule, the ODD part plays an important role. Because the low volume products have a tight target due date, the ODD rule tries to process them as fast as possible. But the WIP balance approaches reduce the weight of due date control, therefore, it has a positive effect on the high volume products. While the low volume products naturally get no benefit but lose cycle time and tardiness performance. In this case, not only the low volume products but also the normal products like ‘C5F’ are influenced. Our first assumption becomes true that the due dates of low volume products are sacrificed by the WIP balance of high volume products. To compensate for the low volume products, the due date control approaches are incorporated into IFD+W, represented as IFD+W+D. Because the low volume products acquire the chance to speed up when they are close to the due date or already tardy despite of WIP balance. This saves cycle time and increases the on-time delivery for low volume products. In contrast, the cycle time and tardiness of other products degrade a little bit, because they share the cost that the low volume products benefit. Since the high volume products have enough time (loose target due date) to spend in the fab, the cost for high volume products is reasonable and acceptable. Moreover, the cycle time and tardiness of the whole fab still maintain the same level compared to IFD+W. The second assumption is confirmed by the fact that due date control overruling WIP balance achieves positive effects for low volume products.

Case 1: 99.5% Fab Loading									
	Avg. Cycle Time (day)			***Percent Tardy Lot (%)***			***Avg. Tardiness for Tardy Lot (day)***		
Product	***IFD***	***IFD+W***	***IFD+W+D***	***IFD***	***IFD+W***	***IFD+W+D***	***IFD***	***IFD+W***	***IFD+W+D***
C6N3	25.2	25.3	25.0	8.8	9.6	0	0.10	0.17	0
B6HF	28.8	29.1	28.7	20.4	28.8	7.1	0.15	0.22	0.06
C4PH	19.6	20.1	19.4	59.7	67.8	45.6	0.48	0.82	0.40
C6N2	28.8	28.3	28.4	0.5	0	0	0.03	0.01	0
OX2	31.2	28.4	28.6	0	0	0	0.005	0.002	0
C5F	34.7	35.0	35.1	5.2	12.0	13.3	0.06	0.26	0.28
C5P	30.4	29.4	29.5	18.2	4.6	9.4	0.14	0.06	0.10
C5PA	33	32.4	32.6	0	0	0	0	0	0
B5C	42.5	41.8	41.9	0	0	0	0	0	0
Fab	30.5	30.0	30.0	12.5	13.6	8.4	0.11	0.17	0.10

IFD + W: IFD rule combines with WIP balance approaches,
IFD + W + D: IFD rule combines with WIP balance and due date control approaches.

Table 5.3.2: Three performance measures of each products for Case 1

Next the same simulation experiments are carried out for Case 2 and the results are showed in Table 5.3.3. The low volume products have an extremely tight target due date (due date flow factor 1.3). Even if the IFD rule is applied, the low volume products are tardy. We realize that it is not possible to achieve non-tardiness, and what we desire is to reduce the tardiness as much as possible. The performance result is quite clear and similar to Case 1 when the WIP balance approaches are introduced to IFD. The cycle time and tardiness of low volume products degrade, although the cycle time of the whole fab is reduced. The due date approaches prove again that they can effectively improve the cycle time and tardiness of low volume products with small cost to the cycle time of other products.

Case 2: 99.4% Fab Loading									
	Avg. Cycle Time (day)			***Percent Tardy Lot (%)***			***Avg. Tardiness for Tardy Lot (day)***		
Product	***IFD***	***IFD+ W***	***IFD+ W+D***	***IFD***	***IFD+ W***	***IFD+ W+D***	***IFD***	***IFD+ W***	***IFD+ W+D***
C6N3	31.9	31.5	31.7	0	0	0	0	0	0
B6HF	36.8	36.4	36.3	0	0	0	0	0	0
C4PH	23.3	22.5	22.8	0	0	0	0	0	0
C6N2	25.6	25.7	25.4	0	0	0	0	0	0
OX2	24.5	24.0	24.2	0	0	0	0	0	0
C5F	30	29.4	29.7	0	0	0	0	0	0
C5P	15.2	15.7	15.0	63.9	72.1	50.1	0.28	0.41	0.18
C5PA	17.1	17.3	17.0	12.7	28.7	10.4	0.14	0.26	0.12
B5C	22.3	22.5	22.1	17.3	22.9	8.2	0.46	0.62	0.25
Fab	25.2	24.9	24.9	10.4	13.8	7.6	0.09	0.14	0.06

IFD + W: IFD rule combines with WIP balance approaches,
IFD + W + D: IFD rule combines with WIP balance and due date control approaches.

Table 5.3.3: Three performance measures of each products for Case 2

5.3.3.3 Cycle Time Distribution Comparison

In this section we change the focus to cycle time distributions to find out how the low volume products are affected by WIP balance and compensated by due date control. Figures 5.3.2, 5.3.3 and 5.3.4 show the cycle time distributions of ‘C4PH’, ‘C5P’ and all products for Case 1, respectively. We have a clear picture that WIP balance approaches increase the cycle time of low volume product ‘C4PH’ and decrease the cycle time of high volume products ‘C5P’. Whereas, due date control approaches shift the cycle time back to the left to support ‘C4PH’ without costing much for the ‘C5P’. Although due date control approaches focus on accelerating low volume products, the high volume products can share the cost from them. In addition, the low volume products contribute less than the high volume product to the fab cycle time. Thus, when due date control approaches are applied, the cycle time distribution of the whole fab stays approximately the same as when applying WIP balance approaches. One important thing we can see from the shape of the cycle time distributions is that WIP balance approaches lead to a sharp distribution, while due date control approaches has a smoother and narrower shape because of better pace of lot movements. The cycle time distributions of ‘C4PH’, ‘C5P’ and all products of case 2 have similar shape as Case 1, and are showed in Figures 5.3.5, 5.3.6 and 5.3.7.

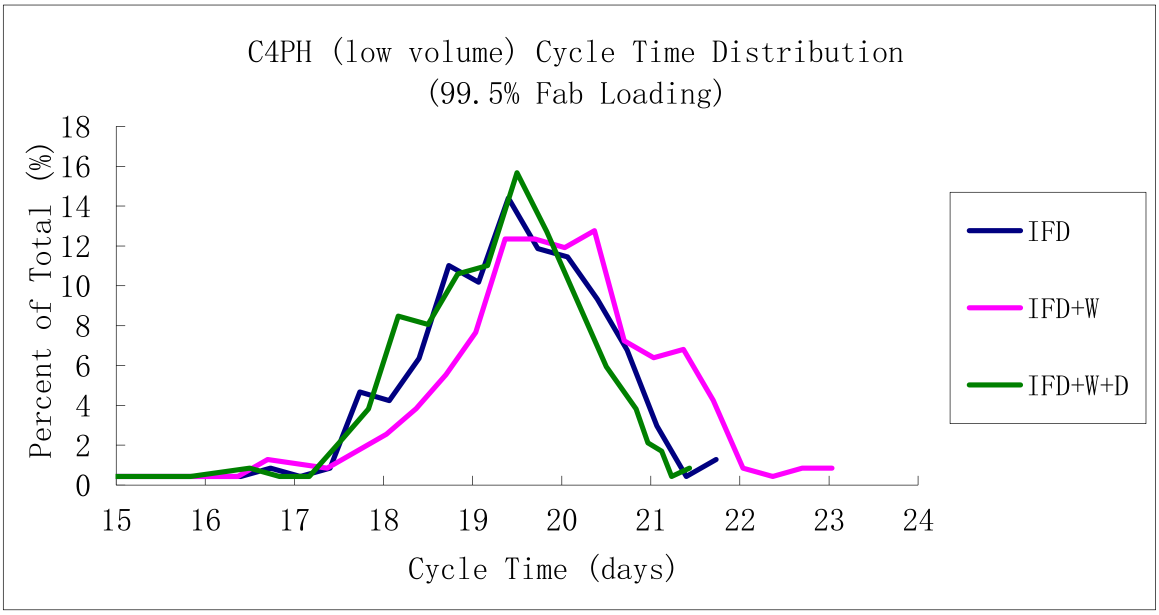


Figure 5.3.2: Cycle time distribution of low volume product ‘C4PH’ in case 1

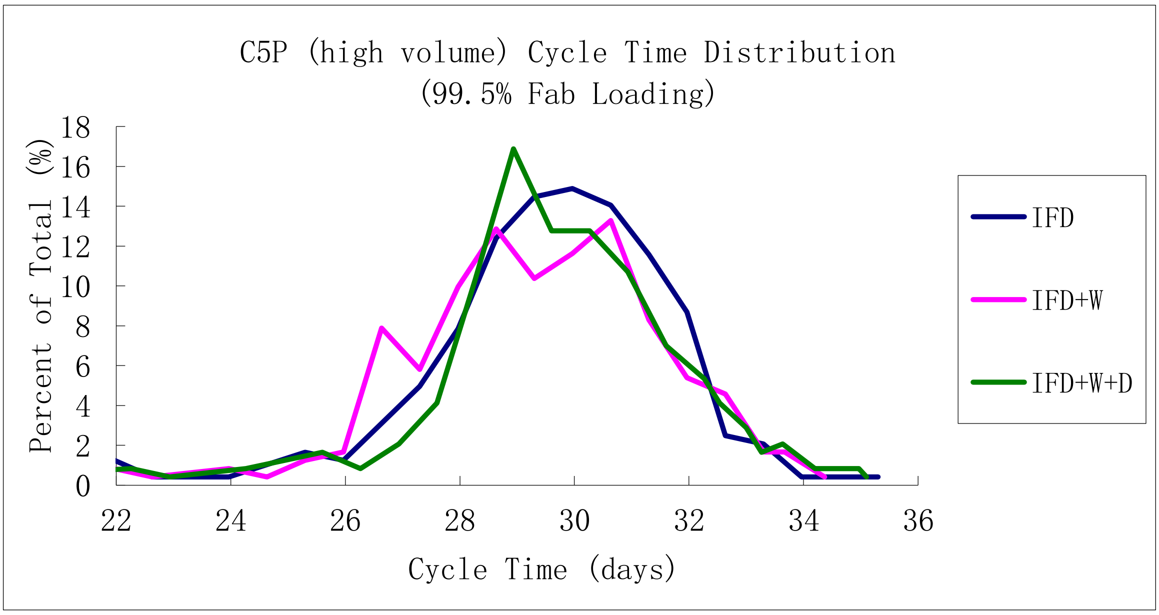


Figure 5.3.3: Cycle time distribution of high volume product ‘C5P’ in case 1

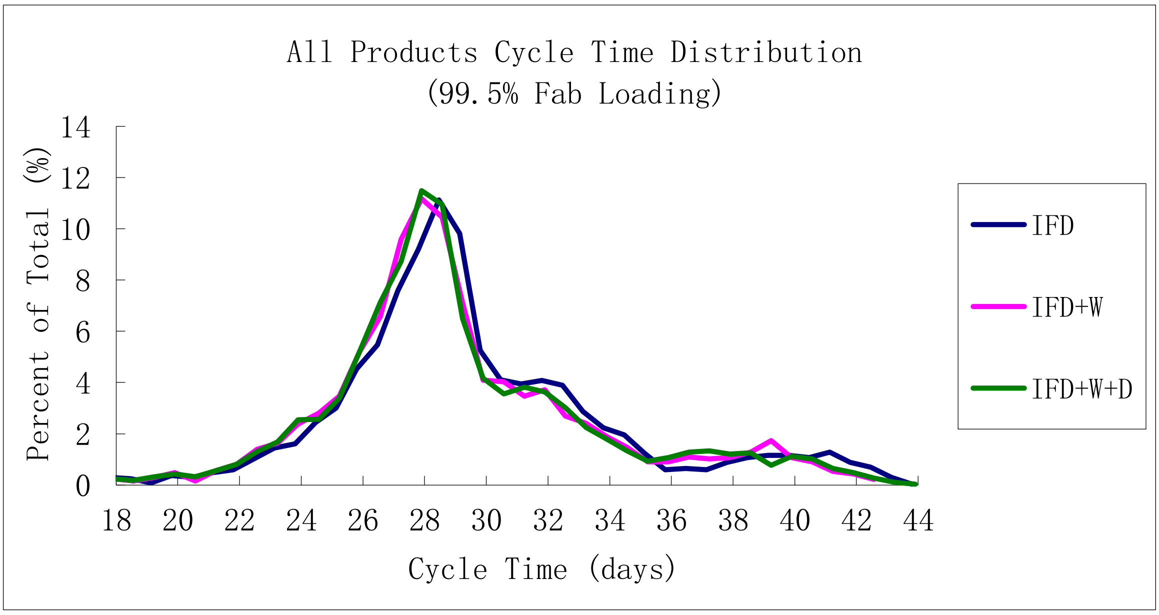


Figure 5.3.4: Cycle time distribution of all products in case 1

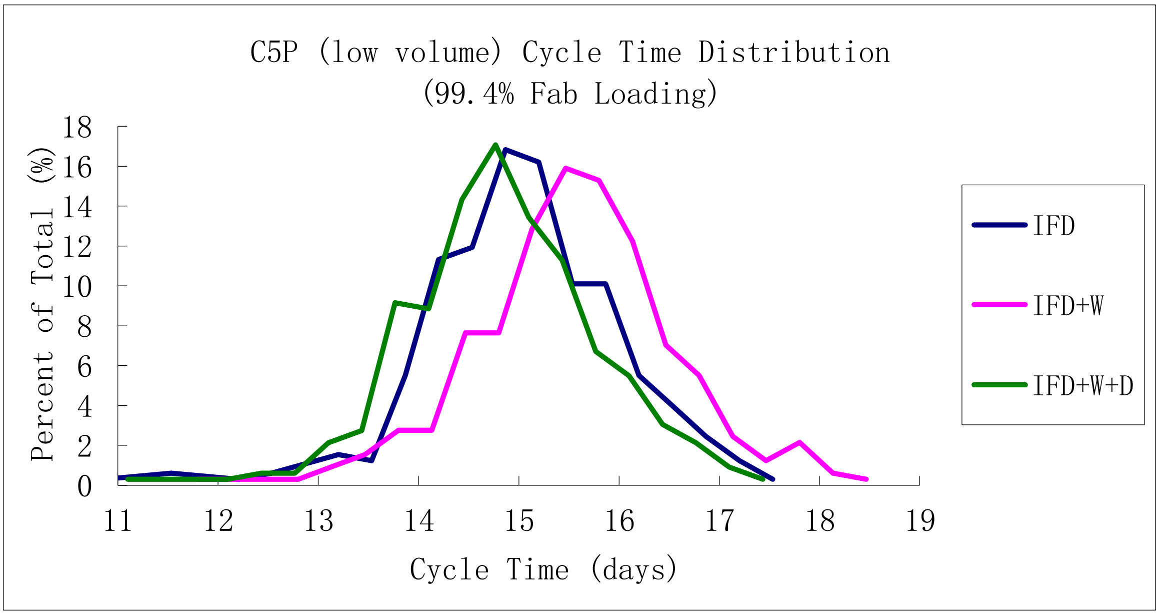


Figure 5.3.5: Cycle time distribution of low volume product ‘C5P’ in case 2

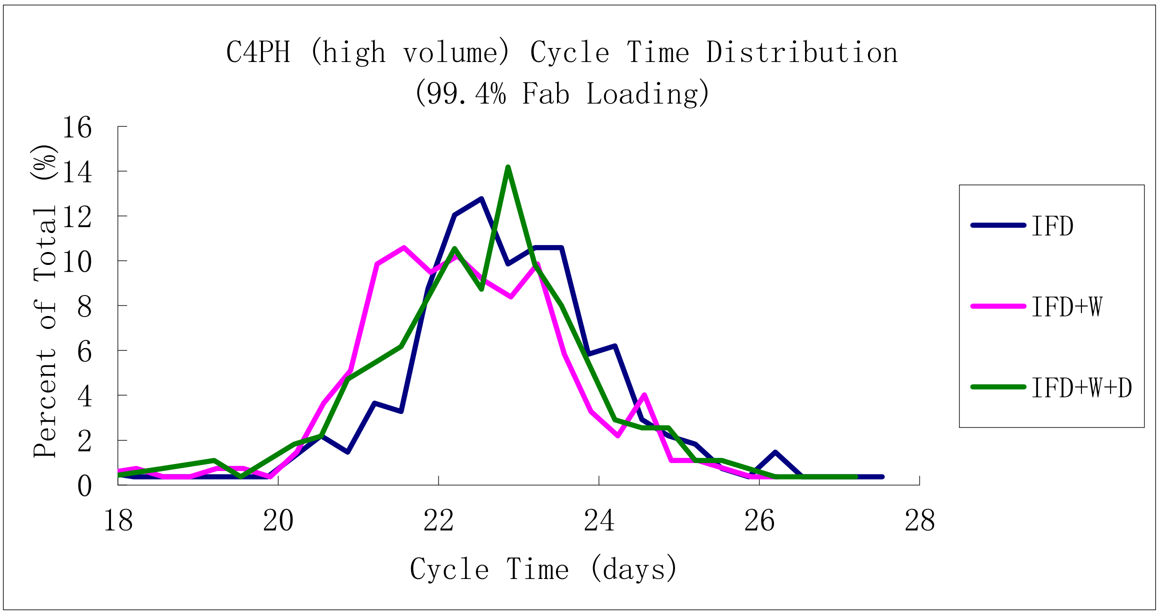


Figure 5.3.6: Cycle time distribution of high volume product ‘C4PH’ in case 2

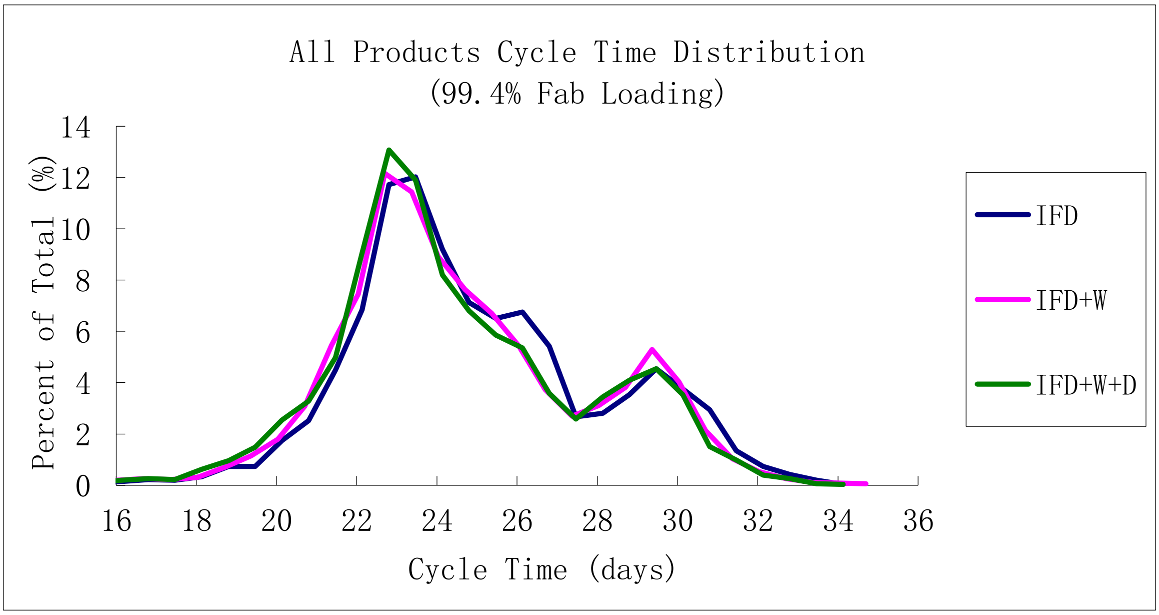


Figure 5.3.7: Cycle time distribution of all products in case 2

5.3.4 Conclusions

After we received an insight into the interaction between WIP balance and due date control, this section concentrated on applying this insight to a practical issue that is how to make a good compromise between low and high volume products.

As demonstrated in Figure 5.1.1 (P.179), WIP balance seems to reduce the weight of due date control. In reality, this theory is reflected by low and high volume products in a wafer fab. Low volume products like test and sample lots have a low release rate, while high volume products like common commodity type lots have a high release rate. The low volume products normally have tight due dates and are more critical than high volume products with regard to delivery reliability. With respect to WIP balance and due date control, there are two main issues for the low volume products. The first issue is that high volume products seem to take advantage of WIP balance, whereas, low volume products are sacrificed. Therefore, the second issue is the due dates of low volume products are more desired than the WIP balance of high volume products. We developed customized WIP balance and due date control approaches which were integrated into a due date oriented global rule called IFD rule. Firstly the WIP balance approaches were combined with IFD to validate the first issue. Then the due date approaches were added to IFD as well to solve the second issue.

From the simulation results, we can conclude that:

- From product viewpoint, the high volume products obtain benefits at the cost of increased cycle time of low volume products from WIP balance. From the fab viewpoint, since the low volume products contribute less than the high volume products, the average cycle time of the whole fab is still improved by WIP balance;

- Due date control approaches can effectively accelerate low volume products, while the cost can be shared by the high volume products. Thus, the average cycle time of the whole fab can maintain the same level for only applying WIP balance.

Actually, the main perspective of this study is not to assess the customized WIP balance and due date control approaches, instead, the priority-based hierarchical dispatching. IFD rule offers us an example on how to integrate customized methods into existing manufacturing rules. In reality, each wafer fab has its own Manufacturing Execution System (MES) [McClellan 2001] using various global rules to make dispatching decision. As the importance of WIP balance is noticed, this priority-based hierarchical dispatching offers a chance to explore the feasibility of incorporating the customized optimization methods into the existing rules, to find out the positive and negative effects and the corresponding methods to solve the problem. For instance, in general, there are two ways to accelerate the low volume products which are to assign either tight due dates or high priorities. Because low volume products already have tight due dates, applying due date control approach is the only way left to speed them up. To support the low volume products, we give higher priorities to the due date control approach to make sure that the due dates of low volume products overrule the WIP balance efforts of high volume products. This hierarchical dispatching is easy to understand and to implement.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

This chapter describes the achievements of this dissertation. We intended to carry out an in-depth study on two important performance indicators WIP and due date in wafer fabs. Our goal was to solve eight issues (described in Abstract) relating to WIP balance and due date control that attracted most attention from Infineon Technology, Dresden Germany. The MIMAC6 model was used as simulation environment to test our proposed approaches. Based on the observation and analysis of the simulation results from MIMAC6 model, we summarize this dissertation in accordance with the following eight issues.

Issue one: Work-center oriented WIP balance

One symptom of WIP imbalance is the starvation and congestion of work-centers. In particular, we observed that lots spend extraordinary long queue times in front of some critical work-centers under high fab loading (Table 3.1.1, P.46; Table 3.1.2, P.47), which causes serious variability in wafer fabs. To correct for this WIP imbalance, we proposed work-center oriented WIP balance to avoid work-center starvation and congestion, so as to reduce WIP variability. The MWVS approach in Section 3.1 and the WI approach in Section 4.1 were the examples of this trend. From the simulation results, we realized that the significant improvement of work-center oriented WIP balance depends on several factors:

(1). Fab loading: We divided the fab loading into three levels which are low (75%), medium (85%) and high (95%) for the simulation experiments. In comparison with MIVS and FIFO, work-center oriented WIP balance achieved promising cycle time reduction under high fab loading (Table 3.1.8, P.58; Table 3.1.9, P.60; Table 3.1.11, P.63; Figure 3.1.5, P.64; Table 4.1.1, P.142; Figure 4.1.4, P.145). It is understandable that the higher the fab loading is, the more serious the variability becomes in the fab and the easier the work-centers suffer from WIP imbalance.

(2). Preventing long queues of work-centers: The MWVS and WI approaches utilized different mechanisms (MWVS used target WIP, WI used real global WIP information) to measure the pull request from downstream work-centers. Nevertheless, they had the same objective to prevent long queue piling up in front of the work-centers, which is of particular importance to reduce the queue times of lots (Table 3.1.11, P.63; Figure 3.1.5, P.64; Table 4.1.2, P.144).

(3). Machine breakdowns, setups and batches: One advantage to manage WIP from the viewpoint of work-centers is to incorporate machine breakdowns, setups and batches into dispatching decision making. The simulation results from Section 4.1 (Table 4.1.1, P.142) showed that the WI approach achieved considerable cycle time reduction when these three aspects were taken into account.

Issue two: Fast but poorly paced lot movement of WIP balance

Conventional WIP balance focuses on reducing cycle time through fast lot movement. However, as the complexity of wafer fabs increase day by day, fast lot movement is no longer the only criterion to judge the performance of WIP balance. In this dissertation, we raised WIP balance to a higher level that is fast along with rhythmic lot movement, in other words, it represents as low cycle times with low variance. Section 3.2 described three cycle time variance minimization methods. The simulation results (Table 3.2.1, P.76; Table 3.2.2, P.77; Table 3.2.3, P.78) demonstrated that they are able to improve the poorly paced lot movement arising from MWVS approach, which has significant effect on forecast, capacity planning and so on. The lower and smoother WIP evolution curve of Matrix Table in Section 4.2 (Figure 4.2.5, P.161) also presented high level of WIP balance. In reality, the WIP needs to be traced and anticipated to make sure the unforeseeable exceptions like machine breakdowns would not cause serious trouble for the smooth manufacturing process. Lot movements with better pace (smooth WIP curve) supports WIP traces and forecasts with less difficulties and more accuracy.

Moreover, as more and more wafer fabs change the manufacturing fashion from make-to-stock to make-to-order, one more reason for low cycle time variance is to increase the ability to meet due dates reliably. Section 5.1 was an extension to Section 3.2 by taking due dates into consideration. The expanded cycle time distributions of WIP balance becomes a potential problem when due dates are involved. The simulation results in Section 5.1 (Table 5.1.2, P.183) showed us the fact that a low average cycle time in combination with a low cycle time variance could improve on-time delivery and reduce tardiness remarkably. We realized that the inherent characteristic of due date control can overcome the drawback arising from WIP balance, especially, we highlighted the remarkable minimized cycle time variance of ODD rule.

Thereby, the **first contributions** of this dissertation are:

- We addressed the potential problem of WIP balance and raised WIP balance to a higher level that can be represented by MIVS+ODD, MWVS_1+ODD (Table 3.2.1, P.76) and Matrix Table (Figure 4.2.5, P.161);
- We found out the key point to solve the conflict between WIP balance and due date control which is the cycle time variance minimization. Thereby, we proposed to use due date oriented rules as a connection point to link WIP balance and due date control (Table 5.1.2, P.183; Figure 5.1.2, P.185; Figure 5.1.3, 5.1.4, P.187; Figure 5.1.5, 5.1.6, P.188);
- We suggested to use the FF rule to replace ODD rule, if it is hard to determine the due date tightness for ODD in the fab, since FF requires no due date information and performs as well as ODD (Table 3.2.1, P.76).

Issue three: How to acquire target WIP for work-centers

As we mentioned in issue one, the performance of MWVS relies on target WIP, one straightforward question is how to acquire the target WIP for MWVS. Section 3.3 introduced three ways to determine the target WIP. At the beginning of our simulation study, the target WIP derived from the average WIP of work-center applying FIFO as dispatching rule. As this method was questioned for its accuracy, we were motivated to explore alternatives like queuing models and neural networks which are considered as standard but sophisticated procedures.

The main perspective of Section 3.3 is not to discuss which method can lead to the most accurate target WIP, since they all have pros and cons. On the contrary, from the simulation results, we were aware that: (1) Preventing congestion is more important than avoiding starvation under high fab loading. (2) Lots are proved to pile up in front of high utilized work-centers. Hence, the major role of target WIP at the highly utilized work-centers is to prevent long queues under high fab loading. When we made a comparison among those three target WIP estimation methods, we noticed that the target WIP of the top 10 highly utilized work-centers from queuing models is relatively higher than in the cases of other two methods (Table 3.3.5, P.93; Table 3.3.6, P.250). This is the reason why the average cycle time performance from queuing models is outperformed by other two methods. It also indicates that an overestimated target WIP for highly utilized work-centers might increase the possibility of long queue times..

Except for the insight into target WIP for highly utilized work-centers, another outcome is that we were motivated to develop a WIP balance approach without the requirement of target WIP by noticing the difficulties in applying and acquiring target WIP in Section 3.3.

One limitation of this dissertation is that we did not spend much effort to carry out an intensive study about what is ‘appropriate’ or ‘accurate’ target WIP. Actually, the appropriate target WIP depends on many factors, such as product type, production system, capacity planning, management capability and so on, into which we cannot get insight, not to mention that there are hundreds of work-centers in the fab. Some experiences from industry are to divide target WIP into several sub levels, for instance, the actual WIP lower than the minimum target WIP means hungry, between the minimum and maximum target WIP is normal, higher than maximum target WIP is crowded. The engineers on site can better understand the meaning of appropriate target WIP levels and identify the status of work-centers by this elaborated target WIP.

Issue four: WIP balance without the requirement of target WIP

As more and more complaints from industry were made about the challenges and difficulties in applying target WIP (see introduction of Section 4.1), we were motivated to find a possibility to achieve WIP balance without the requirement of target WIP. The WI approach and the simulation experiments in Section 4.1 showed a feasible way to meet our goal. To abandon target WIP, first of all we have to understand the role of target WIP in WIP balance. The target WIP plays the role in preventing congestion and avoiding starvation which can be viewed as measuring the pull request from downstream work-centers from the viewpoint of the push/pull philosophy.

Since the dispatching decisions can be different from different standpoints to measure the pull request precisely. We suggested to employ large sets of information, which are upstream/downstream WIP information, setup and batch requirements and operation due date information, to achieve optimal dispatching decision in the WI approach. The simulation results showed that the WI approach succeeded in balancing the WIP as significant as MWVS.

The **second contribution** of this dissertation is:

- The WI approach introduces a new scenario to handle WIP balance without target WIP (Table 4.1.1, P.142; Table 4.1.2, P.144; Figure 4.1.4, P145).

The WI approach is effective because it is a global rule utilizing information both within and outside the domain of the neighborhood of the decision point in space and time. The global rule shows a considerable advantage in overcoming the constraints of using target WIP. As a matter of fact, using global information is the trend of WIP balance in the future as more and more authors address that look-ahead combined with look-back strategies can achieve the best performance of interests effectively in comparison to utilizing local information.

Issue five: work-center status information vs. lot status information

The information used for dispatching in wafer fab can be classified into work-center status and lot status. The work-center status information is described as how many lots or wafers are in the queue for the MWVS rule or how many production hours are left for the WI approach. The lot status information is viewed as ahead of/on/behind schedule, precisely speaking, it is due date information. We noticed that the reason why WIP balance and due date control have negative effect to one another is because they only focus on their preferable information while disregarding the others. Thus, we recommended to take both work-center status and lot status information into account concurrently. The WI approach in Section 4.1, the Matrix Table in Section 4.2 and the two-layer hierarchical dispatching scheme in Section 5.1 showed successful approaches of this idea.

Actually, issue five is an extension of issue four. The WI approach mainly utilized the workload information of work-centers, while the Matrix Table considered both information in parallel and achieved a trade-off between WIP balance and due date control. The WIP evolution curve of the Matrix Table indicated that a smooth WIP flow without serious fluctuation was achieved (Figure 4.2.5, P.161; Figure 4.2.6, P.163), which is considered as a good example of high level WIP balance (low cycle time with low variance).

Issue six: General performances of due date oriented rule

Now that we found out that due date oriented rules are the key point to solve the conflict between WIP balance and due date control, Section 3.4 provided a comprehensive study about 10 due date rules from the literature. Because nowadays many wafer fabs are still controlled by due date rules, we highlight three aspects as a general guideline:

(1). In the MIMAC6 model, we found out that it was difficult to apply lot-based due date rules (that are EDD, MDD, LST and CR) under high capacity loading (Figure 3.4.2, P.112; Figure 3.4.7, P.116). In contrast, operation-based due date rules (that are ODD, A/OPN, MOD, LOST, S/OPN and OCR) performed quite well. Therefore, the rules utilizing operation-based due dates are more effective than the ones utilizing lot-based due dates.

(2). Most of the rules achieved higher cycle time under tight due dates than in the cases of medium and loose due dates. In other words, due date rules are proved to suffer from WIP imbalance under tight due dates because of overemphasizing due date control, which is considered as one drawback of due date control. When the fab is running with tight due dates products under high capacity loading, in order to prevent high WIP, either using operation-based due date rules or using composite rules like MOD to break the dominance of due date control would be preferable (Figure 3.4.2, P.112; Figure 3.4.7, P.116). The

Matrix Table in Section 4.2 and the new composite rule in Section 5.2 were motivated by this fact.

(3). The excellent cycle time variance performance of ODD is addressed (Figure 3.4.3, P.113).

Issue seven: WIP imbalance detection and calibration

The reason why we need to correct the WIP imbalance is because WIP imbalance is time dependent and occurs anytime and anywhere in the fab, despite the effort we spend to achieve WIP balance. Therefore, WIP imbalance detection and calibration is highly recommended since the small WIP imbalance can self-enlarge to become a serious problem if it cannot be stopped in time. Section 4.2 addressed these two aspects. Firstly, the WIP evolution cure of the Matrix Table was a good example to show time dependent WIP imbalance (Figure 4.2.5, P.161). Secondly, the proposed WIP detection and calibration approach made it clear that it could correct WIP imbalance (Figure 4.2.7, P.166). The performance was as good as the one using target WIP to calibrate in literature (Figure 4.2.9, P.170; Table 4.2.2, P.170).

- Detecting throughput decrease as a symptom of WIP imbalance, we utilized the throughput decrease as trigger event to monitor WIP imbalance. Moreover, in order to abandon target WIP, we proposed to apply WIP position analysis combining with other dispatching methods to correct WIP imbalance. This novel WIP imbalance detection and calibration approach without the need of target WIP is the **third contribution** of this dissertation.

We also underlined that it can be embodied into the existing control system to enhance the intelligence of automatic manufacturing, as the simulation results demonstrated that it is able to calibrate WIP imbalance for various typical dispatching rules under different loadings (Figure 4.2.8, P.168).

With regard to the evaluation of calibration performance, the simulation results suggested that it depends on how often we monitor the WIP imbalance occurrence and whether we take action to calibrate. In this study, we monitored the fab status every *X* hours. The interval *X* achieving the best calibration performance showed that neither too short nor too long (Figure 4.2.7, P.166; Table 4.2.1, P.167). On one hand, if the interval is short, we have more chances to capture the imbalance phenomenon. However, the calibration may take effect for some fake imbalance cases that can be self-calibrated. On the other hand, if the interval is too long, the opportunities to capture and calibrate are missed, which leads to enlarged WIP imbalance that may be too late to be corrected. Thus, in practice, we suggest that the WIP imbalance phenomenon has to be monitored anytime. However, whether taking action to the calibration procedure is another story. A full consideration has to be made according to the fab situation to figure out the positive or negative impacts brought by calibration, which requires more elaborated work in addition to simulation experiments.

Issue eight: WIP balance combining with due date control

Chapter 5 included three sections to present different ways to manage WIP balance and due date control when low cycle time as well as good on-time delivery performances are desired simultaneously.

We were aware that unless WIP balance could achieve significant cycle time reduction which forces all lots to finish before due date despite of the expanded cycle time distribution (Figure 5.1.1, P.179). Otherwise we have to face the potential problem that some lots will have excessive tardiness. Section 5.1 that extended from Section 4.2 by introducing due date performance showed an example that we could solve this problem via reducing cycle time variance for WIP balance. The simulation results suggested that the priority-based two-layer hierarchical dispatching scheme incorporating WIP balance and due date control are promising.

- Indeed, as WIP balance and due date control only concentrate on their own target, this two-layer hierarchical dispatching succeeds in taking both targets into account. This is a major benefit in comparison to applying WIP balance or due date control individually (Figure 5.1.2, P.185; Figures 5.1.3, 5.1.4, P.187; Figures 5.1.5, 5.1.6, P.188), which is viewed as the **fourth contribution** of this dissertation.

Differentiating from the two-layer hierarchical dispatching, Section 5.2 presented another way to deal with WIP balance and due date control. The LWNQ rule (WIP balance) and ODD rule (due date control) took effect in parallel inside a proposed composite rule, and their contribution to the composite rule under different due date flow factors was determined by the scaling parameters. This composite rule might be questioned about practical applicability since it is not easy to determine the scaling parameters (Table 5.2.1, P.197; Table 5.2.3, P.199).

- However, the main perspective of this study is to point out that we have to reduce the due date control effect under tight due dates by WIP balance, otherwise overemphasizing due date control causes WIP imbalance (Figures 5.2.1, 5.2.2, P.201, Figures 5.2.3, 5.2.4, P.202; Figure 5.2.5, P.203). Furthermore, this proposed composite rule with multiple objectives does not exist in literature, thus, we consider it as the **fifth contribution** of this dissertation.

We discussed the compensating effects of WIP balance and due date control and the possibility whether they can coexist. Obviously, as the wafer fab is full of variations, it is hard to say whether WIP balance takes precedence over due date control, and vice versa. It depends on the real situation and objectives. We also felt that, although there are various WIP balance and due date control approaches including those we proposed, they are all based on the relation depicted in Figures 1.3.2 (P.13), 4.4.1 (P.101) and 5.1.1 (P.179). To further understand the interaction between WIP balance and due date control, Section 5.3 presented a case study which showed a typical problem in a customer oriented wafer fab. For those wafer fabs applying due date rules as operational control, we noticed that WIP balance sacrifices the cycle time of products with tight due date. In this case, if the tight due date products are expected to leave the fab as fast as possible, at least the tardiness performance has to be minimized. The only way to achieve this is to apply due date control overruling WIP balance to accelerate tight due date products. This is particularly important to the products like hot lots, samples, engineering lots and so on. For a make-to-stock wafer fab (since due date is not a big concern), assigning higher priority to those products is the way to speed them up, however, it will cause irregularity of the WIP flow. This situation can be changed if those products are assigned due dates and accelerated by due date control, because they can be processed at a better pace which gives larger room to adjust the dispatching objective, for instance, by considering the negative effect on the loose due date products.

6.2 Recommendations for Future Research

Several aspects could be considered as potential research directions to make the proposed WIP balance and due date control more applicable and feasible in the operational control in wafer fabs. For example:

- Incorporating setups and batches into WIP balance approaches

As hot topics, various setup and batch strategies have been well studied from academic and industrial researchers. The MWVS could be extended as a global rule which is enhanced by various local setup and batch strategies. Actually, it is the trend of WIP balance in the future since setups and batches have a strong influence on cycle time reduction.

- Due date assignment rules

Section 3.4 only focuses on due date dispatching rules. Indeed, it could be extended by including due date assignment rules, to make it more comprehensive. The conclusion of the performance of due date oriented rules could be more accurate, reliable and convincible via comparison among due date dispatching rules working under different due date assignment rules.

- Replacement of ODD by FF

Section 5.1 might be extended by replacing ODD by FF for MWVS and MIVS rules. The purpose is to find out whether FF rule is more suitable for the fab running with low mix and high volume products (loose target due dates), and ODD rule fits for the fab running with high mix and low volume products (tight and medium target due dates).

- Local WIP balance vs. global due date control and local due date control vs. global WIP balance

The preliminary study in Section 5.3 could go further to examine the interaction between WIP balance and due date control in more detail. To better understand the trade-off between them, we should pay attention to how every local WIP balance decision affects global due date control target, and the opposite case that how local due date control decision influences global WIP balance target. This ‘local vs. global’ experiment will exactly tell us whether it is worth to make WIP balance or due date control decisions.

Glossary

Term	Description
ASIC	Application specific integrated circuits
ATC	Apparent tardiness cost
ATCS	Apparent tardiness cost with setups
AWDL	Acceptable WIP deviation level
A/OPN	Allowance per operation
BMW	Balanced machine workload
BPNN	Back-propagation neural network
CONLOAD	Constant load
CONWIP	Constant WIP
CR	Critical ratio
CT	Cycle time
DDFF	Due date flow factor
EDD	Earliest due date
FF	Flow factor
FIFO	First in first out
FX	Factory explorer
IFD	Infineon dispatching
JIT	Just-in-time
KPI	Key performance indicator
LAWDL	Lower-limit acceptable WIP deviation level
LB	Line balance
LOST	Least operation slack time
LST	Least slack time
LWNQ	Least work at next queue
MAPE	Mean-absolute-percentage-error
MES	Manufacturing execution system
MDD	Modified due date
MIMAC	Measurement and improvement of manufacturing capacities
MIVS	Minimum inventory variability scheduling
MOD	Modified operation due date

MRP	Material requirements planning
MTBF	Mean time between failures
MTTR	Mean time to repair
MWVS	Minimum workload variability scheduling
MWVS_1	Minimum workload variability scheduling with target WIP
MWVS_2	One-step-ahead and one-step-back minimum workload variability scheduling without target WIP
OCR	Operation critical ratio
ODD	Operation due date
OTD	On-time delivery
PPP	Push-pull point
PW	Product weight
Q+Acc.CT	Longest queue time plus accumulated cycle time
RMSE	Root-mean-squared-error
SA	Starvation avoidance
SPT	Shortest processing time
S/OPN	Slack per operation
TOC	Theory of Constraint
UAWDL	Upper-limit acceptable WIP deviation level
Wafer Fabs	Wafer fabrication facilities
WI	Workload indicator
WI(1): L	The first approach which only considers the workload of local work-center
WI(2): U+L+D	The second approach which extends the first approach by considering the workload of upstream and downstream work-centers
WI(3): Mod(U+L+D)	The third approach which incorporates the modification factor into the second approach
WI(4): Mod(U+L+D)+ODD	The fourth approach which combines ODD rule to the third approach to override the WI
WIP	Work-in-process
WIPCT	WIP control table
WR	Workload regulation

References

Atherton LF, Atherton RW (1996) Wafer fabrication: factory performance and analysis. Springer 1996 edition

Baker KR, Bertrand JWM (1981) A comparison of due-date selection rules. **AIIE Transactions** 13:123-13

Baker KR, Bertrand JWM (1981) An investigation of due date assignment rules with constrained tightness. **Journal of Operations Management** 1:109–120

Baker KR, Kanet JJ (1983) Job shop scheduling with modified due dates. **Journal of Operations Management** 4:11-22

Baker KR (1984) Sequencing rules and due date assignments in a job shop. **Management Science** 30:1093-1104

Baker KR, Trietsch D (2009) The principle of sequencing and scheduling. Wiley, 1 Edition

Bertrand JWM (1983) The use of work load information to control job lateness in controlled and uncontrolled release production systems. **Journal of Operational Management** 3:79-92

Bonvik AM, Couch CE, Gershwin SB (1997) A comparison of production-line control mechanisms. **International Journal of Production Research** 35:789-804

Burman DY, Gurrola-Gal FJ, Nozari A, Sathaye S, Sitarik JP (1986) Performance analysis techniques for IC manufacturing lines. **AT&T Technical Journal** 65:46-57

Chambers M, Mount-Campbell CA (2002) Process optimization via neural network metamodeling. **International Journal of Production Economics** 79:93-100

Chung J, Jang J (2009) A WIP balance procedure for throughput maximization in semiconductor fabrication. **IEEE Transactions on Semiconductor Manufacturing** 22:381-390

Collins DW, Palmeri V (1997) An analysis of the “k-step ahead” minimum inventory variability policy using sematech semiconductor manufacturing data in a discrete-event simulation model. **6th International Conference on Emerging Technologies and Factory Automation Proceedings** 520-527

Connors DP, Feigin GE, Yao DD (1996) A queuing network model for semiconductor manufacturing. **IEEE Transactions on Semiconductor Manufacturing** 9:412-427

Dabbas RM, Fowler JW (2003) A new scheduling approach using combined dispatching criteria in wafer fab. **IEEE Transactions on Semiconductor Manufacturing** 16:501-510

Demeester L, Tang CS (1996) Reducing cycle time at an IBM wafer fabrication facility. **Interfaces** 26:34-39

Domaschke J, Robinson J, Leibl F (1998) Effective implementation of cycle time reduction strategies for semiconductor back-end manufacturing. **In Proceedings of the 1998 Winter Simulation Conference** pp.985-992

Elvers DA (1973) Job shop dispatching rules using various delivery date setting criteria. **Production and Inventory Management** 4:62-70

Fowler JW, Robinson J (1995) Measurement and improvement of manufacturing capacities (MIMAC): final report.” Technical Report 95062861A-TR, SEMATECH, Austin

Fowler JW, Hogg GL, Mason SJ (2002) Workload control in the semiconductor industry. **Prod uction Planning & Control** 13:568-578

Fredendall LD, Ojha D, Patterson JW (2010) Concerning the theory of workload control. **European Journal of Operational Research** 201:99-111

Glassey CR, Resende MGC (1988) Closed-Loop job release control for VLSI circuit manufacturing. **IEEE Transactions on Semiconductor Manufacturing** 1: 36-46

Glassey CR, Weng W (1991) Dynamic batching heuristic for simultaneous processing. **IEEE Transactions on Semiconductor Manufacturing** 4:77-82

Goldratt EM. (1984) The goal. Great Barrington, MA

Gupta AK, Ganesan VK, Sivakumar AI (2009) Cycle time variance minimization in dynamic scheduling of single machine systems. **The International Journal of Advanced Manufacturing Technology** 42:544–552

Guo RS, Chiang DM, Pai FY (2007) A WIP-based exception-management model for integrated circuit back-end production processes. **International Journal of Advanced Manufacturing Technology** 33:1263-1274

Ham M, Fowler JW (2007) Balanced machine workload dispatching scheme for wafer fab. **Advanced Semiconductor Manufacturing Conference** pp.390-395

Hsieh BW, Chang SC, Chen CH, Chang MC (2003) Efficient composition of good enough dispatching policies for semiconductor manufacturing. **IEEE International Symposium Semiconductor Manufacturing** pp. 67-70

Huang CL, Huang YH, Chang TY, Chang SH, Chung CH, Huang DT, Li RK (1999) The construction of production performance prediction system for semiconductor manufacturing with artificial neural network. **International Journal of Production Research** 37:1387-1402

Ho CM, Chen TC, Heih P, Chu C, Houn E, Su KJ, Wang PM, Yew WC, Sun SW (2000) Simultaneous cycle-time reduction and ouput enhancement in a fully loaded foundry wafer fab. **In Proceeding of ISSM Conference 2010** pp.63-66

Hopp WJ, Spearman ML (2011) Factory physics. 3 Edition, Waveland Pr Inc

http://www.wwk.com/ [www1, 28.02.2014]

http://code.google.com/p/encog-java/ [www2, 28.02.2014]

Kalisch S, Ringel R, Weigang J (2008) Managing wip and cycle time with the help of loop control. **In Proceedings of the 2008 Winter Simulation Conference** pp. 2298-2304

Keskinocak P, Tayur S (2004) Due date management policies. in Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era, **International Series in Operations Research and Management Science** pp. 485–553

Kim S, Lee YH, Yang T (2008) Robust production control policies considering WIP balance and setup time in a semiconductor fabrication line. **International**

Journal of Advanced Manufacturing Technology 39:333-343

Kumar PR (1993) Re-entrant lines. **Queuing Systems Theory and Applications** 13:87-110

Kuo CJ, Liu CM, Chi CY (2008) Standard WIP determination and WIP balance control with time constraints in semiconductor wafer fabrication. **Journal of Quality** 15:409-423

Leachman RC, Kang J, Lin V (2002) SLIM: short cycle time and low inventory in manufacturing at Samsung electronics. **Interfaces** 32:61-77

Lee YH, Bhaskaran K, Pinedo M (1997) A heuristic to minimize the total weighted tardiness with sequence dependent setups. **IIE Transactions** 29**:**45-52

Lee YH, Kim T (2002) Manufacturing cycle time reduction using balance control in the semiconductor fabrication line. **Production Planning and Control** 13:529-540

Lee YH, Park J, Kim S (2001) Experimental study on input and bottleneck scheduling for a semiconductor fabrication line. **IIE Transactions** 34:179-190

Li S (1991) Equi-variability graph approach for modeling of manufacturing systems. Invited paper, **in Proceeding 29th Annu. Allerton Conference**, Allerton, IL

Li S (1993) Innovative method in planning and scheduling in semiconductor manufacturing. Invited paper, **in Proceeding Semiconductor Manufacturing Technology Workshop**, cosponsored by National Taiwan University and Taiwan Industrial Technology Research Institute

Li S, Tang T, Collins DW (1996) Minimum inventory variability schedule with applications in semiconductor fabrication. **IEEE Transactions on Semiconductor Manufacturing** 9:1-5

Lin YH, Lee CE (2001) A total standard WIP estimation method for wafer fabrication. **European Journal of Operation Research** 131:78-94

Little JDC (1992) Are there ‘Laws’ of manufacturing. Manufacturing Systems: Foundations of World-Class Practice pp.180-188

Lu SC, Ramaswamy D, Kumar PR (1994) Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. **IEEE Transactions on Semiconductor Manufacturing** 7:374-388

Marek RP, Elkins DA, Smith DR (2001) Understanding the fundamentals of kanban and conwip pull systems using simulation. **In Proceedings of the 2001 Winter Simulation Conference** pp. 921-929

McClellan M (2001) Introduction to manufacturing execution systems. **MES Conference and Exposition**, Baltimore, Maryland, US.

Monden, Yasuhiro (1981) What makes the toyota production system really tick. **Industrial Engineering** 13:36-46

Muhlemann AP, Lockett AG, Farn CI (1982) Job shop scheduling heuristics and frequency of scheduling. **International Journal of Production Research** 20:227–241

Narendra KS (1996) Neural networks for control: theory and practice. **Proceedings of IEEE** 84:1385-1406

Nemoto K, Akcali E, Uzsoy R (1996) Quantifying the benefits of cycle time reduction in semiconductor wafer fabrication. **Electronics Manufacturing Technology Symposium** pp.130-136

Pai FY (2004) WIP management model for semiconductor back-end manufacturing. **Journal of American Academy of Business** 5:357-363

Panwalker SS, Iskandar WW (1977) A survey of scheduling rules. **Operations Research** 1:45-61

Perdaen D, Armbruster D, Kempf K, Lefeber E (2008) Controlling a reentrant manufacturing line via the push-pull point. **International Journal of Production Research** 46: 16 4521-4536

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

Robinson JK (2002) Understanding and improving wafer fab cycle times. Technical Report, Fab Time Inc

Robinson J, Fowler JW, Bard J (1995) The use of upstream and downstream information in scheduling semiconductor batch operations. Taylor and Francis Ltd, 1995

Rose O (1999) Conload - a new lot release rule for semiconductor wafer fabs. **In Proceedings of the 1999 Winter Simulation Conference** pp. 850-855

Rose O (2002) Some issues of the critical ratio dispatch rule. **In Proceedings of the 2002 Winter Simulation Conference** pp. 1401-1405

Rose O (2003) Comparison of due-date oriented dispatch rules in semiconductor manufacturing. **In Proceedings of the 2003 Industrial Engineering Research Conference** pp.18-20

Spearman ML, Woodruff D, Hopp W (1990) CONWIP: a pull alternative to kanban. **International Journal of Production Research** 28:879-894

Spearman ML, Zazanis MA (1992) Push and pull production systems: issues and comparisons. **Operations Research** 40:521-532

Strum R, Frauenhoffer F, Dorner J, Kirschenhofer O, Reisinger T (1999) Advance WIP control for make-to-order wafer fabrication. **Advanced Semiconductor Manufacturing Conference** pp.31-36

Sze SM, Lee MK (1985) Semiconductor devices: physics and technology. Wiley 3 Edition

Tang T (1993) Simulation model for minimum inventory variance policy practiced in semiconductor manufacturing plants. Master of Technology Research Project, Department of Manufacturing and Industrial Technology, Arizona State University

Toba H (2000) Segment-based approach for real-time reactive rescheduling for automatic manufacturing control. **IEEE Transactions on Semiconductor Manufacturing** 13:264-272

Tu YM, Chao YH, Chang SH, You HC (2005) Model to determine the backup capacity of a wafer foundry. **International Journal of Production Research** 43:339-359

Uzsoy R, Lee CY, Martin-Vega LA (1993) A review of production planning and scheduling models in the semiconductor industry part I: system characteristics performance evaluation and production planning. **IIE Transactions** 24:47-60

Vargas-Vilamil FD, Rivera DE, Kempf KG (2003) A hierarchical approach to production control of reentrant semiconductor manufacturing lines. **IEEE Transaction on Control Systems Technology** 11:578-587

Ventura J, Weng MX (1972) Minimizing single machine completion time variance. **Manage Science** 41:1448–1455

Vepsalainen APJ, Morton TE (1987) Priority rules for job shops with weighted tardiness costs. **Management Science** 33:1035-1047

Waikar AM, Sarker BR, Lal AM (1995) A comparative study of some priority dispatching rules under different shop loads. **Production Planning & Control** 6:301-310

Wein LM (1988) Scheduling semiconductor wafer fabrication. **IEEE Transactions on Semiconductor Manufacturing** 1:115-130

Wein LM (1991) Due-date setting and priority sequencing in a multiclass M/G/1queue. **Management Science** 37:834-850

Whitt W (1993) Approximation for the G1/G/m queue. **Production and Operations Management** 2:114-161

Wight OW (1970) Input/output control: a real handle on lead time. **Production & Inventory Management Journal** 11:9-31

Yeh TM, Pai FY, Tsou CS (2008) The construction of a real-time WIP exception monitoring system for semiconductor industry. **IEEE 4th International Conference on Wireless Communications, Networking and Mobile Computing** pp.1-4

Yu CA, Huang HP (2002) On-line learning delivery decision support system for high product mixed semiconductor foundry. **IEEE Transactions on Semiconductor Manufacturing** 15:274-278

Zhou Z, Rose O (2010) A pull/push concept for tool group workload balance in a wafer fab. **In Proceedings of the 2010 Winter Simulation Conference** pp. 2512-2516

Appendix

	Products								
Work-centers	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
10121_DNS-PUD	0.13	0.13	0.10	0.14	0.10	0.11	0.10	0.09	0.10
10123_DNS-3	0.13	0.13	0.10	0.13	0.09	0.11	0.11	0.10	0.11
10130_DNS-HP	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.50
10151_DNS-1	0.00	0.20	0.20	0.00	0.20	0.00	0.20	0.20	0.00
10152_DNS-2	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00
10330_DUV	0.25	0.25	0.00	0.00	0.00	0.25	0.00	0.00	0.25
10341_CLOVEN	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00
11021_ASM_A1_A3_G1	0.12	0.00	0.16	0.24	0.16	0.12	0.08	0.08	0.04
11022_ASM_A2	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.50
11024_ASM_A4_G3_G4	0.14	0.09	0.05	0.05	0.09	0.14	0.18	0.18	0.09
11025_ASM_B1_H2	0.00	0.20	0.60	0.00	0.00	0.00	0.00	0.00	0.20
11026_ASM_B2	0.22	0.09	0.04	0.17	0.17	0.09	0.09	0.09	0.04
11027_ASM_B3_B4_D4	0.13	0.06	0.06	0.13	0.13	0.13	0.13	0.13	0.13
11029_ASM_C1_D1	0.14	0.07	0.07	0.14	0.14	0.14	0.14	0.14	0.00
11030_ASM_C2_H1	0.17	0.11	0.06	0.22	0.11	0.11	0.11	0.11	0.00
11031_ASM_C3	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.00	0.50
11032_ASM_C4	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.33
11122_ASM_D2	0.50	0.00	0.25	0.00	0.00	0.25	0.00	0.00	0.00
11125_ASM_E1_E2_H4	0.13	0.17	0.09	0.13	0.09	0.13	0.09	0.09	0.09
11127_ASM_E3_G2_H3	0.18	0.12	0.06	0.12	0.12	0.12	0.12	0.12	0.06
11128_AMS_E4	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00
11129_ASM_F1_F2	0.11	0.00	0.22	0.22	0.11	0.11	0.11	0.11	0.00
11132_ASM_F4_D3	0.14	0.29	0.00	0.14	0.14	0.14	0.00	0.00	0.14
11227_ASM_H3	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11524_MAX1+2_AL-TEMP	0.16	0.16	0.05	0.11	0.11	0.16	0.05	0.05	0.16
11732_AST_2000	0.00	0.40	0.00	0.00	0.00	0.00	0.20	0.20	0.20
11822_WJ_999	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
12021_AUTO-CL_undot	0.14	0.09	0.09	0.15	0.12	0.09	0.11	0.11	0.09
12022_AUTO-CL_dot	0.14	0.22	0.07	0.11	0.09	0.09	0.08	0.08	0.10
12031_FSI_S1+S2	0.09	0.14	0.14	0.18	0.00	0.09	0.14	0.14	0.09
12035_FSI_A1	0.00	0.00	0.67	0.00	0.33	0.00	0.00	0.00	0.00
12040_FINESONIC	0.25	0.00	0.00	0.25	0.25	0.25	0.00	0.00	0.00
12131_RINS-DRY_1	0.08	0.16	0.16	0.08	0.08	0.08	0.14	0.10	0.12
12221_HF-DIP-5_B	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12222_NF-1_D	0.00	0.00	0.33	0.67	0.00	0.00	0.00	0.00	0.00
12223_HF-DIP-4_C	0.13	0.19	0.00	0.13	0.13	0.13	0.13	0.13	0.06

12224_HF-DIP-3_N	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12225_MOD.POLY_A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
12226_NF-2	0.16	0.11	0.13	0.13	0.11	0.08	0.08	0.08	0.13
12228_HF-DIP-1_M	0.00	0.20	0.00	0.00	0.00	0.00	0.40	0.40	0.00
12229_NF-3_E	0.14	0.00	0.14	0.21	0.14	0.21	0.07	0.07	0.00
12230_NF-VIA	0.20	0.00	0.00	0.20	0.20	0.20	0.00	0.00	0.20
12232_NF-5_G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
12233_NF-1INSI_H	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.33
12240_NIT-ETCH	0.22	0.11	0.00	0.11	0.11	0.11	0.11	0.11	0.11
12255_HF-DIP-2_L	0.00	0.00	0.00	0.80	0.20	0.00	0.00	0.00	0.00
12331_RST100_1+2	0.13	0.13	0.13	0.13	0.08	0.13	0.13	0.13	0.04
12531_SH	0.14	0.11	0.11	0.13	0.10	0.10	0.09	0.09	0.13
12552_S106_MET	0.04	0.26	0.00	0.04	0.04	0.04	0.30	0.22	0.04
12553_POSI_GP	0.18	0.14	0.14	0.18	0.09	0.09	0.05	0.05	0.09
12820_F1	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
12825_F2	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
12831_DISCO	0.13	0.13	0.13	0.00	0.13	0.13	0.13	0.13	0.13
12840_NMP-B	0.14	0.14	0.00	0.14	0.00	0.14	0.14	0.14	0.14
13021_AME_1+3_AlSiCu	0.17	0.00	0.17	0.17	0.17	0.17	0.00	0.00	0.17
13024_AME_4+5+7+8	0.11	0.16	0.09	0.13	0.07	0.07	0.16	0.13	0.07
13121_LAM_490B_1+2	0.25	0.00	0.00	0.13	0.13	0.13	0.13	0.13	0.13
13124_LAM_4400_Poly	0.11	0.16	0.05	0.21	0.05	0.11	0.11	0.11	0.11
13128_LAM_4500_Oxid	0.04	0.13	0.00	0.13	0.13	0.13	0.17	0.13	0.13
13215_P5E_ALU1+2	0.00	0.38	0.00	0.00	0.00	0.00	0.38	0.25	0.00
13226_P5000-E_6B_P	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.25	0.25
13521_CDE_8_1+2	0.14	0.07	0.14	0.14	0.14	0.14	0.04	0.04	0.14
13621_IPC_3200	0.12	0.11	0.11	0.10	0.11	0.10	0.13	0.12	0.10
14021_AMC-EPI_1+2	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.33
14131_AMT-PREC_1+3	0.09	0.18	0.09	0.09	0.09	0.09	0.18	0.09	0.09
14134_A4A-WOLF_1+2	0.00	0.38	0.00	0.00	0.00	0.00	0.38	0.25	0.00
14137_AMT-PREC_7	0.11	0.16	0.05	0.11	0.11	0.11	0.16	0.11	0.11
14521_MS6200_ET1	0.09	0.09	0.00	0.09	0.09	0.09	0.18	0.18	0.18
14531_ULVAC	0.17	0.00	0.17	0.17	0.17	0.17	0.00	0.00	0.17
14551_ENDURA_1	0.06	0.24	0.06	0.06	0.06	0.06	0.24	0.15	0.06
14821_DNS-SOG_1	0.10	0.20	0.00	0.10	0.10	0.10	0.20	0.10	0.10
15121_LTS_3	0.13	0.11	0.09	0.12	0.10	0.11	0.12	0.09	0.10
15122_LTS_1	0.13	0.13	0.12	0.13	0.10	0.10	0.11	0.10	0.10
15123_LTS_2	0.13	0.14	0.07	0.12	0.11	0.11	0.12	0.11	0.10
15131_LZZZZ	0.15	0.10	0.10	0.15	0.09	0.11	0.08	0.08	0.11
15322_ASTEP200_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
15326_RUD_AUTOEL	0.16	0.16	0.12	0.17	0.09	0.12	0.07	0.07	0.05

15327_AUTO_EL3	0.05	0.68	0.05	0.05	0.00	0.05	0.05	0.05	0.05
15421_SURF_2	0.33	0.00	0.00	0.17	0.22	0.28	0.00	0.00	0.00
15523_OMNI_RS_50	0.13	0.00	0.13	0.25	0.13	0.13	0.13	0.13	0.00
15627_HIT_S6000	0.15	0.19	0.07	0.12	0.09	0.11	0.12	0.09	0.07
15932_NICOLET	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.33
16121_IMP-HC_1+2	0.17	0.20	0.06	0.06	0.09	0.11	0.09	0.09	0.14
16221_IMP-MC_1+2	0.14	0.06	0.14	0.16	0.08	0.12	0.14	0.14	0.04
17021_KEITH350	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
17041_KEITH450_+_425	0.08	0.08	0.00	0.08	0.08	0.08	0.25	0.25	0.08
17221_K-SMU236	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17421_HOTIN	0.33	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.33
19021_WAF-MA	0.08	0.17	0.08	0.08	0.08	0.08	0.17	0.17	0.08
19301_DAMAGE	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
20540_CAN_0.43_MII	0.19	0.00	0.17	0.20	0.14	0.16	0.00	0.00	0.14
20543_CAN_0.55_MIV	0.00	0.85	0.00	0.00	0.00	0.00	0.08	0.08	0.00
20550_CAN_0.52_i-line	0.00	0.19	0.00	0.00	0.00	0.00	0.44	0.38	0.00

Table 3.1.12: The Product Weight of each product for all work-centers in MIMAC6 model

Work-center	***Simulation Experiment (Based on FIFO, wafers)***			***Queuing Model*** ***(*** $WIPL_i^{QM}$ ***, wafers)***			***BPNN*** ***(*** $WIPL_i^{BPNN}$ ***, wafers)***		
	75% loading	***85% loading***	***95% loading***	***75% loading***	***85% loading***	***95% loading***	***75% loading***	***85% loading***	***95% loading***
10000_Virtual	0	0	0	0	0	0	0	0	0
10121_DNS-PUD	69.7	81	89.5	90.4	102.1	108.4	84.0	105.3	105.4
10123_DNS-3	51.7	50	68.4	69.9	79.2	84.6	65.0	81.7	82.3
10130_DNS-HP	0.5	0.5	0.6	0.5	0.6	0.7	0.5	0.6	0.6
10151_DNS-1	4.6	6.7	9.4	9.2	10.9	11.5	8.5	11.2	11.2
10152_DNS-2	1.7	2	3.3	2.9	3.8	4.8	2.7	3.9	4.4
10220_CONVAC	0	0	0	0	0	0	0	0	0
10330_DUV	6.8	8.2	8.2	7.9	9.7	13.0	7.3	9.9	12.6
10341_CLOVEN	15.6	21.1	33	25.5	37.5	54.5	23.7	38.7	52.9
11021_ASM_A1_A3_G1	63	65.3	83.8	23.2	26.3	78.5	21.5	27.1	76.3

11022_ASM_A2	6.6	8.1	9.9	2.4	3.1	3.6	2.2	3.2	3.5
11024_ASM_A4_G3_G4	65.4	92	100.1	30.4	34.2	104.8	28.2	35.3	101.9
11025_ASM_B1_H2	6.8	7.3	9	2.2	2.5	2.7	2.0	2.5	2.6
11026_ASM_B2	100.8	124	375.9	75.2	145.6	404.1	69.9	150.1	383.1
11027_ASM_B3_B4_D4	87.3	110	118.3	30.8	34.0	120.5	28.6	35	112.2
11029_ASM_C1_D1	80.1	104.7	110.7	31.0	37.7	113.4	28.8	38.8	110.3
11030_ASM_C2_H1	42.8	53.3	60.9	17.8	20.8	23.0	16.5	21.4	22.4
11031_ASM_C3	8.4	8.8	12.1	2.5	3.2	3.4	2.3	3.2	3.3
11032_ASM_C4	9.3	10.6	13	3.1	3.4	4.0	2.9	3.5	3.9
11122_ASM_D2	9.6	11	15.4	3.7	4.1	4.5	3.4	4.2	4.3
11125_ASM_E1_E2_H4	48.9	56	83	31.5	36.7	97.4	29.3	37.8	94.7
11127_ASM_E3_G2_H3	36.6	55.3	62.9	18.8	21.1	22.5	17.4	21.7	21.8
11128_AMS_E4	2.7	4	4	2.7	2.7	3.1	2.9	2.8	3.1
11129_ASM_F1_F2	25	31.2	37.6	23.6	27.8	29.0	21.9	28.6	28.2
11132_ASM_F4_D3	17.2	21.3	24.9	8.0	9.1	9.5	7.4	9.4	9.2
11227_ASM_H3	1.4	1.6	1.8	1.4	1.4	1.5	1.2	1.4	1.4
11524_MAX1+2_AL-TEMP	18.8	25.3	26	13.9	15.5	17.0	12.	14.5	16.5
11732_AST_2000	5	7.2	4.5	5.0	6.9	6.5	4.6	5.8	6.2
11822_WJ_999	28.7	35.1	32.3	29.8	46.4	48.3	27.7	44.9	46.9
12021_AUTO-CL_undot	38.2	58.9	72.8	32.7	36.9	39.2	30.4	36.1	38.1
12022_AUTO-CL_dot	30.6	61.7	75.5	34.9	39.5	42.1	32.5	40.7	40.9
12031_FSI_S1+S2	17	20.9	24.2	6.7	7.4	8.1	6.2	7.7	7.8
12035_FSI_A1	4.4	6	5	1.0	1.1	1.3	0.9	1.1	1.3
12040_FINESONIC	5.6	6.3	9.4	8.2	10.1	9.9	7.6	8.4	9.6
12131_RINS-DRY_1	25.8	89.6	64.9	50.9	59.8	67.7	47.3	61.6	65.8

12221_HF-DIP-5_B	0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.1
12222_NF-1_D	1	0.9	1.5	1.1	1.3	1.3	1.0	1.2	1.2
12223_HF-DIP-4_C	14.8	24.7	24.5	16.6	21.1	25.1	15.4	21.8	24.4
12224_HF-DIP-3_N	0.5	0.9	0.6	0.6	0.6	0.7	0.5	0.6	0.6
12225_MOD.POLY_A	0.7	0.6	1.7	0.6	0.7	0.9	0.5	0.7	0.9
12226_NF-2	17.2	20.5	29.6	23.1	27.0	30.0	21.4	27.8	29.1
12228_HF-DIP-1_M	1.5	1.5	2.2	1.8	2.2	2.3	1.6	2.3	2.2
12229_NF-3_E	14.4	14	23.3	15.5	17.5	17.8	14.4	16.2	17.2
12230_NF-VIA	3.2	5.2	6.4	4.3	4.8	6.3	3.9	4.9	6.1
12232_NF-5_G	0.6	1	1.4	0.7	0.9	1.2	0.6	1	1.1
12233_NF-1INSI_H	1.1	0.9	1.4	1.0	1.1	1.2	0.9	1.1	1.1
12240_NIT-ETCH	16.1	26.1	26.5	22.0	30.8	31.5	20.5	29.7	30.6
12255_HF-DIP-2_L	3.2	3.2	5.5	4.6	5.9	5.6	4.3	5.2	5.4
12331_RST100_1+2	74.7	144	200.4	85.5	153.6	233.5	79.5	158.4	217.1
12531_SH	38.6	52.3	82.4	41.6	51.3	98.3	38.7	52.9	95.6
12550_PROP_MET	0	0	0	0	0	0	0	0	0
12552_S106_MET	25.4	46.7	53.2	37.3	41.8	61.0	34.7	43.1	59.3
12553_POSI_GP	46.3	60.2	170.5	47.6	74.5	174.9	44.2	76.8	156.1
12820_F1	3.6	4.8	6.1	3.2	3.5	3.8	2.9	3.2	3.7
12825_F2	8.6	5.2	9.8	6.2	6.0	6.9	5.8	5.5	6.7
12831_DISCO	8.1	10.5	10.2	12.8	15.0	15.1	11.9	13.1	14.6
12840_NMP-B	7.8	9.3	10.9	15.6	18.7	20.2	14.4	19.2	19.6
13021_AME_1+3_AlSiCu	40.2	58.1	92.1	51.4	64.5	89.9	47.8	66.5	87.5
13024_AME_4+5+7+8	68.5	89.3	180.3	59.1	69.3	185.6	54.9	71.4	180.5
13121_LAM_490B_1+2	9.5	9.6	9.9	12.9	14.1	14.9	12.0	14.1	14.4
13124_LAM_4400_Poly	25.4	26.4	36.8	35.2	44.5	46.3	32.7	43.9	45.0

13128_LAM_4500_Oxid	21.4	24.7	29.1	31.4	36.6	38.4	29.2	36.7	37.3
13215_P5E_ALU1+2	5.1	8.2	9.8	16.2	18.2	20.5	15.0	18.7	19.9
13226_P5000-E_6B_P	3.9	5.3	6.5	6.5	8.1	8.6	6.0	8.3	8.4
13521_CDE_8_1+2	21	27.1	44.3	30.8	35.2	38.3	28.6	36.3	37.2
13621_IPC_3200	40.5	89.9	134.9	101.8	127.4	146.9	94.7	131.4	142.8
14021_AMC-EPI_1+2	13.2	15.8	17.2	24.7	27.9	30.2	22.9	28.7	29.4
14131_AMT-PREC_1+3	18.7	27.3	24.8	16.7	18.4	21.9	15.5	19	21.3
14134_A4A-WOLF_1+2	10.1	14.8	18.9	29.7	35.8	43.9	27.6	36.9	42.6
14137_AMT-PREC_7	34	49.4	47.4	25.9	29.2	31.1	24.8	30.1	30.3
14521_MS6200_ET1	17.4	18.1	24.3	26.6	31.6	34.7	24.7	32.6	33.7
14531_ULVAC	21.5	25.4	29.8	19.8	23.7	29.4	18.4	24.4	28.5
14551_ENDURA_1	36	46.8	87.2	51.9	60.6	79.9	48.3	62.5	77.6
14821_DNS-SOG_1	20.3	44.6	33.1	21.6	24.6	31.6	20.1	25.4	30.7
15121_LTS_3	108.1	172.1	231.4	100.1	163.1	287.7	93.0	168.2	279.8
15122_LTS_1	50.2	60	93.7	94.2	111.9	123.7	87.6	115.5	120.3
15123_LTS_2	48.6	65.8	74.2	70.2	81.9	88.7	65.3	84.5	86.2
15131_LZZZZ	40.1	66.4	120	56.2	81.0	91.5	52.3	83.6	88.9
15240_MICRO8_10_EPL1	0	0	0	0	0	0	0	0	0
15322_ASTEP200_2	0.5	0.3	0.9	0.3	0.3	0.5	0.3	0.3	0.4
15326_RUD_AUTOEL	6.2	7.2	8.2	7.9	8.9	9.4	7.4	8.2	9.1
15327_AUTO_EL3	2.4	3	2.7	4.2	4.6	4.9	3.9	3.8	4.7
15420_SURF_1	0	0	0	0	0	0	0	0	0
15421_SURF_2	2.3	2.5	3.9	2.8	3.2	3.4	2.6	3.3	3.3
15523_OMNI_RS_50	1.9	1.6	2.6	2.1	2.4	2.3	1.9	2.5	2.2

15627_HIT_S6000	45.2	58.3	97	69.8	89.6	102.0	64.9	92.4	99.2
15932_NICOLET	1.2	0.9	1	0.3	0.3	0.4	0.3	0.3	0.3
16121_IMP-HC_1+2	43.3	73.3	69	29.4	33.5	36.4	27.3	34.5	35.3
16221_IMP-MC_1+2	65.8	127.9	117.3	38.3	47.3	98.3	35.5	48.8	95.6
17000_	0	0	0	0	0	0	0	0	0
17021_KEITH350	1.3	1.8	1.9	2.3	2.6	3.7	2.1	2.6	3.5
17041_KEITH450_+_425	46.3	59.1	66.3	66.6	76.7	84.6	61.9	79.1	82.3
17221_K-SMU236	8.1	12.7	18.1	20.9	24.0	25.2	19.4	24.7	24.5
17421_HOTIN	40.2	67.9	73.5	18.3	34.7	60.0	3.1	3.9	58.3
17900_	0	0	0	0	0	0	0	0	0
19021_WAF-MA	5.6	4.2	4.4	4.6	5.1	5.4	4.2	5.3	5.2
19301_DAMAGE	0.5	0.6	0.9	0.7	0.8	0.9	0.7	1	0.
20540_CAN_0.43_MII	78.2	165.9	324.8	146	190	350	146	190	350
20543_CAN_0.55_MIV	3.2	3.8	4.2	4.8	5.3	5.6	4.4	5.5	5.4
20550_CAN_0.52_i-line	22.7	31.1	37.7	36.4	41.6	47.6	33.8	40.8	46.2
31001_LAM_TCP	0	0	0	0	0	0	0	0	0
31002_AME_5000_SACVD	0	0	0	0	0	0	0	0	0
31003_AMT_5500_TiSi	0	0	0	0	0	0	0	0	0
31004_WESTECH_CMP	0	0	0	0	0	0	0	0	0

Table 3.3.6: Average WIP level for each work-center from three different approaches

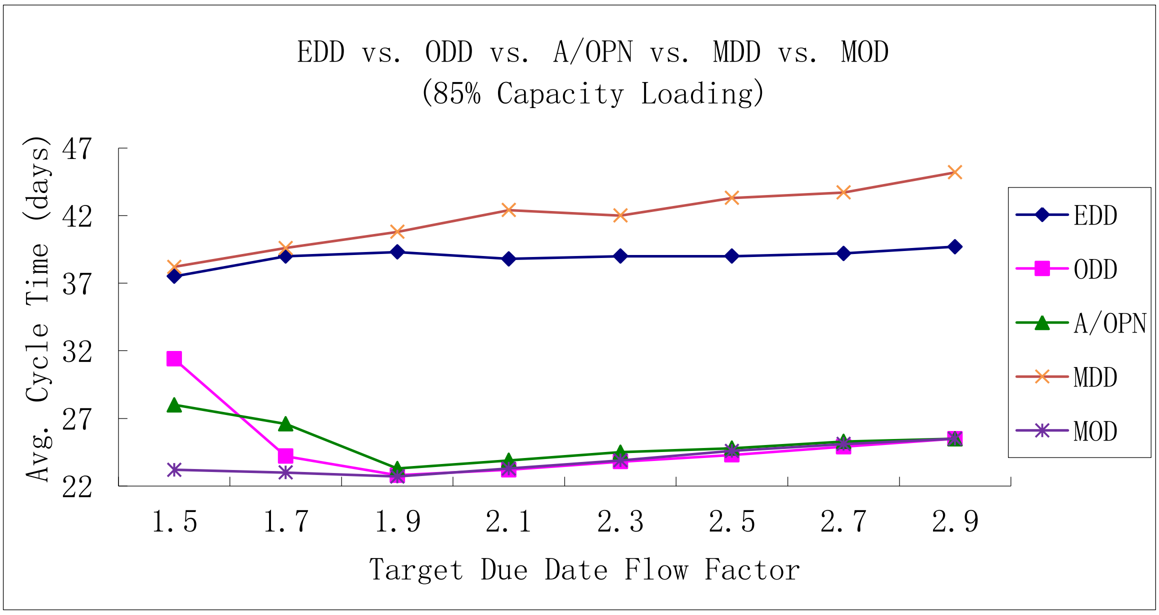


Figure 3.4.12: Average cycle time comparison of allowance-based rules vs. composite rules

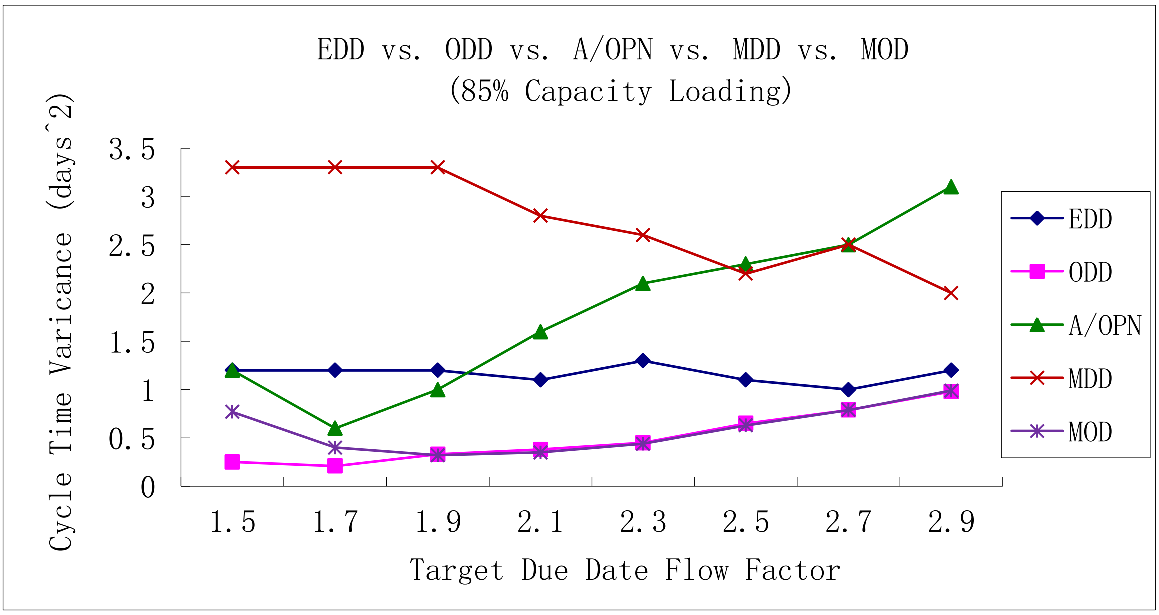


Figure 3.4.13: Cycle time variance comparison of allowance-based rules vs. composite rules

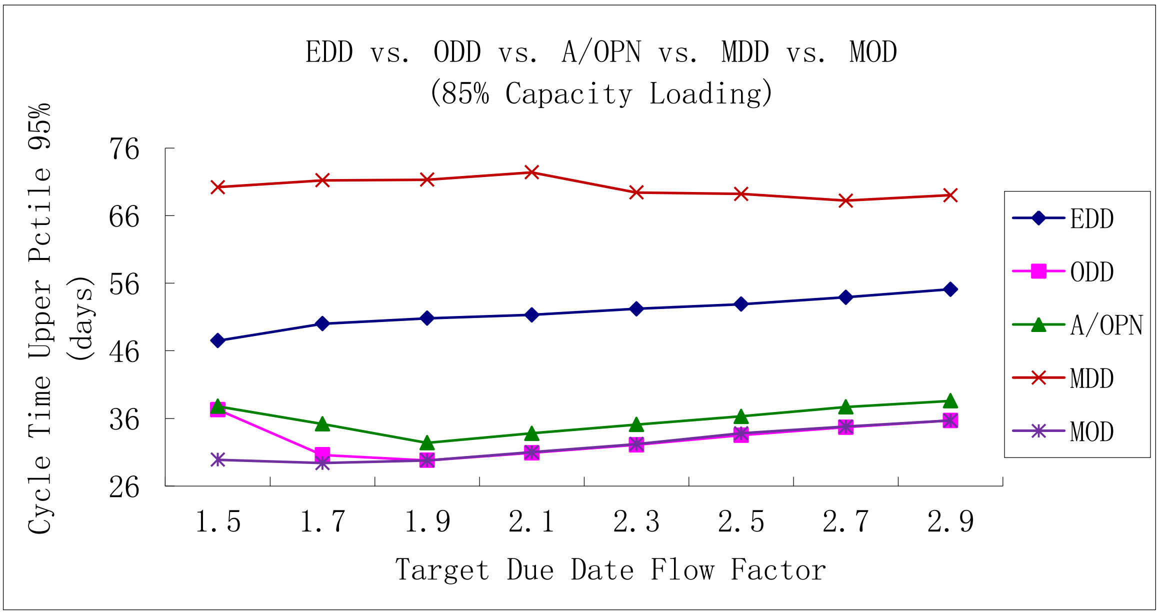


Figure 3.4.14: Cycle time upper 95% percentile comparison of allowance-based rules vs. composite rules

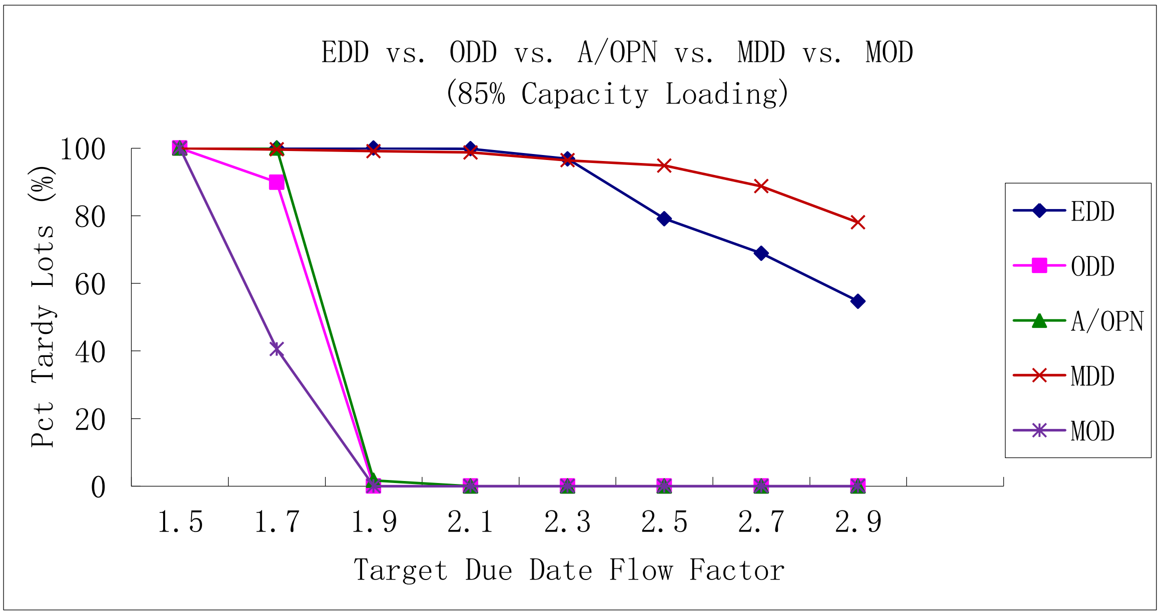


Figure 3.4.15: Percent tardy lots comparison of allowance-based rules vs. composite rules

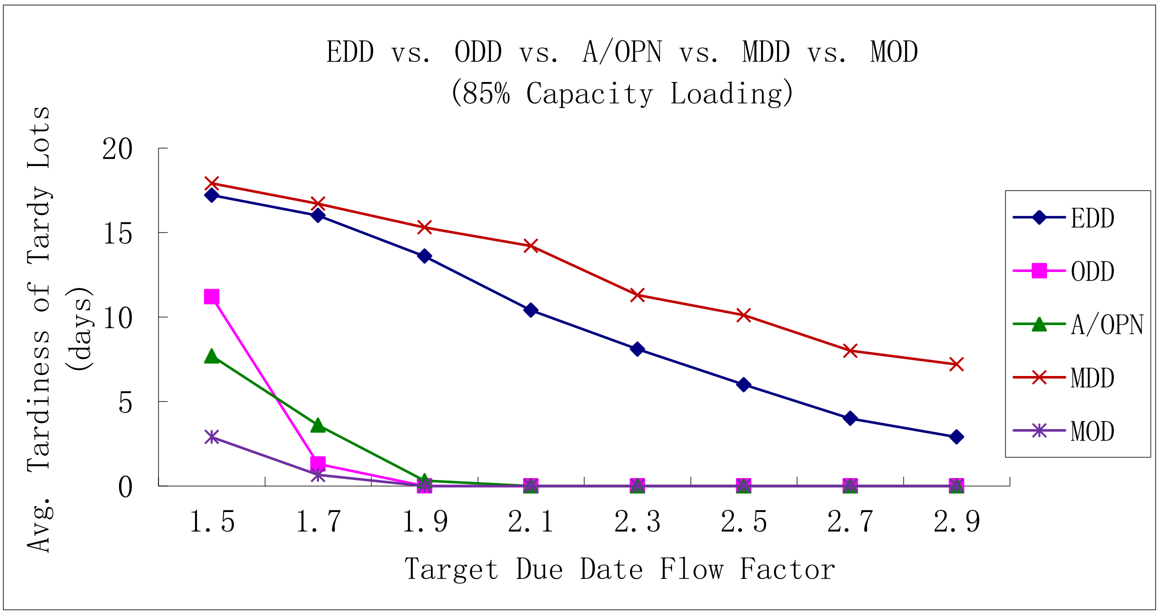


Figure 3.4.16: Average tardiness of tardy lots comparison of allowance-based rules vs. composite rules

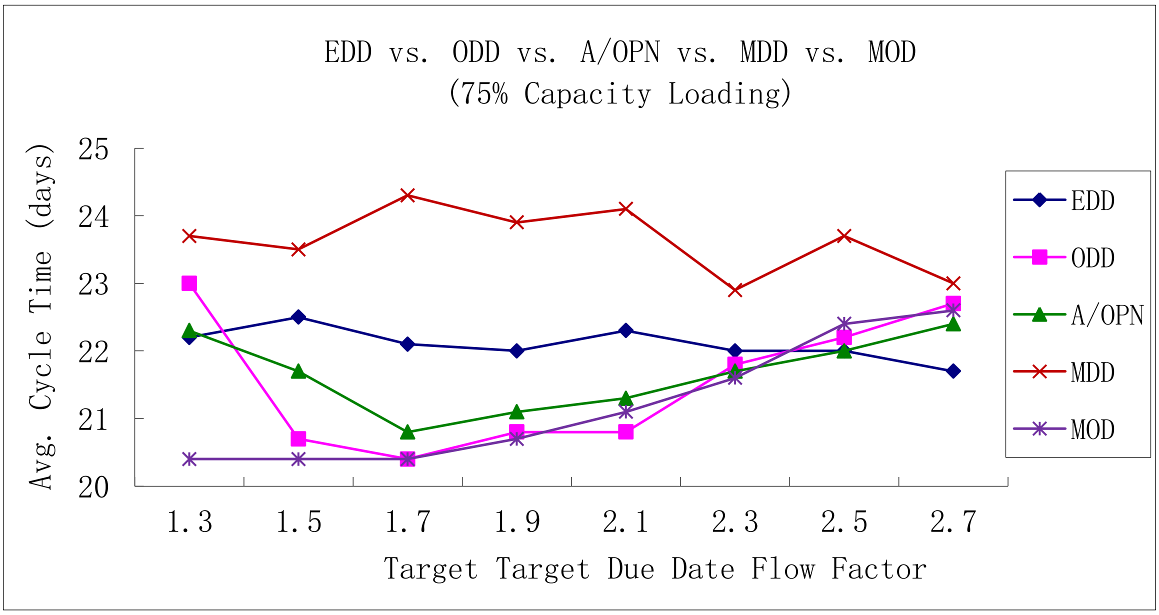


Figure 3.4.17: Average cycle time comparison of allowance-based rules vs. composite rules

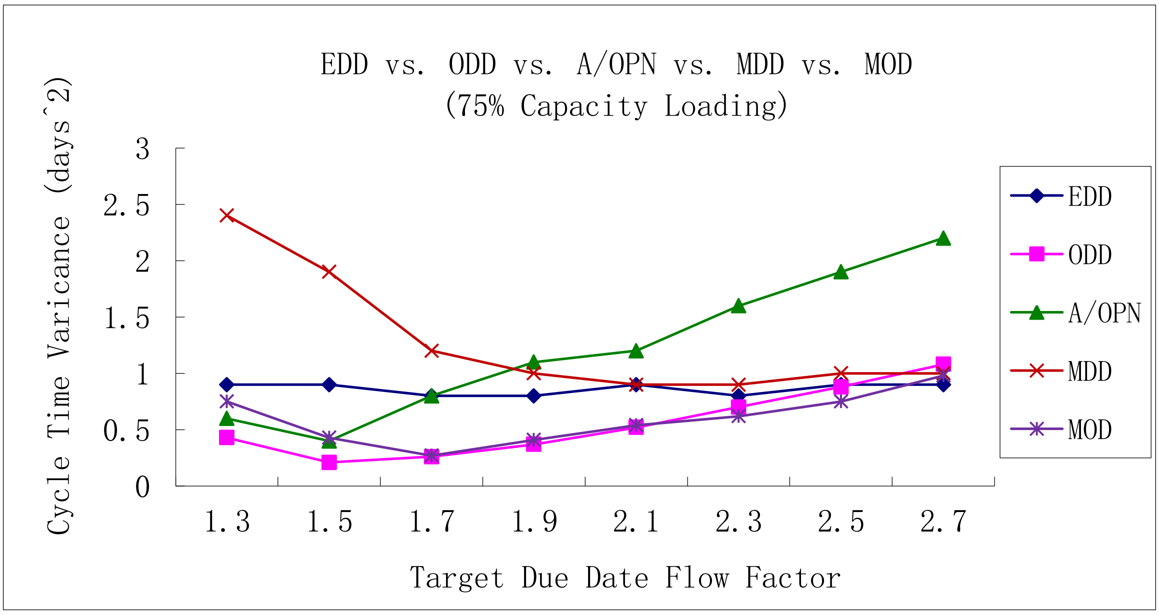


Figure 3.4.18: Cycle time variance comparison of allowance-based rules vs. composite rules

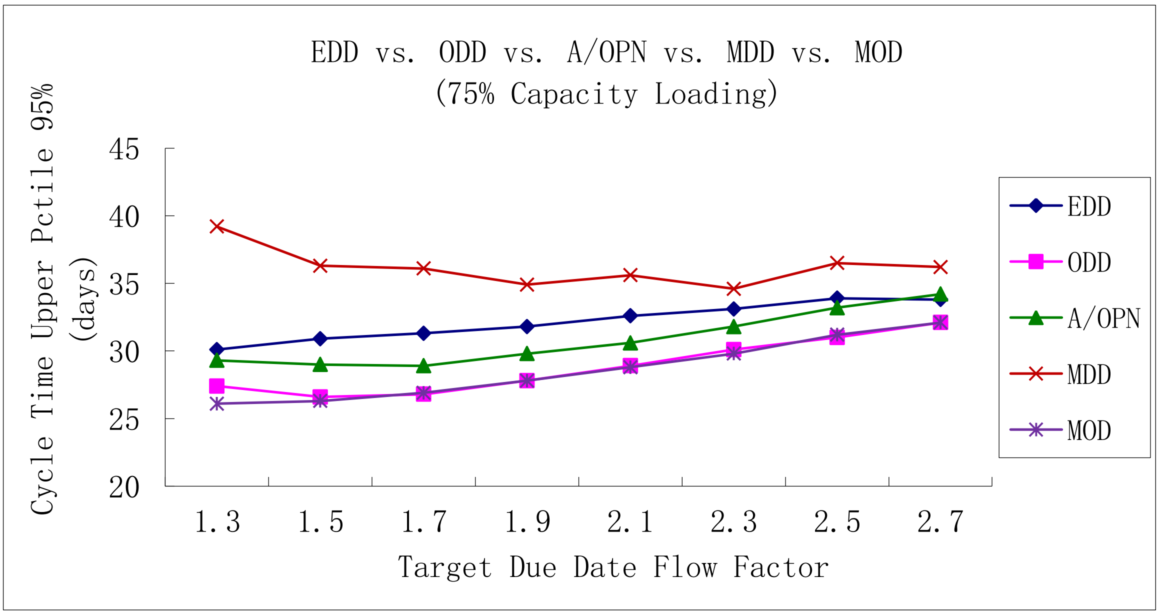


Figure 3.4.19: Cycle time upper 95% percentile comparison of allowance-based rules vs. composite rules

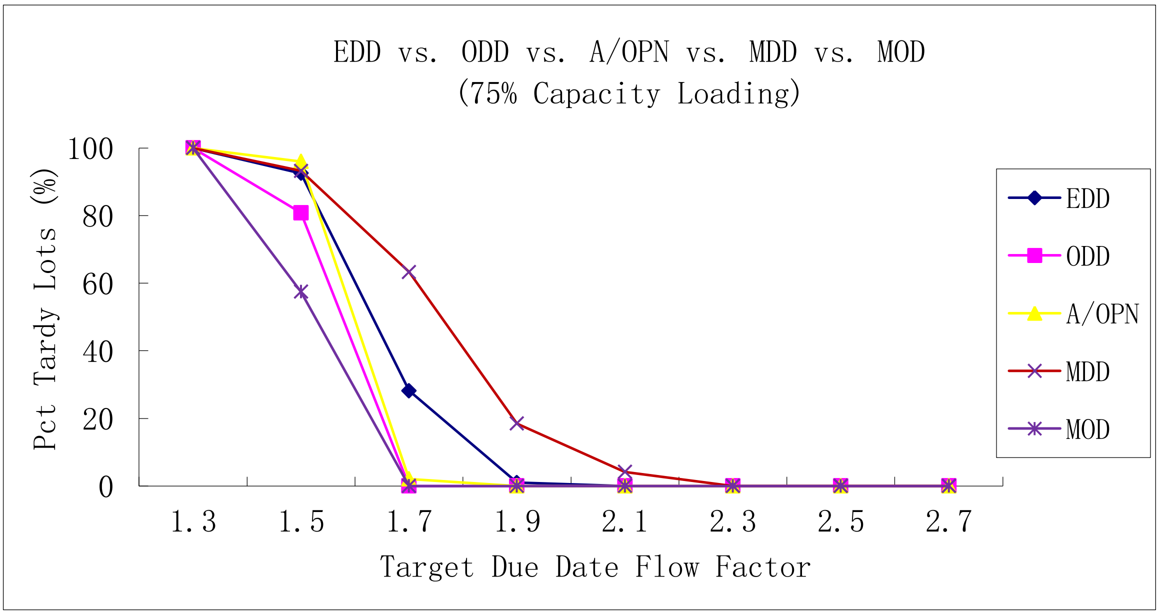


Figure 3.4.20: Percent tardy lots comparison of allowance-based rules vs. composite rules

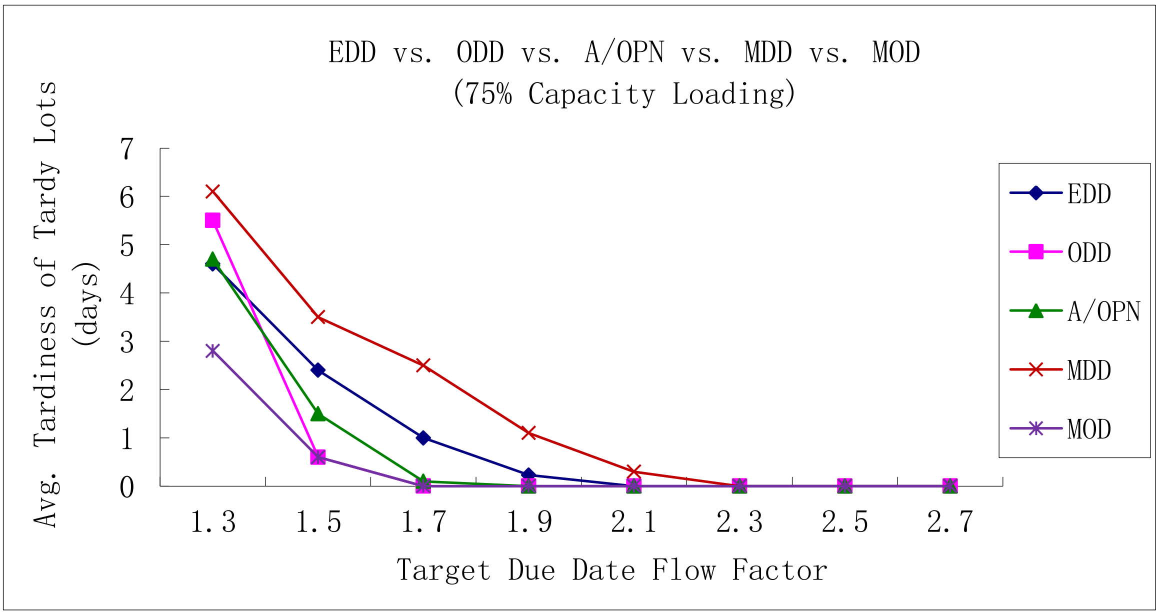


Figure 3.4.21: Average tardiness of tardy lots comparison of allowance-based rules vs. composite rules

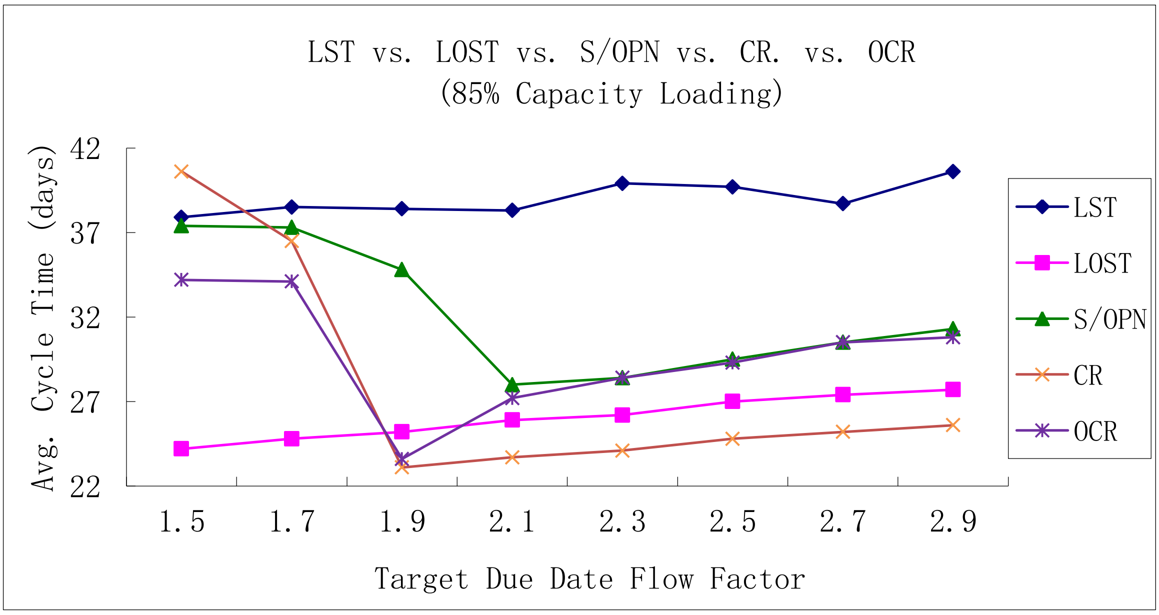


Figure 3.4.22: Average cycle time comparison of slack-based rules vs. ratio-based rules

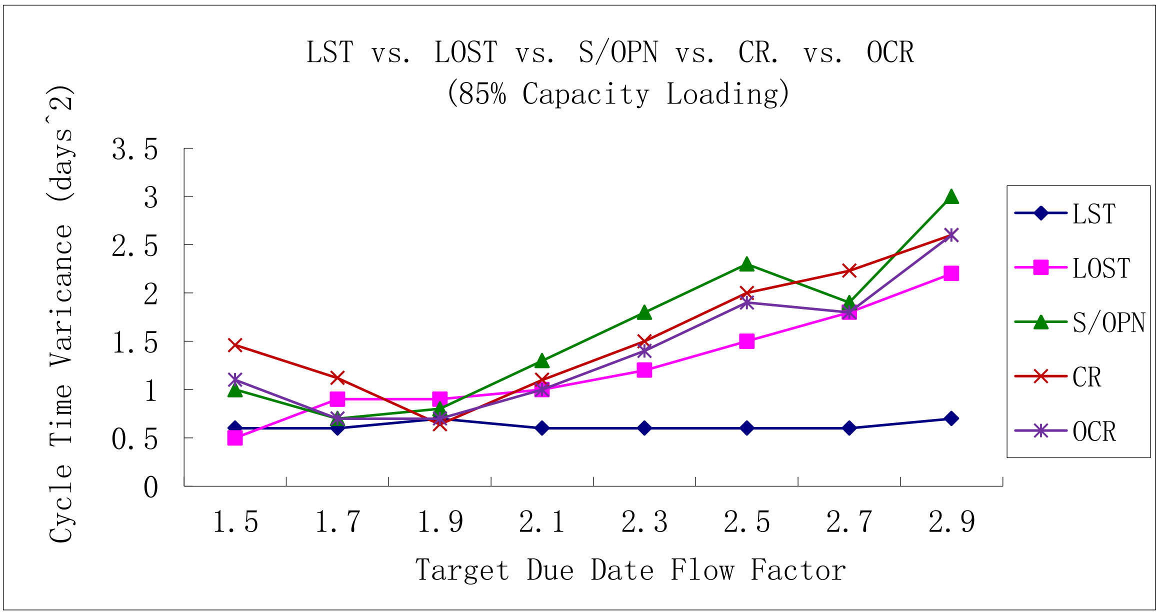


Figure 3.4.23: Cycle time variance comparison of slack-based rules vs. ratio-based rules

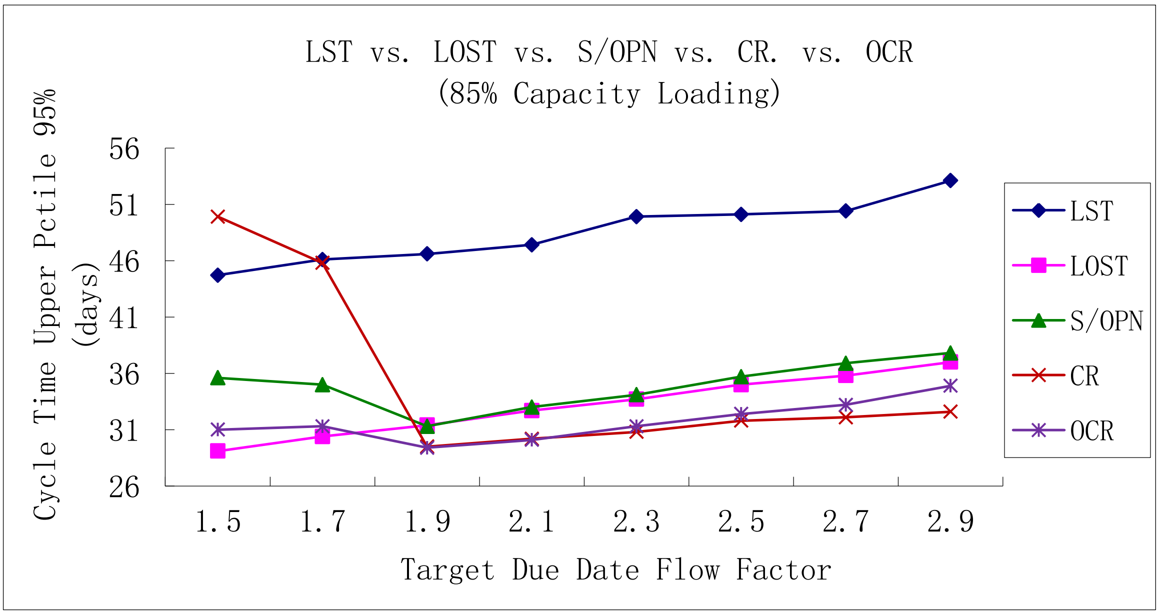


Figure 3.4.24: Cycle time upper 95% percentile comparison of slack-based rules vs. ratio-based rules

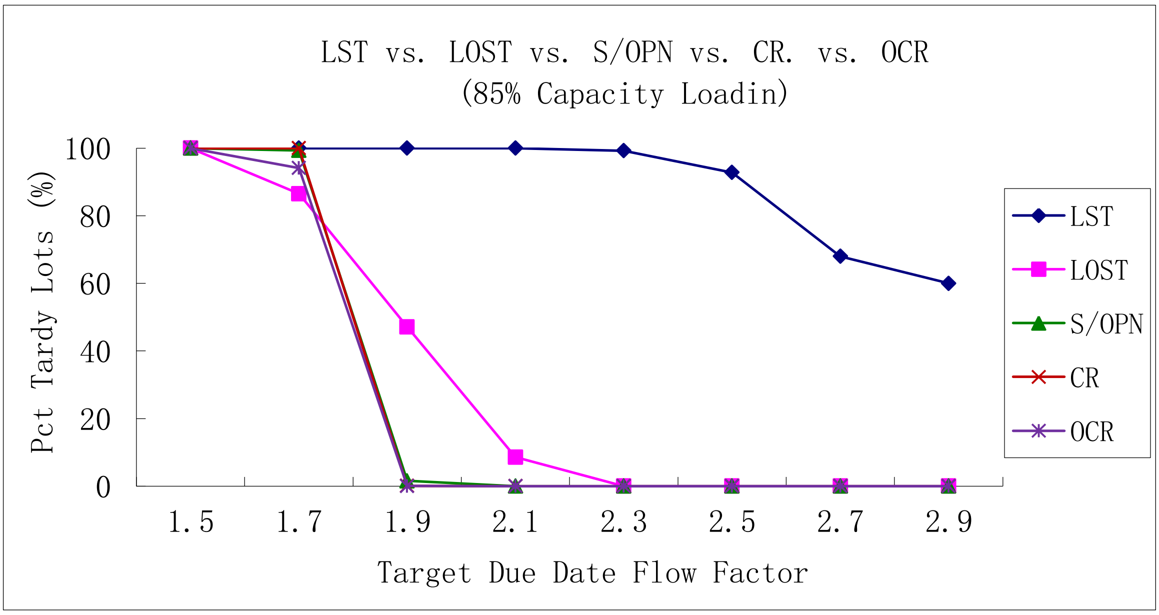


Figure 3.4.25: Percent tardy lots comparison of slack-based rules vs. ratio-based rules

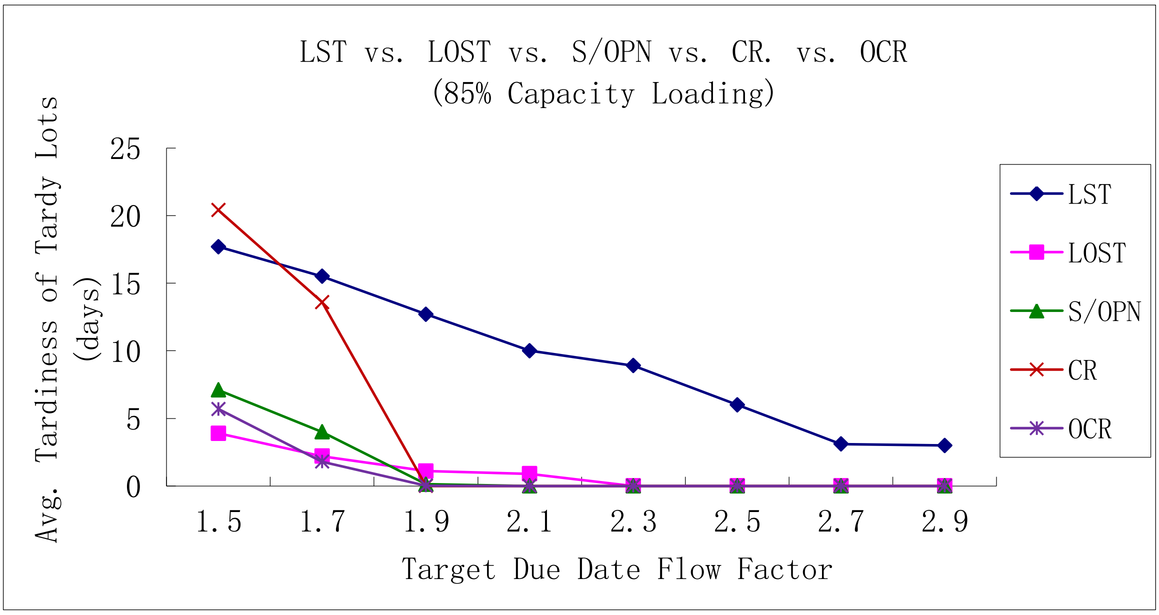


Figure 3.4.26: Average tardiness of tardy lots comparison of slack-based rules vs. ratio-based rules

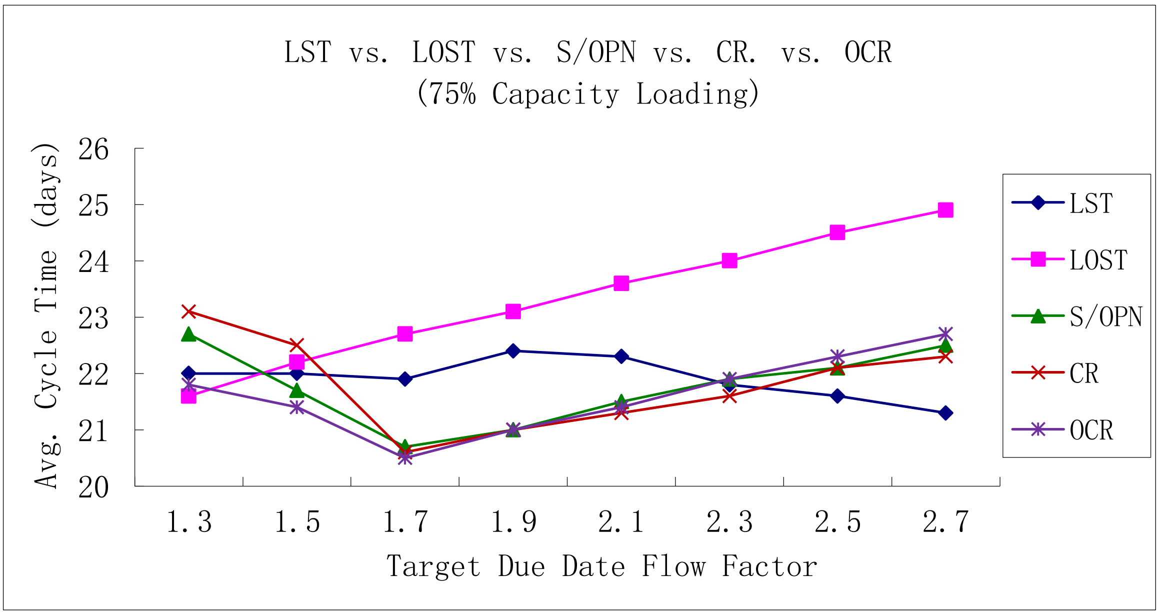


Figure 3.4.27: Average cycle time comparison of slack-based rules vs. ratio-based rules

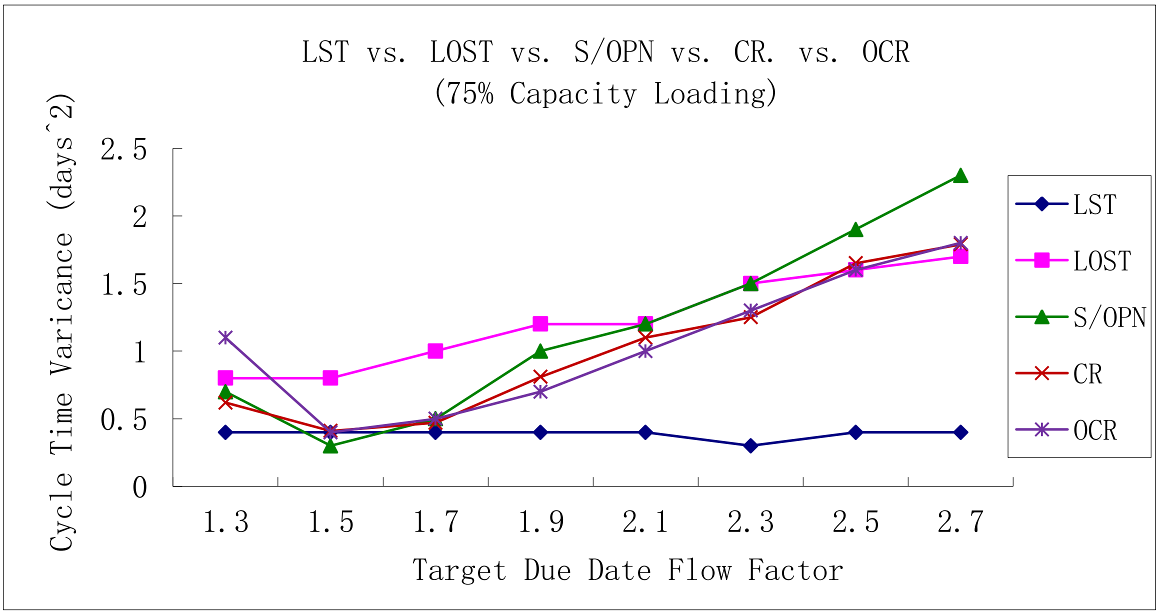


Figure 3.4.28: Cycle time variance comparison of slack-based rules vs. ratio-based rules

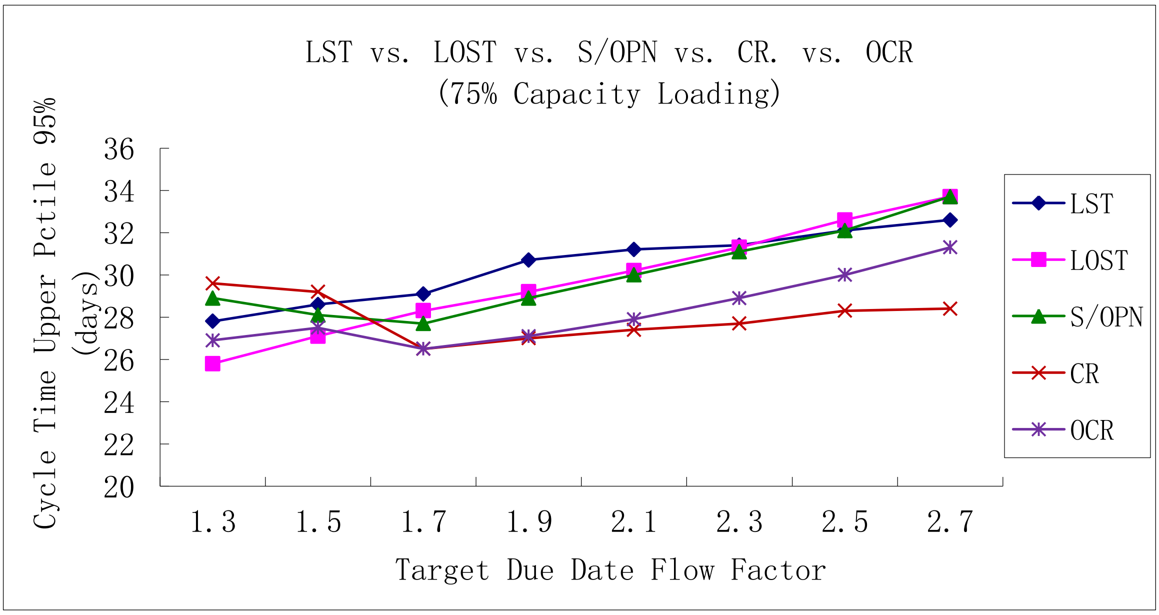


Figure 3.4.29: Cycle time upper 95% percentile comparison of slack-based rules vs. ratio-based rules

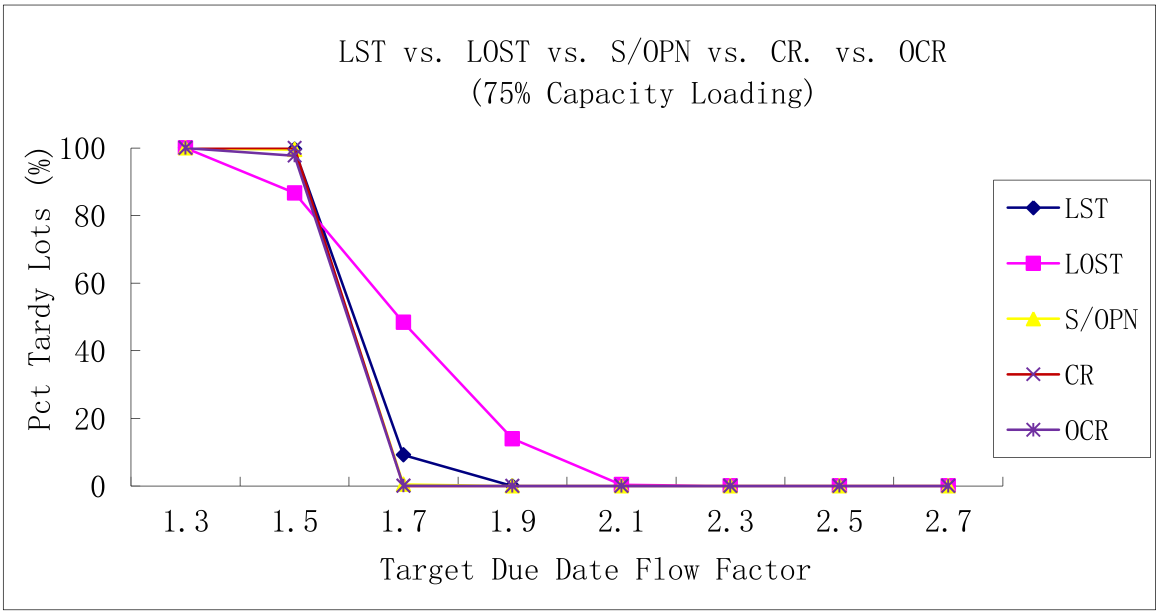


Figure 3.4.30: Percent tardy lots comparison of slack-based rules vs. ratio-based rules

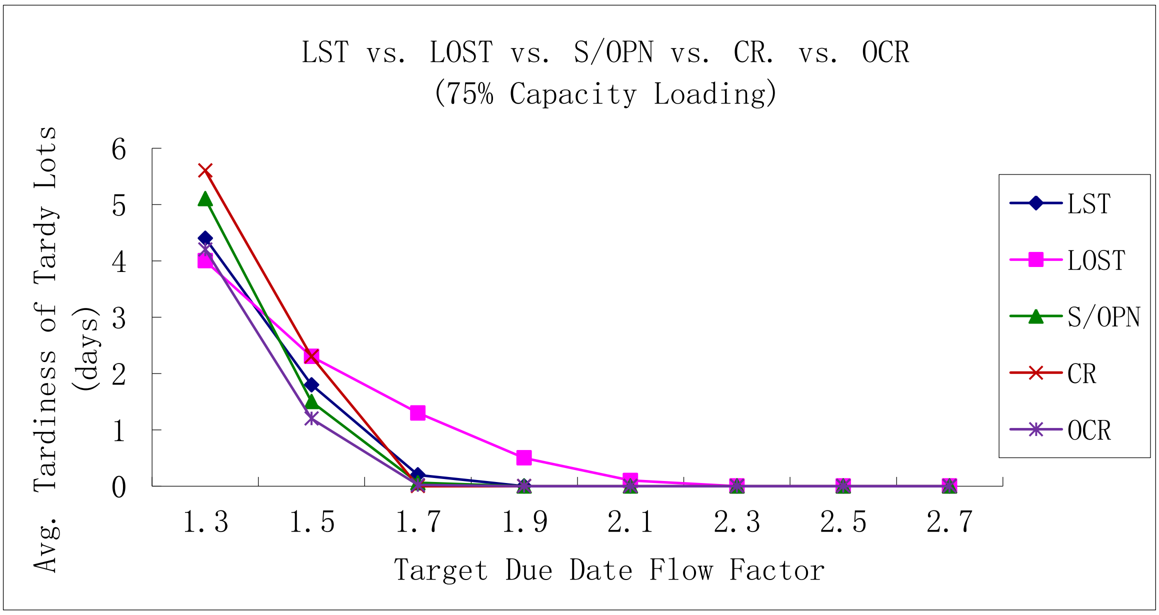


Figure 3.4.31: Average tardiness of tardy lots comparison of slack-based rules vs. ratio-based r