SIMULATION STUDIES OF PUSH, CONWIP AND MODIFIED CONWIP SYSTEMS IN A WAFER FABRICATION ENVIRONMENT

by

NUR AMALINA BINTI MUHAMMAD

Thesis submitted in fulfillment of the requirements

for the degree of

Master of Science

September 2014

ACKNOWLEDGEMENT

First and foremost, I would like to express my appreciation and wholehearted sense of gratitude to my supervisor Dr. Chin Jeng Feng for his enthusiastic guidance, valuable suggestions and persistent supervision which were indispensable for the completion of this research work. I am indebted to him for his constant encouragement and meticulous efforts in correcting errors and suggesting improvements

To the staff in Silterra (M) Sdn Bhd, Mr. Azizi, Mr. Rosli, Mr. Azrin, Mr. Zafir, Mdm. Shafaq, Mr. Goh, Mr. Badrul and Mr. Syahril, that have assisted me in conducting the case study, thank you for your warm welcome and support. The cooperation and comments were crucial for the success of this research.

I would like to thank to all friends, colleagues and staff within Universiti Sains Malaysia. I want to thank specially for their continuous support, happy moments together and wonderful friendship to Ernnie, Syazwani, Joshua Chan, Joshua Prakash and Shukri.

My special thanks to my parents, Muhammad and Fatimah, and my family members, Nuzura, Saipul, Kashaf, Hafiza, Shuhidan, Sara, Faiz, Lukman and Nazirah for their unconditional love along with their support and encourages me.

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

CHAPTER 2 – LITERATURE REVIEW

CHAPTER 3 – RESEARCH METHODOLOGY

CHAPTER 4 – THEORETICAL MODEL DEVELOPMENT

CHAPTER 5 – SIMULATION RESULTS

CHAPTER 6 – DISCUSSIONS

CHAPTER 7 – CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

APPENDIXES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CN	Total CONWIP cards
NC	Percentage of Nested cards
ST	Setup time
OD	Ratio of order distribution
TL	Total layer
TO	Total output
TCO	Total critical output
AvgFT	Average flow time
SDFT	Standard deviation flow time
AvgWT	Average waiting time
MUtil	Machine utilization
PS-B	Push system with batching system
PS-NB	Push system with non-batching system
CS-B	CONWIP system with batching system
CS-NB	CONWIP system with non-batching system
CONLOAD-B	CONLOAD system with batching system
CONLOAD-NB	CONLOAD system with non-batching system
M-CONWIP-B	M-CONWIP system with batching system
M-CONWIP-NB	M-CONWIP system with non-batching system
Nested-CS-B	Nested CONWIP system with batching system
Nested-CS-NB	Nested CONWIP system with non-batching system
WIP	Work-in-process

LIST OF SYMBOLS

p_n	Job type in production, n=1, 2, 3 or 4.
m_n	Machine operates in production, n=1, 2, 3, 4, 5 or 6.
b_n	Buffer between machine, n=1, 2, 3, 4, 5, or 6.
n_{CJ}	Total critical job completed for the whole duration of analysis
n_{NCJ}	Total non-critical job completed for the whole duration of analysis
a_{CJi}	Time critical job i arrives in b_1, i=1, 2... TCO
a_{NCJi}	Time non-critical job i arrives in b_1, i=1,2... n_{NCJ}
b_{CJi}	Time critical job i is completed, i=1,2... TCO
b_{NCJi}	Time non-critical job i is completed$_1$, i=1,2... n_{NCJ}
h_{CJi}	Time critical job i spends in the system (exclude process time)
h_{NCJi}	Time non-critical job i spends in the system (exclude process time)
$(t_{leave_{i,M_j}})_{CJi}$	Time critical job i leaves machine m_j, i=1,2... TCO and j=1,2...6
$(t_{enter_{i,B_j}})_{CJi}$	Time critical job i enters b_j, i=1,2... TCO and j=1,2...6
$(CT_{i,M_j})_{CJi}$	Cycle time critical job i at m_j, i=1,2... TCO and j=1,2...6
$(t_{leave_{i,M_j}})_{NCJi}$	Time non-critical job i leaves machine m_j, i=1,2... n_{NCJ} and j=1,2...6
$(t_{enter_{i,B_j}})_{NCJi}$	Time non-critical job i enters b_j, i=1,2... n_{NCJ} and j=1,2...6
$(CT_{i,M_j})_{NCJi}$	Cycle time non-critical job i at m_j, i=1,2... n_{NCJ} and j=1,2...6
t_{busy}	Time duration of machine M perform the process
t_{actual}	Total time duration of the whole system is performed

KAJIAN SIMULASI BAGI SISTEM TOLAK, CONWIP DAN CONWIP YANG TERUBAH SUAI DALAM PERSEKITARAN FABRIKASI WAFER

ABSTRAK

Fabrikasi wafer ialah pembuatan semikonduktor bahagian depan. Ramai penyelidik berminat untuk memperbaiki perancangan dan kawalan pengeluaran disebabkan oleh sifatnya yang kompleks. Kajian ini ialah penyelidikan asas untuk mendapatkan pemahaman yang lebih besar daripada aspek-aspek asas sistem berasaskan kerja-dalam-proses yang tetap (CONWIP), dalam persekitaran pembuatan wafer. Perbandingan ini dicapai dengan lima sistem kawalan pengeluaran dalam persekitaran pembuatan yang menyamai kekompleksan fabrikasi wafer. Ini termasuk sistem tolak, CONWIP dan CONWIP yang terubah suai (CONLOAD, M-CONWIP dan CONWIP tersarang). Kecuali CONWIP tersarang, sistem-sistem yang lain ialah kaedah yang diadakan dalam industri yang berbeza. CONWIP tersarang ialah kaedah yang baru terkandung dalam penyelidikan ini untuk membenarkan pelepasan awal set kad pelengkap untuk mempengaruhi pemilihan kerja. Sifat yang menarik minat berkaitan dengan fabrikasi wafer ialah aliran proses masukan semula, penyediaan kerja bersandar, pencampuran kerja, keutamaan dan pengelompokan. Ukuran prestasi yang dipertimbangkan ialah jumlah keluaran (TO), jumlah pengeluaran yang kritikal (TCO), purata masa aliran (AvgFT), sisihan piawai masa aliran (SDFT), purata masa tunggu (AvgWT) dan penggunaan mesin (MUtil). Kajian ini dilakukan dengan menggunakan simulasi peristiwa yang diskrit dan analisis statistik. Pemilihan parameter, pemboleh ubah dan ukuran prestasi dalam simulasi secara relatifnya berdasarkan kepada satu kajian kes fabrikasi wafer. Untuk analisis statistik, analisis varians (ANOVA) telah menentukan hubungan antara pemboleh ubah dan ukuran prestasi setiap sistem. Kaedah gerak balas permukaan (RSM) telah menilai pengertian hubungan, diikuti dengan perbandingan grafik antara sistem dan mencari nilai hampir yang optimum bagi setiap pemboleh ubah yang memberikan prestasi yang lebih baik kepada sistem. Keputusan simulasi menunjukkan bahawa sistem tolak telah menghasilkan TCO yang paling tinggi, namun dengan keputusan AvgFT, SDFT dan AvgWT yang paling rendah. Sistem CONLOAD telah mengatasi sistem tolak untuk kesemua prestasi ukuran. Sistem CONWIP menunjukkan keputusan yang sederhana untuk kesemua prestasi ukuran. Sistem M-CONWIP telah mencapai MUtil yang tertinggi, tetapi hanya pembaikan yang sederhana dalam prestasi ukuran yang lain. Sistem CONWIP tersarang telah mencapai AvgFT, SDFT dan AvgWT yang unggul berbanding kesemua sistem yang lain. Mengaplikasikan sistem pengumpulan telah mengatasi sistem yang tidak pengumpulan.

SIMULATION STUDIES OF PUSH, CONWIP AND MODIFIED CONWIP SYSTEMS IN A WAFER FABRICATION ENVIRONMENT

ABSTRACT

Wafer fabrication is a front-end semiconductor manufacturing. Notoriously complex, it has attracted voluminous researchers to improve its production planning and control. This study is a fundamental research with the aim to obtain greater understanding of the basic aspects of CONWIP-based systems in the wafer fabrication environment. This comparison was achieved by five production control systems in a manufacturing environment emulating the complexity of wafer fabrication. These include push, Constant Work-In-Process (CONWIP) and modified CONWIP systems (CONLOAD, M-CONWIP and Nested CONWIP). Except Nested CONWIP, the other systems are methods established in different industries. Nested CONWIP is a newly conceived method in this research to allow early release of complementary cards to influence job selection. Characteristics of interest relevant to wafer fabrication are the re-entrant process flow, job dependent setup, job mix, priority and batching. The performance measures considered are total output (TO), total critical output (TCO), average flow time (AvgFT), standard deviation flow time (SDFT), average waiting time (AvgWT) and machine utilization (MUtil). The study was performed by using the discrete event simulation and statistical analysis. The selection of parameters, variables and performance measures in the simulation was relatively based on a wafer fabrication case study. For statistical analysis, analysis of variance (ANOVA) determined the relationship between variables and performance measures for each system. Response surface methodology (RSM) evaluated the significant of the relationships, graphical comparison between systems and obtained values approximate to the optimal for each variable which gives better performance of the system. The simulation results show that the push systems produced the highest TCO, nonetheless with the poorest result of AvgFT, SDFT and AvgWT. CONLOAD systems outperformed push systems for all performance measures. CONWIP systems show moderate improvements for all performance measures. M-CONWIP systems attained the highest MUtil, but only moderate improved other performance measures. Nested CONWIP achieved excellent AvgFT, SDFT and AvgWT than all other systems. Applying batching system consistency outperformed the non-batching system.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Consisting of six sections, this chapter provides general ideas of the research. First, research background presents the theoretical foundations of this research and elaborates production planning and control systems. Then, problem statements, research objectives, research limitations, significance of the research and thesis outlines are illustrated in following subsections.

1.2 Research background

Most companies in manufacturing industries gain their customers' confidences through their capabilities to fulfill the customer requirements. These requirements commonly include on-time job deliveries and minimum production costs. Manufacturing companies with complex manufacturing processes are more difficult to determine delivery time because of control difficulty in work-in-process (WIP) and flow times (Swaminathan, 2000 and Maass, 2008). Flow time refers to the time a job spends from the job release till its completion in the system (Chincholkar and Herrmann, 2008). Reducing the flow time improves the ability to meet the delivery dates. For example, WIPs management and flow time reduction are carried out in wafer fabrication by process simplification, line balancing, effective scheduling and dispatching (Akcalt et al., 2001; Wang et al., 2007). Also, an accurate flow time forecast is significant for on-time delivery, which can minimize WIPs in the system (Chen et al., 2012).

WIPs management and flow time reduction are always associated with the production planning and control. Production planning includes planning and job release, while in the production control; there are scheduling and dispatching (McKay and Wiers, 2003 and Mönch et al., 2013). Anticipated demand is a vital input to any production planning. A typical planning decision is the job quantities to be released or completed within a certain time in order to optimize the revenue and the finite manufacturing capacity (Jonsson and Mattsson, 2003). Job release refines the decision made at the planning level by disaggregating the quantities in time and space, into individual jobs. Each job contains a production order with specific quantity. A typical job release decision results in a set of jobs that have to be launched into the production line at a certain point in time (Stevenson et al., 2005). Scheduling is the process of allocating scarce resources over time to optimize one or more objectives in a decision-making process (Fowler et al., 2006). Scheduling can be performed for jobs on a single machine, work centre, work area, and finally for all machines on the production line (Phanden et al., 2011). Dispatching assigns the next job to be processed from a set of jobs awaiting service on a machine. Each job is prioritized by using dispatching rules. The production planning and control system hierarchy is shown in Figure 1.1.

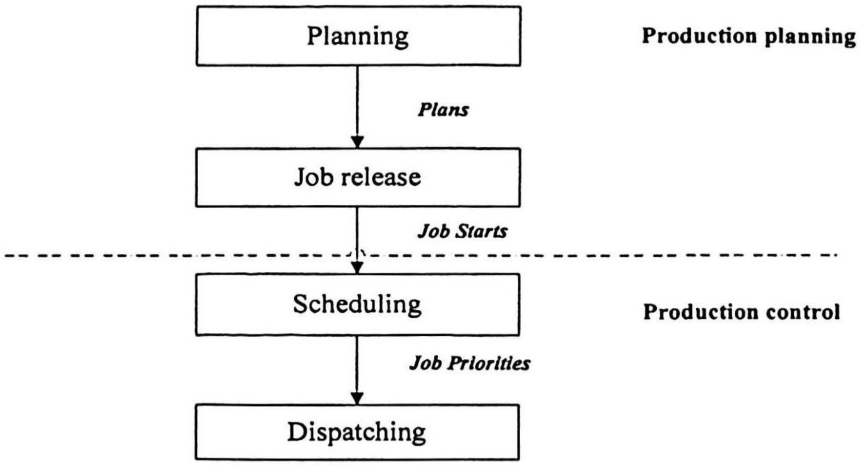

Figure 1.1: The approach in production planning and control. *Source: (Mönch et al., 2013)*

Crudely classified, production controls are either characterized as push or pull systems. In a push system, the movements of jobs are not controlled and hence, the jobs immediately move to the downstream machine after completing all the upstream processes (Kumar and Panneerselvam, 2007). There are no references of the previous researches regarding the application push system in the wafer fabrication (Lozinski and Glassey, 1988). But, push system still being used during comparison with other systems and integrated with other systems as a hybrid system. On the other hands, in a pull system, the movements of jobs are controlled, e.g. with cards and only move to the downstream machines once being triggered to do so (Hopp and Spearman, 2004). The selection of an appropriate production significantly reduces the flow time and the WIPs level in the system (Sacks et al., 2010). There are many references found of the previous researches regarding the application pull systems in the wafer fabrication as will be discussed in the following chapter.

The advantages of pull systems in capping WIPs by using cards receives great attention from many researchers (Grosfeld-Nir and Magazine, 2005; Germs and Riezebos, 2010; Li et al., 2012). Two most common types of pull systems are

Kanban and CONWIP. In Kanban, the control occurs at each machine by authorizing a number of cards per machine. The total WIPs are referred by the accumulative number of all these cards. A job is triggered to move by an available card on each machine. In CONWIP, the control occurs at an early stage of line. The total WIPs are also referred by the number of cards authorized by the system. The job is triggered to move once the process for other jobs has completed and the card is returned to the front line.

In contemporary literature, the pull systems have been modified to improve performances in environments of the system. The modification of the pull systems, especially in CONWIP systems is based on several configuration criteria such as card allocation (Yang et al., 2011) and dispatch rule (Bahaji and Kuhl, 2008). Card allocation contains of shared and dedicated cards, while dispatch rule contains of static and dynamic rules (Prakash and Chin, 2014).

1.3 Problem statements

A substantially high investment cost and fierce global competitions compel wafer fabrication company for effective production planning and control, primarily to meet two objectives. The first is to maximize the overall machine efficiency while duly fulfilling customer requirements. The second is to reduce and regulate flow time to minimize overall inventory level. Nevertheless, the complexities of the manufacturing are overwhelming; easily hamper efforts to achieve the above objectives.

In more details, wafer fabrication commonly undertakes more than hundreds individual steps and revisiting many processes multiple times (Ibrahim et al., 2003; Ibrahim et al., 2010). The typical flow time of a job from start to finish can take up to about few months. Independent fabs have production predominantly 'build to order' with small batch sizes or single discrete items, high job varieties and unidirectional process routings. For example, in an arbitrary time, a significant number of heterogeneous jobs are loaded, each varies in terms of numbers of steps, flow times, processing routes and re-entrant to the same machines. Changing job mix constantly moves the bottlenecks across production line.

In the wafer fabrication, improper approaches to control the jobs enter the production system increases the number of WIPs in the system. A higher WIPs level directly upsurges the production, material and total costs of the wafer fabrication. A high number of WIPs and longer time for WIPs to spend in the system will raise total production cost to complete the overall process. In addition, the increment price per wafer per year directly proportional to the increasing production cost and it becomes problems when the excessive WIPs be present in the system (AlixPartners, 2014; GlobalSemiconductorAlliance, 2014). For an example, China's growing semiconductor company also faced the problem of over-capacity in the global market, which due to the uncontrollable number of WIPs. It is resulted to the decreasing of revenue growth up to 3% per year for the China's semiconductor industry (PricewaterCoopers, 2013). Thus, the revenue cost is declining albeit the production costs are increasing (Heck and Pinner, 2007). This situation leads to the future loss of the wafer fabrication company. This becomes a severe problem for the management of the wafer fabrication company. Solid State Technology (2014) outlined that the major challenge faced by the wafer fabrication industry was the fabrication cost. Additionally, fab cost alone accelerated 40% per year. Wang and Chiu (2014) suggested that reduction of production cost is important to increase the profit of the company. Minimizing the number of WIPs in the system helps the management to reduce the production cost (Hussain and Drake, 2011).

The problems of high WIPs in the wafer fabrication are consequences of the poor decision-making during production planning and control (Pfund et al., 2006). Each decision must be considered every point of view, such as the job roles, scheduling and dispatching. The company must regulate the WIPs in the system by maintaining the number of output produced and at the same, reducing the waiting time per job. A longer waiting time leads to higher flow time and late delivery to the customers. Thus, in this research, several production control systems are selected and compared in a simulated wafer fabrication environment especially on their performance of the stated problem faced by the company. The research also examine leading variables to few contemporary production control systems. According to Cua et al. (2001), the determination of accurate variables in the simulation model is essential in production planning and control studies.

1.4 Research objectives

The objectives of this research are:

- To investigate and compare the production performances of application push, CONWIP and modified CONWIP systems in the wafer fabrication environment by using discrete event simulations.
- To explore the variables that gives influence to each abovementioned system and determines their significant relationships to the performance measures.

1.5 Research limitations

- Wafer fabrication is one of the most complex semiconductor manufacturing processes. Simulation of a real world in wafer fabrication contributes to the time constraints to develop a complex system configuration and takes long execution times to statistically generate valid estimate results. This limitation also faced by Nazzal and McGinnis (2007) and Kiba et al. (2009).
- Application of production control systems in the wafer fabrication especially push and Kanban can be traced up until a certain years only due to lack of researches done regarding these systems.
- The reviews are focused on the application of modified CONWIP systems in the wafer fabrication environments. Therefore, several reviews relating modified CONWIP systems are not cited, which are applied in the other area of applications.
- The manufacturing simulation software used was WITNESS 2008. The results obtained from this software were not verified using other software. In fact, the use of WITNESS 2008 as the only simulation software is applied in many researches, which include Calinescu (2002) and Lu (2009).

1.6 Significance of the research

This research offers novelty knowledge about the performance of five different types of production control system and the superiorities between those systems based on the wafer fabrication environment. This research is significant for two categories, which are researchers and wafer fabrication company.

To the researchers, the novelty knowledge can be a reference for them for future research in order to minimize the problem and improve the performance of production planning and control in the wafer fabrication. To the wafer fabrication company, the novelty knowledge of the research can be a guidelines or references during selecting the suitable type of production control system to be applied in the production line. The significance variables understudied also helps them in considering the suitable magnitude for each variable involved during the production.

1.7 Thesis outline

Chapter 1 introduces the research background of the research, containing the division of production planning and control system. Then, the problem statements are discussed regarding the wafer fabrication environment. The research objectives, research limitations and significance of the research are also presented in the following subsections.

Chapter 2 provides a literature review on production planning and control, focusing on the push and pull systems. The general background of wafer fabrication is presented along with the issues of complexity factors in the wafer fabrication. Then, the literature of adopting pull systems and its existing limitations in the wafer fabrication are discussed in order to identify the need for further investigation.

Chapter 3 describes the methodology used in the research, including detailed descriptions on each step taken in building the relevant simulation models. The concept of analysis of variance (ANOVA) in relating the simulation result within each variable and performance measure, and the concept of response surface methodology (RSM) in providing the polynomial equation of simulation model are also discussed.

Chapter 4 presents the background of the wafer fabrication company and follows by the concept of Nested CONWIP system. The simulation model building for the simulation is discussed as well.

Chapter 5 presents results from the five sets of experiments with push, CONWIP, CONLOAD, M-CONWIP and Nested CONWIP system. The results of the simulation are plotted in order to analyze the performance of each case study by varying the number of variables.

Chapter 6 details the discussion regarding the results and findings obtained from the experiments.

Chapter 7 presents the conclusions of this research work and recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter first explains the overview of the manufacturing process in wafer fabrication, followed by discussion on the main factors that lead to the complexity in the wafer fabrication. The production control systems (push, Kanban, CONWIP, hybrid, CONLOAD and M-CONWIP systems) are reviewed and the comparisons between those systems are presented. Next, several configuration criteria for modified CONWIP systems in the wafer fabrication are discussed. Finally, the literature findings are explained.

2.2 Overview of wafer fabrication

The best means for semiconductor industry is the process of wafer fabrication into a complete integrated chip (IC). It makes electronic products to be functioning in the computer, communication devices, robotics, weaponry and automobiles (Kalpakjian and Schmid, 2010). The wafer is defined as the thin slice of silicon crystal and IC is defined as a small silicon wafer that can pack from tens, hundreds, thousands or millions electronic components such as transistors, resistors and capacitors and a small area of chips (Sze and Kwok, 2006).

The following process steps have to be performed in wafer fabrication starting with receiving raw wafer as outlined in Figure 2.1 and summarized in Table 2.1.

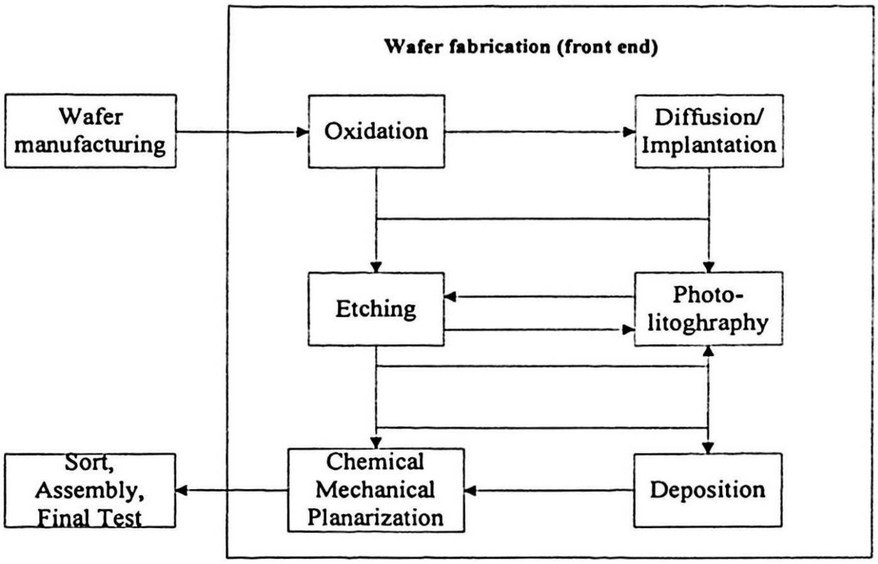

Figure 2.1: Manufacturing process flow in the wafer fabrication. (Mönch et al., 2013; Mönch et al., 2011; Bates, 2000)

Table 2.1: Manufacturing processes in the wafer fabrication. (May and Spanos, 2006; Mönch et al., 2011)

Process	Descriptions
Oxidation	Refers to the growth of a thin layer of oxide on the surface of a wafer, as a result of the reaction of oxygen with the substrate material. Their roles are effective in "gate oxide" material, dopant masking and device isolation.
Implantation	The electrical character of the doping types and concentrations is altered and it is similar to oxidation except using a different gas other than oxygen. Ion implantation is a much more intensive process and it is accomplished by accelerating the ions through a high-voltage filed as much as one million electron volts into the wafers.
Photolithography	Uses a reticle with a pattern of the chip deposited onto the wafer and being uniformly coated with photoresist like a thick light-sensitive liquid. Every wafer going through this step more than 40 times because the circuits are made up of layers.
Etching	Works to remove material from the wafer surface immediately after photolithography process. Uncovered areas by photoresist are then removed from the wafer.
Thin film metal and thin film dielectric	The thin-film metal and dielectric process adds materials to the wafers. The process machines in this area operate on one wafer at a time. The machines in this area are equipped with two load ports to allow concurrent loading in of two lots. The dispatching policy in this area requires lot cascading and queue time reduction.
Chemical Mechanical Planarization (CMP)	The step to clean and levels the wafer surface. Chemical slurry is distributed to a wafer and equalizes wafer surface. It resulted in diminishing the thickness of the wafers before adding a new layer.

2.2.1 Complexity factors of wafer fabrication

Transformation of semiconductor wafer from a bare silicon wafer into complete IC with millions of transistor circuits undergoes a wide range of process levels. Such process levels include manufacturing process, production planning and production controls. The fabrication of the wafer is the most complex manufacturing process when it undergoes a wide range of process steps before producing a complete IC (Kim et al., 2003). The complexity factors are confounded by many factors, including job mix, routes with hundreds process steps and large number of machines involved (Chou and Hong, 2000).

The first contributor to the complexity factors of wafer fabrication is the mix of different job types (Mönch et al., 2003). Commonly, it takes more than hundreds individual steps during fabrication and typical flow time from start to completed job can take up to about few months with back and forth iterations through the previous mentioned six processes (Chik et al., 2010; Balakrishna et al., 2011). Different job types are having different route of steps. Since the routes of steps in wafer fabrication cannot be simply changed, these complexities must be handled properly in terms of job release, scheduling and dispatching. Therefore, the wafer can be processed smoothly without any delay and the company can tolerate with the aggressive due dates decided by the customers (Kim et al., 1998; Kim et al., 2001).

Most manufacturing systems will have straightforward cycle processes without any revisiting the same machine except for the rework job (Fowler et al., 2002). In the wafer fabrication, re-entrant process flow is essential. Re-entrant process flow means that the repetitive process of jobs (a lot of wafers) on the same machine (Lee et al., 2002; Sun, 2009), due to the layer structures on the wafer and limited number of machines a fab could have. This situation will lead to the long waiting time on the job being processed. In fact, jobs waiting time is the biggest contribution to the lead time of the wafer fabrication and become challenging in wafer fab while dealing with the aggressive due dates from the customers (Zhou and Rose, 2012).

The bottleneck is another challenge in wafer fabrication (Gupta et al., 2006). The restrictions for the machines that can be used simultaneously will increase the number of wafers queued in front of the machines and typically treated as a bottleneck (Wang et al., 2007). Photolithography process is the most frequent process being visited by a job in order to place several layers of pattern on the wafer surface (Tanjong, 2011). The machine that performed this process will have a high bottleneck in front. Arisha and Young (2004) explained that bottleneck happened due to the re-entrant flow and unscheduled breakdown. In addition, variability of the inter-arrival rate of jobs also can be a factor of bottleneck to happen in the wafer fabrication (Wu, 2005).

Sequence-dependent setup times can be another issue in wafer fabrication, which contributes to the increasing waiting time and bottleneck (Philip, 2005; Oey and Mason, 2001). Sequence-dependent setup times occur in certain areas and are related to gas pressure, temperature changes, metal composition, etc. (Mönch et al., 2013). For an example, dopants have to be changed frequently in the ion implantation work area. This effort depends on the predecessor dopant. If the setups are not treated correctly by the production control staff, the corresponding machines can become bottlenecks. Alternatively, the jobs are processed in batches in order to reduce setup time, costs and improve quality. Batching system is a technique used in manufacturing, in which the jobs with the same manufacturing process are grouped together (Cooney, 2002). On the other hands, non-batching system is a technique used in manufacturing, in which the machine processes an individual job based on the sequence in the buffer (Mason et al., 2005).

2.3 Production control systems

Production monitoring has to be ubiquitous in wafer fabrication, thus without proper planning of production control will lead to the shortage, high flow time and late delivery to the customers (Olhager, 2003; Krishnamurthy et al., 2004; Dolgui and Prodhon, 2007). Production control system is a function to control the movement of jobs through the manufacturing process from raw material until producing the completed jobs (Ballard, 2000). As mentioned in Section 1.2, two forms of production control systems are push and pull systems. The terms of push and pull systems refer to the means to release job into the production system (Hopp and Spearman, 2011). Spearman and Zazanis (1992) explained the behavior of push and pull systems in details. While the literature review in both systems is comprehensive, more emphasis is placed in the semiconductor industry in particular the wafer fabrication.

2.3.1 Push systems

Push system schedules the release of jobs based on the forecast demand (Hopp and Spearman, 2000) and a job starts to enter the production as scheduled in Master Production Schedule (MPS). It releases a job into a production process purely based on a schedule and the release time is not changed according to event in the process itself (Zijm, 2000). Commonly used push systems are Material Requirement Planning (MRP) and Manufacturing Resource Planning (MRP II) (Kim et al., 2002).

MRP is a flow-control system designed to make sure materials (parts or components) with dependent demand available when they are needed. The information used by MRP needs 100% accuracy of inventory records so that the material can be supplied as needed (Enns, 2002). The MRP II is an extension of MRP and incorporates feedback of data (through shop floor control system) on various aspects of operating performance so that corrective action can be taken in a timely manner (Groover, 2007). MRP and MRP II are called push systems since they push items (materials or resources) in the production chain (Bidgoli, 2010).

The general workflow of push system is depicted in Figure 2.2 and summarized as follows (Bonney et al., 1999; Ghrayeb at al., 2009; Weitzman and Rabinowitz, 2003). A push system works to enter the jobs into the production as in the schedule or forecast to start the process. As each step of the process is completed, the job is moved to the downstream machine or buffer, regardless the preparedness of the next operation to receive the job. Eventually, the job will be completed, without necessarily following the initially plan. Nevertheless, push systems are popular mainly because it is able to enhance utilization of the capacity for each machine, often at the expense of excessive WIPs.

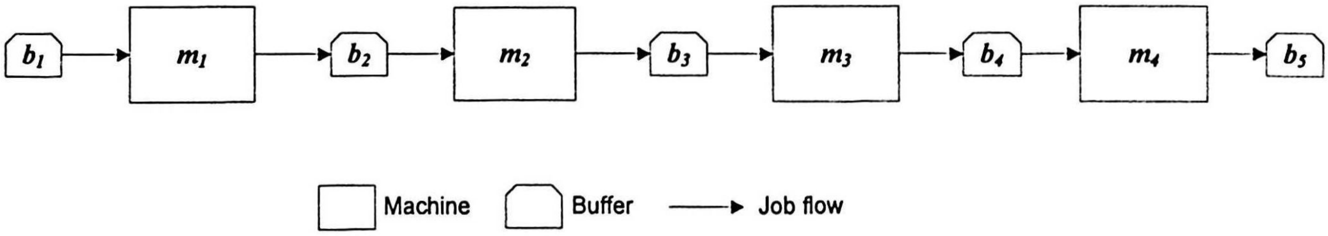

Figure 2.2: The workflow in push system (Ghrayeb at al., 2009)

Lozinki and Glassey (1988) reported that there are no references solely focused on push system or MRP being used in the wafer fabrication. They suggested that MRP may not be suitable for wafer fabrication. Nevertheless, push system still compared with other systems in wafer fabrication (Rose, 1999; Lee et al., 2002; Bahaji and Kuhl, 2008). Also, the concept of push system is widely used in hybrid systems applied to wafer fabrication. The hybrid systems integrate push system with the other systems (Lin et al., 2006; Müller et al., 2012; Gonçalves et al., 2005). The literature of hybrid push systems will be explained in the section 2.3.3.

2.3.2 Pull systems

A pull system emerges from the core concept of lean manufacturing called just-in-time (JIT). Lean manufacturing seeks to eliminate the waste during production. In concept, pull system authorizes the release of job based on the system status by using cards (authorization signals), resulted in materials are produced only when requested by the customers (Liker, 2004). Thus, in pull system, the production is scheduled based on customer demand, in contrast to push systems, where the production is scheduled based on the forecast (Gastermann et al., 2011). Pull system allows a job into the production when an unused card is available, which shows the readiness to start another process. Typically, the authorization signals are the result of the completion of job or voids in a stock level at some point in the line (Seidman and Holloway, 2002). Hopp and Spearman (2004) characterized several benefits of pull system such as lower WIPs level, shorter flow times, smoother production flow, improve quality and reduced cost.

Common examples of pull systems are Kanban system and CONWIP system. Both are extensively used in industry (Gaury et al., 2000). Both systems will be revealed in details in the following sections.

2.3.2.1 Kanban systems

The workflow of Kanban system is depicted in Figure 2.3 and summarized largely based on Tardiff and Maaseidvaag (2001). Kanban system is applied by circulating cards between the machine and the downstream buffer. The machine must have an available card to trigger the operation. Then, job can be withdrawn from the first buffer or upstream machine to perform the operation. The finished job is attached to the available card and relocated to the downstream or final buffer. Once the machine picks up the job to perform the operation, it detaches the card from the job. A signal is given when the card is sent back to the upstream machine to perform next process. The flow is continuous until the job reaches the final process, where a completed job is ready.

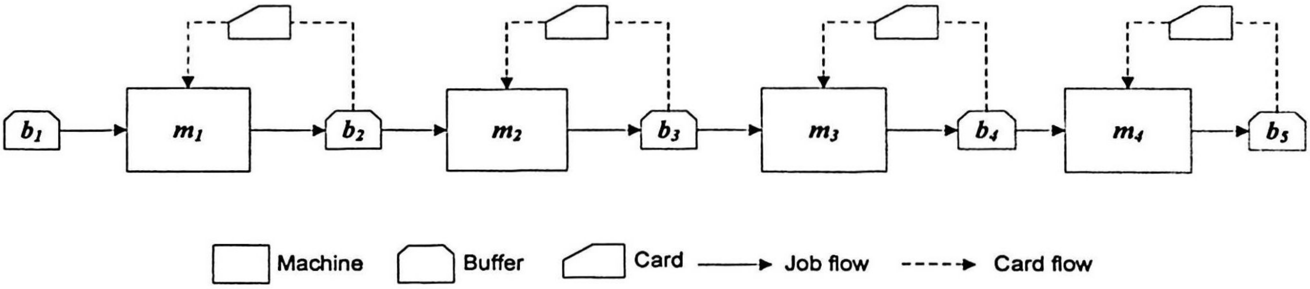

Figure 2.3: The workflow in Kanban system (Tardiff and Maaseidvaag, 2001)

Several literatures discuss Kanban system in the wafer fabrication (Otenti, 1991; Kraft, 1992; Schömig, 1994; Qiu et al., 2002; Lee et al., 2002). Nevertheless, literatures for Kanban system in the wafer fabrication can only be traced up to year 2002.

Otenti (1991) successfully modified and implemented a Kanban system in a wafer fabrication facility to prevent the line unevenness that led to increasing flow times, failures and poor predictability. The objective of Kanban system was to minimize the queue time by quick delivering WIPs when the operation is ready to perform the process. The Kanban system was implemented stage by stage from the first third station of the line until to the last third station of the line.

Kraft (1992) used Kanban in the wafer fabrication as a basic manufacturing inventory management. The objectives are minimum flow time and WIPs, balance distributed line and maximum process yields. Besides using cards, dedicated visual Kanban was applied to trigger the production signal such as static and flashing red light, static and flashing green light and static yellow light. The dedicated visual Kanban was also installed to monitor and display dynamic factory information. The result shows that the approach is more effective to reduce inventory with a balanced line.

Schömig (1994) presented the Kanban system in a wafer fabrication environment organized as a job shop with a great number of re-entrant machines. The Kanban systems with Late-Buffer First Served (Kanban-LBFS) were compared with several conventional sequencing rules and inventory regulation mechanisms such as Earliest Due Date (EDD), Shortest Remaining Processing Time (SRPT) and First-In First-Out (FIFO). The simulation results demonstrate that most conventional sequencing rules are superior than Kanban-LBFS rule, and the shortcomings of discrete event simulation that may influence the temporal behavior of these systems.

Qiu et al. (2002) presented a well-distributed and a more manageable WIPs management system for an advanced wafer fabrication system. The concept of virtual production lines (VPL) system by using Kanban system was proposed. By taking advantaged of both systems containing VPL and Kanban, the findings show that the flexibility, scalability, and responsiveness of a shop floor information system have improved.

Lee et al. (2002) mentioned in their paper that some attempts to implement a Kanban system in wafer fabrication based on a pull system of JIT approach. But, there are certain conditions must be strictly followed when Kanban is implemented. The manufacturing must defined when jobs are started based on available machine, not push material past the machine capacity load factor and all materials should run FIFO. Unfortunately, a strict implementation of the Kanban system driven to be difficult in many cases, such as the presence of dynamic line balancing, bottleneck, and starvation.

2.3.2.2 CONWIP systems

The workflow of the CONWIP system (Spearman et al., 1990) is depicted in Figure 2.4 and summarized as follows (Marek et al., 2001). An available card is attached to the job waiting in the first buffer to start the production process. Once the process is completed at an upstream machine, the part is pushed into the downstream buffer and then to the downstream available machine to start a new process. Upon completing the final process, the completed job is pushed to the final buffer and the card is first detached and then, sent to the first buffer to attach to the other waiting jobs to start a new round of production.

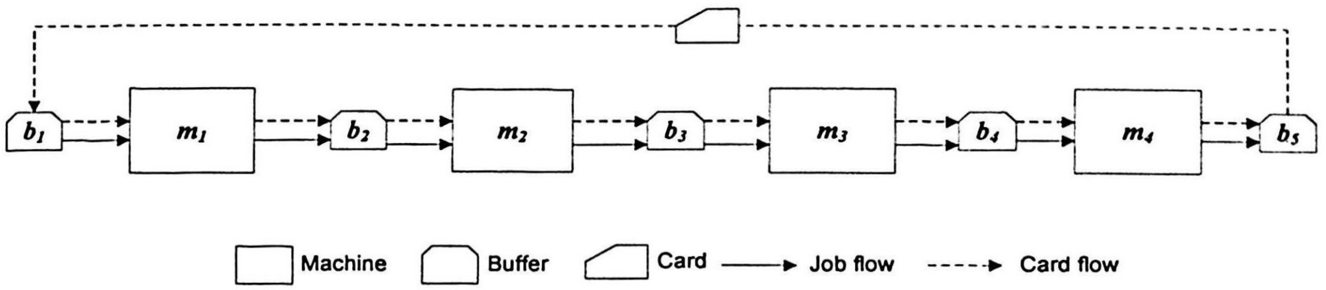

Figure 2.4: The workflow of CONWIP systems (Spearman et al., 1990)

Numerous literatures discuss CONWIP system in wafer fabrication (Sha et al., 2006; Qi et al., 2008; El-Kilany, 2011; Eng and Sin, 2013; Yoon and Kim, 2013).

Sha et al. (2006) adopted a CONWIP system in the photolithography area as the release strategy during establishing a set of dispatching rules. The rework of defective jobs was considered during the production to coordinate and balance the workload of the entire production line. The integration of proposed dispatching rule - Reworking the Defective Wafer (Rw-DR) into CONWIP system has fasten the manufacturing procedure to completion, reduce the machine workload of the waiting line and increase the output quantity.

Qi et al. (2008) proposed job release method, namely WIPLOAD control (WIPLCtrl) in the wafer fabrication and compared the performance of proposed hybrid approach with the Shift Release (SR) and CONWIP. WIPLCtrl releases the next jobs when the remaining processing time is lower than the theoretical processing time, and while SR releases jobs into the system at the beginning of every shift. The comparisons were done in terms of standard deviation of flow time, average flow time, average lateness, total output and WIPs inventory. For all comparisons, the WIPLCtrl was superior to CONWIP and SR, while CONWIP outperformed SR.

El-Kilany (2011) evaluated the impact of different job release policies such as CONWIP, continuous WIPs review and periodic WIPs review on the performance of wafer fabrication facilities. Simulation of the Intel Mini-Fab model compared the proposed job release policies by using different dispatching rules: FIFO and SRPT. It concluded that using periodic WIPs review can be simpler and more pragmatic compared to CONWIP that demands continuous monitoring of WIPs levels and also synchronized job releases with job completions.

Eng and Sin (2013) performed CONWIP as production control strategy and simulated the processes at the back-end of semiconductor manufacturing. They used a single universal card by introducing a single loop control in the system and multi loop CONWIP control mechanisms which uses three different cards for each loop. The results showed that the CONWIP system was more effective to minimize WIPs levels and to meet throughput targets compared to the current systems. The results also demonstrated that corresponding improvement inflow times. The multi loop control accomplishes better performance and more robust than the single loop control.

Yoon and Kim (2013) proposed a CONWIP system to offset the limitation of advanced Operation Due Date (OPNDD) in a wafer fabrication facility. OPNDD caused excessive queue lengths in the bottleneck machines. The simulation results showed that the adaptive CONWIP system improved the standard deviation of flow times.

2.3.3 Hybrid systems

Hybrid system combines the advantage of push system (dynamic production of make-to-order (MTO) environment) and pull system (WIPs control for make-to-stock (MTS) environment) (Smalley, 2009). A perusal of literatures shows that the hybrid system is not new in wafer fabrication (Lee and Lee, 2003; Lin et al., 2006; Perdaen et al., 2008; Kwang and Chin, 2008).

Lee and Lee (2003) proposed three types of control approaches in thin film transistors and liquid crystal display (TFD-LCD), which are push, pull and push-pull systems. The push system aims to maximize the production volume as being driven by the throughput, while the pull system is driven by the production target of unit period or demand. Push-pull system is the mixture of these two policies. The production target is accomplished by using a linear programming model for a number of planning periods. It showed that a pull system offers stable throughput and delivery satisfaction at a smaller cost and fewer productions. The push system is superior to the other two counterparts in the flow time and throughput.

Lin et al. (2006) simulated and analyzed automated material handling (AMHS) for a wafer fabrication by considering the consequences of the dispatching rules. A hybrid push/pull was proposed by integrating with the dispatching rules like the first-encounter-first-served (FEFS) and shortest distance with the nearest vehicle (SD_NV). The simulation results revealed that the resultant AMHS achieved lower WIPs levels and flow times.

Perdaen et al. (2008) applied push system in the first stage of line and pull system at the final stage of the line. This system is namely as the push-pull point (PPP). The PPP policy (either pure push, pure pull or coupled with CONWIP), deals with short-term fluctuations in demand did not shows any improvement. However, when the PPP policy is combined with a CONWIP system in a high demand with high variance, the improvements are shown, where the WIPs level and mismatch between demand and target were reduced.

Kwang and Chin (2008) adopted a push-pull sampling methodology to improve the efficiency of a test job in a wafer fabrication to improve its efficiency. The push-pull methodology in the automated sampling controlled and propagated the correct sampling data for the operators to run test jobs with the exact samples. The work has given creation of a robust controllable system which improves manufacturing capacity and shortens the lead time to end customer. An effective control is achieves to reduce non-value added activities and improved quality.

2.3.4 CONLOAD systems

Constant Load (CONLOAD) system is a job release rule for wafer fabrication presented by Rose (1999). It works as a CONWIP system; limits the movements of cards between the first machine and the bottleneck. The intentions are to maintain the bottleneck utilization at a desired level and to deliver a smooth evolution of the WIPs. The workflow of the CONLOAD system is depicted in Figure 2.5 and summarized as follows: An available card is attached to the job waiting in the first buffer to start the production process. Once the process is completed at the upstream machine, the job is pushed into the buffer and moved into the available machine at downstream for the subsequent process step. If the machine is busy, it waits until the machine is vacant. The process is continuous until the job reaches the bottleneck machine, and the card is detached and sent to the first buffer to start a new production.

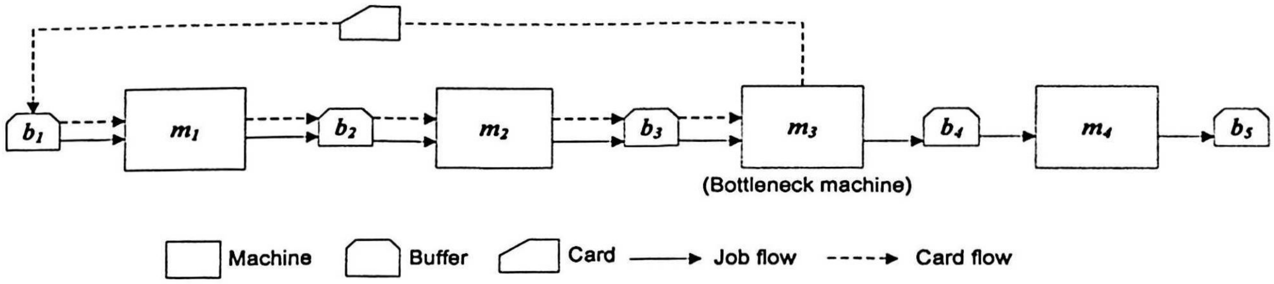

Figure 2.5: The workflow of CONLOAD system (Rose, 1999)

Literatures on CONLOAD system in the wafer fabrication can be found in Rose (1999), Rose, (2001) and Mönch, (2005). Rose (1999) presented CONLOAD to control the average bottleneck load and to smooth the job release into the system. To release a job into the fab, CONLOAD considers the capacity of work at the machines by this particular job. The performance of CONLOAD was compared with CONWIP, Constant Work (CONWORK) and push systems in terms of bottleneck utilization and WIPs over time. The simulation results showed that CONLOAD was superior to the other rules and achieving desired level of bottleneck utilization. Also, the CONLOAD reduced the fluctuations in flow times and stabilize the job release into the system.

Rose (2001) performed four CONWIP-like (classic CONWIP, Layerwise CONWIP, CONLOAD and TOTAL CT) job releases on flow times and WIPs levels of a wafer fabrication facility subjected to the changes of dynamic load. Classic CONWIP controlled WIPs thresholds; Layerwise CONWIP controlled the progress of each layer, CONLOAD controlled the average bottleneck load and TOTAL CT controlled the remaining flow times. All the approaches reduced the variability of the WIPs and flow time of the fab compared to the uncontrolled fab.

Mönch (2005) investigated several job release strategies (push, CONWIP and CONLOAD) applied to a distributed shifting bottleneck heuristic (DSBH) approach.