行政院國家科學委員會專題研究計畫 成果報告

晶圓製造廠在時間限制下之產能規劃與現場管控決策模式(II)

研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 96-2221-E-216-038-
執行期間：96年08月01日至97年10月31日
執行單位：中華大學工業工程與系統管理學系

計畫主持人：杜瑩美
共同主持人：張盛鴻
計畫參與人員：博士班研究生-兼任助理人員：陳欣男
博士班研究生-兼任助理人員：盧俊偉

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 97年10月23日

行政院國家科學委員會補助專題研究計畫 **V 成果報告**
□期中進度報告

晶圓製造廠在時間限制下之產能規劃與現場管控決策模式
(II)

計畫類別：☑ 個別型計畫　□ 整合型計畫
計畫編號：NSC 96－2221－E－216－038－
執行期間： 96 年 8 月 1 日 至 97 年 10 月 31 日

計畫主持人：杜瑩美　中華大學 工業工程與系統管理學系
共同主持人：張盛鴻　明新科技大學 工業工程與管理學系
計畫參與人員：陳欣男 盧俊偉 中華大學 科技管理研究所

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：中華大學 工業工程與系統管理學系

中 華 民 國 97 年 10 月 23 日

一、 中英文摘要

中文摘要

近年來由於半導體製程日趨精密、生產步驟越來越多，進而延長了在製品在加工等候區中的等候時間。為避免在製品等候加工的時間過長而成為不良品，工程師會在某些加工機台前之等候線設立等候時間限制，藉以確保最終產出的良率。然而，等候時間限制的設立也必定會影響生產線的流暢與產出。

在前一年度的研究計畫中致力於等候時間限制下產能決策模式之研究。因此，本年度之研究將著眼於現場的管控模式方面。雖然在產能決策模式中已加入保護性產能的觀念並將當機模式考量於等候模型中，然而，現場的作業不時地受到不確定之動態因子所干擾。因此，如何在這些不確定因素發生時給予適當的管控處理是不可或缺的。

關於等候時間限制問題下之現場管控方法，吾人認為，其解決方法必須包含下列二個方向，等候時間區段內的在製品水準與機台的投料控制。，本研究提出以存貨理論中的最大存貨水準與最小安全存量的概念，輔以動態批量決策的手法，針對以上各種不同之等候時間限制問題環境提出有效之投料與派工法則，以提供管理者進行現場管控時之依據。

關鍵詞：等候時間限制；晶圓製造；現場管控；安全存量法則

Abstract

Recently, wafer fabrication has become more complicated and lengthening the product queue time. To ensure final product yield, engineers need to set up queue time limits for particular machines. There is no doubt that time constraint will influence the manufacturing efficiency which is a new challenge to wafer fabrication. In such a situation, the overall solution is necessary to overcome these difficulties.

In the previous research, our study focuses on capacity planning under time constraint. The follow-up of our research will concentrate on shop floor control under time constraint. Although the concept of protective capacity has been applied and the queuing model is modified by adding the average machine failure behavior, the uncertainties always attack the processes of shop floor. Therefore, how to manage and control the shop floor under time constraints is indeed very important.

The concept of (S,s) policy will be introduced in our shop floor control rules. Furthermore, there will be three situations of time constraints took into it, a segment of processes, sequential time constraints and batch process. By our rules, the WIP will be protected from over queue time by the upper bond of inventory, and the machine will avoid starvation by lower bond of it.

By the result of our research, the fabs are getting effective and reasonable approaches to solve the problem of time constraints.

Keywords: Time Constraints; wafer fabrication; shop-floor control; [S,s] policy

二、　報告內容

1. 前言

隨著半導體製程日趨進步，半導體產業的作業管理也越來越困難。在進步到 0.13 微米以下製程之後，生產步驟已高達數百道至數千道之多，再加上回流式加工(Reentrant process)的特性，使得生產流程較其他產業複雜許多。此外，由於半導體生產設備價格極為昂貴，使得晶圓廠的產能設置會較為緊縮，再加上煩長且複雜的製程特性，致使生產半成品(WIP)在加工等候線上的等候時間必定會較為延長。過長的等候時間必定會增加晶圓在生產過程中受到污染的機率【1】，導致產品的良率降低、有效產出變少；更嚴重的，因為報廢導致必須重新投料以補足顧客需求，此舉勢必會增加生產週期時間，最後嚴重影響交期。在顧客滿意度至上的今日，無法達到顧客要求的交期，對廠商來說，此結果將嚴重影響其競爭力。有鑑於此，為避免產品因在等候線上等候時間過長而成為不良品，工程師在製造現場會根據良率考量以及個人主觀因素訂下一個時間來做為等候時間(Queue Time, QT)的限制。晶圓如未能在此一時間限制內送達下一加工站內加工，就必須將在製品處以特殊記錄、重工、或直接報廢的處理，而此主觀訂定的時間就稱為等候時間限制(Time constraints, TC)【2】。

對於晶圓製造的管理者而言，必須解決等候時間限制問題是一個既存在且無法避免的事實。由過去文獻中【3、4、5】可發現，面對等候時間限制問題的思考方向可從多方面加以考慮，包括最初的產能規劃與現場管控模式都必須將等候時間限制加以考慮。因此，在前一年度的計畫(NSC 95-2221-E-216-014)中即針對等候時間限制問題下之產能決策提出一套完整的解決模式，吾人透過保護性產能的概念，解決在等候時間限制下批量工作站以及後段製程中的連續型等候時間問題的產能規劃模式。

然而，等候時間限制的影響並不止於產能規劃的層面。即便擁有充足的產能，任何可能對產出造成波動的行為、任何可能影響在製品在等候線中等待時間的決策，都可能使在製品超過時間限制的問題發生。因此，在等候時間限制的影響之下這些因子都必須重新加以考慮，特別是在現場管控方面必須多加琢磨，否則即使有良好的產能規劃亦無法獲得較佳的產出。然而，晶圓製造的排程是非常困難的，除了上述之流程數量龐大與再回流的特色之外，還有批量加工機台(Batch Tools)以及隨機當機(Random Failures)的特性使這樣的系統變得無比複雜。Russ 與 John【6】認為是今日所遇到最複雜的排程問題之一。目前大部分晶圓製造現場文獻中所謂的「排程」，在實際上都僅能做到「派工」的階段，而並非真正的排程。而且，大部分的法則僅針對單一績效指標作考量，也很少考慮過晶圓因等候過久招致污染或氧化之現象。因此，若只打算以現有的派工或排程手法改善等候時間限制問題仍難有令人滿意之結果。

實務上，目前現場存有幾種型態的等候時間限制：(1)單一時間限制機台；(2) 一生產線段之間具有時間限制且生產線內之機台皆為單批作業機台，在該段起始加工之後便一直存在時間限制，必須跨越多個工作站加工之後時間限制才會消除，而通常此線段末端為瓶頸機台；(3) 含批量機台之連續型時間限制機台，其中批量機台及後續之單批機台皆為時間限制機台。在此三類問題中，除了第一類型之問題已有學者對其加以研究外，較困難的第二、三類型至今都尚無一較佳的管控模式來協助管理者在複雜的晶圓製造中達到縮短週期時間、避免報廢及提高產出之目的。

從上述背景與問題描述中可得知，隨著製程的進步，等候時間限制的問題勢必越來越嚴重。有鑑於此，本研究希望提出一系列的方法解決在等候時間限制下晶圓製造廠之現場管控問

題。本研究將以存貨理論中最大存貨水準與最小安全存量的觀點(S,s)，發展一套有效的派工與投料法則，期望透過最高存水準來避免 WIP 過長的等候時間，並以最小安全存量來避免機台挨餓。除在製品存貨水準制訂之外，本研究更進一步針對派工法則進行研究。吾人認為，晶圓廠為一多產品且生產型態複雜之系統，因此，一套動態派工與生產批量決策手法是必須的。本研究將透過上、下游系統最大處理能力與 WIP 超過時間限制之緊急程度之間的交互關係，來決定機台加之工產品種類與生產批量，以因應生產系統中的各種情況。

2. 文獻探討

在本章節中，文人將針對上述三類等候時間限制問題之相關文獻進行討論。

(1)單一時間限制機台：Lee 和 Jung【4】針對等候時間限制的系統提出了利用現場管理的方法使系統達到最佳化，他們利用分散式現場管理排程的技術，減少了工件重新加工的機率。Wolfgang 與 Joerg【5】則是提出 JIT 的看板(Kanban)派工方法來改善時間限制的影響。此研究是針對晶圓製造廠中，濕蝕刻與爐管兩加工站間的時間限問題，提出一套有效的派工方法來達到減少週期時間與增加機台利用率兩個目標。其中，Wolfgang 與 Joerg 將機台當機與到達率不同兩要素考慮在內，並加入 JIT 即時生產的觀念，以減少存貨水準的增加。

(2)一生產線段之間具有時間限制且生產線內之機台皆為單批作業機台，在該段起始加工之後便一直存在時間限制，必須跨越多個工作站加工之後時間限制才會消除，而通常此線段末端為瓶頸機台：由於管理者無法透過現有方法得知此生產線段內所能容納之存貨上限，普遍存在在製品超過時間限制的疑慮，導致投料的不足，因而容易產生末端之瓶頸機台閒置，造成產能的浪費。另一方面，工廠中總是存在著各種不確定因素，例如：機台當機、統計波動以及報廢隨機等等。在製品存在的主要目的便是為了預防上述各種因素發生，而其中又以機台當機最為重要【7】，加上晶圓廠背負著高昂機器設備折舊與成本回收的強大壓力，各廠無不努力增加其機台利用率，盡可能地讓機器不要閒置造成浪費，因此生產線上總是存在著大量在製品。但在時間限制之生產線段下，卻是不容許現場無限制的堆積在製品，因為此舉將使機台超出負荷，容易導致大量等候中的在製品報廢。而且 Goldratt 清楚指出在工廠中，瓶頸資源產出才是決定整廠產出的重要關鍵【8】，本不該讓在製品任意堆置於線上。其主要做法是設置緩衝(Buffer)，以防止瓶頸挨餓，但對於 Buffer 大小之決定，卻是以長時間觀察日後空洞之填補情形作逐步修正【9】。對於如何明確、快速地決定 Buffer 值則並未深入作探討，更未曾考慮於類似時間限制條件下之特殊狀況。

(3)含批量機台之連續型時間限制機台，其中批量機台及後續之單批機台皆為時間限制機台：批量工作站之產品除了需要等候批量工作站加工完畢外，還必須為了集批而增加額外的等候時間。在過去針對批量機台現場管控模式之探討，若以對未來產品到達的資訊的預測能力區分，主要可以分為兩大類：一、.資訊未知，二、資訊已知。而這些現場管控模式主要目的是研究如何增加晶圓產出與降低產品週期時間或使等候線長度最短。在產品到達的資訊未知方面，最早提出的學者 Neuts【10】發展出最小批量法則(Minimum Batch Size, MBS)，主要是利用現場管控模式來計算出爐管區的批量大小，但此研究是針對單一產品單一機台所提出的派工法則。Bala 與 Gunvant【11】認為批量加工，會面臨著成批工件下載和生產系統規畫的問題。學者對此分別提出了啟發式方法和模擬法來加以探討，其中啟發式方法能幫助批量機台決定各批次的優先權，並提高機台的利用率。Rulkens *et al.*【12】則是以動態模擬的方式，找出特定的生產情況下最佳的批量大小，從縮短批量機台的加工週期（Cycle Time），進而縮短整個生產系統的加工週期。而後隨著晶圓廠技術的進步，批量機台的現場管控模式已經不能滿足於單一產品單一機台的環境下，逐漸地走向多樣產品單

一機台甚至往多樣產品多台機台的發展模式，但其環境還是侷限於已知未來到達 lot 資訊的環境模式下，探討其策略及排程方式。Weng *et al*.【13】發展出 MCR (Minimum Cost Rate heuristic)該研究試圖降低等候線內等候加工晶圓批的數目，其目的使單位持有成本最小化。此法則較 MBS、DBH、NACH，有較低之等候線長度標準差。

在產品到達的資訊已知方面，即為目前所發展的前瞻策略（look-ahead strategy），它是使用預測的方法預先將在製品做出最佳排序，以判斷出最佳決策。Glassey *et al*.【14】學者為了彌補 MBS 無法預測下一批 lot 到達資訊，發展出動態成批法則（Dynamic Batch Heuristic, DBH），此法則強調平均等候時間最小化之績效。Fowler *et al*.【15】發展出下一批來到決策法則（Next Arrival Control Heuristic, NACH），NACH 法則是修正 DBH 法則並導入滾動區間（Rolling Horizon）之觀念，強調在各決策時點只考慮下一個 lot 抵達時間。同時將單一產品單一機台擴展至多樣產品單一機台的環境，並運用 WSPT(Weighted Shortest Processing Time）之觀點，使利用 NACH 法則雖然在預測 lot 抵達時間資訊中有誤差因子存在，但依然具有一定的穩定性。而儘管有這麼多學者發展出批量機台之現場管控模式，然而這些管控模式都沒有考慮到等候時間的限制以及後續之連續時間限制機台。因此，在當今的環境下也就無法達到現場管控的目的。

3. 研究方法

有關於晶圓廠在等候時間限制下之現場管控的問題，先前已有學者在單站等候時間限制上以提出一套以 JIT 為基礎之派工法則，本研究在此就不再贅述。因此，吾人將以下列兩類問題進行探討研究：A)生產線段之時間限制問題。B)含批量機台之連續型時間限制問題。

關於等候時間限制問題下之派工方法，吾人認為，其解決方法必須包含下列二個方向，等候時間區段內的在製品水準與機台的投料控制。因此，本研究將以存貨理論中(S,s)法則為概念，設立等候時間限制區段內的合理在製品水準與區段源頭機台的停止加工點與再加工點，藉以達成現場管控的目標。

3.1 生產線段之時間限制問題

生產線段之時間限制問題亦即：在一生產線段之間存在著時間限制，當工件在該線段之起始站加工之後便開始累計時間，必須在時間限制內通過多個加工站，並且在通過此線段之最終加工站後此時間限制才會消除。此為在半導體廠中存在已久的時間限制環境，此外，此線段之最終站通常為瓶頸機台。其環境如圖 1 所示。

在現場管控的方法上，本研透過用安全存量控制([S,s] policy)的邏輯控制生產線段初始站的投料。當生產線段內的在製品數量超過安全上限時，初始站隨即停止投料動作；反之，當線段內在製品水準低於安全下限時，初始站則恢復投料。透過此控制法則將生產線段內之在製品水準控制在合理範圍之內，一方面能降低在製品超過實踐限制的機率，另一方面可避免瓶頸機台因統計波動而造成挨餓。

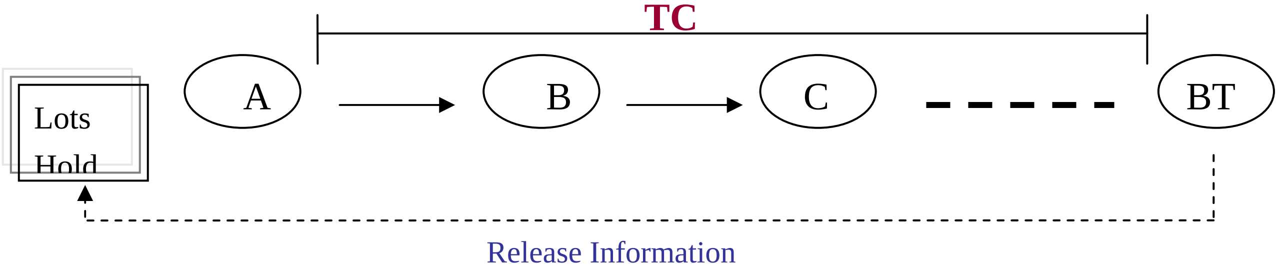

圖 1. 生產線段之時間限制

3.1.1 存貨下限之計算

設置存貨下限(s)之目的在於避免瓶頸機台閒置，首先必須推估上游機台之當機損失時間，進而轉換為損失產能，最後將損失產能賺換成瓶頸機台所需之緩衝時間。其計算邏輯如下：

$$s = \left(\frac{\sum_{i=1}^{n} EC_i - X_i}{\mu_{BT}} - \frac{EC_{BT}}{\mu_{BT}}\right) + \sum_{i=1}^{n} PT_i \tag{1}$$

$$X_i = \left(\sum_{j=1}^{m_i} \mu_{ij} \times A_{ij} - \sum_{J=1}^{m_{BT}} \mu_{BT_j} \times A_{BT_j}\right) \times TR_i \tag{2}$$

$$EC_i = \sum_{j=1}^{m_i} CL_{ij} \times \mu_{ij} \tag{3}$$

$$CL_{ij} = (1 - A_{ij}) \times TR_{ij} \tag{4}$$

3.1.2 存貨上限之計算

存貨上限(S)的設置目的為避免在製品超過時間限制，具有時間限制的生產線段內，其合理的 WIP（轉換為 time buffer）上限應保持在時間限制以下。然而，在機台當機所造成產能損耗的影響之下，此上限值應透過上游各站之產能損耗因子進行修正。其計算邏輯如下：

$$S = \left(TC - \sum_{i=1}^{n-1} PT_{ij}\right) - EC' \tag{5}$$

$$EC' = \frac{\sum_{i=1}^{n} EC_i - X_i}{\mu_{BT}} - \frac{EC_r - X_r}{\mu_{BT}} \tag{6}$$

3.2 含批量機台之連續型時間限制問題

在含批量加工站之連續型時間限制問題方面，本研究主要針對晶圓製造廠中爐管加工區進行討論，此加工區包含爐管機台（批量加工）以及其上、下游工作站，而其中爐管機台與其下游工作站前之等候線皆設有時間限制。加工環境如下圖所示：

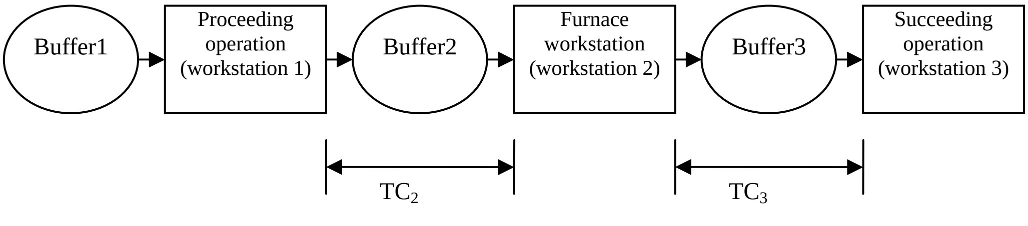

圖 2. 含批量加工機台之連續行等候時間問題

在此本研究針對此一環境所提出之控管法則方面，包含爐管機台加工批量決策、上游機台停工與復工、以及產品選擇。其管控邏輯與步驟分述如下：

I.決策時點：當產品進入或離開爐管加工站或下游加工站時，必須進行管控決策

II.上游機台之停工與復工：

i. 當爐管加工站前之 WIP 量(B_2)大於最大安全存量(S_2)，或是爐管加工站之最小加工批量數(MB)大於下游機台等候線之最大可容忍值(Tol)時，上游機台即應

停止加工。

$$S_i = (TC_i \times m_i - \sum_{j=1}^{m_i} TR_{ij}) \times \mu_i \tag{7}$$

$$s_i = S_2 - C_i \tag{8}$$

$$MB = \frac{B_2}{(TC_2 / PT_2) \times m_2} \tag{9}$$

$$Tol = (TC_3 \times \mu_3 \times m_3 - B_3) - (\mu_2 \times m_2 - \mu_3 \times m_3) \times PT \tag{10}$$

ii. 當爐管加工站等候線上之 WIP 數(B2)小於安全存量下限(s2)時，上游機台即應恢復加工。

$$s_i = S_2 - C_i \tag{2}$$

III.爐管機台管控法則：

i. 加工產品種類之選取：爐管機台前等候線上之各類在製品中，具有最高緊急性指標(I_x)之產品應優先被加工。

$$I_x = W_{x1} \times \sum_{l=1, i \in s_1}^{N_{S1}} \left(\frac{1}{TC_2 - T_{xl}}\right) + W_{x2} \times \sum_{k=1, k \in S2}^{N_{S2}} \left(\frac{1}{T_{xk} - TC_2}\right), \quad x = 1, \ldots, p \tag{11}$$

ii. 加工批量數之決策：爐管機台之當次加工批量數(Q)可由下列邏輯決定之：

$$Q = \begin{cases} 0, & Tol < 0 \\ Min(C, Tol, q_{2,x}) & Tol \geq 0 \end{cases} \tag{12}$$

4. 結論與建議

面對半導體晶圓廠中的時間限制問題是一項艱鉅的挑戰，管理者必須在良率（降低 WIP 水準以防止超過時間限制）與產出（提高 WIP 水準以防止瓶頸機台閒置）之間做出取捨。然而，在缺乏系統化的解決方法之下，管理者大多只能憑藉經驗法則找出合理的在製品水準。而本研究所提出之時間限制下現場管控法則，透過存貨管理的概念計算生產系統內合理之在製品水準，一方面能控制在製品的等候時間，另一方面亦能顧及瓶頸機台的使用率。相信在此管控法則的幫助之下，管理者定能更有效地解決晶圓廠內的時間限制問題。

在半導體產業之中，由於過高的機台設置成本使然，生產系統內的各工作站通常會被要求盡可能維持在高使用率。一般來說，除非使用率高於 95%，否則幾乎無法添置機台。然而，綜觀吾人前後二年之專題計畫，過高的機台使用率即為解決時間限制問題之瓶頸所在。在時間限制的影響之下，半導體產業的產能配置策略是否仍須如此緊繃，是一件值得思考問題。半導體製造技術日趨進步，在進入奈米級製程之後，時間限制問題越趨普遍，時間限制也急遽縮短，利用率動則 90%以上的產能配置早已無法應付如此艱鉅的問題。後續相關研究可進行檢討普遍存在於半導體產業的產能配置策略，並找出最佳的產能配置邏輯。

三、　參考文獻

1. Cunningham, J. A. (1990), "The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing," *IEEE Transactions On Semiconductor Manufacturing*, **3**(2), 60-71.
2. Robinson, J. K. and Giglio, R. (1999), "Capacity planning for semiconductor wafer fabrication with time constraints between operations," *1990 Winter Simulation Conference,* Phoenix, AZ, USA.
3. Kim, J. H., Lee T. E., Lee H. Y. and Park D. B. (2003), "Scheduling Analysis of Time-Constrained Dual-Armed Cluster Tools," *IEEE Transactions on Semiconductor Manufacturing*, **16**(3), 521-534.
4. Lee, S.H. and Jung, M.Y. (2003), "Timing constraints' optimization of reserved tasks in the distributed shop-floor scheduling," *International Journal of Production Research*, **41**(2), 397-410.
5. Wolfgang, S. and Joerg, D. (1999), "Implementation of modeling and simulation in semiconductor wafer fabrication with time constraints between wet etch and furnace operations," *1999 International Symposium on Semiconductor Manufacturing Conference*, Santa Clara, CA, USA.
6. Russ, M. D. and John, W. F. (2003), "A New Scheduling Approach Using Combined Dispatching Criteria in Wafer Fabs," *IEEE Transactions on Semiconductor Manufacturing*, **16**(3), 501-510.
7. Hopp, W. J. and Spearman, M. L. (1995), *Factory Physic – Foundations of Foundations of Manufacturing Management,* Chicago, IL:IRWIN.
8. Goldratt, E. M. and Cox, J. (1992), *The Goal*, 2nd Edition, North River Press.
9. 吳鴻輝、李榮貴(2001)，「限制驅導式現場排程與管理技術」，全華書局。
10. Neuts, J. D. C. (1967), "A General Class of Bulk Queues with Poisson Input," *Annals of Mathematical Statistics*, **38**(3), 759-770.
11. Bala R., and P. Gunvant (1998), "Modelling Furnace Operations Using Simulation and Heuristics," 1998 *Winter Simulation Conference*, Washington DC, USA.
12. Rulkens, H. J. A., E. J. J. van Campen , J. van Herk, and J. E. Rooda (1998), "Batch Size Optimization Of A Furnace and Pre-clean Area By Using Dynamic Simulations," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, Boston, MA, USA.
13. Weng, W. W. and R. C. Leachman (1993), "An Improvement Methodology for Real-time Production Decision at Batch-process Workstations," *IEEE Transactions on Semiconductor Manufacturing*, **16**(3), 219-225.
14. Glassey, C. R. and W. W. Weng (1991), "Dynamic Batching Heuristic for Simultaneous Processing," *IEEE Transactions on Semiconductor Manufacturing*, **4**(2), 77-82.
15. Fowler, J. W., D. T. Phillips, and G. L. Hogg (1992), "Real-time Control of Multi-product Bulk-Service Semiconductor Manufacturing Process," *IEEE Transactions on Semiconductor Manufacturing*, **5**(2), pp.158-163.

四、　計畫成果

就本質而論，本計畫之研究成果同時具有實務及學術價值。在實務方面，本計畫之成果提供晶圓製造廠一套在等候時間限制問題下之系統化現場管控方法；在學術上，本研修正安全存貨管理的邏輯用以解決時間限制問題。此外，本研究亦已將主要成果發表於國際學術研討會以及學術期刊之中。

本研究之主要成果分述如下：

1. 現行派工法則之探討：針對半導體產業常見之派工法則加以研究，並探討各項法則之優缺點。
2. 安全存量模型之修正：修正傳統安全存量模型，使其概念可應用於時間限制下之現場管控法則。
3. 現場管控法則之建立：針對不同的環境，分別提出系統化的現場管控模式，幫助管理克服時間限制問題的挑戰。
4. 研究成果彙整與發表：包含上一年度之計畫，吾人已將研究成果彙整成為四篇期刊論文(三篇已陸續發表於 *Journal of the Chinese Institute of Industrial Engineers*(二篇)以及 *International Journal of Production Research*，一篇仍在審查中)，以及三篇研討會論文(陸續發表於 2008 Global Business & International Management Conference, 19th Annual Conference of the Production and Operations Management Society, 17th International Conference on Pacific Rim Management)

附錄-符號表

1. In section 3.1

A_{ij}	工作站 i 之機台 j 之可用率
μ_{ij}	工作站 i 之機台 j 之服務率
m_i	工作站 i 之機台數
TR_{ij}	工作站 i 之機台 j 之當機平均修復時間
PT_i	工作站 i 之平均加工時間
BT	瓶頸加工站
r	生產線段之投料站（初始站）
n	生產線段內之工作站數(不含瓶頸站)
TC	時間限制長度
EC_i	工作站 i 之期望損失產能
X_i	工作站 i 之產能負荷修正因子

2. In section 3.2

m_i	工作站 i 之機台數
μ_i	工作站 i 之服務率
Bi	工作站 i 目前的在製品水準 (in time buffer formation)
TR_{ij}	工作站 i 之機台 j 之平均當機修復時間
C_i	工作站 i 之最大(滿批)批量數
TC_i	等候線 i 之時間限制長度
PT	爐管機台當前之加工時間
q_{ix}	工作站 i 之等候線上產品 x 之數量
W_{x1}	尚未超過時間限制之產品 x 之權重
W_{x2}	已超過時間限制之產品 x 之權重

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

97 年 8 月 14 日

報告人姓名	杜瑩美	服務機構及職稱	中華大學 工業工程與系統管理學系 副教授
會議 時間 地點	自 2008 年 7 月 31 日至 2008 年 8 月 2 日 美國波特蘭	本會核定補助文號	**NSC 96-2221-E-216-038**
會議名稱	（中文）2008 全球企業與國際管理學術研討會 （英文）2008 Global Business & International Management Conference		
發表論文題目	（中文） （英文）Capacity Planning for Batch-serial Processes with Time Constraints in Wafer Fabrication		

報告內容應包括下列各項：

一、參加會議經過

2008 Global Business & International Management Conference was hold Portland, USA. The conference scope is multidisciplinary. It will publish research and applied articles from all areas of Business Management. The conference will also consider a variety of methodological approaches. In the conference, I presented a paper entitled “Capacity Planning for Batch-serial Processes with Time Constraints in Wafer Fabrication” and the topic attracted the attention of attendants because the issue has not been researched a lot in the past. In addition, some other topics about management have been presented and they were all impressed me very much.

二、與會心得

The conference will serve as an important forum for the exchange of ideas and information to promote understanding and cooperation among the global businesses and international management. This year's conference theme is “Human Resource Management & Cross-Culture Issue, Global Economy, Global Industry & Accounting and Global Issues & Business Education”. Special focuses will be placed on management human resource, international industries, organizational change and global economy to exchange programs and opportunities for practitioners from North America, Taiwan, Korea, Malaysia, Canada, and other places. Due to the price of crude oil continuously increasing and financial crisis of house loan in US, the global business and economy are getting worse. Hence, there are many papers discussed these issues and proposed many ideas and methodologies for government and industries. This is a rich and colorful trip not only in the research field but also to find a history and charm city, Portland. In the finally, I would like to thank the budgets support from National Science Council and Chung-Hua University.

三、考察參觀活動(無是項活動者省略)

None.

四、建議

International Conference is a good way not only to get new ideas quickly but also to face to discuss with the authors. However, the funding is a big problem for us. Even through our school and National Science Council fund us, the funding is still limited. Besides, the upper bond of airfare is unreasonable in the high oil price ages. For example, the airfare from Taiwan to Portland is around NTD60,000 but the upper bond is NTD39,000. This amount is far away from real price. It means we should offer the price difference. Moreover, the daily allowance also should be updated by present situation. Therefore, I suggested updating the relative data to fit actual state.

五、攜回資料名稱及內容

1. Conference Proceeding: 2008 Global Business & International Management Conference
2. CD of the proceedings.
3. J.I.M.S. Journal of International Management Studies
4. The Journal of Global Business Management --- Special Edition

六、其他

Capacity Planning for Batch-Serial Processes with Time Constraints in Wafer Fabrication

Ying-Mei Tu, Department of Industrial Engineering and System Management
Chung Hua University, Taiwan

ABSTRACT

Over the past decade, wafer fabrication has become more complicated. To ensure final product yield, engineers need to set up queue time limits, named "time constraint", for particular machines during wafer processing. Time constraint problems are more serious when the lots released from a batch workstation which often causes the peak workload on the downstream serial workstation. Under the time constraints environment, the peak workload will be a critical problem and hurt the performance of downstream workstation seriously.

This work applied GI/G/m queuing network to develop a capacity planning model for batch-serial processes. By this queuing network model, the expected waiting time between batch-serial processes can be estimated. Managers can also determine the capacity through the setting of expected rate of exceeding time constraints. Furthermore, how the arrival smoothing of the upstream batch workstation will impact on the downstream serial workstation is also analyzed. The result showed that it can effectively decrease the waiting time on the downstream serial workstation through increasing the upstream batch machine numbers and decreasing the batch size. It also implied that increasing the number of downstream serial machines is not the only way under the batch-serial process with time constraint environment. Therefore, the investment function of batch-serial process equipment with time constraints can be established and support managers for investment decision.

Keywords: *Capacity Planning, Time constraints, Batch-serial processes, Queuing theory, Arrival smoothing*

INTRODUCTION

Recently, wafer fabrication has become more complicated and tough to maintain high yield rate. To prevent copper film oxidation or fluorine precipitation of undesirable interface on wafer surface, the technology development engineers set up a period of time between processes to identify the lots that cannot exceed over the time limit which is called time constraint.

As we know that there are many types of products and process flows in a wafer foundry. In order to keep high competitiveness, the factories have to achieve short cycle times, to keep low levels of inventory, to get high throughput under the lowest investment, and to maintain the flexibility of the production line. Therefore, the capacity planning is difficult under such variant environments. From this point of view, different product flows, low volume products, batching, hot lots, time constraints, and so on will generate a lot of problems for shop floor control especially in the areas with batching and un-batching processes. Generally, it will take long time to complete a full batch for batch tools and when the lots released from a batch workstation which often causes the peak workload on the downstream serial workstation. Regarding to batching process, Tu and Liou (2006) proposed a capacity planning to determine the number of batch tools under time constraints. However, it only took batching process into account. Generally, the waiting time of the downstream serial workstation will increase as the batch number increasing. It means that the batch number and number of batch tools will influence the performance of downstream serial workstation especially under time constraints. If the peak workload is heavy, the un-batching process will be difficult within time constraints. On the contrary, if managers intend to keep the queue time of lots within time constraints, then the batch size of batching process should be minimized or the number of downstream workstations should be increased. Minimized batch number will result in throughput reduction and impact on delivery. Increasing workstations number will raise the production cost. They both hurt the profit of company. Therefore, in order to achieve the production goal and keep the productivity, a good capacity planning which including batch-serial tools is imperative.

Sivakumar and Chong (2001) pointed out that cycle time will be reduced by arrival and service rate smoothing. Therefore, Tu and Chen (2006) applied this concept in the furnace area to analyze how the arrival smoothing will impact on the batch-serial process, such as waiting time and throughput. The results showed that it can decrease the peak workload and the waiting time on the un-batch workstation through increasing the batch machine numbers and decreasing the batch size. It also implied that increasing the number of un-batch machines is not the only way to solve the time constraint issue. Generally, the un-batch machine is more expensive than the batch machine in the furnace area. If the investment function of batch-serial processes equipment with time constraints can be established, it will be helpful for investment decision, particularly in diffusion area.

This paper applied GI/G/m queuing theory to develop a capacity planning model for batch-serial process. By this queuing network model, the expected waiting time between batch-serial processes can be estimated. Managers can also determine the capacity through the setting of expected rate of non-exceeding time constraints. At the end, conclusions and recommendations for capacity planning in the ramping or fully scale stage are presented.

LITERATURE REVIEW

There are some papers that addressed queuing theory for serial-batch process. Chung and Huang (1999) developed a block-based cycle time (BBCT) estimation algorithm to estimate the product cycle time, consists of the waiting time due to batching and loading factor in every block. The corresponding waiting time was estimated using the M/M/c queuing model and batch factor flow time (BFFT) estimation algorithm.

Louw and Page (2004) described an open queuing network modeling approach to estimate the time buffer in production systems. A GI/G/m queuing network is applied to model the real production environment, such as batch processing, reentrant customers, and machine failure. Nevertheless, this research did not consider the duration of machine failure (mean time to repair; MTTR). By simply using availability to represent machine behavior was inadequate under time constraints.

Fowler et al. (2002) applied $G/G^{(bp)}/c$ queue to study the optimal batching size that minimize the expected cycle time of batch-processing operations for a real-world semiconductor manufacturer. This model, denoted by $G/G^{(bp)}/c$, represents multiple products, multiple servers, batch-processing, incompatible products and unequal batch service size queues. But this paper also ignored the impact of performance measure by time constraints.

Besides, there are many papers addressed the time constraint problems and queuing theory. Robinson and Giglio (1999) first described the specific problem of time constraint for semiconductor manufacturing. They developed an approximation based on M/M/c queuing formulas to predict the probability of reprocessing when lots exceed the time limit. Machine characteristics like batch tools, machine failures and different arrival rates were not considered.

Wolfgang Scholl and Joerg Domaschke (2000) studied the time constraints issue between furnace and wet etch by simulation, they applied JIT in the experiment, and the upstream machine will not be processed until the downstream machine is available. The purpose of this research was to achieve cycle time target and maximize machine utilization, but the method of shop floor control was not described.

Tu and Liou (2006) applied GI/G/m queuing theory to develop a capacity determination model for serial-batch process with time constraints in the furnace area of semiconductor fabrication. They created a check table for different utilizations, time constraints, and the probability of expected waiting time over time constraints. The check table can be easily derived from the approximation formula for the fully scale stage. But machine failures were not considered.

In addition, there are fewer papers addressed the batch-serial process problems. Louw and Page (2004) defined that when batches arrival at the downstream workstation can be viewed as three parts: a queue of whole batches, a queue of a partial batch, and a server. (Figure 1) This research calculated the average waiting time and the average waiting number of the whole batches and the partial batch separately. Chung and Huang (1999) addressed the block-based cycle time estimation algorithm to calculate the cycle time of the batch-serial process. Park *et al.* (2000) addressed the mean value analysis (MVA) to calculate the cycle time of re-entrant line with batch machines and multi-class jobs. But this research overestimated the cycle time, and it was a rough estimate in the cycle time of the batch-serial process.

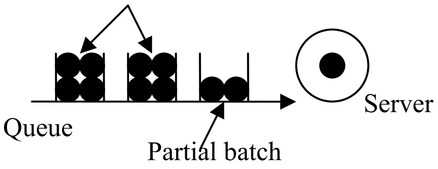

Figure 1. Batches arrival at the un-batch workstation

METHODOLOGY

To take the factor of time constraints into capacity planning, this paper applies the concept of queuing theory. The expected waiting time obtained from queuing systems can work out the problem of time constraints. To choose the most suitable queuing model for this study is very important. Therefore, the model denoted by GI/G/m, represents multiple products, multiple servers, the arrival rate, and service rate obtained by any given distribution.

Moreover, the un-batching of batch needs to be taken into consideration in the queuing systems. Therefore, there are two parts in our queuing systems. The first part is whole batches waiting for un-batching; the second part is a partial batch wait for processing. In addition, we view that the second part and the un-batch workstation as a fictitious batch machine. Note that there is one fictitious batch machine, so we can derive the total expected waiting time in GI/G/1 queuing system. The parameters and performance measures of GI/G/1 queue are modeled in Figure 3.

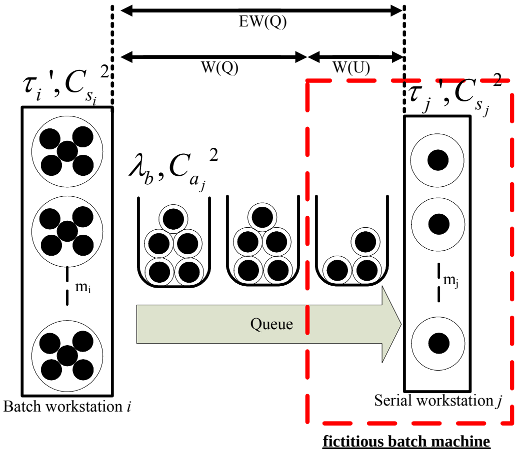

Figure 3. Parameters and performance measures of GI/G/1 queue

From Figure 3, W(Q) is the expected waiting time of any product by batch type to wait for un-batching. W(U) is the expected waiting time of any product to wait for processing. Summation of W(Q) and W(U) can derive the total expected waiting time in GI/G/1 queuing system denoted by EW(Q). Finally, the critical concept to consider time constraint problems is to apply the probability formulas to estimate the probability of expected waiting time, which is over time constraints.

Data input

The following data were required for the capacity planning model.

f	different end products,
λ_{j_p}	total arrival rate (lot/time-unit) at workstation *j*,
λ_{kj_p}	mean arrival rate (lot/time-unit) of product *k* at workstation *j*,
λ_{j_b}	total arrival rate (batch/time-unit) at workstation *j*,
b	batch size,
p_{kj}	the ratio of product *k* at workstation *j*,
τ_{kj}	mean service time of product *k* at workstation *j*,
$C^2_{s_{kj}}$	the SCV of service time of product *k* at workstation *j*,
ρ_j	traffic intensity at workstation *j*,
m_j	machine numbers at workstation *j*,
λ_b	total arrival rate (batch/time-unit) of fictitious batch machine,
$C^2_{a_j}$	the SCV of batch arrival rate of fictitious batch machine,
n	total number of the upstream workstation of workstation *j*,

$MTTR_{jl}$	mean time to repair of machine *l* of workstation *j*,
$MTBF_{jl}$	mean time between failures of machine *l* of workstation *j*,
$C_{sd_{jl}}^2$	the SCV of downtime of machine *l* of workstation *j*
TC	time constraint

The modified GI/G/1 queuing model

Step 1: The Initial Capacity Determination

In this section, the minimum capacity that can meet the system's basic requirement is determined, which is defined as "initial capacity."

Step1.1: Aggregating mean arrival rate

The first step of parameter calculation is aggregating arrival rates of individual product into the mean arrival rate (λ_{jp}) of the workstation *j*. The formulas are represented as follows:

$$\lambda_{j_p} = \sum_{k=1}^{f} \lambda_{kj_p} \tag{1}$$

$$\lambda_{j_b} = \frac{\lambda_{j_p}}{b} \tag{2}$$

Step1.2: Aggregating mean service time and service time variation

After the mean arrival rate is accumulated, the service time and squared coefficient of variation (SCV) of service for individual products also have to be aggregated into the mean service time (τ_j) and SCV of service time (C_{sj}^{2}) of the workstation *j*. The following are the equations for aggregation:

$$\tau_j = \frac{\sum_{k=1}^{f} \lambda_{j_p} p_{kj} \tau_{kj}}{\lambda_{j_p}} \tag{3}$$

$$C_{s_j}^2 = \frac{\sum_{k=1}^{f} \lambda_{j_p} p_{kj} {\tau_{kj}}^2 (C_{s_{kj}}^2 + 1)}{\lambda_{j_p} {\tau_j}^2} - 1 \tag{4}$$

Step1.3: The initial capacity determination

From the definition of Queuing Theory, the traffic intensity (ρ) must be smaller than one to keep the system at steady-state. Therefore, the initial capacity would be the smallest integer m that is greater than the arrival rate divided by the service rate, and can be presented as follows:

$$\rho_j = \frac{\lambda_{j_p} \times \tau_j}{m_j} < 1 \tag{5}$$

Therefore,

$$m_j = \left\lfloor \lambda_{j_p} \tau_j \right\rfloor + 1 \tag{6}$$

Step 2*: Adjust Mean Arrival Rate and Inter-arrival Time Variation of Fictitious Batch Machine

$$\lambda_b = \lambda_{j_b} \tag{7}$$

$$C_{a_j}^2 = \alpha + \sum_{i=1}^{n} \beta \, C_{a_i}^2 \tag{8}$$

(See appendix A to obtain the equations of α and β)

Step 3*: Adjust Mean Service Time and Service Time Variation of Fictitious Batch Machine under Machine Failure

This section introduces a novel scheme to reveal the effect of machine failure which takes into consideration availability and duration of downtime. In this work, the machine failure was regarded as irregular customers whose arrival rate and mean service time will be 1/(MTTR+MTBF) and MTTR, respectively. We assumed that the machine failures were operation dependent; therefore, it will only affect the service data of the system. (Tu and Chen, 2006) To reveal the effect of machine interrupts, the mean service time and SCV of service time were modified.

First, the service time for individual products of fictitious batch machine has to be aggregated into the mean service time (τ_b). The following is the equation for aggregation:

$$\tau_b = \frac{\tau_j \times b}{m_j} \tag{9}$$

Second, we put the duration of τ_b to use, the adjusted mean service time (τ_j') and SCV of service time ($C_{sj}^{2}{}'$) of the workstation j under machine failure are presented as follows:

$$\tau_j' = \frac{\tau_j \times b + \sum_{l=1}^{m_j} \frac{\tau_b \times MTTR_{jl}}{MTTR_{jl} + MTBF_{jl}}}{b + \sum_{l=1}^{m_j} \frac{\tau_b}{MTBF_{jl} + MTTR_{jl}}} \tag{10}$$

$$C_{s_j}^2{}' = \frac{b\tau_j^{\,2}(C_{s_j}^2 + 1) + \sum_{l=1}^{m_j} \frac{\tau_b \times MTTR_{jl}^{\,2}}{MTTR_{jl} + MTBF_{jl}}(C_{sd_{jl}}^2 + 1)}{(b + \sum_{l=1}^{m_j} \frac{\tau_b}{MTBF_{jl} + MTTR_{jl}})(\tau_j')^2} - 1 \tag{11}$$

Third, we adjust mean service time of fictitious batch machine under machine failure (τ_b'). The formula is represented as follow:

$$\tau_b' = \frac{\tau_j' \times b}{m_j} \tag{12}$$

Finally, we find that the SCV of service time of the fictitious batch machine is equal to the SCV of service time of the workstation j (C_{sb}^2) from the following formulas.

$$E(\tau_b') = E(\frac{\tau_j' \times b}{m_j}) = \frac{b}{m_j} E(\tau_j') \tag{13}$$

$$Var(\tau_b') = Var(\frac{\tau_j' \times b}{m_j}) = (\frac{b}{m_j})^2 Var(\tau_j') \tag{14}$$

$$C_{s_b}^2 = \frac{Var(\tau_b')}{E(\tau_b')^2} = \frac{(\frac{b}{m_j})^2 Var(\tau_j')}{(\frac{b}{m_j} E(\tau_j'))^2} = C_{s_j}^2{}' \tag{15}$$

Step 4*: System Performance Estimation

The expected waiting time was referred to the revision queuing model which was modified from EW(M/M/m) approximation formula to GI/G/1 model by Whitt (1993)：

$$W(Q) = (\frac{C_{a_j}{}^2 + C_{s_b}^2}{2}) \times EW(M/M/1) \tag{16}$$

$$EW(M/M/1) = \tau_b' \frac{\rho_b}{1 - \rho_b} \tag{17}$$

For un-batching behavior, the first lot of a batch arrives to the fictitious batch machine has not waited for processing. The last lot has to wait τ_b'. Thus, the average waiting time of any product to un-batch is:

$$W(U) = \frac{0 + \tau_b'}{2} = \frac{\tau_b'}{2} \tag{18}$$

Therefore, the total expected waiting time of the GI/G/m queue is:

$$EW(Q) = W(Q) + W(U) \tag{19}$$

From the obtained parameters, the probability of expected waiting time over time constraints can be calculated. The approximation method was referred to Whitt (1993). It can be obtained as follows:

Because W(U) is a constant, so we assume $\mathrm{TC}' = TC - W(U)$ (20)

$$P\left(EW(Q) > \mathrm{TC}\right) = P\left(W(Q) > \mathrm{TC'}\right) \approx \delta e^{-\eta \cdot \mathrm{TC'}} \tag{21}$$

$$\eta = 2 \times 1\left(1 - \rho_b\right) / (C_{a_j}^2 + C_{s_b}^2) \tag{22}$$

$$\delta \approx \eta \times W(Q) \tag{23}$$

Step 5*: The Capacity Determination under Time Constraint

From step 1* to step 4*, we can get the probability of expected waiting time over time constraints (P), and see it whether it under the probability (β*) set by the manager. If P > β*, we can increase the un-batch machine number; If P < β*, the capacity determination under time constraint can be obtained.

Capacity determination model

Based on the concept of queuing theory and the application of GI/G/1 formula, the derived capacity determination model can provide an analysis platform for decision support. Managers can forecast the accurate capacity requirement and determine the robust equipment investment plan in semiconductor foundry by this model. The framework of the capacity determination model is shown in Figure 4.

From this model, the most complex characteristics such as multi-products and un-batch operation were considered into arrival rate and service rate. The time constraint problems were also considered into the probability formula of GI/G/m queue model. Finally, the decision maker could easily use the information to decide the investment plan of equipments.

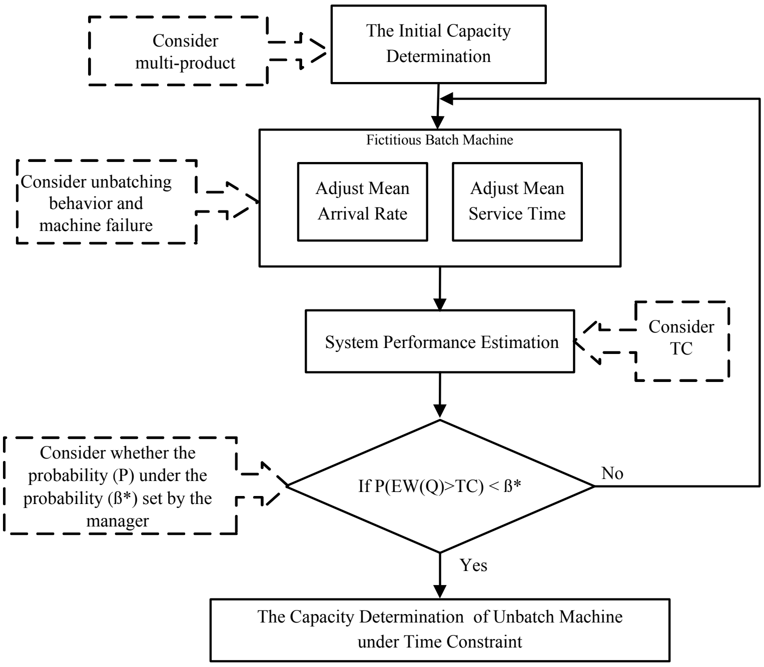

Figure 4. Capacity determination model

ARRIVAL SMOOTHING

The Definition and Basic Concept of Arrival Smoothing

In a steady-state environment, the arrival rate will be equal to the output rate. Arrival smoothing is also output smoothing. From this viewpoint, the output of batch machine can be smoothed by decreasing its batch size. For example, the original batch size of a batch machine is 6 lots, and the processing time is 1 hour, we can decrease the batch size to 2 lots and increase the number of machine to 3, the output rate is still 6 lot/hr, but the output get smoother. (Figure 5)

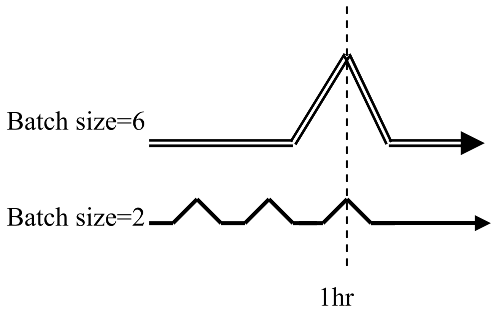

Figure 5. The peak workload of different batch size.

The influence of arrival smoothing on system performance

According to Tu and Chen (2006), the peak workload and the waiting time on the un-batch workstation will decrease through arrival smoothing. Nevertheless, the authors proposed different solution in this research, increasing the upstream batch machine to smooth the workload of downstream machine. The results showed that increasing the number of upstream batch machines not only decreasing the peak workload but also making the system performances better. It also implied that increasing the number of downstream serial machines is not the only way under the time-constraint environment. Based on this concept, if the investment function of batch-serial process equipment with time constraints can be established, it will be helpful for investment decision, particularly in diffusion area.

CONCLUSIONS

This research applied GI/G/m queuing theory to develop a capacity planning model for batch-serial process with time constraints. Then it can be combined with the capacity determination model developed by Tu and Liou (2006) to develop the complete capacity determination model in the furnace area of semiconductor fabrication.

This research also applied the concept of arrival smoothing in the batch-serial process with time constraints in the furnace area of semiconductor fabrication. The results showed that the smoother arrival rate, the better system performances. It also implied that increasing the number of downstream serial machines is not the only way under the time-constraint environment. Generally, the downstream serial machine is more expensive than the upstream batch workstation in the furnace area. If the investment function of batch-serial process equipment with time constraints can be established, it will be helpful for investment decision, particularly in diffusion area.

Acknowledgement. The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 95-2221-E-216-014

REFERENCES

Chung, S.H., and Huang, H.W. (1999). The block-based cycle time estimation algorithm for wafer fabrication factories. International Journal of Industrial Engineering, 6(4), 307-316.

Fowler, J.W., Phojanamongkolkij, N., Cochran, J.K. and Montgomery, D.C. (2002). Optimal batching in a wafer fabrication facility using a multi-product G/G/c model with batch processing. International Journal of Production Research, 40(2), 275-292.

Lin, C. W. (2004). Capacity determination model for sequential time constraints in wafers fab. Master Thesis, CHU, Hsin-Chu, Taiwan.

Louw, L., and Page, D. C. (2004). Queuing network analysis approach for estimating the sizes of the time buffers in Theory of Constraints-controlled production systems. INT. J. PROD. RES., 42(6), 1207-1226.

Park, Y., Kim, S. and Jun, C.H. (2002). Mean value analysis of re-entrant line with batch machines and multi-class jobs. Computers & Operations Research, 29(8), 1009-1024.

Robinson, J. K., and Giglio, R. (1999). Capacity Planning for Semiconductor Wafer Fabrication with Time Constrains between Operations. Proceedings of the 1999 Winter Simulation Conference, 1(7), 880-887.

Sivakumar, A.I. and Chong, C.S. (2001). Simulation-based analysis of cycle time distribution, and throughput in semiconductor backend manufacturing. Journal of Computers in Industry, 45(1), 59–78.

Tu, Y.M., and Chen, C.L. (2006). The influence of arrival smoothing between batch and serial processes on system performance. The 7th Asia Pacific Industrial Engineering and Management Systems Conference 2006, Bangkok, Thailand, 1919-1924.

Tu, Y.M., and Chen, H.N. (2006). Waiting time approximation with resource failure. International Conference of Pacific RIM Management, Hawaii, USA.

Tu, Y.M., and Liou, C.S. (2006). Capacity determination model with time constraints and batch processing in semiconductor wafer fabrication. Journal of the Chinese Institute of Industrial Engineers, 23(3), 192-199.

Whitt, W. (1993). Approximations For The GI/G/m Queue. Production and Operations Management, 2(2), 114-161.

Wolfgang, S. and Joerg, D. (2000). Implementation of modeling and simulation in semiconductor wafer fabrication with time constraints between wet etch and furnace operations. IEEE Transactions on Semiconductor Manufacturing, 13(3), 273-277.

APPENDIX

In equation (8): $C^2_{a_j} = \alpha + \sum_{i=1}^{n} \beta\, C^2_{a_i}$

$$\alpha = 1-\omega_j$$

$$+\omega_j \sum_{i=1}^{n} p_{ij}(q_{ij}\rho_i'[1+m_i^{-0.5}(\max\{C^2_{s_i}',0.2\}-1)]+(1-q_{ij})^2 C^2_{s_j}') \tag{A1}$$

$$\beta = \omega_j p_{ij} q_{ij}[(1-\rho_i'^2)+(1-q_{ij})] \tag{A2}$$

$$\omega_j = [1+4(1-\rho_i')^2(\nu_j-1)]^{-1} \tag{A3}$$

$$\nu_j = [\sum_{i=0}^{n} p_{ij}^2]^{-1} \tag{A4}$$

$$p_{ij} = {\lambda_{ij}}\big/{\lambda_j} \tag{A5}$$

$$\lambda_j = \sum_{i=1}^{n} \lambda_i \tag{A6}$$

Where

$C^2_{a_i}$	the SCV of batch arrival rate at *i*th operation
p_{ij}	the proportion of arrivals to workstation *j* that come from workstation *i*
q_{ij}	the proportion of production that go from workstation *i* to workstation *j*
λ_j	the total arrival rate at workstation *j*
λ_{ij}	the total arrival rate from workstation *i* to workstation *j*

Because there is no outside arrival in this research, so equation (A5) is equal to 1, and the above equations can be simply as follows:

$$p_{ij} = {\lambda_{ij}}\big/{\lambda_j} = 1$$

$$\Rightarrow \nu_j = [\sum_{i=0}^{n} p_{ij}^2]^{-1} = 1$$

$$\Rightarrow \omega_j = [1+4(1-\rho_i')^2(\nu_j-1)]^{-1} = 1$$

$$\Rightarrow \begin{cases} \alpha = 1-\omega_j+\omega_j \sum_{i=1}^{n} p_{ij}(q_{ij}\rho_i'[1+m_i^{-0.5}(\max\{C^2_{s_i}',0.2\}-1)]+(1-q_{ij})^2 C^2_{s_j}') \\ \quad = \rho_i'[1+m_i^{-0.5}(\max\{C^2_{s_i}',0.2\}-1)] \\ \beta = \omega_j p_{ij} q_{ij}[(1-\rho_i'^2)+(1-q_{ij})] = (1-\rho_i'^2) \end{cases}$$

(A8)

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

97 年 10 月 22 日

報告人姓名	陳欣男	服務機構及職稱	中華大學 科技管理研究所 博士班研究生
時間 會議 地點	2008/5/9 至 2008/5/12 La Jolla，San Diego，CA，USA	本會核定 補助文號	**NSC 96-2221-E-216-038**
會議 名稱	（中文）第 19 屆生產與作業管理學會年度會議 （英文）19th Annual Conference of the Production and Operations Management Society		
發表 論文 題目	（中文）等候時間限制問題下之批量機台現場管控模式 （英文）Shop-Floor Control Model of Batch Machine under Time Constraints		

報告內容應包括下列各項：

一、參加會議經過

19th Annual Conference of the Production and Operations Management Society was held in La Jolla, USA. The conference served as important forum for the exchange of ideas and information to promote understanding and cooperation among the operation and production management. In the conference, I presented a paper entitled “Shop-Floor Control Model of Batch Machine under Time Constraints” and the topic attracted the attention of attendants because the issue has not been researched a lot in the past. In addition, some other topics about management have been presented and they were all impressed me very much.

二、與會心得

The international conference provided good opportunities for exchanging information and ideas among scholars in same research fields. Because there’s almost no Chinese conferee attended in this conference, so, this conference could advance me on skills of presentation and communication in English. I think I had a very good experience in attending POMS 2008 conference.

三、考察參觀活動(無是項活動者省略)

None.

四、建議

Form this conference, I found the international conference is a good activity for scholars. It can gather most scholars with same research field to share their ideas and experiences. Furthermore, it can promote the research mood. In addition, the business of tourism can also be flourishing. Therefore, I suggest encouraging the university to hold the international conference.

五、攜回資料名稱及內容

1. Conference Program: 19th Annual Conference of the Production and Operations Management Society
2. CD of the proceedings.

1. 其他

None.

Abstract Number: 008-0213

Shop-Floor Control Model in Batch Processes of Wafer Fabrication with Time Constraints

YING-MEI TU[*1] and HSIN-NAN CHEN[2]

1: Department of Industrial Engineering & System Management, Chung Hua University, Taiwan
No.707, Sec.2, WuFu Rd., HsinChu, Taiwan 300, R.O.C

2: Graduate Institute of Management of Technology, Chung Hua University, Taiwan
No.707, Sec.2, WuFu Rd., HsinChu, Taiwan 300, R.O.C

Email: YING-MEI TU, amytu@chu.edu.tw
*: Corresponding Author

POMS 19th Annual Conference

La Jolla, California, U.S.A.

May 9 to May 12, 2008

Abstract. Time constraint is a queue time boundary set between particular operations to ensure final product yield. Due to the dilemma of increasing machine efficiency or decreasing the queue time of WIP, the issues of time constraint become more complex in batch processing.

This work proposed a shop floor control policy in serial-batch-serial processes with time constraints. The concept of safety stock ([S,s] policy) is introduced to control the WIP level for avoiding machines idle and wafers exceeding time constraints simultaneously. Length of time constraints, MTTR of machines and service rate of workstations are adopted to determine the batch size and boundaries of WIP level. The job hold/release policy is addressed to control the situations of excessively high WIP level. Furthermore, the performance of proposed model is compared with DJAH and MBSX rules. The results indicated that the proposed model could control the batch processing with time constraints more effectively.

Keywords: time constraints, wafer fabrications, batch processing, shop floor control, batch size

1. Introduction

A time constraint (TC) is a time window set between two specific operations to prevent undesirable copper film oxidation or fluorine precipitation on wafer surfaces (Robinson and Giglio 1999, Tu and Liou 2006). Wafers will be reworked or scrapped if they exceeded TC, which will increase cycle time and decrease productivity. To control workstations with TCs effectively is an essential task for semiconductor industry. However, as the result of high production volume, expensive equipment, and time consuming processes, machines in wafer fabrication are usually required to operate at a high utilization level. Extremely high utilization of workstations will cause the difficulty in resolving issues of TC.

The most familiar TC issue is set between wet etch and furnace operations, which wafers will be reworked at wet etch operation while they exceeded TC. In this stage, batch processing and long processing time will cause the dilemma of the management. Machine efficiency and require time for forming batches become a trade-off issue while determining batch size. Low machine efficiency or long batching forming time will both increase the waiting time and also increase the rate of exceeding TC (Rulken et al. 1998). To determine the batch size more effectively will be crucial for conquering the issues of TC in this stage.

There were many studies related to batch size determination in wafer fabrications (Fowler et al. 1992, Glassry and Weng 1991, Weng and Leachmen 1993). These studies have attempted to optimize the performance of batch-processing workstation, for instance, maximize throughput and minimize cycle time. However, the succeeding operation of furnace workstation is still with TC. To optimize the output of furnace workstation will increase the loading as well as the rate of wafers exceeding TC in succeeding operation. Hence, the shop-floor control model in batch process with TC issue should involve a rule to take succeeding operation into consideration.

Accordingly, the purpose of this work is to develop a shop floor control policy to resolve the TC issues between furnace workstation and its preceding and succeeding operations. The proposed model determines the batch size of furnace operation dynamically by considering TC interval and WIP level. Furthermore, the proposed model could also control the WIP at a safe level by adopting safety stock ([S,s]) policy. Finally, the performance comparison among the proposed model, MBS and DJAH was performed.

2. Literatures Review

Previous studies related to batch size determination and shop-floor control rule in batch processes of wafer fabrications are reviewed in this section. The common methodology for determining batch size can be classified into two categories, Minimum Batch Size rule (MBS) and Look-ahead strategy. MBS also called threshold policy, which determines the batch size based on mean arrival rate and service rate of workstation but without future customers arrival information (Neuts, 1967, Deb and Serfozo 1973, Rulken et al. 1998). With MBS rule, machines start operation while batch size is greater than the threshold, on the other hand, they will keep waiting. Weng and Leachmen (1993) introduced multi-product into MBS and proposed MBSX rule. In this policy, the customer with longest waiting time will have the highest priority in the queue.

The look-ahead strategy was first addressed by Glassey and Weng (1991). They proposed the Dynamic Batch Heuristic (DBH) model with time horizon and take future arrival information into consideration to determine batch size dynamically. In this study, they proved that DBH can perform better performance than MBS. Fowler et al. (1992) introduced the concept of rolling horizon into DBH and proposed Next Arrival Control Heuristic (NACH) policy. NACH emphasize that the start of operation should be decided at every customers arriving and leaving. Van Der Zee et al. (1997) introduced multi-product and multi-server into NACH and proposed Dynamic Job Assignment Heuristic (DJAH). This study indicated that the cost considered should be unit time per item of batch. MBSX and DJAH are both methodologies common used in batch size determination in wafer fabrications. However, these methods are without considering TC issues.

There were some studies related to shop floor control policies with TC. Lee and Jung (2003) proposed distributed shop floor scheduling to decrease the rate of wafers exceeding TC. They found the optimal schedule to meet all the timing and other constraints. Wolfgang and Joerg (2000) developed a Kanban dispatching rule to reduce the influence of TC issues. They considered both machine breakdown and difference of arrival rate between each product to reduce the WIP level. However, these studies did not consider the determination of batch size.

3. Dynamic Batch Job Control with Time Constraints (DBCTC)

In this section, a Dynamic Batch Job Control with Time Constraints (DBCTC) is proposed. The system described in this model involves three workstations, three queues, and two TCs. The configuration of the system is presented as figure 1.

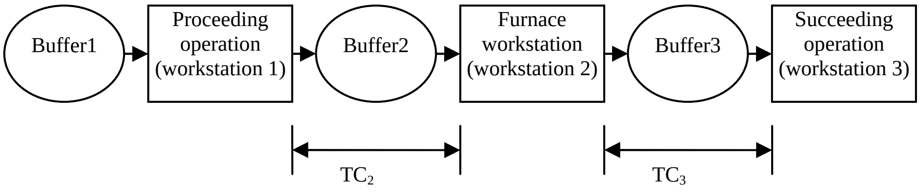

Figure 1. The furnace operations with TC

3.1 Notations

The following terms were required for the DBCTC.

m_i	Number of machines of workstation i
μ_i	Service rate of workstation i
B_i	Current time buffer of workstation i

TR_{ij}	Mean time to be repaired of machine *j* at workstation *i*
C_i	Maximum batch size of workstation *i*
TC_i	Time length of TC in buffer *i*
PT	Processing time of current batch at furnace workstation
q_{ix}	The number of WIP in the queue of product *x* at workstation *i*
W_{x1}	The weight of product x which has not exceeded TC
W_{x2}	The weight of product x which has exceeded TC

3.2 Safety WIP level [S_i,s_i] of workstations

Setting of maximum WIP (S_i) in front of workstation with TC is to prevent wafers exceeding TC. Therefore, S_i should be maximum number of customers that workstation *i* can service within TC interval. The upstream operation should suspend processing if WIP level of workstation *i* is greater than S_i. Moreover, to prevent machine idle, the WIP level should keep at a sufficient quantity to handle machine breakdown at upstream operation. The equations are given below:

$$S_i = (TC_i \times m_i - \sum_{j=1}^{m_i} TR_{ij}) \times \mu_i \qquad (1)$$

$$s_i = S_2 - C_i \qquad (2)$$

3.3 Minimum batch size of furnace workstation

The minimum batch size (*MB*) is set to ensure that WIP can be processed within TC interval. Therefore, minimum batch size should be WIP level divided by the maximum batches that furnace can serve. The equation can be presented as follows.

$$MB = \frac{B_2}{(TC_2 / PT_2) \times m_2} \tag{3}$$

3.4 Tolerance of succeeding operation

The tolerance of succeeding operation (*Tol*) is defined as maximum number that the WIP level of succeeding operation can be increased. The *Tol* is the extra WIP quantity except already in the queue that can be processed within TC interval on succeeding operation. The formula is represented as:

$$Tol = (TC_3 \times \mu_3 \times m_3 - B_3) - (\mu_2 \times m_2 - \mu_3 \times m_3) \times PT \tag{4}$$

3.5 Emergency index of product

The emergency index (I_x) is a weight of product selection for processing at furnace workstation. The I_x is the sum of reciprocals of remaining time to exceed TC and of time length that has exceeded TC. Furthermore, there is a negative-correlation between waiting time and product yield. The weight of WIP which already exceeded TC is negative-correlated with their waiting time.

$$I_x = W_{x1} \times \sum_{l=1, i \in s_1}^{N_{S1}} (\frac{1}{TC_2 - T_{xl}}) + W_{x2} \times \sum_{k=1, k \in S2}^{N_{S2}} (\frac{1}{T_{xk} - TC_2}), \qquad x = 1, \ldots, p \tag{5}$$

Where,

Tx_l is the waiting of the lth lot remaining under TC

Tx_k is the waiting of the kth lot over TC

N_{s1} is the numbers of lot of product x under TC in the queue

N_{s1} is the numbers of lot of product x over TC in the queue

3.6 Shop-floor control rule

I. ***Decision points.*** The decisions should be made at the time of lots arrival or departure from furnace workstation or succeeding operation.

II. ***Decisions of suspending and reinstating proceeding operation***

  i. The proceeding operation should be suspended when B_2 is greater than S_2 or MB is greater than Tol.

  ii. When B_2 is smaller than S_2, the proceeding operation should be reinstated.

III. ***Control rules for furnace workstation***

  i. ***Selection of product.*** The product with largest emergency index should be selected into workstation.

  ii. ***Determination of batch size***. The batch size (Q) should be determined by following rule:

$$Q = \begin{cases} 0, & Tol < 0 \\ Min(C, Tol, q_{2,x}) & Tol \geq 0 \end{cases} \tag{6}$$

4. Simulation Experiments

In this section, simulation experiments were performed to compare the performance performed by proposed model, MBS and DJAH. The simulation model was designed to explore the characteristics of finance workstation and its preceding and succeeding operations. In simulation model, the TCs were set in front and back of furnace workstation. Wafers exceeded the TC set in front of furnace operation will be reworked, but will be marked and continue their processes if they exceed the TC set in back of furnace operation. In the simulation model, there were four products in the system. The monthly demand rate of each product were 960, 1920, 2880 3840 lots (with reentry wafers). Table 1 shows the detailed data of each workstation.

	Number of machines	***TC (hr)***	***MTTR (hr)***	***Availability***	***Maximum batch size***
Preceding operation	3	-	2	95%	1
Furnace workstation	8	9	4	95%	5

Succeeding operation	7	4	2	95%	1

Table 1. Detailed data of workstations

The simulation program used in this research was eM-Plant version 7.0. The running horizon for each simulation was set at 360 days, 24 hours a day. The first 30 days comprised a warm-up period; therefore, the results are for the remaining 330 days. Each treatment was run 30 times to obtain average results.

4.1 Results comparison

In this section, the defective output (output wafers with ever exceeding TC) and average time length of exceeding TC of DBCTC, MBSX and DJAH were compared. Figure 2, 3 and 4 present the results of comparison.

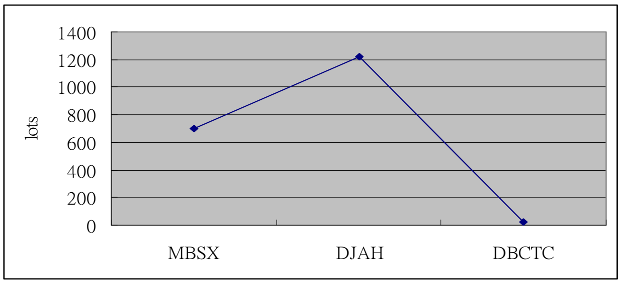

Figure 2. Quantity of wafers exceeded TC set in front of furnace workstation (lots)

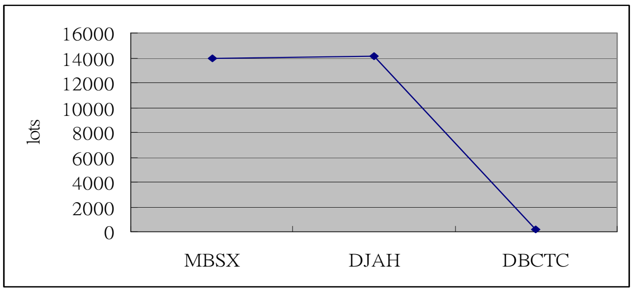

Figure 3. Quantity of wafers exceeded TC set in back of furnace workstation (lots)

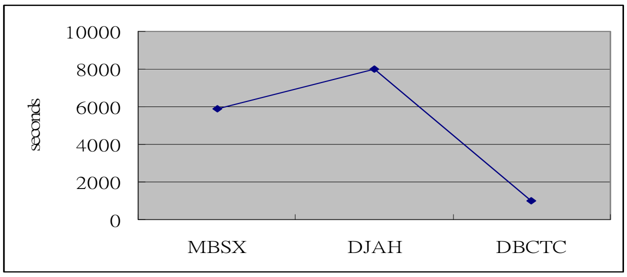

Figure 4. Average time length of exceeding TC

These figures supported that the proposed model could perform better performance in issues of TC with batch processing. Moreover, the results indicated that we should consider the succeeding operation with TC in resolving TC issues at furnace workstation. To maximizing output of furnace workstation will also increasing the loading of succeeding operation. The objective of MBSX and DJAH are both only optimizing the furnace operation, therefore, the performance performed at succeeding operation became undesirable. The proposed model considered furnace workstation and succeeding operation simultaneously, hence, could control the rate of wafers exceeding TC effectively.

5. Conclusions

In this work, a dynamic shop-floor control policy for batch process with TC was proposed. With considering furnace workstation and succeeding operation simultaneously, the rate of wafers exceeding TC could be controlled effectively. Moreover, by adopting concept of [S,s] policy, managers can suspend operation or decreasing batch size dynamically to controlled the WIP level. The simulation experiments support that DBCTC could perform better performance than MBSX and DJAH in TC issues.

The setup time is another critical factor for batch size determination. The required time for machine setting up will affect the result significantly. For instance, long setup time will result in larger batch size to reduce setup times. Future studies should address setup time into the model.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 96-2221-E-216-038.

Reference

1. Deb, R.K. and Serfozo, R.F., 1973. Optimal control of batch service queue. *Applied Probability*, **5**, 340-361.
2. Fowler, J.W., Philips, D.T. and Hogg, G.L., 1992. Real-time control of multi-product bulk-service semiconductor manufacturing process. *IEEE Transactions on Semiconductor Manufacturing*, **5**(2), 158-163.
3. Glassey, C.R. and Weng, W.W., 1991. Dynamic batching heuristic for simultaneous processing. *IEEE Transactions on Semiconductor Manufacturing*, **4**(2), 77-82.
4. Lee, S.H. and Jung, M.Y., 2003. Timing constraints' optimization of reserved tasks in the distributed shop-floor scheduling. *International Journal of Production Research*, **41**(2), 397-410.
5. Neuts, J.D.C., 1967. A general class of bulk queues with poisson input. *Annals of Mathematical statistics*, **38**, 759-770.
6. Robinson, J. K. and Giglio, R., 1999. Capacity planning for semiconductor wafer fabrication with time constraints between operations. *Proceedings of 1999 Winter Simulation Conference*, **1**, 880-887.
7. Rulken, H.J.A., Van Campen, E.J.J., Van Herk, J., and Rooda J.E., 1998. Batch

size optimization of furnace and pre-clean area by using dynamic simulations. *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 439-444.

8. Tu, Y.M. and Liou, C.S., 2006. Capacity determination model with time constraints and batch processing in semiconductor wafer fabrication. *Journal of the Chinese Institute of Industrial Engineers*, **23**(3), 192-199.
9. Wolfgang S. and Joerg D., 2000. Implementation of modeling and simulation in semiconductor wafer fabrication with time constraints between wet etch and furnace operations. *IEEE Transactions on Semiconductor Manufacturing*, **13**(3), 273-277.
10. Weng, W.W. and Leachman R.C., 1993. An improved methodology for real-time production decisions at batch-process work station. *IEEE Transactions on Semiconductor Manufacturing*, **6**(3), 219-225.