行政院國家科學委員會專題研究計畫　成果報告

具綁機特性之晶圓代工廠的光罩派工

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-009-083-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學工業工程與管理學系(所)

計畫主持人：巫木誠

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 95 年 8 月 21 日

行政院國家科學委員會補助專題研究計畫 ■ 成果報告
□期中進度報告

具綁機特性之晶圓代工廠的光罩派工

計畫類別：■ 個別型計畫 □ 整合型計畫
計畫編號：NSC 94－2213－E009 －083 －
執行期間： 2005 年 8 月 1 日至 2006 年 7 月 31 日

計畫主持人：巫木誠
共同主持人：
計畫參與人員： 邱紹傑、陳振富、黃友祿、楊桂豐

成果報告類型(依經費核定清單規定繳交)：■ 精簡報告 □完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、
列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學工業工程與管理學系

中華民國 95 年 8 月 18 日

中文摘要

過去光罩派工的研究大部分應用於存貨生產之晶圓廠，該類型的晶圓廠著重產出量的績效表現。本研究針對接單生產之晶圓廠，發展以高達交率為目標之光罩派工法則。此派工法則包含三種基本概念：生產線平衡、避免飢餓以及家族式派工。透過模擬的方式，本研究所提法則與過去方法在不同長短之光罩設置時間下進行比較。實驗的結果顯示本研究所提法則在大部分情況下皆優於過去方法。

Abstract

This paper develops a dispatching algorithm to improve *on-time delivery* for a make-to-order semiconductor wafer fab with two special characteristics: mask setup and machine dedication. A new algorithm is proposed for dispatching series workstations. Simulation experiments show that the algorithm outperforms the previous methods both in on-time delivery rate, cycle time, and only slightly less than the best benchmark in throughput. The experiments are carried out in ten test scenarios, which are created by the combination of two product-mix-ratios and five mask-setup-times.

Keywords: semiconductor dispatching, make-to-order, on-time delivery rate, machine-dedication, mask setup.

1. Introduction

Semiconductor manufacturing is a complex process. Hundreds of operations are required to produce a wafer. A semiconductor factory (also called a fab) typically involves several dozens workstations. A workstation is a group of functionally identical machines that process several operations on the same wafer. A job (also called a lot), which is a cassette that typically carries 25 wafers, may have to enter a workstation several times. Due to the *reentry* characteristics, we may have a great many types of WIP (work-in-process) waiting before a workstation for dispatching—a decision for determining which job should be processed first while a machine is available. The dispatching decision for a semiconductor fab is very important because it could significantly affect the fab performance such as on-time delivery, cycle time and throughput.

Dispatching decisions for a wafer fab have been extensively studied in the literature. Most studies aimed to develop dispatching rules to reduce cycle time and increase throughput (Lu *et al.* 1994, Li *et al.* 1996, Yoon and Lee 2000). Some intend to reduce the tardiness (Lu and Kumar 1991, Kim *et al*. 2001). Some others aimed to improve the on-time delivery, in addition to the improvement of cycle time and throughput (Kim *et al*. 1998b, Lee *et al*. 2002, Dabbas and Fowler 2003). Yet, most previous studies assume that there is no mask setup time for a stepper.

Steppers are very important machines in a fab, which essentially perform the *exposure operations*. An exposure operation is to “photo-print” a circuit pattern onto a wafer by light projection through a mask, which records the circuit pattern. Different exposure operations require different masks. The change of mask on a stepper requires a setup time. *Mask setup* for a stepper should therefore be considered in wafer dispatching. Yet, only a few studies on semiconductor dispatching (Kim *et al.* 1998a, Chern and Liu 2003) concern the mask setups of steppers.

In an up-to-date fab, one of the distinguished features is *machine-dedication*. It demands a job being dedicatedly processed by a particular machine while it enters a workstation. That is, machines in a workstation cannot be taken as *exactly* identical for some critical operations. With the *machine-dedication* characteristic, the capacity of a workstation would be reduced because its machines cannot mutually support in capacity when processing the critical operations. Imposing a significant constraint on workstation capacity, the machine-dedication characteristic is therefore indispensable while developing dispatching decisions for an up-to-date fab.

Steppers in a fab can be categorized into two types: *high-resolution* and *low-resolution*. In a fab, the machine-dedication constraint is imposed *only* on *high-resolution steppers*. That is, once a wafer has been processed by a high-resolution stepper, its remaining exposure operations have to be processed by the same stepper. Other high-resolution steppers, even with same specification, cannot process the wafer. The purpose of machine-dedication is to ensure good manufacturing quality because any two machines in reality cannot be “completely identical”; there always exist slight differences. A low-resolution stepper workstation, to the contrary, has no machine-dedication feature. Therefore, any two steppers in such a workstation can support each other in capacity.

In the research literature, *mask setup* and *machine dedication* are either only partially dealt with or both dealt with in an “old-technology” context where the setup time is much longer than the state-of-art technology. Kim et al. (1998a) considered mask setup but ignored machine-dedication. Chern and Liu (2003) considered both mask setup and machine dedication. However, their work was developed for a relatively old-technology fab, in which each mask setup time takes about 6 minutes while an up-to-date stepper in 2005 takes only about 1.5 min. Significant change in setup time may affect the performance of dispatching policies. Therefore, the performances of dispatching algorithms should be compared under various mask setup times.

In Chern and Lius’ work (2003), their algorithm was essentially developed for *make-to-stock* (MTS) fabs, which are with *high-volume* and *low-variety* characteristics, such as those making DRAM (dynamic random access memory). In contrast, the products manufactured by *make-to-order* (MTO) fabs are usually with *low-volume* and *high-variety* characteristics. The main performance of an MTS fab is *throughput*, while that of a MTO fab is *on-time-delivery*. Therefore, dispatching algorithms that are effective for an MTS fab may not perform as well in an MTO fab.

Due to the *high-volume* and *low-variety* characteristics, an MTS fab is usually equipped with *multiple-masks* for each exposure operation. In contrast, an MTO fab, with *low-volume* and *high-variety* characteristics, usually adopts *one-mask* policy in order to reduce manufacturing cost. In comparing the performance of dispatching algorithms for an MTO fab, we have to adopt one-mask policy instead of multiple-task policy.

Considering the requirement of mask setup, this research aims to develop dispatching methods for a fab that has the following two features: *make-to-order* and *machine-dedication*. Three performance metrics are considered, which involve *on-time delivery rate*, *throughput* and *cycle time*. Of these three, MTO fabs are most concerned with *on-time delivery* in order to retain or attract customers. A dispatching algorithm *LBSA-F* that utilizes ideas of line-balance (LB), starvation avoidance (SA), and family-based dispatching (F) has been developed.

Simulation experiments based on the data provided by a real MTO fab are performed to evaluate the proposed algorithms. In the simulation experiments, one-mask policy is adopted to reflect the characteristic of MTO fabs. And to justify the robustness of LBSA-F, scenarios with various mask setup times are compared.

Four benchmark algorithms are used to compare with the LBSA-F algorithm. These four include the LBSA-I (line-balance starvation-avoidance individual-based) algorithm, the SDA-F algorithm by Chern and Liu (2003), the LWL-F algorithm by Kim et al (1998a), and the FCFS-F (first-come-first-serve family-based) algorithm. Results show that the LBSA-F algorithm outperforms the four benchmarks in terms of on-time delivery and cycle time; and is only slightly less than the best benchmark in terms of throughput.

The remainder of this paper is organized as follows. Section 2 describes the two dispatching decisions that we focus on. Section 3 presents the proposed dispatching algorithm *LBSA-F*. Simulation experiment results are given in Section 4. The reasons why LBSA-F would perform well are discussed in Section 5. Concluding remarks are presented in the last section.

2. Research problem

The various decisions associated with the shop floor control of a fab are first described. Among these decisions, only two—the *dispatching of dedicated and non-dedicated workstations* are investigated in this research. The other decisions, not a focus of this research, are dealt with by some existing methods in our simulation experiments.

2.1. *Releasing Decisions*

Releasing decisions are to determine *when to release* a job to a fab, and determine *which job to release*. Methods for releasing decisions can be classified into two types: open-loop control and closed-loop control. Open-loop control denotes that a job is released to a fab based on a predetermined schedule, which is independent of the current status of the fab. Uniform releasing policy, a typical method of open-loop control, releases jobs "uniformly" (Glassey and Resende 1988a). That is, the release rate and release pattern on each day is identical. Closed-loop control denotes that the time when a job is released depends on the current WIP status of the fab. Along the line of closed-loop control, Glassey and Resende (1988b) developed a starvation avoidance (SA) algorithm; Wein (1988) developed a workload regulation (WR) algorithm; Bechte (1988) used a queueing model to compute the WIP threshold for releasing new jobs; Spearman *et al.* (1990) proposed a CONWIP (constant WIP) method. This research adopts the uniform releasing policies in the simulation experiments.

In a fab with *machine-dedication,* at the time point of releasing a job, a *stepper-assignment decision* must be made. That is, the job should be assigned to a *high-resolution* stepper for processing the critical exposure operations of the job. In this research, the decision is based on the *accumulated load* of each *high-resolution* stepper. That is, at a job releasing time point, the job to be released is assigned to the high-resolution stepper that is the lowest in terms of accumulated load. The main idea of this *stepper-assignment* decision is to keep each high-resolution stepper balanced in load from a *long-term* perspective.

The stepper-assignment or machine-assignment decision can also be intricately formulated as a linear programming (LP) program if the cycle time between any two subsequent operations on a high-resolution stepper is certain and available (Gamila and Motavalli 2003, Liaw 2004). However, the cycle time in a MTO fab is usually with stochastic behavior; the adoption of such LP formulations needs to be further justified. In the simulation experiments, we adopt the heuristic of balancing accumulated-load because it is widely used in practice.

2.2. *Dispatching Decisions*

Dispatching is to determine which one to process among the jobs waiting before a workstation. Different types of workstations need various dispatching methods. Workstations in a fab in general can be classified into two types: batch workstation, and series workstation. A *batch machine* processes several jobs at a time; for example, a furnace machine may process six jobs (150 wafers) simultaneously to reduce processing cost. In contrast, a *series machine* (e.g., a stepper machine) processes one wafer at a time until all the wafers in the job has been completed.

Many algorithms for the dispatching of batch workstations have been published (Weng and Leachman 1993, Kim *el at.* 1998a). Among these, the most commonly used one in industry is the minimum batch size (MBS) method. The MBS method denotes that the batch size (the number of jobs simultaneously processed) should exceed a predefined threshold, which can be determined by a queuing model (Neuts 1967, Phojanamongkolkij *et al.* 2002). While two or more batches meet the MBS threshold, the first-in-first-out (FIFO) rule is applied to break the tie in determining dispatching priorities.

High-resolution steppers are usually the *bottleneck* of a fab because they are very expensive and relatively limited in quantity. In a fab, only the high-resolution steppers have machine-dedication feature, while the others (either series or batch machines) do not have the characteristics. Since high-resolution steppers are a type of series machines, we therefore classify the *series workstations* into two types: *dedicated* and *non-dedicated*. A typical dedicated workstation includes several high-resolution steppers, which are accommodated in a particular area but cannot support each other in capacity due to the constraint imposed by machine-dedication.

This research focuses on developing the dispatching algorithms for two types of *series workstations*, by assuming that the MBS dispatching algorithm has been applied to the *batch workstation*. The main objective is to maximize the *on-time delivery*, and the other two performance criteria are *throughput* and *cycle time*. A semiconductor product (also called IC, integrated circuit) is a component of a consumer product such as cell phone and computer. Late delivery of *make-to-order* ICs would postpone the delivery of the consumer product, whose assembly needs many other components. As a result, the effect of IC delay would be amplified and lead to an abundant increase in the inventory of non-IC components. Therefore, on-time delivery is the most concerned objective in this research.

3. Dispatching algorithms

As stated, series workstations involve two types: *dedicated* and *non-dedicated*. The proposed dispatching algorithm for each type is presented, where a series workstation may be simply called a workstation in short.

3.1 *Dispatching for Dedicated Workstation*

A dedicated workstation involves only high-resolution steppers, which require a mask in processing an operation. Different operations require different masks. A group of jobs that use the same mask is called a *job-family*. At a time that needs to change mask, two dispatching decisions have to be made: (1) *choosing a job-family,* and (2) *prioritizing jobs in the chosen job-family.*

Research on mask change involves two main approaches—*individual-based* and *family-based* (Chern and Liu 2003). The *family-based* approach tends to keep processing the same job-family. That is, the current mask will not be changed unless it has no job to process. In contrast, the *individual-based* approach requests that a mask-changing decision must be made whenever an operation is completed.

Adopting the family-based approach, this research develops a *line-balanced (LB)* method (Ignall 1965, Yamada and Matsui 2003) in dispatching job-families. In the LB approach, the process route is decomposed into many *process segments*. One each process segment, its last operation is processed by the dedicated workstation while the others are by non-dedicated workstation (Fig. 1).

The fab of interest produces a single product family that involves *I* products. Each product, with the same process route but different in operation times, has *J* segments. Whenever a mask needs to be changed, the number of job-families to be chosen is (*I***J*)-1.

The procedure for dispatching a dedicated-stepper is described below. To undergo the procedure, a pre-simulation has to be performed to determine CT_{ij}, which denotes the mean cycle time required to complete *all* the operations in segment *j* of product *i*. The estimation of CT_{ij} in the simulation assumes that the fab adopts the FCFS-F (first-come-first-serve family-based) dispatching algorithm.

Procedure *Dispatching_ Dedicated_Workstations*

Step 1: Compute the flow rate (ν_{ij}) for each job-family as below, where WIP_{ij} denotes the number of jobs for the job-family of product *i* at segment *j*.

$$\nu_{ij} = \frac{WIP_{ij}}{CT_{ij}}$$

Step 2: Compute the normalized flow rate (λ_{ij}) as below, where R_i denotes the ratio of release rate (jobs per unit of time) for product *i*.

$$\lambda_{ij} = \frac{\nu_{ij}}{R_i} = \frac{WIP_{ij}}{CT_{ij} \cdot R_i}$$

Step 3: Select the job-family that has the maximum normalized flow rate.

$$(i^*, j^*) = Arg\ max(\lambda_{ij})$$

Step 4: Use CR (critical ratio) rule to prioritize the jobs in the selected family.

The main idea of the above procedure is *line-balancing*. Consider an *ideally* line-balanced production line where the flow rate (jobs per day) of each product on each segment (ν_{ij}) can be so well controlled that it always equals its release rate R_i. Then the fab output rate equals the release rate R_i. That is, in the ideally line-balanced case, $\lambda_{ij} = \frac{\nu_{ij}}{R_i} = 1$ for each i and j. The deviation of λ_{ij} from 1 indicates the degree of unbalance for product i on segment j.

While ideally line-balanced, the standard WIP level for segment j of product i is $Std\_WIP_{ij} = CT_{ij} \cdot R_i$. That is, $\lambda_{ij} = \frac{WIP_{ij}}{CT_{ij} \cdot R_i} = \frac{WIP_{ij}}{Std\_WIP_{ij}}$. The job-family with the highest λ_{ij} should be first processed in order to smooth the WIP distributions among segments and head for line balancing.

In Step 4, to maximize on-time delivery rate, we use CR (critical ratio) method to prioritize the jobs in the selected job-family. The CR of a job denotes the ratio of its remaining time over its remaining processing time, which is intended to measure the possibility of on-time delivery. The lower the CR value, the lower is the possibility of on-time delivery and should be processed first. Other dispatching rule such as SRPT (shortest remaining processing time) might be a good heuristic for other performance criteria such as throughput (Walrand 1988). However, this research concerns more on on-time delivery; therefore CR is proposed.

3.2. *Dispatching for Non-dedicated Workstations*

For a non-dedicated workstation, its number of job-families can be $I * J * K$, where I denotes the number of products, J denotes the number of route segments, and K denotes the number of dedicated-steppers. Likewise, there are two decisions for the dispatching of non-dedicated workstations: (1) choosing job-family, and (2) prioritizing the jobs for the chosen job-family.

This research uses the concept of starvation avoidance (SA) (Glassey and Resende 1988b) to choose the job-family. As stated, the dedicated-steppers are bottleneck (Lee 2002) in a fab; therefore, it is important to supply enough jobs to each dedicated-stepper to prevent it from being starved.

The procedure for dispatching non-dedicated workstations is presented below, where N denotes the workstation for making the dispatching decision and D denotes the dedicated-stepper workstation (Figure 1). To undergo the procedure, a pre-simulation has to be carried out in order to determine CT_{ijk}, which denotes the mean cycle time of the job-family (product i on segment j assigned to dedicated-stepper k) from workstations N to D.

Procedure *Dispatching_Non-dedicated_ Workstations*

Step 1: Compute the flow rate (ν_{ijk}) for each job-family as below, where WIP_{ijk} denotes the WIP level of the job-family (product i on segment j assigned to dedicated-stepper k) from workstations N to D.

$$\nu_{ijk} = \frac{WIP_{ijk}}{CT_{ijk}}$$

Step 2: Compute the *normalized* flow rate (λ_{ijk}) as below where R_i denotes the ratio of release rate for product i.

$$\lambda_{ijk} = \nu_{ijk} / R_i ,$$

Step 3: Select the job-family that has the minimum normalized flow rate.

$$(i^*, j^*, k^*) = Arg\min(\lambda_{ijk})$$

Step 4: Use CR (critical ratio) to prioritize the jobs in the selected job-family.

3.3. *Comparison of the Two Dispatching Algorithms*

Of the above two dispatching algorithms, the one for *dedicated-steppers* is to balance the *throughput among segments*, and is called a *line-balancing* (LB) dispatching. The other one, for non-dedicated workstations, is to prevent dedicated-steppers from being “starved”, and is called a *starvation-avoidance* (SA) dispatching.

The LB dispatching is designed from the perspective of controlling the *output mix of* *job-families* that leave from dedicated-steppers (bottlenecks). In contrast, the SA dispatching is designed from the perspective of controlling the *input mix of job-families* that arrive to dedicated-steppers. The output control aims to produce a product at a rate as close as possible to its release rate. The input control aims to provide enough WIPs to dedicated-steppers for them to effectively realize the output control.

4. Simulation experiments

4.1. *Data, Assumptions, and Benchmarks*

The proposed dispatching algorithms are compared with four benchmark methods by discrete-event simulation. The test data of process route and processing times are provided by an MTO fab in industry. The fab involves 60 workstations, of which 9 are batch workstations and 51 are series workstations, and the workstations in total involve 262 machines. The MTBF (mean time between failure) and MTTR (mean time to repair) of machines are also available, with the assumption of exponential distributions.

The fab—an MTO fab adopts the one-mask policy. A single product family, involving five logical products, is produced. Each product has the same process route, which involves 12 segments and 344 operations (Table 1). Taking a particular product as a standard, the processing time of the other four products is modeled by multiplying the standard by a uniform distribution, UNIF(0.95, 1.05). The exposure operation time for a lot (25 wafers) is 1.66 hours = 100 min.

<< Insert Table 1 about here>>

Each job or lot has 25 wafers. The due date of lot *k* is defined by $\delta_k = a_k + u \cdot pt_k$, where δ_k denotes the due date, a_k denotes the release time, and pt_k denotes the total processing time, and *u* denotes a scale factor for defining due date (Kim *et al.* 1998b). Note that $u \cdot pt_k$ is also called *committed cycle time*, which indicates the cycle time committed to customers. Suppose the *production cycle time* of a lot is longer than $u \cdot pt_k$, the lot will be late. Since the fab delivers the wafer lots to customer once a day, δ_k (in unit of day) is rounded up to an integer.

Three performance metrics, *on-time delivery rate, throughput*, and *mean cycle time* are to be compared. Of the three, *on-time delivery rate* is most critical for a competitive MTO fab to retain or attract customers. Adopting a uniform-releasing policy, the fab releases 31 lots per day in total.

Two product mix ratios, which are described by R_A = (1:1:1:1:1) and R_B = (1:2:3:1:2), are used to evaluate the dispatching algorithms. Five cases of mask setup time (*s* = 0, 30, 90, 180, 360, in unit of sec.) are evaluated, where *s* = 90 is the current practice of an up-to-date fab. In total, ten cases—the combination of two product mixes and five mask setup times are tested.

Each simulation experiment is performed with 20 runs; each run is with a different random seed. The time horizon for a simulation run is 270 days; the first 90 days is taken as "warm-up" time because after which the WIP and throughput have reached a steady state (Fig. 2). The output data of the subsequent 180 days is collected for analysis. Simulation programs, coded in eM-plant (http://www.tecnomatix.com), are run on a personal computer equipped with AMD-3000$^+$ CPU.

The proposed dispatching algorithm is designated as *LBSA-F*, where *LB* denotes line-balance, *SA* denotes starvation-avoidance, and *F* denotes family-based approach for mask dispatching. Four algorithms are compared with the LBSA-F. The first one—*LBSA-I*, also developed by us, is a variation of *LBSA-F*, with *I* denoting the use of the individual-based approach for mask dispatching. The second one—*SDA-F* denotes the algorithm proposed by Chern and Liu (2003). The third one—*FCFS-F*, widely used in industry, denotes the first-come-first-serve algorithm with family-based approached for mask dispatching. The fourth one—*LWL-F* (Loop Workload Leveling family-based) denotes the algorithm proposed by Kim *et al.* (1998a). A comparison of these algorithms is summarized in Table 2.

4.2 *Experiment Results for s= 90*

As stated, the experiments involves five mask setup time (*s* = 0, 30, 90, 180, 360 in unit of sec.). The case of *s* = 90 is most concerned because it is the current practice of an up-to-date fab. Experiment results for the two product mixes (R_A and R_B) at *s* = 90 are analyzed below.

Table 3 shows the mean and standard deviation of each performance metric under various dispatching algorithms. An analysis of variance (ANOVA) is carried out to justify the effects of the dispatching rules (Montgomery 1991). The ANOVA results (Table 4) showed that the dispatching rules had significant effect on each performance metric (at the significance level of 0.01) in each product mixes. The Duncan's multiple range tests were also performed to categorize the dispatching rules based on their performances and the results are given in Table 5.

<< Insert Table 3 about here>>

<< Insert Table 4 about here>>

<< Insert Table 5 about here>>

From these results, we could conclude that LBSA-F outperforms the four benchmarks in terms of *on-time-delivery rate* and *cycle time*, in each product mixes (R_A and R_B). Yet, in terms of throughput, LBSA-F performs the best in product mix R_A while ranks the third in product mix R_B.

The reason why LBSA-F in R_B does not perform as well as that in R_A is analyzed below. Comparing to R_A = (1:1:1:1:1), the production volume of each product in R_B = (1:2:3:1:2) is less uniform. Using the normalized flow rate (λ_{ij}) as the main dispatching criterion, the LBSA-F tends to make the on-time-delivery rate of each product as close as possible. Therefore, in dealing with *small-volume* products, masks have to be changed more frequently. This implies the increase of total mask setup time, which consequently leads to the decrease of bottleneck utilization and fab throughput. The above analysis is supported by the experiment results of LBSA-F, which indicated that the average utilization of dedicated-stepper in R_A is 99.38% while that in R_B is 99.25%.

Comparing with LBSA-I, LBSA-F performs better in each performance metric. This finding seems reasonable because the LBSA-I, an individual-based algorithm, tends to change mask more frequently and consequently reduce throughput. Since both LBSA-F and LBSA-I use the normalized flow rate (λ_{ij}) as the main dispatching criterion, the reduction of throughput in LBSA-I tends to reduce its on-time-delivery. This finding indicates that *s* = 90 sec. is a substantial amount in terms of mask setup, and cannot be ignored in developing dispatching algorithms.

4.3 *Experiment Results for Various Cases of s*

Over the years, the mask setup time has been progressively reduced due to the advance of technology. To justify the performance of *LBSA-F* algorithm in various fabs, from *a traditional fab* to a *future one*, simulation experiments are performed for ten test cases, which are the combination of five mask setup times (*s* = 0, 30, 90, 180, 360) and two product mixes (R_A and R_B).

Figure 3 and Figure 4 respectively show the experiment results in product mix R_A and R_B. In the two figures, the performance of FCFS-F is taken as a *baseline* for comparison. That is, the performance difference between a dispatching algorithm and FCFS-F is revealed in the figures. The trends of the two figures appear quite consistent. Therefore, we refer to Figure 3 in analyzing the experiment results.

Part (a) in Figure 3 indicates that, in terms of on-time-delivery rate, *LBSA-F* outperforms the other four algorithms for each *s*. The comparison between LBSA-F and SDA-F indicates that their difference in performance is increased while *s* becomes smaller. It reveals that SDA-F performs well in the case of *s* = 360 but no so while *s* = 90 or smaller. Seemingly, the smaller the mask setup time, the higher is their performance differences. This indicates that the variation of mask setup time indeed affects the performance of dispatching algorithms and cannot be ignored.

Part (b) in Figure 3 reveals that *LBSA-F* also outperforms the other four algorithms for each *s*, in terms of mean cycle time. In the experiments, the due date of each job has been predetermined. Therefore, the shorter the production cycle time, the higher is the on-time-delivery rate. The finding about on-time-delivery and that about cycle time appear quite consistent.

Part (c) in Figure 3 reveals that LBSA-F performs well for each *s*, in terms of throughput. The performance of LBSA-F only slightly differs from the best benchmark in each *s*. As shown in the figure, the throughput of LBSA-I performs well in the cases of *s* = 0, 30, 90, but drops significantly in the cases of *s* = 180 and 360. This implies that family-based dispatching algorithms are preferred in the cases of requiring long mask setup time.

5. **Discussion**

As stated, the fab of interest involves two distinguished features: *make-to-order,* and *machine-dedication*. We attempt to explain why LBSA-F would perform well in such a fab.

The *make-to-order* feature would lead to *a high variety of job-families* waiting before the bottleneck (a dedicated-stepper). In the test examples, the process route contains 12 segments in which 11 segments involve dedicated-operations (Table 1). That is, a dedicated-stepper has to process a product 11 times with 5 products simultaneously produced. This implies that 55 types of job-family would be waiting before a dedicated-stepper. In practice, this number could be ten times bigger.

The *machine-dedication* feature would lead to a *significant reductio*n of *the total WIPs* waiting before a dedicated-stepper. Consider a workstation that involves 11 steppers and is having Q jobs waiting for processing. If the steppers are non-dedicated, the total WIPs available for a particular stepper is Q. While the steppers become dedicated, the total WIPs available for a particular stepper on average reduces to $Q/11$.

The above analysis indicates that a *make-to-order* fab with *machine-dedication* feature would yield such a result—the WIPs waiting before a dedicated-stepper are with *high-variety* and *low-volume* characteristics. The characteristics also hold for the *non-dedicated workstations* on the upstream of the dedicated-stepper. By contrast, in the case of *make-to-stock* fabs without *machine-dedication* feature, the WIPs waiting before a workstation are relatively with *low-variety* and *high-volume* characteristics.

The main performance metric for a make-to-order fab is *on-time-delivery*. To maximize on-time-delivery, LBSA-F attempts to smooth *the normalized flow rate of each job-family* at each segment. That is, each segment of a product is urged to have the same *output rate*, preferably as close to its *release rate* as possible. This tends to reduce the segment flow rate variation, which leads to the reduction of output rate variation, and consequently increases the on-time delivery.

In LBSA-F, the dispatching for *non-dedicated* workstations has considered the *downstream machine-dedication constraint*. By contrast, most previous algorithms ignored this constraint and tend to render the WIP profile of the dedicated-stepper unbalanced. As a result, their performances in on-time-delivery would be reduced.

6. Concluding remarks

This research considers the requirement of mask setup and develops dispatching algorithms for a make-to-order fab with machine-dedication feature. The dispatching algorithms are evaluated by three performance metrics: on-time-delivery rate, cycle time, and throughput. Of the three, on-time-delivery rate is most critical to a make-to-order fab in order to retain and attract customers.

We proposed a dispatching algorithm—*LBSA-F*, which uses the idea *line-balancing* (LB) to control the output pattern of bottleneck, the idea of *starvation avoidance* (SA) to control the input pattern of bottleneck, and the idea of *family-based* in mask dispatching.

Simulation experiments have been performed in ten test cases that are the combination of two product mixes and five mask setup times. Four benchmark algorithms are used to compare with *LBSA-F*. Experiment results show that *LBSA-F* outperforms the four benchmarks both in on-time delivery rate and cycle time, and is slightly less than the best benchmark in throughput.

Some extensions of this research may be investigated. First, the effects of other shop floor control decisions on *LBSA-F* may be examined. These shop control decisions include the determination of job releasing time, the assignment of jobs to dedicated-machines at the time of job releasing, and the dispatching of batch machines. Second, dispatching algorithms for fabs with machine-dedication features in the context of producing both MTO and MTS products may be studied. As stated, the main performance metric for MTO is on-time delivery and that for MTS is throughput. At such a hybrid production environment, the two performance metrics are both very important. To ensure a good performance in each performance metric, the dispatching priorities between MTO and MTS products may have to be dynamically determined. This initiates a need for enhancing LBSA-F to perform well in such a hybrid product environment.

Acknowledgement

The authors are grateful to Taiwan Semiconductor Manufacturing Corp. for providing data for simulation experiments. This research is partially supported by a research project NSC90-2622-E009-001.

References

Bechte, W., 1988, Theory and Practice of Load-Oriented Manufacturing Control. *International Journal of Production Research*, **26**, 375-395.

Chern, C. C. and Liu, Y. L., 2003, Family-based scheduling rules of a sequence -dependent wafer fabrication system. *IEEE Transactions on Semiconductor Manufacturing*, **16**(1), 15-25.

Dabbas, R. M. and Fowler, J. W., 2003, A new scheduling approach using combined dispatching criteria in wafer Fabs. *IEEE Transactions on Semiconductor Manufacturing*, **16**(3), 501-510.

Gamila, M. A. and Motavalli, S., 2003, A modeling technique for loading and scheduling problems in FMS. *Robotics and Computer Integrated Manufacturing*, **19**, 45-54.

Glassey, C. R. and Resende, M. G. C., 1988a, A scheduling rule for job release in semiconductor fabrication. *Operations Research Letters*, **7**(5), 213-217.

Glassey, C. R. and Resende, M. G. C., 1988b, Closed-loop Jop shop release control for VLSI circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, **1**(1), 26-46.

Ignall, E. J., 1965, A review of assembly line balancing. *The Journal of Industrial Engineering*, **16**(4), 244-254.

Lee, Y. H., Park, J. and Kim, S., 2002, Experimental study on input and bottleneck scheduling for a semiconductor fabrication line. *IEEE Transactions on Semiconductor Manufacturing*, **34**(2), 179-190.

Li, S., Tang, T., and Collins, D. W., 1996, Minimum inventory variability schedule with applications in semiconductor fabrication. *IEEE Transactions on Semiconductor Manufacturing*, **9**(1), 145-149.

Liaw, C. F., 2004, Scheduling two-machine preemptive open shops to minimize total completion time. *Computers & Research*, **31**, 1349-1363.

Lu, S. H. and Kumar, P. R., 1991, Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, **36**(12), 1406–1416.

Lu, S. C. H., Ramaswamy, D. and Kumar, P. R., 1994, Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. *IEEE Transactions on Semiconductor Manufacturing*, **7**(3), 374-388.

Kim, Y. D., Lee, D. H. and Kim, J. U., 1998a, A simulation study on lot release control, mask scheduling, and batch scheduling in semiconductor wafer fabrication facilities. *Journal of Manufacturing Systems*, **17**(2), 107–117.

Kim, Y. D., Kim, J. U., Lim, S. K. and Jun, H. B., 1998b, Due-date based scheduling and control policies in a multiproduct semiconductor wafer fabrication facility. *IEEE Transactions on Semiconductor Manufacturing*, **11**(1), 155-164.

Kim, Y. D., Kim, J. G., Choi, B. and Kim, H. U., 2001, Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates. *IEEE Transaction on Robotics and Automation*, **17**(5), 589-598.

Montgomery, D. C., 1991, *Design and Analysis of Experiments* (New York: Wiley).

Neuts, M. F., 1967, A general class of bulk queue with Poisson input. *Ann. Math. Stat.*, **38**, 759–770.

Phojanamongkolkij, N., Fowler, J. W. and Cochran, J. K., 2002, Determining Operating Criterion of Batch Processing Operations for Wafer Fabrication. *Journal of Manufacturing Systems*, **21**(5), 363-379.

Spearman, M. L., Dabid, L. W. and Wallace, J. H., 1990, CONWIP: A Pull Alternative to Kanban. *International journal of Production Research*, **28**(5), 879-894.

Walrand, J., 1988, *An introduction to queueing networks*. (Reading: Prentic Hall).

Wein, L. M., 1988, Scheduling semiconductor wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing*, **1**, 115-130.

Weng, W. W. and Leachman, R. C., 1993, An Improved Methodology for Real-Time Production Decisions at Batch-Process Work Station. *IEEE Transaction on semiconductor manufacturing*, **6**(3), 219-225.

Yamada, T. and Matsui, M., 2003, A management design approach to assembly line systems. *Int. J. Production Economics*, **84**, 193-204.

Yoon, H. J. and Lee, D. Y., 2000, A control method to reduce standard deviation of flow time in wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing*, **13**(3), 389-392.

Http://www.tecnomatix.com

Table 1: Process route and processing times of the test fab.

Segment	Number of Operations	Processing Time (Hours)
Seg 1	12	27.97
Seg 2	67	87.91
Seg 3	89	78.15
Segt 4	19	17.81
Seg 5	15	11.23
Seg 6	19	19.74
Seg 7	15	11.23
Segt 8	19	19.74
Segt 9	15	11.23
Seg 10	19	19.74
Seg 11	15	11.23
Seg 12	40	39.09
Total	344	355.07

Table 2: A comparison of dispatching algorithms

Dispatching algorithm	Dedicated workstation	Non-dedicated workstation	
	Steppers	Steppers	Non-steppers
LBSA-F	LB-F	SA-F	SA
LBSA-I	LB-I	SA-I	SA
SDA-F	SDA-F	SDA-F	FCFS
FCFS-F	FCFS-F	FCFS-F	FCFS
LWL-F	LWL-F	LWL-F	FCFS

Table 3: Experiment results for s = 90 sec. (a) product mix R_A , (b) product mix R_B

(a)

R_A						
Dispatching Algorithm	On-time delivery rate		Cycle Time		Throughput	
	Mean	St. dev.	Mean (day)	St. dev (day)	Mean (lot)	St. dev. (lot)
LBSA-F	88%	3.2%	23.4	1.3	5523.4	9.6
LBSA-I	78%	2.4%	23.8	2.7	5507.0	11.3
SDA-F	9%	1.8%	28.8	2.8	5341.9	27.3
FCFS-F	48%	10.1%	25.2	1.4	5504.7	12.9
LWL-F	49%	6.4%	25.3	1.9	5520.2	13.2

(b)

R_B						
Dispatching Algorithm	On-time delivery rate		Cycle Time		Throughput	
	Mean	St. dev.	Mean (day)	St. dev (day)	Mean (lot)	St. dev. (lot)
LBSA-F	89%	2.7%	23.4	1.3	5512.9	8.0
LBSA-I	82%	1.3%	23.6	3.6	5507.5	9.3
SDA-F	21%	4.6%	27.6	2.9	5338.9	14.7
FCFS-F	70%	6.0%	24.5	1.1	5541.7	10.9
LWL-F	55%	5.4%	25.0	1.7	5533.1	17.4

Table 4: ANOVA for s = 90sec. (a) product mix R_A, (b) product mix R_B

Throughput					
	SS	Deg. of	MS	F	p
Dispatching Rules	4.78E+05	4	1.20E+05	457	0
Error	2.48E+04	95	2.62E+02		
Cycle Time					
	SS	Deg. of	MS	F	p
Dispatching Rules	360.74	4	90.18	1418	0
Error	6.04	95	0.06		
On-time delivery rate					
	SS	Deg. of	MS	F	p
Dispatching Rules	7.62618	4	1.90655	585.761	0
Error	0.30921	95	0.00325		

(a)

Throughput					
	SS	Degr. of	MS	F	p
Dispatching Rules	5.63E+05	4	1.41E+05	892	0
Error	1.50E+04	95	1.58E+02		
Cycle Time					
	SS	Degr. of	MS	F	p
Dispatching Rules	229.79	4	57.45	1070	0
Error	5.1	95	0.05		
On-time delivery rate					
	SS	Degr. of	MS	F	p
Dispatching Rules	5.71755	4	1.42939	743.15	0
Error	0.18272	95	0.00192		

(b)

Table 5: Duncan's multiple range test for s = 90:
(a) product mix R_A, (b) product mix R_B

Rule	Throughput	Results	Rule	CT	Results	Rule	On-time delivery rate	Results
LBSA-F	5523.400	A	LBSA-F	23.4481	A	LBSA-F	0.877284	A
LWL-F	5520.200	A	LBSA-I	23.8143	B	LBSA-I	0.780176	B
LBSA-I	5507.000	B	FCFS-F	25.1950	C	LWL-F	0.492436	C
FCFS-F	5504.700	B	LWL-F	25.2575	C	FCFS-F	0.476221	C
SDA-F	5341.850	C	SDA-F	28.8196	D	SDA-F	0.088803	D

(a)

Rule	Throughput	Results	Rule	CT	Results	Rule	On-time delivery rate	Results
FCFS-F	5541.700	A	LBSA-F	23.40683	A	LBSA-F	0.888542	A
LWL-F	5533.100	B	LBSA-I	23.58496	B	LBSA-I	0.815768	B
LBSA-F	5512.900	C	FCFS-F	24.48898	C	FCFS-F	0.700032	C
LBSA-I	5507.500	C	LWL-F	25.00969	D	LWL-F	0.548924	D
SDA-F	5338.900	D	SDA-F	27.61668	E	SDA-F	0.213940	E

(b)

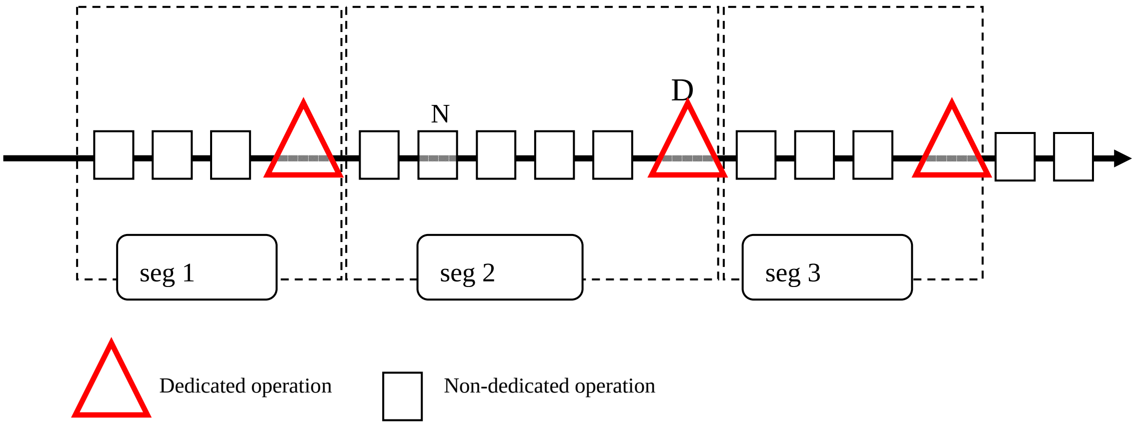


Fig. 1 Segments in a process route

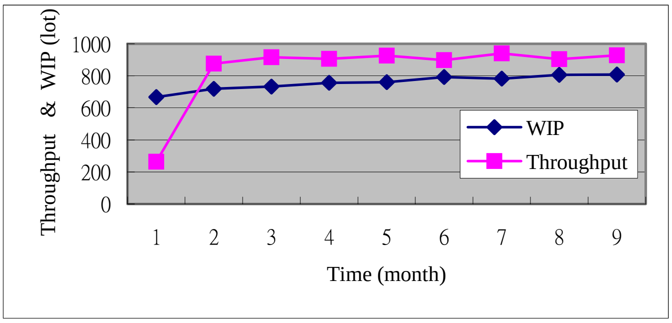
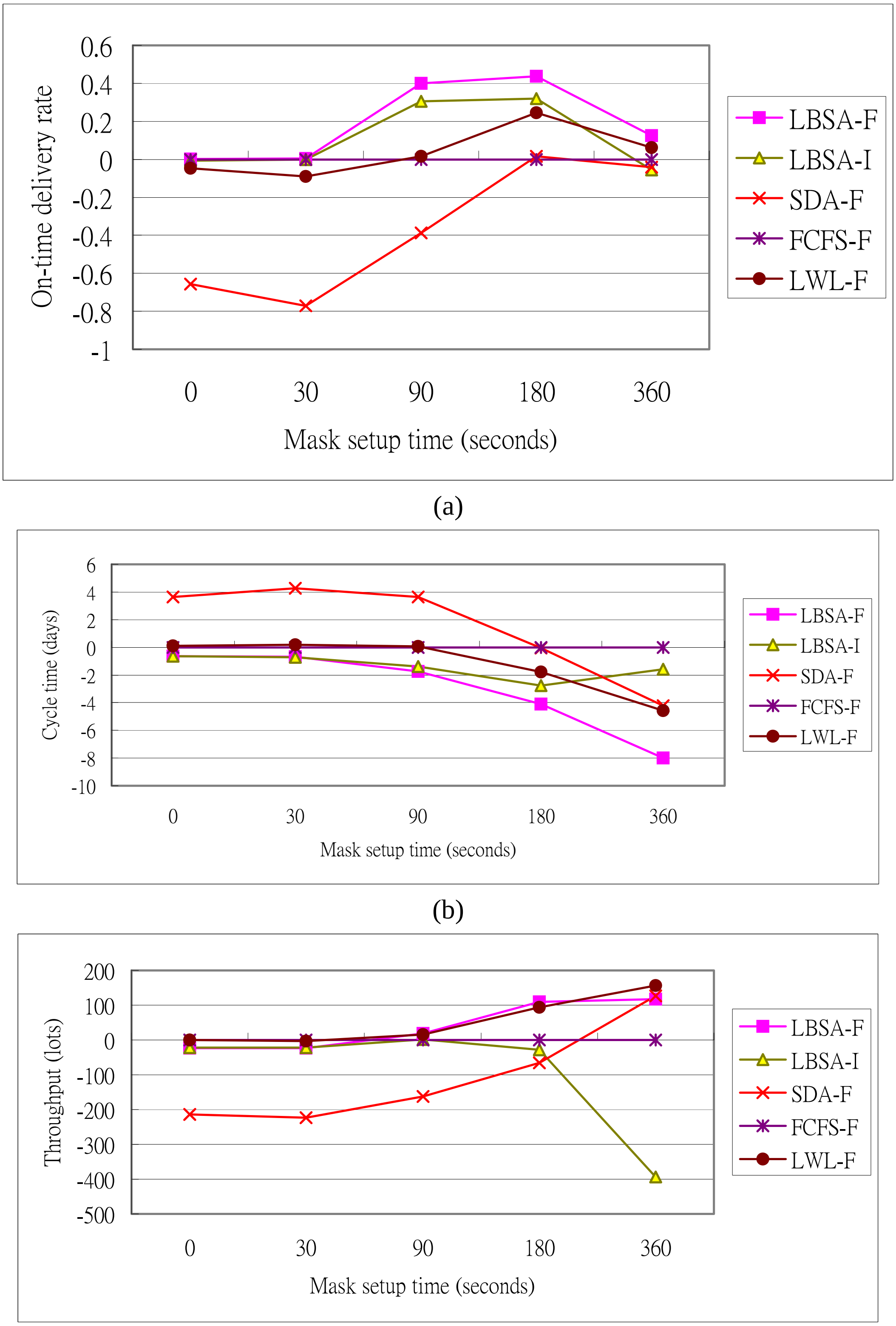


Fig. 2 Time plots of throughput and WIP

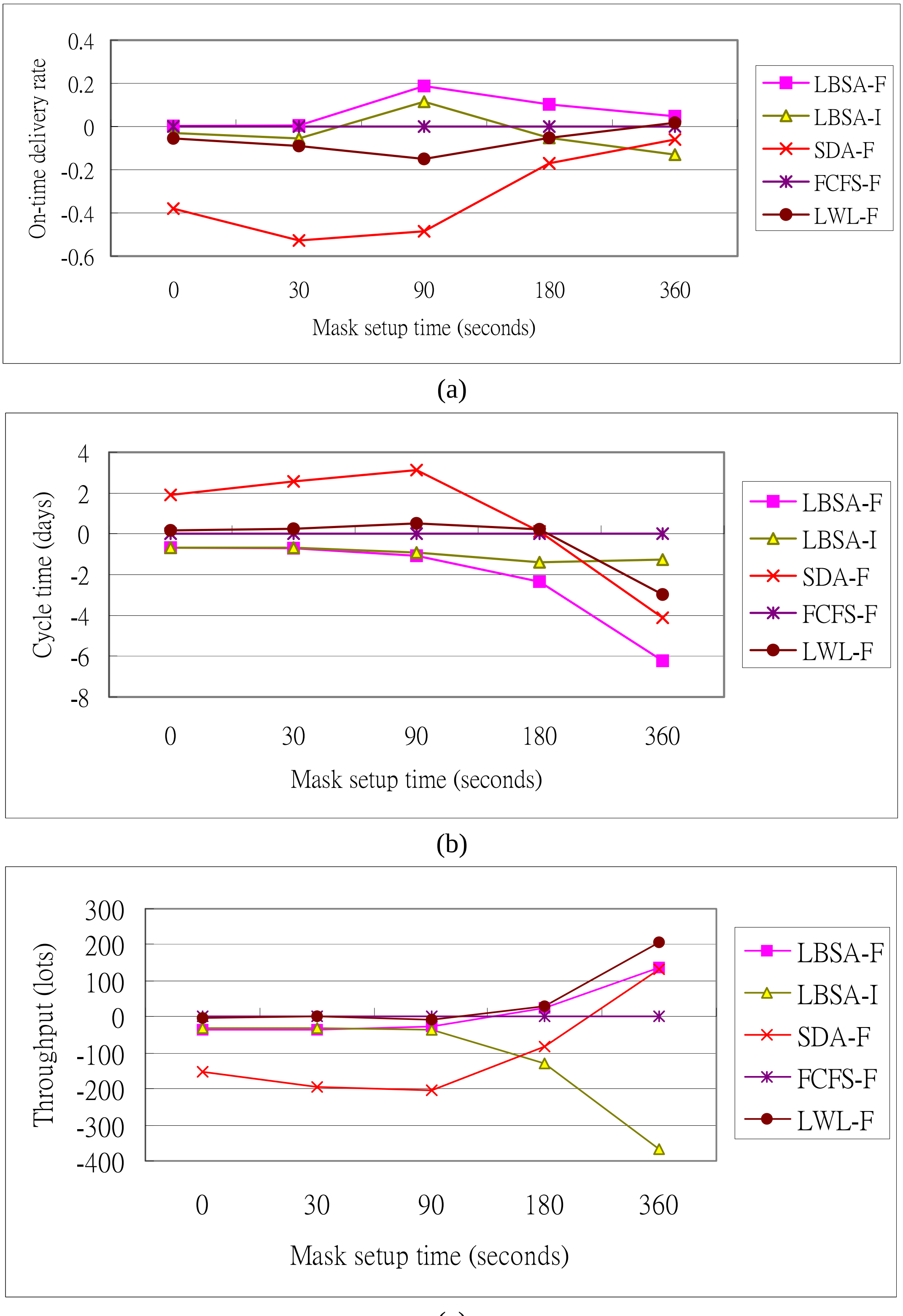


(a)

(b)

(c)

Fig. 3 Performance comparison in various mask setup times, with product mix R_A, (a) on-time delivery rate, (b) cycle time, (c) throughput.



(a)

(b)

(c)

Fig. 4 Performance comparison for various mask setup times, with product mix R_B, (a) on-time delivery rate, (b) cycle time, (c) throughput.

2005/8-2006/7 巫木誠參加學術會議心得

學術會議心得（一）

(1) 基本資料

- 會議名稱：International Conference on Logistics and Supply Chain Management (ICLSCM 2006)
- 大會主題：Logistics Strategies and Technologies for Global Business
- 主辦單位：香港大學工業與製造系統系(Department of Industrial and Manufacturing System Engineering)
- 時間：2006, 1/5-1/7
- 地點：香港大學

(2) 心得：

- Keynote Speech: 講員（劍橋大學教授）提到一個重要的概念，只要製造的供應鏈整合的好，亞太企業仍然可以有世界級的競爭力，他以台灣鴻海公司爲例來說明，專注於核心競爭力的重要。他提出一個新觀點，亞太代工企業並不需一定要急於建立自有品牌，也可先強化核心的製造競爭力，等實力充足時，再購併品牌。他預測鴻海未來有購併國際級品牌的可能性。
- 學術論文：供應鏈的研究用了很多新興的工具，譬如 meta-heuristics (GA, SA, Tabu, Particle Swarm), Data mining, DEA。

學術會議心得（二）

(1) 基本資料

- 會議名稱：2006 INFORMS (Institute for Operations Research and Management sciences) International 2006
- 主辦單位：香港科技大學工業工程與物流管理學系 (Department of Industrial Engineering and Logistics Management)
- 時間：2006, 6/25-6/28
- 地點：Sheraton Hotel & Towers Hong Kong

(2) 心得：

- Keynote Speech: Prof. Hau Lee（史丹佛大學教授）提到二個供應鏈管理新的研究議題：第一：新興國家(emerging countries)的供應鏈管理，新興國家的資訊、運輸架構、文化與已開發國家不同。如何將已開發國家的供應鏈管理議題移植到供應鏈管理。第二：綠色供應鏈：環保越來越重要，如何將環保因素納入供應鏈管理的研究。

- 學術論文：財務工程(financial engineering)在此次會議有一些論文發表，香港科技大學工業工程系也開始聘任財務工程教授，看起來，如何利用 OR/MS 等技術於財務領域，是一個值得注意的發展方向。