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(I)
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2221-E-216-014-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：中華大學工業管理學系

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：陳欣男
碩士班研究生-兼任助理：陳秋玲

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 96 年 10 月 18 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

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（I）

計畫類別：∨ 個別型計畫　□ 整合型計畫
計畫編號：NSC　95-2221-E-216-014
執行期間：　2006 年　8 月　1 日至　2007 年　7　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 陳秋玲 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　96　年　10　月　15　日

一、 中英文摘要

中文摘要

半導體晶圓廠中，解決等候時間限制問題已是不可避免的挑戰。而在晶圓製造的後段製程中，晶圓在製品會經過一連串具有等候時間限制的加工步驟，並且，這一連串加工程序會迴流多次，此種問題稱之為『連續型等候時間限制』。設置適當大小的保護性產能為解決等候實踐限制問題的根本，本研究利用 GI/G/m 等候網路模型分析晶圓在製程間超出時間限制的機率，並探討在可接受的機率下該設置多少保護性產能。此外，並針對晶圓製造廠的特性進行傳統等候模型的修正，特別是針對機台當機行為與批量加工特性加以著墨。透過修正後的模型，產能規劃的結果將更準確與可行。

關鍵詞：產能規劃，等候時間限制，等候模型，晶圓製造

Abstract

Time constraints related issues are unavoidable tasks in wafer fabrications especially for back-end copper-interconnect process while the wafers are processed through sets of continuous operations associate with various time constraints. We shall address ourselves to capacity planning in order to eradicate the difficulties of sequential time constraints. This study proposed a queuing approach capacity planning procedure to calculate the required capacity level ensuring pre-defined target yield which could be achieved. The workstations with time constraint are modeled as GI/G/m queuing network, and the analysis of probabilities of wafers exceeding queue time is adopted to determine the required capacity. Furthermore, virtual customers and were introduced into the model to adjust system service rate to character the machine failure. Moreover, a concept of suppositional machine is proposed to represent the batch/un-batch operations.

Keywords: *capacity planning; sequential time constraint; queuing model; wafer fabrications*

二、 報告內容

1. Introduction

Time constraint (TC) is a time window set between two specific operations to prevent undesirable copper film oxidation or fluorine precipitation on the wafer surface [4, 7]. In the back-end process, TC can be observed by sets of continuous processes which is named as STC (Sequential Time Constraints) in particular. These time windows in the back-end will be squeezed significantly, especially in the copper-interconnect process. The issues of STC are more complicated than of TC because it can not be resolved by controlling the WIP level. Furthermore, the solutions for issues of sequential time constraints could cover which of single TC.

Previous studies have indicated that effective capacity planning is the foundation for effectively managing time constraints [1, 4, 7]. The capacity planning models refer to issues of TC will concentrate on how many protective capacities should be established for controlling the queue time. In semiconductor fabrication, queuing models have generally been applied for capacity planning [2, 3, 8]. However, the factors including reentry flow line, varied product mix, unstable machines, and batch processing, would increase the complexity of the system and aggrandize the

difficulties in capacity planning as well. Therefore, for wafer fabrications, queuing models must be adjusted by factors stated above.

Thus, the purpose of this work is to develop a capacity planning model with sequential time constraints in the back-end. The workstation with time constraint was modeled as GI/G/m queuing system; hence, the back-end process will be modeled as a queuing network. Using the pre-defined target yield (ratio of non-exceeding time constraint, Y_e) to be the limitation of the system, the model can calculate the numbers of each workstation which is required to achieve the goal. Furthermore, novel schemes to modify queuing models in which particular factors of wafer fabrications are proposed, especially unstable machines and batch processing. Hence, we could determine the capacity more effectively in a wafer fabrication.

2. Capacity Planning for Sequential Time Constraints

The objective of this model is to establish necessary number of machines at each workstation with TC in this stage, so that enables managers to ensure the target yield. The capacity determination model proposed by this work applies the GI/G/m queuing network model to solve the sequential time constraints problem.

2.1 The logic of capacity determination model

Figure 1 illustrates a queuing system with time constraint and it reveals that the product queue time could not exceed the time window. In the queuing model, the probability function P(W> x) represents the probability that a customer waits longer than a time period *x* [10]. Let *x* equal the queue time limit and then P(W>x) indicates the probability that wafers exceed the time constraint. Furthermore, if the number of machines (*m*) was increased, the probability value would decrease when the products' arrival rate and mean service time were fixed. Hence, the ratio of over to under TC could be controlled by governing the capacity.

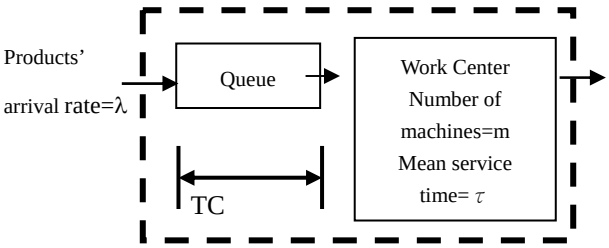

Figure1. Queuing system with time constraint

Capacity determination model for sequential time constraints

Step1. Setting the Target Yield of Each Product

Managers would set an acceptable target yield instead of a perfect ratio. The initial step of this model involves setting an expected target yield (Y_e) for the whole system. It represents that managers expect the ratio of lots that have exceeded TC to those that have not could be controlled under this target. Then, this target will be decomposed to be stage targets for each product at each workstation (Y_f).

$$Y_f = \sqrt[r_f]{Y_e} \quad (1)$$

Step2. Calculate the Parameters of the System

In this step, the mean service time(τ_j), service time variation(Cs_j^2), and arrival rate(λ_j) of each workstation are calculated through equations developed by Whitt in 1983 [9]. To solve the phenomenon of reentry to the production line, the parts of different production steps of individual product classes have to be regarded as different types of customers at each workstation.

Step.3 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" (m_j). 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:

$$m_j = \lfloor \lambda_j \tau_j \rfloor + 1 \tag{2}$$

Step4. Adjust mean service time and service time variation

In this work, virtual customers, whose arrival rate and mean service time will be 1/(MTTR+MTBF) and MTTR respectively, are introduced to characterize the behavior of machine failure. Furthermore, a concept of suppositional machine which processed batched customers is proposed to solve the issues of batch-serial processing.

$$\tau_j^{'} = \frac{\frac{\lambda_j}{b}\tau_j + \sum_{l=1}^{m_j} \frac{\tau_b \times TR_{jl}}{TB_{jl} + TR_{jl}}}{\frac{\lambda_j}{b} + \sum_{l=1}^{m_j} \frac{\tau_b}{TB_{jl} + TR_{jl}}} \tag{3}$$

$$C_{sj}^{2^{'}} = \frac{\lambda_j \tau_j^2 (C_{S_j}^2 + 1) + \sum_{l=1}^{m_j} \frac{\tau_b \times TR_j^2}{TB_j + TR_j} (C_{a_{jl}}^2 + 1)}{(\frac{\lambda_j}{b} + \sum_{l=1}^{m_j} \frac{\tau_b}{TB_{jl} + TR_{jl}}) \tau_j^{'2}} - 1 \tag{4}$$

$$\tau_b = \frac{\tau_j \times b}{m_j} \tag{5}$$

Step5. Compute the variation of inter-arrival time

After mean arrival rate, adjusted mean service time, and SCV of service time are obtained, the SCV of inter-arrival time can be calculated based on the equations proposed by Whitt in 1993 [10].

Step6. Obtain the probability function $P(EW_j \le x)$

From the parameters obtained, the probability of the event in which customer waiting time is smaller than time period x can be calculated.

$$P(EW_j \le x) \approx 1 - \alpha e^{-\eta x} \tag{6}$$

$$EW_j = \frac{C_{aj}^2 + C_{sj}^{2^{'}}}{2} \times \frac{\tau_j^{'}(\rho_j^{\sqrt{2m_j+1}-1})}{m_j(1-\rho_j)} \tag{7}$$

Step7. Determine the required capacity for time constraints machine

From the yield target set above, the low bound of probability that the wait time for each product f exceeds the time constraint can be obtained. If x is equal to TC_{jf}, the probability $P(EW_j \leqq x)$ would be the stage yield of product f at workstation j. The required capacity with regard to

product f (m_{jf}) would be the capacity which satisfies its own stage yield. Finally, the required capacity of workstation j ($m_j^{'}$) would be the maximum number out of all $m_{jf}^{'}$.

$$m_j^{'} = MAX(m_{jf}) \text{ ; for all } f \tag{8}$$

$$m_{jf} = MIN\{m : PW(EW_j \le TC_{jf}) \ge Y_f\} \quad | \quad m \in Integer; \quad m \ge m_j \tag{9}$$

3. Conclusion

In this work, a capacity determination model for sequential time constraints based on a GI/G/m queuing network is proposed (**has already published in *17th International Conference of Pacific Rim Management,* and submitted in *International Journal of Advanced Manufacturing Technology***). The analysis of probabilities of wafers exceeding queue time is adopted to determine the required capacity to reach the target yield. Furthermore, a novel method to address machine failure and batch operations is introduced in this model (**has published in *The 7th Asian Pacific Industrial Engineering and Management Systems Conference***). The behavior of service interrupts and batch/un-batch processing could be observed more accurately and easily through this method.

Sufficient protective capacity is the key factor for resolving issues of time constraint. By allocating exceeding capacity, it can effectively control the queue time of wafers. Nevertheless, the shop-floor control policies would also influence the queue time significantly. For following study, to discuss the shop-floor control policies is the most crucial task for resolving issues of time constraint.

三、 參考文獻

1. Christie, Robert M.E. and Wu, S.D., Semiconductor capacity planning: stochastic modeling and computational studies, *IIE Transactions*, **34**, 2002, 131-143.
2. Connors, D.P., Feigin, G.E. and Yao, D.D., A Queueing network Model for Semiconductor Manufacturing, *IEEE Transaction on Semiconductor Manufacturing*, 9, 1996, 412-427.
3. Louw, L. and Page, D.C., Queuing network analysis approach for estimating the size of the time buffers in Theory of Constraints-controlled production systems, *International Journal of Production Research*, 42, 2004, 1207-1226.
4. Robinson, J. K. and Giglio, R., Capacity planning for semiconductor wafer fabrication with time constraints between operations, *Proceedings of 1999 Winter Simulation Conference*, 1, 1999, 880-887.
5. Tu, Y.M. and Chen H.N., "Model to Determine the Capacity of a Wafer Foundry with Sequential Time Constraints," *17th International Conference of Pacific Rim Management*, 2007, Las Vegas, USA.
6. Tu, Y.M. and Chen C.L., "The Influence of Arrival Smoothing between Batch and Serial Processes on System Performance," *The 7th Asian Pacific Industrial Engineering and Management Systems Conference*, 2006, Bangkok, Thailand.
7. Tu, Y.M. and Liou, C.S., Capacity Determination Model with Time Constraints and Batch

Processing in Semiconductor Wafer Fabrication, *Journal of the Chinese Institute of Industrial Engineers*, 23(3), 2006, 192-199.

8. Uzsoy, R., Lee, C.Y. and Louis, A.M., A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part1: System Characteristics, Performance Evaluation and Production Planning, *IIE Transactions*, 24, 1992, 47-60.
9. Whitt, W., The Queueing Network Analyzer, *The Bell System Technical Journal*, 62, 1983, 2779-2815.
10. Whitt, W., Approximations for the GI/G/m Queue, *Production and Operations Management*, 2, 1993, 114-161.

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場面對等候時間限制問題時產能設置之依據；在學術上，本研究提供一套等候理論應用於晶圓製造廠之修正模型與設置保護性產能之概念。此外，本研究亦已將主要成果發表於國際學術研討會以及國際學術期刊之中。

本研究之主要成果分述如下：

1. 收集半導體製造與等候時間限制相關文獻
   針對關於等候模型、產能規劃、派工法則應用於晶圓製造等相關研究進行分析與整理，總計三十餘篇。
2. 等候理論於時間限制問題之修正
   本研究針對半導體晶圓廠內機台當機、批量加工、迴流製程等主要製造特性進行等候模型之修正，並成功的將這些特性於等候模型中描述。
3. 保護性產能決策模式之建立
   透過本計畫所提出之產能決策模式，能有效地決定保護性產能設置的大小，並且經由保護性產能的設置，達成控制晶圓超過等候時間限制的機率的目標。
4. 簡易晶圓廠模擬環境之建立
   本計畫利用 eM-Plant 7.0 呈現半導體晶圓廠之製造過程與特性，此模擬模型並能提供後續研究之分析與驗證之依據。

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

96 年 7 月 24 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
時間 會議 地點	自 2007 年 7 月 12 日至 2007 年 7 月 14 日 美國拉斯維加斯	本會核定 補助文號	**NSC 95-2221-E-216-014**
會議名稱	（中文）2007 亞太地區管理學術研討會 （英文）2007 International Conference of Pacific Rim Management		
發表論文題目	（中文）晶圓代工廠中連續型時間限制機台之產能決策模式 （英文）Model to Determine the Capacity of a Wafer Foundry with Sequential Time Constraints		

報告內容應包括下列各項：

一、參加會議經過

2007 International Conference of Pacific Rim Management was hold Las Vegas, USA. The conference served as important forum for the exchange of ideas and information to promote understanding and cooperation among the various businesses. In the conference, I presented a paper entitled “Model to Determine the Capacity of a Wafer Foundry with Sequential 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 conference will serve as an important forum for the exchange of ideas and information to promote understanding and cooperation among the various businesses. This year's conference theme is “ Innovation, Sustainable Management, and Global Concern. ” Special focuses will be placed on management innovations, technology transfer, and educational exchange programs and opportunities for educators and practitioners from North America, Taiwan , Hong Kong , China , and other places. A total of 138 papers from different countries around the world were presented in the conference. The conference covered forty-eight important management tracks in the areas like Human Resources, Accounting, E-commerce/E-business, Marketing, Healthcare, Production Technology and Industrial Management, Supply Chain Management, and Finance. The conference featured two distinguished keynote speakers: Mr. C. K. Lee, Chairman, Nevada National Bank, U.S.A. with topic “Commercial Real Estate Investment and Hospitality Industry in the United States,” and Dr. Edwaed H. Chow, Dean, College of Commerence, Nation chengchi University, Taiwan with topic “The Challenge of Management Education in Taiwan.” In addition, a panel discussion with topic “Pursuing High Quality Teaching and Research in Management College” was hold by Dr. Karen Bowerman, Dean, California State University, San Bernadino,CA, U.S.A.. This is a rich and colorful trip not only in the research field but also in the holding process of an international conference. 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

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: The XVII ACME International Conference on Pacific Rim Management.
2. CD of the proceedings.

六、其他

C3902.doc

Model to Determine the Capacity of a Wafer Foundry with Sequential Time Constraints

Ying-Mei Tu
Department of Industrial Engineering & System Management, Chung Hua University
No.707, Sec.2, WuFu Rd., HsinChu, Taiwan 300, R.O.C, amytu@chu.edu.tw, +886-3-5186067
Hsin-Nan Chen
Graduate Institute of Management of Technology, Chung Hua University,
No.707, Sec.2, WuFu Rd., HsinChu, Taiwan 300, R.O.C., d9203014@chu.edu.tw

Abstract: Recently, wafer fabrication has gotten more complicated and lengthened the product queue time significantly. To ensure production line yield, engineers have to set up queue time limits to particularly sequential machines during wafer processing, we name it as "sequential time constraints." This issue can be discovered at back-end of copper-interconnect process. Moreover, wafer can only be scraped or given a mark but can not be reworked when it exceeds queue time limits at these stages.
To eradicate difficulties with sequential time constraints, capacity planning must be addressed. Therefore, this research presents a capacity determination model, employing the GI/G/m queuing model, to eliminate the difficulties causing by sequential time constraints. The probability of waiting time exceeds queue time constraints under certain capacity level can be calculated by this model. Hence, by setting the target of non-exceeding time constraint ratio in advance, we can get the required capacity which can achieve the objective.
Due to the capacity of each workstation will be set according to the acceptable scrap ratio, the level of wafers scrapped resulting from excess queue time can be controlled. Moreover, to make the queuing model more relevant to the wafer fabrication environment, the arrival and service parameters were modified by reentry property of production routing, product scraped and machine failure.

1. INTRODUCTION

As wafer fabrication has gotten more and more complicated, to ensure high yield of products becomes more difficult. Furthermore, in order to prevent copper film oxidation or fluorine precipitation of undesirable interface on wafer surface, technology development engineers will set a time window between specific processes to restrain the wafers cannot wait over the time limit which called time constraint (TC) (Robinson and Giglio, 1999, Tu and Liou, 2006). Originally, time constraint was set between the wet etch and furnace operations. By progress of technology generations, time constrains can also be discovered at a set of continuous operations in back-end process. It is named "sequential time constraints." Wafers will be processed through a series of operations with time constraint, and repeat these steps several times in the back-end of manufacturing process. The wafers, generally, will be scraped or given a noticeable mark rather than be reworked when it exceeded the queue time limits in this stage. Furthermore, in these processes, time constraints will be shrunk acutely especially in copper-interconnect process.
The major difficulty of sequential time constraints in back-end process would be "no refuge anymore." In the front-end, managers can hold the wafers at the queue without queue time limits to control the WIP level of workstation with time constraint. Nevertheless, in the back-end, wafers could not be held at any stage if it has been processed by any operation in the sequential time constraints interval. Therefore, the management of sequential time constraints is more difficult than past. Moreover, as a result of high production volume, expensive equipment, and time consuming process, machines in a wafer fabrication are usually requested to keep at a high utilization level. When resolving issues of sequential time constraint, however, extremely high utilization of workstations might cause the disasters of management.
Previous studies have indicated that an effective capacity planning is the foundation for conquering time constraints (Christie et al., 2002, Robinson and Giglio, 1999, Tu and Liou, 2006). These researches developed capacity planning models to determine the required capacity with time constraint by applying Queuing Theory. Nonetheless, these researches only addressed the issues in the front-end of wafer fabrication, which is much simpler than in the back-end. Furthermore, different ways or dispatching rules for handling the wafers exceeded time constraint would influence the capacity allocation. To be scraped or to be set to lowest priority when wafers exceeded time constraint, for instance, would require lower capacity because of the change of arrival rate. Previous studies did not discuss these impacts particularly.
As proved efficient and accurate schemes, queuing models have been usually applying for capacity planning and cycle time estimation in production systems (Connors et al., 1996, Uzsoy et al., 1992, Louw and Page, 2004). Nevertheless, applying for wafer fabrications, queuing models must be modified by many factors to fulfill semiconductor manufacturing characteristics, particularly in reentry lines, products reworked or scraped as well as unstable machines (Connors et al., 1996). According to investigation, the machine breakdown is the key issue of sequential time constraints. To address machine failure, previous researches had modified queuing models by machine availability (Connors et al., 1996, Louw and Page, 2004). In these studies, the available capacity could be considered as raw capacity multiplied by machine availability. However, they did not consider the impact of varied downtime duration at fixed machine

availability. Schoemig has pointed out the downtimes frequency had a significant impact on variability while the availability of the machines was fixed (Schoemig, 1999). The downtime duration, hence, must be considered in the modification of machine failure.
The purpose of this paper, thus, is to develop a capacity planning model with sequential time constraints by applying GI/G/m queuing model. Through the target yield (ratio of non-exceeding time constraint, Y_e) set by managers, the model could calculate how much capacity of each workstation was required to achieve the goal. Furthermore, a novel scheme which applying both of availability and downtime duration was proposed to address machine breakdown. By this approach, the impact of downtime duration and downtime frequency at fixed machine availability can be observed clearly. Moreover, the impacts of different ways and dispatching rule for handling wafers exceeded time constraints will be discussed.

2. LITERATURE REVIEW

There are many studies addressed queuing models applied for capacity planning (Adersson and Olsson, 1998, Chen et al., 1988, Hung and Leachman, 1996, Leachman and Carmon, 1992). According to Lazowska, queuing network models can be expected to be accurate within 90~95% for throughput rates estimation and within 70~90% for cycle time estimation separately (Lazowska et al., 1984). These factors have always played as important roles in production systems (Chung and Huang, 2002). However, many unreasonable assumptions existed in conventional queuing models, for example, no server failure was assumed. In a manufacturing system, the behavior of machine failure will be a critical factor in capacity allocation (Uzsoy et al., 1992).
To address machine failure, many researches have strived to modify the queuing models. Louw and Page proposed a cycle time estimation model using Queuing Theory and modified GI/G/m queuing model by applying machine availability (Louw and Page, 2004). Mitrany and Avi-Itzhak have proposed a multi-server queuing system with service interruptions based on M/M/s queuing model and modified by using the Markov-Chain (Mitrany and Avi-Itzhak, 1968). Hopp has proposed a queuing model with server failure which modified by machine availability and MTTR at the same time (Hopp, 1996). However, this model only addressed single machine in the system, it was too simple to present complicated environments. These studies on machine failure, the duration of downtimes either was not taken into account, or was too complicated for practical applications.
There were many studies discussed capacity planning to solve issues of time constraints. Robinson and Giglio proposed a capacity planning methodology based on M/M/c queuing scheme to solve the time constraint between the wet etch and furnace operations (Robinson and Giglio, 1999). They calculated the rework probability and recomputed the required capacity of the wet etch workstation. In this model, however, they did not consider the capacity of the furnace operation which directly influenced the probability of rework. Fundamentally, it should take a good planning on succeeding workstation for solving issues of time constraints.
Tu and Liou developed a capacity determination model based on GI/G/c queuing approach to address time constraints with batch processing (Tu and Liou, 2006). They determined required capacity of furnace workstation based on length of time constraint. Furthermore, the relationship among the over time constraint probability machine utilization, and length of time constraint was discussed. In this study, however, issues in the back-end were not addressed. Machines were assumed no failure occurred; this could result in inaccuracy of capacity allocation.
Capacity allocation is the critical for conquering time constraints, hence, capacity planning must be considered first. For the back-end of wafer manufacturing, it is important to develop a capacity determination model with time constraints and a novel approach to address machine failure.

3. CAPACITY DETERMINATION MODEL FOR SEQUENTIAL TIME CONSTRAINTS

In this section, the capacity determination model for the back-end of wafer fabrication with sequential time constraints will be described. As stated above, there are many ways to handle the wafers which exceeded time constraints. Given a noticeable mark is the most common way among them, moreover, it has the lowest impact on capacity allocation. Therefore, the basic model was based on the way of given a noticeable mark on the wafers which exceeded time constraints.
The capacity determination model proposed by this work applies the GI/G/m queuing model to solve the sequential time constraints problem. The objective of this model is to establish necessary capacity level that enables managers to ensure the target yield.

3.1 The Logic of Capacity Determination Model

Figure 1 illustrates a queuing system with time constraint. It reveals the product queue time could not exceed the time window. In the queuing model, the probability function $P(W> x)$ represents the probability of a customer waits longer than a time period x (Whitt, 1993.) Let x equal queue time limit, $P(W>x)$ indicates the probability of wafers exceed the time constraint. Furthermore, if the number

of machines (m) was increased, the probability value would decrease when the products' arrival rate and mean service time were fixed. The ratio of over TC, hence, could be controlled by dominating the capacity.

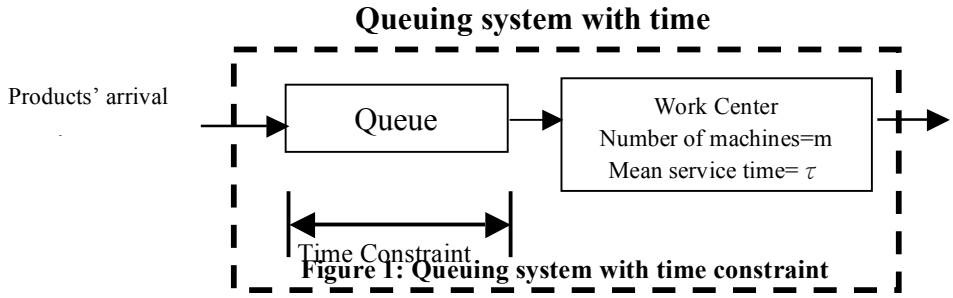

Figure 1: Queuing system with time constraint

3.2 Capacity Determination Model for Sequential Time Constraints

The previous section describes the capacity determination methodology for a system with time constraint. This conception could be adopted when managers deal with the issues of sequential time constraints. By this methodology, the target yield set by managers could be achieved through allocating necessary capacity.

3.2.1 Notations

The following data were required for the capacity determination model.

arrival data

λ_{of} demand rate of product *f*

service data

τ_{fk} service time of product *f* at its *k*th operation step

$C^2_{s_{fk}}$ the SCV of service time of product *f* at its *k*th operation step

TB_{lj} mean time between failures (MTBF) of machine *l* of workstation *j*

TR_{lj} mean time to repair (MTTR) of machine *l* of workstation *j*

$C^2_{d_{lj}}$ the SCV of downtime of machine *l* of workstation *j*

m_j number of machines at ~~a~~workstation *j*

routing data

s_{fk} the workstation visited by product *f* at its *k*th operation step

t_f total number of operation steps of product *f*

r_f number of workstations with time constraint on the production route of product *f*

TC_{jf} time constraint of product *f* at workstation *j*

n total number of workstation in the system

3.2.2 The capacity determination Procedure

Stpe1. Setting the Target Yield of Each Product

From the perspective of time constraint, pervious study indicated that the marginal cost for decreasing the over time constraint

probability would be getting greater and greater (Tu and Liou, 2006). Managers would set an acceptable target yield instead of perfect ratio. The initial step of this model would be setting an expected target yield (Y_e) for whole system. It represents that mangers expect the ratio of lots have exceeded TC could be controlled under this target. Then, this target will be decomposed to be stage targets of each product at each workstation.

This model assumed that within the sequential time constraints system, the target yield of each product will be the same. Based on the visiting times of time constraint machines of products on their production routing, the stage yield of product *f* (Y_f) could be calculated. The equation is given below:

$$Y_f = \sqrt[r_f]{Y_e} \tag{1}$$

Step2. Calculate the Parameters of the System

In this step, the mean service time, service time variation, and arrival rate of each workstation are calculated. To solve the phenomenon of reentry production line, the parts of different production step of individual product classes have to be regarded as different type of customers at each workstation. The detailed process is stated as follows:

Step2.1 Aggregating mean arrival rate

The first step of parameters calculation is aggregating arrival rates of individual product into the mean arrival rate (λ_j) of the workstation. The formula is represented as follows.

$$\lambda_j = \sum_f \sum_{\{k|s_{fk}=j\}}^{t_f} \lambda_{of} \tag{2}$$

Step2.2 Aggregating mean service time and service time variation

After mean arrival rate was accumulated, the service time and squared coefficient of variation (SCV) of service of individual products also have to be aggregated into mean service time (τ_j) and SCV of service time (C^2_{sj}) of the workstation. Follows are the equations for aggregation:

$$\tau_j = \frac{\sum_f \sum_{\{k|s_{fk}=j\}}^{t_f} \lambda_{of} \tau_{fk}}{\lambda_j} \tag{3}$$

$$C_{sj}^2 = \frac{\sum_{f=1} \sum_{\{k|s_{fk}=j\}}^{t_f} \lambda_{of} \tau_{fk}^{\ 2} (C_{sfk}^2 + 1)}{\lambda_j \tau_j^{\ 2}} - 1 \tag{4}$$

Step.3 The Loading Capacity Determination

In this section, the minimum capacity that can meet system's basic requirement was determined, which was defined as "loading capacity". From the definition of Queuing Theory, the traffic intensity (ρ) must be smaller than one to keep the steady-state of the system. Therefore, the loading capacity would be the smallest integer *m* that greater than arrival rate divided by service rate, it can be presented as follows:

$$\rho_j = \frac{\lambda_j \tau_j}{m_j} < 1 \tag{5}$$

Therefore,

$$m_j = \lfloor \lambda_j \tau_j \rfloor + 1 \tag{6}$$

Step4. Adjust mean service time and service time variation under machine failure

This section will introduce a novel scheme to reveal the effect of machine failure which adopted availability and downtime duration. 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. To reveal the effect of machine interrupts, mean service time and SCV of service time were modified. The adjusted mean service time ($\tau_j^{'}$) and SCV of service time ($C_{sj}^{2'}$) are presented as follows:

$$\tau_j^{'} = \frac{\lambda_j \tau_j + \sum_{l=1}^{m_j} \frac{TR_{jl}}{TB_{jl} + TR_{jl}}}{\lambda_j + \sum_{l=1}^{m_j} \frac{1}{TB_{jl} + TR_{jl}}} \tag{7}$$

$$C_{sj}^{2'} = \frac{\lambda_j \tau_j^2 (C_{S_j}^2 + 1) + \sum_{l=1}^{m_j} \frac{TR_j^2}{TB_j + TR_j} (C_{d_{jl}}^2 + 1)}{(\lambda_j + \sum_{l=1}^{m_j} \frac{1}{TB_{jl} + TR_{jl}}) \tau_j^{'2}} - 1 \tag{8}$$

Step5. Compute the variation of inter-arrival time

After mean arrival rate, adjusted mean service time and SCV of service time are obtained, the SCV of inter-arrival time can be calculated. It can be obtained as follows (Whitt, 1983):

$$C_{aj}^2 = \alpha + \sum_{i=1}^{n} \beta C_{ai}^2 \tag{9}$$

Where α and β were referred to (Whitt, 1983)

Step6. Obtain the probability function $P(EW_j \le x)$

From the parameters obtained, the probability which customers waiting time smaller than time period x can be calculated. It can be obtained as follows:

$$P(EW_j \le x) \approx 1 - \alpha e^{-\eta x} \tag{10}$$

$$\eta = 2m_j(1-\rho_j)/(c_{aj}^2 + c_{sj}^{2'}) \tag{11}$$

$$\alpha \approx \eta EW_j \tag{12}$$

$$EW_j = \frac{c_{aj}^2 + c_{sj}^{2'}}{2} \times \frac{\tau_j^{'}(\rho_j^{\sqrt{2m_j+1}-1})}{m_j(1-\rho_j)} \tag{13}$$

Step7. Determine the required capacity for time constraints machine

From the yield target we set above, the low bound of probability of each product *f* waits exceeded time constraint can be obtained. Let x equal to TC_{jf}, the probability $P(EW_j \le x)$ would be the stage yield of product *f* at workstation *j*. The required capacity regards with product *f* (m_{jf}) would be the capacity which satisfying its own stage yield. Finally, the required capacity of workstation *j* ($m_j^{'}$) would be the maximum number of all $m_{jf}^{'}$. The equations were shown as follows:

$$m_j^{'} = MAX(m_{jf}) \text{ ; for all } f \tag{14}$$

$$m_{jf} = MIN\{m : PW(EW_j \le TC_{jf}) \ge Y_f\} \quad | \quad m \in Integer; \quad m \ge m_j \tag{15}$$

From the procedures proposed in this work, required capacity for sequential time constraints can be determined. The probability of wafers over queue time limit can also be controlled below the target set by managers. In the following section, we will discuss the influences of different handling methods of wafers which exceed time constraints on capacity allocation.

4. DISCUSSION OF DIFFERENT SCENARIOS

The capacity determination model proposed in above section was based on the scenario that given a noticeable mark to the wafers when its queue time exceeded time constraint. As stated, different action would influence the capacity allocation. However, should the capacity determination model be re-defined or just be adjusted somehow in different scenario? This is an interesting question. Some common ways to handle wafers exceeding time constraint were discussed, furthermore, the adjusted model will be proposed in this section.

4.1 Wafers Exceeding TC Are Set to Be Lower Priority in the Queue

In this scenario, when the wafers have ever exceeded TC in its production route, it will be set lower priority than normal ones. This control policy is based on the goal of minimizing total amount of lots ever exceeded time constraints. From some managers' opinions, they would rather prefer wafers exceeded time constraints in each workstation are always the same lots, so that could decrease the probability of over TC of other lots. The total amount of lots ever over TC, hence, could be decreased.
In this policy, the mean arrival rate and mean service time of each workstation would not be changed. Consequently, the probability function $P(EW_j > x)$ was not shifted and the proposed model does not have to be modified in this scenario. However, because wafers exceeded TC were always the same lots, the accumulation effect of over TC ratio will be disappeared. Hence, the stage yield of product should be adjusted. The equation (1) could be modified as follows:

$$Y_f = Y_e \tag{16}$$

While the target yield is smaller than one, the stage yield of product would smaller than presented before. In this scenario, it will require smaller capacity for handling time constraints.

4.2 Wafers Exceeding TC Are Set to Be Higher Priority in the Queue

In the back-end of wafer fabrication, if wafers exceeded TC many times, it will lead to the lower yield finally. When the lot has been over TC once, it will be processed as soon as possible in the remaining processes. In this scenario, the lot will be set as high priority if over TC. This control policy would aspire to minimize the times of over TC of each lot.
In this scenario, same as above, the parameters were not changed and the probability function was not shifted as well. The capacity determination procedure does not to be modified accordingly. Furthermore, the lots wait over queue time limit in each workstation were varied, the accumulation effect exists. The capacity allocation procedure would not need to be modified at all. Nevertheless, this model will result in the largest amount of lots ever exceeding TC, and the least times of over TC in these lots. In other words, this scenario would be the worst case of the model presented in section 3. Even so, the determined capacity could achieve the target yield.

4.3 Wafers Exceeding TC Are Scraped

No matter how many times or how long the lots exceed time constraints, the yield of these lots will be impacted. Furthermore, some final products have to be scraped after completion. However, this situation could waste the capacity, which was the most precious resource in a fab. It will result in lower throughput, longer lead-time, and lower on time delivery. So, managers would scrape WIP rather than final products. Some managers were cautious with this problem. They decide to scrape the wafers when over TC to avoid the wastes of capacity
In this scenario, because of some wafers were scraped, input volume were increased to fill the demand. Furthermore, arrival rate of each workstation would be varied. The ratio of lots passed to the succeeding workstation would remain the stage yield of the workstation. Hence, the equation (2) could be modified to be:

$$\lambda'_{of} = \lambda_{of} / Y_e \tag{17}$$

$$\lambda_j = \sum_f \sum_{\{k|s_{fk}=j\}}^{t_f} Y_f^{h_k} \lambda'_{of} \tag{18}$$

Where, h_k= number of workstation with TC that the product f has visited on its kth operation

Furthermore, because a part of wafers in the queue will leave the system without operation, hence, the mean service time of the system will be changed. The scraped lots could be the customers with service time equal to zero, thus, the equation (3) become:

$$\tau_j = \frac{\sum_f \sum_{\{k|s_{fk}=j\}}^{t_f} \lambda_{of} Y_f \tau_{fk}}{\lambda_j} \tag{19}$$

After these parameters have been adjusted, we could determine the required capacity under such a situation. However, because of the number of machines must be an integer value, it might cause a discrepancy between stage yield and result of equation 10. The adjusted demand rate and arrival rate as above would be incorrect. This phenomenon will result in over-estimating of the required capacity in the proposed model. To solve this problem, it should replace the stage yield by the result of equation 10 and re-determine the required capacity again. Nonetheless, a deviation still exists between the newly estimated yield and the previous one. Therefore, the procedure stated above should be repeated by newly yields superseding old ones. This process would be repeated until the newly determined capacity is equal to last one.

5. CONCLUSION

In this work, a capacity determination model for sequential time constraints based on GI/G/m queuing model was proposed. The required capacity for reaching the target yield set by managers can be determined. Furthermore, a novel method to address the machine failure was introduced in this model. The behavior of service interrupts could be observed more accurate and easier. Moreover, different ways for treating the wafers which has ever exceeded the time constraints were discussed. The model was also modified to fit these scenarios.

The priority of each lot was given before they arrived each workstation and its priority will be fixed during the waiting process. However, the priority of lots might be changed in the meantime of queuing for some situations in the real world. For instance, a lot could be set to be high priority immediately while it exceeds the queue time limit. For the future study, it could to focus on the models with dynamic priorities.

REFERENCE

1. Andersson, M. and G. Olsson, 1998, A Simulation Based Decision Support Approach for Operational Capacity Planning in a Customer Order Driven Assembly Line, Proceedings of the 1998 Winter Simulation Conference, 2, 935-941.
2. Chen, H., J. M. Harrison, A. Mandelbaum, V. Ackere and L. M. Wein, 1988, Empirical evaluation of a queuing network model for semiconductor wafer fabrication, Operations Research, 36(2), 202-215.
3. Christie, Robert M.E. and S.D. Wu 2002, Semiconductor capacity planning: stochastic modeling and computational studies, IIE Transactions, 34, 131-143.
4. Chung, S.S. and H.W. Huang, 2002, Cycle Time Estimation for Wafer Fab with Engineering Lots, IIE Transactions, 34, 105-118.
5. Connors, D.P., G.E. Feigin and D.D. Yao, 1996, A Queueing network Model for Semiconductor Manufacturing, IEEE Transaction on Semiconductor Manufacturing, 9, 412-427.
6. Hopp, W.J. and M.L. Spearman, 1996, Factory Physics, Irwin McGraw-Hill, New York.
7. Hung, Y. F. and R. C. Leachman,1996, A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations, IEEE Transactions on Semiconductor Manufacturing, 9(2), 257-269.
8. Leachman, R. C. and T. Carmon, 1992, On Capacity Modeling for Production Planning with Alternative Matching Types, IIE

Transactions, 24(4), 62-72.
9. Louw, L. and D.C. Page, 2004, Queuing network analysis approach for estimating the size of the time buffers in Theory of Constraints-controlled production systems, International Journal of Production Research, 42, 1207-1226.
10. Mitrany, I.L. and B. Avi-Itzhak, 1968, A Many-Server Queue with Service Interruptions, Operations Research, 16(3), 628-638.
11. Robinson, J. K. and R. Giglio, 1999, Capacity planning for semiconductor wafer fabrication with time constraints between operations. Simulation Conference Proceedings, Winter, Volume 1.
12. Schoemig, A. K., 1999, On the Corrupting Influence of Variability in Semiconductor Manufacturing, 1999 Winter Simulation Conf. (WSC), 837–842.
13. Tu, Y.M. and C.S. Liou, 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.
14. Uzsoy, R., C.Y. Lee and A.M. Louis, 1992, A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part1: System Characteristics, Performance Evaluation and Production Planning," IIE Transactions, 24, 47-60.
15. Whitt, W., 1983, The Queueing Network Analyzer, The Bell System Technical Journal, 62, 2779-2815.
16. Whitt, W., 1993, Approximations for the GI/G/m Queue, Production and Operations Management, 2, 114-161.

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