This article was downloaded by: [National Chiao Tung University 國立交通大學]
On: 27 April 2014, At: 22:06
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

IIE Transactions
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uiie20

Cycle time estimation for wafer fab with engineering lots

SHU-HSING CHUNG [a] & HUNG-WEN HUANG [a]
[a] Department of Industrial Engineering and Management , National Chiao Tung University , Hsin-Chu, Taiwan, ROC E-mail:
Published online: 17 Apr 2007.

To cite this article: SHU-HSING CHUNG & HUNG-WEN HUANG (2002) Cycle time estimation for wafer fab with engineering lots, IIE Transactions, 34:2, 105-118, DOI: 10.1080/07408170208928854

To link to this article: http://dx.doi.org/10.1080/07408170208928854

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

IIE Transactions (2002) **34**, 105–118

Cycle time estimation for wafer fab with engineering lots

SHU-HSING CHUNG* and HUNG-WEN HUANG

Department of Industrial Engineering and Management, National Chiao Tung University, Hsin-Chu, Taiwan, ROC
E-mail: t7533@cc.nctu.edu.tw

Received November 1999 and accepted June 2001

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

Due to the interaction between the process complexity and equipment diversity in a wafer fab, it is rather difficult to estimate the material flow time of wafer lots. Facing competition, it is common for a wafer fab to produce a certain quantity of engineering lots. However, introducing engineering lots into the factory will increase the complexity of the material flow control and the difficulty in cycle time estimation. The purpose of this paper is to develop cycle time estimation algorithms for a wafer fab by analyzing the material flow characteristics. Simulation results have shown that the algorithm is capable of generating satisfactory cycle time estimations with or without existing engineering lots.

1. Introduction

The cycle time has always played a very important role in production planning and control systems. For production schedule development, due date assignment and production capacity requirement planning, the cycle time of each product type is used as the planning basis.

The difficulty in estimating the cycle time is mainly caused by the uncertain queueing time. Queueing time is the time that a lot waits for an available machine and it is a result of heavy machine loading. However, the queueing time can also be a result of batch accumulation. In a wafer fab, both batch and serial machines are used. Batch machines can process wafers in one or more lots (one lot, in general, is equal to 25 wafers), while serial machines only process wafers equal to or less than one lot at a time. Since the operation time for a batch machine is expected to be longer than that for a serial machine, a restriction for loading "at least the minimum batch size simultaneously" is usually imposed to effectively use the machines limited capacity. As a result, even if there is a batch machine idle, lots may be forced to wait if they do not meet the minimum batch size condition. This kind of queueing time depends on the batch size setting and not on the machine loading.

There are around 300 to 400 process steps in each wafer's production process. Meanwhile, there are more than 80 different machine types and the maximal batch size for each machine can be in lots of one, two, four or six. The re-entry properties of the wafer fabrication process and the equipment diversity are the major factors that make the prediction of the queuing time difficult. In addition, it is very common for a wafer fab to process a certain quantity of engineering lots or technology development lots. Since these lots are for experimental purposes, their process recipes are usually different from that of normal lots. Hence, engineering and normal lots cannot be mixed together. The high operation priority of engineering lots has a great impact on production planning and control, and processing an engineering lot particularly results in the loss of machine capacity because batch machines are not necessarily fully loaded. Hence, the impact on cycle time for the normal lots because of introducing engineering lots into the system must be evaluated.

Whether the cycle time can be predicted accurately depends on the ability to effectively identify the material flow characteristics in a factory. The material flow speed for wafer lots depends on the interactions among the process and equipment properties, product mix and production control policies. Clearly, full recognition of the material flow characteristics in a wafer fab is the first step for cycle time estimation.

The main work of this paper is to develop a cycle time estimation mechanism to predict the cycle time of each product type with the existence of engineering lots. A quick response time and satisfactory estimation are the distinctive features of this mechanism. Since queueing theory is applied in the proposed cycle time estimation mechanism, the utilization rate of each workstation can be calculated.

The cycle time estimation algorithms developed here are based on the following assumptions:

- There are two different lot types, engineering and normal, in the system.

*Corresponding author

0740-817X © 2002 "IIE"

- The First-In-First-Out (FIFO) dispatching rule is applied separately to both engineering and normal lots.
- The minimum batch size restriction is not applicable to engineering lots.
- The engineering lots cannot be processed together with other normal lots.
- Each engineering lot cannot be processed in conjunction with another engineering lot (due to different experimental purposes).
- For each experimental study, only one engineering lot (25 pieces) is introduced.
- The engineering lot operations are non-preemptive.

2. Literature review

Cycle time estimation algorithms can be classified into the following four categories:

1. Simulation. Discrete event simulations are used to simulate the lot cycle time that occurs in a wafer fab. Because of the high complexity of wafer manufacturing, Atherton and Atherton (1995) believed that a simulator is the only tool to describe the dynamic behavior in wafer fab. Glassey and Resende (1988), Wein (1988), Cunningham (1990), Matsuyama and Atherton (1990), Glassey and Weng (1991), Etheshami *et al.* (1992), Fowler *et al.* (1992), Narahari and Khan (1997), Wood (1997) and Kim *et al.* (1998) have adopted the discrete event simulator as a verification tool; however, their major interest is to use the simulation to perform a comparison between varieties of scheduling policies.

2. Statistical analysis method. Regression analysis or some other statistical analysis method is applied to determine the relationship between the cycle time and other related parameters. Raddon and Grigsby (1997) have built a regression model to estimate the lot cycle time. Based on this model, the lot cycle time deviation compared with the actual data is within ±2 days. Since historical conditions are used to forecast the future, the greater the changes in the system, the less accurate are these statistical analysis methods. Thus, Enns (1995) believed that the models developed with such an approach do not have a unified applicability and are only useful for short-term estimations.

3. Analytical method. The analytical method is based primarily on queueing theory or some other mathematical model to derive the lot cycle time and its deviation.

Martin (1998) used basic queuing theory to develop the actual-to-theoretical cycle time ratio, the X-factor. The X-factor formula is expressed as:

$$X\text{-}factor = \frac{1 - \rho/2}{1 - \rho}, \tag{1}$$

where ρ denotes the machine utilization. Furthermore, Martin (1998) developed overall X-factor system estimation (X_{OA}). The X_{OA} is generated with the X-factors of both the bottleneck and non-bottleneck machines, weighted according to their proportion in the total system process time. The planned cycle time will then be equal to X_{OA} multiplied by the total process time.

Conway *et al.* (1967) derived the cycle time estimation formula for a single machine using the Laplace transformation formula, as shown below:

$$E(X) = E(P) + \frac{\lambda E(P^2)}{2(1 - \rho)}, \tag{2}$$

where $E(X)$ is the expected product cycle time, $E(P)$ is the expected actual process time, $E(P^2)$ is the square of the expected process time and λ is the arrival rate of the product.

Su (1998) modified Equation (2) by including the batch sizes of the batch machines into the model. The modified model is shown below (*ABS* denotes the average batch size of the batch machines):

$$E(X) = E(P) + \frac{\lambda E(P^2)}{2(1 - \rho) \times ABS}. \tag{3}$$

Usually current system conditions are reflected by the performance of the lots that have just been processed through the operations. Vig and Dooley (1991) thus applied the dynamic feedback and updated methods to calculate the average cycle time for each process step based on the cycle time of the last three completed lots.

4. Hybrid method. The hybrid method combines different methods to produce a cycle time estimation. For example, by consideration of the lot characteristics and the actual load, Enns (1995) applied analytical methods and simulations to develop a dynamic cycle time estimation method. Moreover, Kaplan and Unal (1993) combined simulation and statistical analysis methods to estimate the cycle time.

Regarding the impact of hot (or engineering) lots in the production system, Atherton and Atherton (1995) believed that the loss in production capacity caused by a hot lot is a result of the existence of a more complicated process, more process steps, a higher re-entry frequency and a longer processing time. Increasing the number of hot lots may cause the bottleneck to shift, which will make production planning and capacity assignments ineffective.

Etheshami *et al.* (1992) proposed another view of the influence that introducing a hot lot into a fab has on other lots. When the proportion of hot lots in a fab increases, the average cycle time of the system will remain a constant but the standard deviation of the system cycle time will increase sharply. An increase in the hot lot ratio also makes the average cycle time and the standard deviation of normal lots increase significantly. Miller (1989) and Fronckowiak *et al.* (1996) have also obtained the same conclusions as Etheshami *et al.* (1992) for a R&D fab.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

In the above studies, simulation was usually used as a tool to point out that the system performance would deteriorate after the introduction of a hot lot. However, since simulation results themselves do not possess any general applicability, the cycle time under a specific system environment still cannot be estimated. Therefore, a cycle time estimation algorithm is developed in this paper in accordance with production systems that have engineering lots.

3. Block-based cycle time estimation algorithm with an engineering lot

3.1. *The basic principle and structure of block-based cycle time estimation algorithm*

Observing the material flow in a wafer fab, we found that the queueing time was determined primarily by the following two factors: (i) the workstation load level; and (ii) the differences in batch size and throughput rate between upstream and downstream workstations. In this paper, we define the above factors as *the load-factor* and *the batching-factor* respectively.

Batch rules and dispatch rules will control the flow speed of wafer lots. A batch rule has a higher priority than a dispatch rule. When a lot arrives at the batch workstation for a specific operation, the system counts the number of lots that belong to the same product type and priority class to check if the minimum batch size condition is met. If the condition is met, the lots will then be regarded as a candidate lot-batch. Each candidate lot-batch is allowed to be processed in that workstation according to its arrival sequence. If the number of lots with the same product type and priority class does not meet that condition, the batch forming behavior will be continued until the minimum batch size condition is satisfied. Furthermore, if the number of lots is greater than the maximum batch size for that batch machine, the excess lots will be separated into a new lot-batch, which will again be verified as to whether the minimum batch size condition is met.

The above phenomenon states that a queueing line in front of a batch workstation can be divided into a *batch-control queue area* and a *priority-control queue area*. Lots located in each area correspond to waiting for batch formation or for machine availability, respectively. Obviously, the queue time that a lot spends in the priority-control queue area has a direct connection to the machine load. Contrarily the queue time that a lot spends in the batch-control queue area is related to the minimum batch size setting and the respective throughput rate of the upstream and downstream workstations. The lower the minimum batch size, the shorter the time required for batch formation. The lower the throughput rate of the upstream workstation, the longer the time required for batch formation.

Based on distinction of batch-control queue area and priority-control queue area, a further analysis on the lot statuses in these queue areas and the respective type of queue time can be made. This concept is shown in Fig. 1.

1. B Type: The lot has a batch formation status or is a temporarily peak loading queue status. The lots in this status come from one of the following two conditions: (i) the machine is in an idle state when the lot arrives at the batch-control queue area; or (ii) a temporarily peak load occurs. Temporarily peak loading happens when the batch size adopted in the upstream workstation is larger than that of the downstream workstation. Thus, the load built on

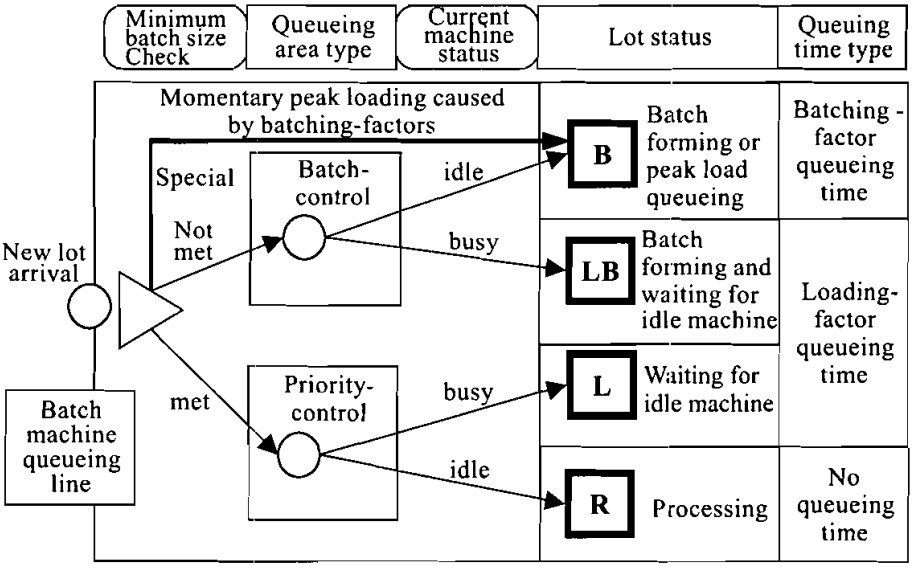

Fig. 1. The classification of lot statuses and queueing time type.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

the downstream workstation will be greater than the throughput capability provided by all machines in that workstation. Since the queue times resulting from the foregoing two conditions are caused due to batch-factors, are classified into the batching-factor type of queue.

2. LB Type: The lot has a batch formation status and the machine is busy. Under this circumstance, the lot is not only waiting for an available machine but also it is queueing for batch formation. Since the queue time for an available machine is related to the load factor, it will be formulated using the loading-factor queue time model.
3. L Type: The lot has a priority-control queue status, and the machine is busy. Under this circumstance, the lot-batch has met the minimum batch size. Hence, this type of queue is caused purely by machine load. Therefore, such a queue time is formulated using the loading-factor queue time model.
4. R Type: The lot is in the priority-control queue area and the machine is idle. If this lot is the only lot-batch in the queue line or if it is the highest priority lot-batch, this lot-batch will be immediately processed. Hence, its queueing time in the priority-control queue area is zero. If this lot-batch do not have the highest priority, then it is classified as L-type instead of R-type.

Next, we will show the impact that results from introducing an engineering lot into the system.

3.1.1. *The queueing time characteristics of engineering lots*

Since an engineering lot is not restricted by the batching rule, it does not need to wait for batch formation. Moreover, since the process of every engineering lot on a batch machine is a single lot operation, no temporarily momentary peak load problem will be caused for downstream machines. Therefore, an engineering lot does not have a batching-factor queue time.

The queue time faced by an engineering lot is caused by the loading-factors. The major two circumstances are:

1. An engineering lot waits for the machine to change from a busy status to an idle status.
2. An engineering lot is forced to wait as another engineering lot books the machine with a higher priority (the FIFO dispatching rule is applied to engineering lots).

Since the product mix ratio of engineering lots in a fab is rather low, the queueing time caused by the above two circumstances will be relatively low. A sketch map of these concepts is shown in Fig. 2.

3.1.2. *The queueing time characteristics of normal lots*

The introduction of an engineering lot will cause an increase in both the batching-factor and loading-factor queue times for normal lots. An increase in the loading-factor queue time is caused by an increase in the probability for a normal lot to wait for an available machine because the normal lot has a lower priority than an engineering lot. An increase in the batch formation time is caused by a decrease in the normal lot flow rate. Clearly, an engineering lot causes a capacity loss on a batch machine because it is not a full load. The available capacity for normal lots is consequently decreased and the respective queue time is increased. A sketch map of the foregoing concepts is shown in Fig. 3.

Based on the above descriptions, the cycle time for a lot-batch to flow through the entire factory includes the following three factors:

1. Queue time resulting from the loading-factors: T_Q. The queue time increases as the workstation load increases.

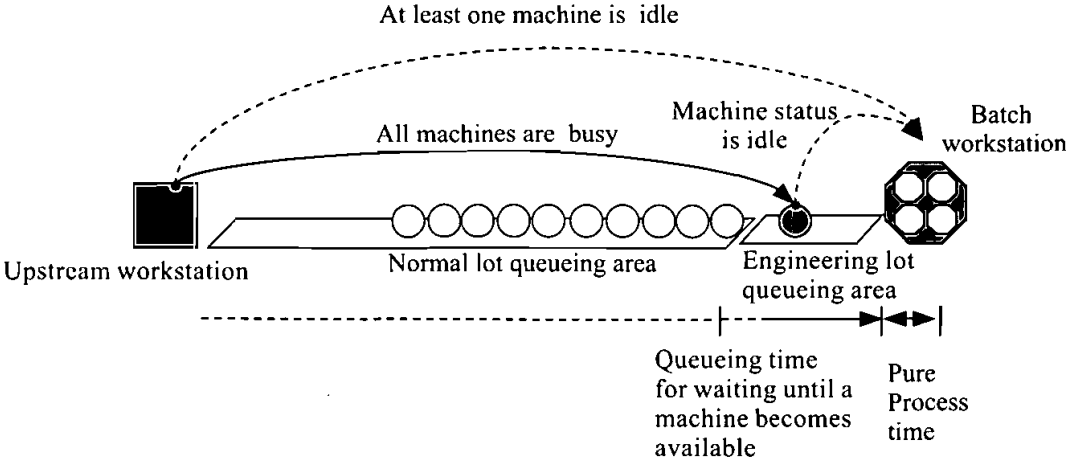

Fig. 2. The formation of the queueing time for engineering lots.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

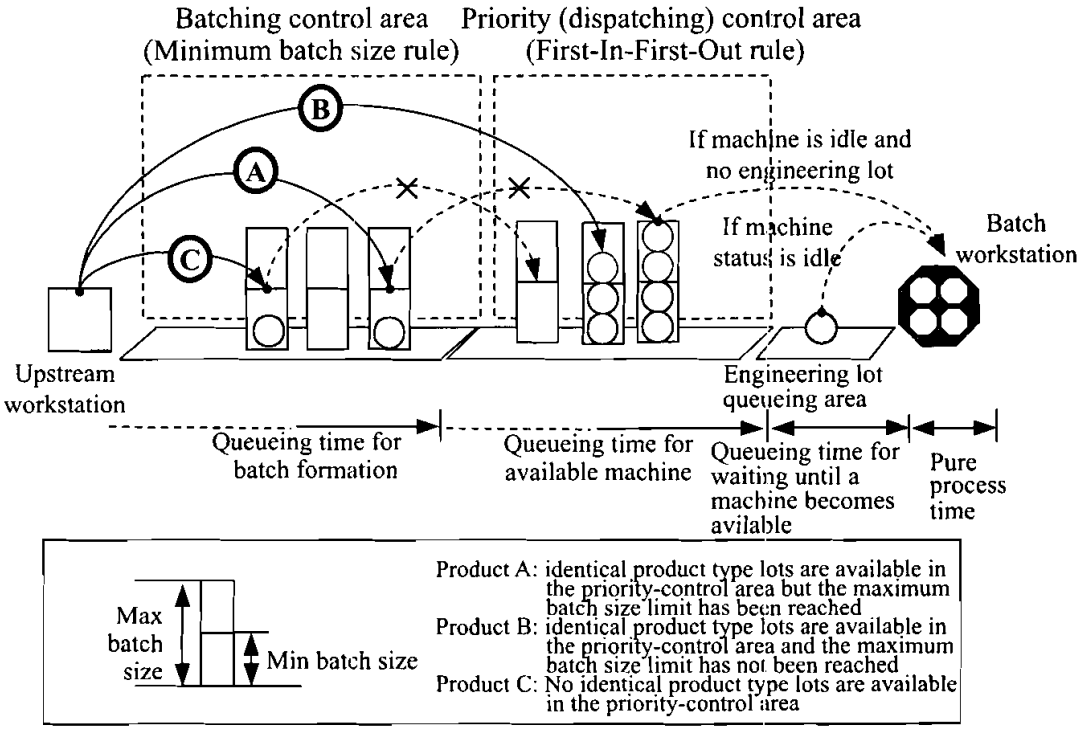

Fig. 3. The formation of the queueing time for normal lots.

2. Queue time resulting from the batching-factors: T'_B. The queue time is increased due to the difference in batch sizes and throughput rates between the upstream and downstream workstations.
3. Theoretical process time: T_P. Theoretical process time includes the pure processing time, and the loading and unloading times.

The non-preemptive priority queue model will be used to estimate T_Q and the Batching-Factor Flow Time estimation algorithm (BFFT) developed by Chung and Huang (1999) will be used to estimate T'_B and T_P. For convenience, we let $T_B = T_P + T'_B$. Then, the total cycle time for a batch of lots to flow through the entire factory, T_T, will be equal to the sum of related T_Q and T_B values of the whole process. The estimation procedure of T_Q and T_B will be described in the next two sections and are depicted in Fig. 4.

3.2. *The estimation method for the loading-factor queue time*

In wafer fabs, an engineering lot that has a higher operation priority than usual is treated as a hot lot, but it does not pre-empt a lot than is currently being processed. This means that no lot that is currently being processed is forced to return to the queue line even if an engineering lot enters the workstation. This kind of queue discipline matches with what the non-preemptive priority queue model (Hiller and Lieberman, 1990; Winston, 1991) defines. Hence, the non-preemptive priority queue model will be adopted to estimate the queue time for each lot in the system.

Before describing the model, all required notations are defined as below.

c_k = the number of available machines for the kth workstation ($c_k = c_k^e + c_k^n$);
c_k^e = the average number of machines being used by engineering lots at the kth workstation;
c_k^n = the average number of machines available for use by normal lots at the kth workstation;
e_k = the operation efficiency of the kth workstation;
$\Re$ = the planned total wafer quantity that will be released to the shop-floor ($\Re = \Re^e + \Re^n$);
$\Re^e$ = the planned total engineering lot quantity;
$\Re^n$ = the planned total normal lot quantity;
T = the length of the planning period;
w_k^r = the system time in a steady-state for lots with the rth priority at the kth workstation (including the process time);
w_k^{rq} = the queue time in the steady-state for lots with the rth priority at the kth workstation;
λ_k = the average lot arrival rate at kth workstation;
λ_k^e = the average engineering lot arrival rate at kth workstation;
λ_k^n = the average normal lot arrival rate at kth workstation;
μ_k = the average service rate at kth workstation;

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

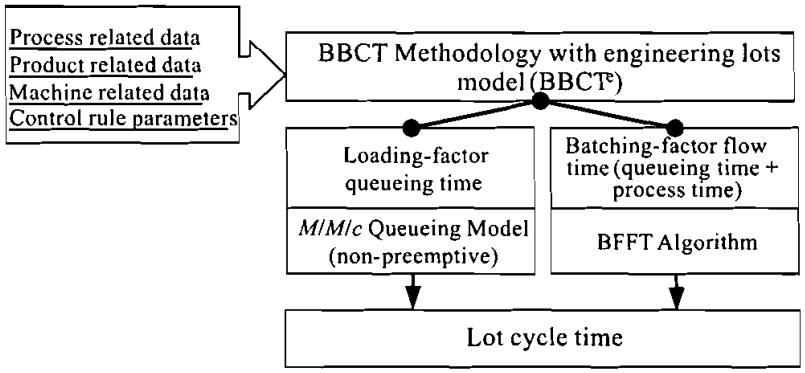

Fig. 4. The BBCT methodology when engineering lots are present in the system.

π_i^n = the product mix ratio for the ith product type among all normal lots;
π_i^e = the product mix ratio for the ith product type among all engineering lots;
f_{ik} = the frequency for the process of product type i flowing through the kth workstation;
γ_k = the average rework probability at the kth workstation;
O_k = the maximum output rate at the kth workstation;
O_k^e = the maximum engineering lot output rate at the kth workstation;
O_k^n = the maximum normal lot output rate at the kth workstation;
$\overline{P}_k^e$ = the average processing time for engineering lots at the kth workstation;
$\overline{P}_k^n$ = the average processing time for normal lots at the kth workstation.

The procedure for estimating loading-factor queue time is stated briefly in the following sections.

3.2.1. *Estimating the average lot arrival rate*

λ_k is equal to the sum of the engineering lot arrival rate and the normal lot arrival rate at the kth workstation. The formulas are shown as below:

$$\Re = \Re^e + \Re^n, \tag{4}$$

$$\lambda_k^n = \Re^n \times \sum_i \pi_i^n f_{ik}(1+\gamma_k), \quad i \in G(n), \tag{5}$$

$$\lambda_k^e = \Re^e \times \sum_{i'} f_{i'k}(1+\gamma_k), \quad i' \in G'(e), \tag{6}$$

$$\lambda_k = \lambda_k^e + \lambda_k^n, \tag{7}$$

where $G(n)$ represents the set of all product types that are normal lots and $G'(e)$ represents the set of all product types that are engineering lots.

3.2.2. *Estimating available machine units for each workstation*

The number of available machines at the kth workstation, c_k, is equal to the total number of machines, n_k, minus the equivalent number of machine breakdowns and maintenance. For a machine m that belongs to workstation k, we assume that the mean time to Preventive Maintenance (PM) of machine m is $MTTPM_{k_m}$, the mean time between PM is $MTBPM_{k_m}$, the mean time between failure is $MTBF_{k_m}$, the mean time to repair is $MTTR_{k_m}$.

$$c_k = \sum_{m=1}^{n_k}\left(1 - \frac{MTTR_{k_m}}{MTBF_{k_m} + MTTR_{k_m}} - \frac{MTTPM_{k_m}}{MTBPM_{k_m} + MTTPM_{k_m}}\right), \quad \text{for each } k. \tag{8}$$

3.2.3. *Estimating the machine units being used by engineering lots at workstation k, c_k^e*

$$c_k^e = \text{Min}\left(\frac{\sum_{i\in G'(e)}\sum_{\{l:M(i,l)=k\}}\left[\Re^e \pi_i^e\right]p_{il}}{T}, c_k\right), \quad \text{for each } k, \tag{9}$$

where p_{il} denotes the processing time for the ith product type at the lth process step. $M(i,l)$ is defined as the workstation type used for product type i at lth process step.

3.2.4. *Estimating the average number of machines available for use by normal lots at the kth workstation, c_k^n*

Since engineering lots have a higher priority for utilizing machine capacity, normal lots are only allowed to consume the remaining capacity.

$$c_k^n = \text{Max}(c_k - c_k^e, 0). \tag{10}$$

3.2.5. *Estimating the average processing time for engineering lots and normal lots at each workstation*

$$\overline{P}_k^e = \frac{\sum_{i\in G'(e)}\sum_{\{l:M(i,l)=k\}}\left[\Re^e \pi_i^e\right]p_{il}}{\sum_{i\in G'(e)}\left[\Re^e \pi_i^e\right]f_{ik}}, \quad \text{for each } k, \tag{11}$$

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

$$\overline{P_k^n} = \frac{\sum_{i\in G(n)} \sum_{\{l:M(i,l)=k\}} [\Re^n \pi_i^n] p_{il}}{\sum_{i\in G(n)} [\Re^n \pi_i^n] f_{ik}}, \quad \text{for each } k. \quad (12)$$

3.2.6. *Estimating the maximum output rate for an engineering lot and for a normal lot, O_k^e and O_k^n, at each workstation*

$$O_k^e = \frac{c_k}{\overline{P_k^e}}, \quad \text{for each } k, \quad (13)$$

$$O_k^n = \frac{c_k^n \times B_{ij}^{\max}}{\overline{P_k^n}}, \quad \text{for each } k, \quad (14)$$

$$O_k = \frac{\Re^e}{\Re} O_k^e + \frac{\Re^n}{\Re} O_k^n, \quad \text{for each } k. \quad (15)$$

3.2.7. *Estimating the average service rate for each workstation*

The average service rate at workstation k, μ_k, is equal to the product of the weighted average output rate for processing engineering lots and normal lots at the kth workstation. The μ_k can be written as:

$$\mu_k = T \times e_k \times O_k. \quad (16)$$

3.2.8. *Estimating the queue time in a steady-state for engineering lots and normal lots at each workstation*

The queueing model is an appropriate method to estimate the queue time incurred by an average workload at a workstation. Because of numerous re-entries to critical machines and different time lengths for arrival and service in each re-entry loop, it is assumed that each workstation is an independent $M/M/c$ queuing system. That is, the service time is exponentially distributed with mean $1/\mu_k$, and the rth priority lots have interarrival times that are exponentially distributed with rate λ_k^r at the kth workstation. According to the non-preemptive priority queue model (Hiller and Lieberman, 1990; Winston, 1991), the queue time for the engineering lots and normal lots can be formulated as shown in Table 1.

The above discussions show the procedure to estimate the loading-factor queue time for normal and engineering lots. The batching-factor queue time for a normal lot will be estimated by modifying the BFFT algorithm so as to fit with the existence of engineering lots. The core concept of the BFFT algorithm and the block definition will be discussed in detail in the next section.

3.3. *The definition and basic concept of a block*

A *block* is defined as those batch and serial-type process steps delimited by the two nearest batch-type workstations according to the process sequence (Chung and Huang, 1999). Since a batch-type workstation is an important source of interference in the material flow, separating the whole process into numerous blocks can help us to analyze and identify material flow characteristics.

Due to the wafer fabrication complexity, there are three additional special block types besides the general block types defined above, as shown in Fig. 5. The basic definitions and the range covered are described below.

General Block Type: BSB

Both start and end process steps of the general block type individually being processed on batch-type workstations

Table 1. Non-preemptive priority queueing model

Parameters	*Engineering lot* (r = e)	*Normal lot* (r = n)
w_k^e and w_k^n	$w_k^e = \frac{\mu_k}{A\times(\mu_k-\lambda_k^e)} + \frac{c_k}{\mu_k}$	$w_k^n = \frac{\mu_k^2}{A\times(\mu_k-\lambda_k^e)\times(\mu_k-\lambda_k^e-\lambda_k^n)} + \frac{c_k}{\mu_k}$
w_k^{eq} and w_k^{nq}	$w_k^{eq} = \frac{\mu_k}{A\times(\mu_k-\lambda_k^e)}$	$w_k^{nq} = \frac{\mu_k^2}{A\times(\mu_k-\lambda_k^e)\times(\mu_k-\lambda_k^e-\lambda_k^n)}$
A_k	$A_k = c_k!(\mu_k-\lambda_k)\left(\frac{\mu_k}{c_k\lambda_k}\right)^{c_k}\sum_{j=1}^{c_k-1}\frac{1}{j!}\left(\frac{c_k\lambda_k}{\mu_k}\right)^j+\mu_k$, where $\lambda_k = \lambda_k^e + \lambda_k^n$	

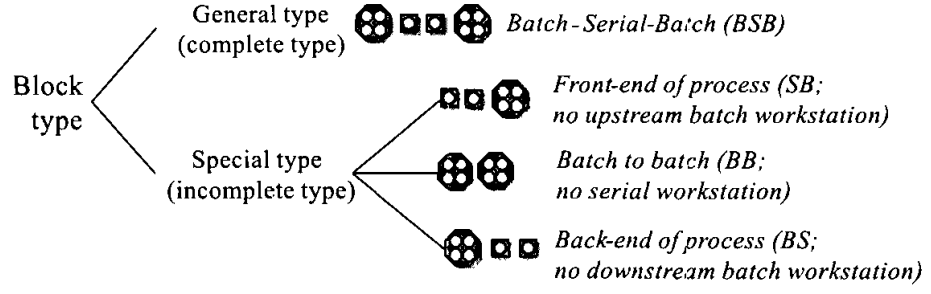

Fig. 5. The definition of a block.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

(B) and the process steps between them correspond to serial-type workstations (S).

Special Block Type: BB, SB and BS
Each of the following block types is an incomplete general-block type.

- The BB type: batch workstation to batch workstation. Both start and end process steps belong to batch-type process (B). There is no serial-type process in between.
- The SB type: serial workstation to batch workstation. The start process step of the block is a serial-type process (S), and the end process step is a batch-type process (B). The SB type only occurs at the beginning of the whole process and it is an exception of the general block type case.
- The BS type: batch workstation to serial workstation. The start process step of the block is a batch-type process (B), and the end process step is serial-type process (S). The BS type only occurs at the end of the whole process plan and is also an exception of the general block type case.

Figure 6 illustrates a wafer process in terms of blocks. It shows that the first block of the ith process is a SB type. It only appears once in the process and covers process steps 1 to 3. The 11th block is classified as a BSB type, a general block type and covers process steps 100 to 103. The 32nd block is classified as a BB type and covers process step 212 to 213. The last block of the process is the 51st block, which is classified as a BS type; it appears only once in the process and covers process steps 298 to 300. Figure 6 shows that each complicated wafer fabrication process is made up of numerous blocks; each block can have different blocks and is categorized into one of BSB, BB, SB or BS types. To effectively estimate the cycle time for a lot flowing through the process, the material flow characteristics for each block type must be clearly confirmed.

To confirm the material flow characteristics in a block, we must analyze the major factors that determine the lot flow rate. Regardless how many process steps are included in each block, the lot flow rate in a block will be constrained by the workstation with the lowest output rate. We define the workstation with the lowest output rate as the "critical workstation". Obviously, the critical workstation is the most important point at which to observe the material flow characteristics in the block and its output rate is the key value. Notice that the critical workstation may be a batch-type or a serial-type. After a lot completes the critical operation step, it will go through downstream process steps in the block at an output rate that is relatively faster than that in the critical workstation.

In addition to the critical workstation, the batch characteristics of the two batch workstations also affects the material flow. For each block, the material flow may be interrupted before the second batch workstation because the minimum batch size condition has not been met. This characteristic of wafer lots waiting for batch formation has made the second batch workstation play the role of a material flow interrupter.

Furthermore, when multiple lots are released from the first batch machine of a block, a temporarily peak load may occur in the downstream workstation. This severely disturbs the smoothness of the material flow. Thus, the first batch workstation also plays the rather important role of material flow disturbance.

Combining the above discussions, the material flow for the block will be affected by: (i) the output rate and the batch size setting of the first and second batch workstations; and (ii) the output rate of the workstation with the lowest output rate among all serial workstations in the block. These three important workstations are defined here as material flow observation points.

The product mix ratio and the throughput target of a planning period will affect the output rate of each workstation. The three observation points must be confirmed whenever the throughput target or the product mix ratio is changed. Once the interactions among them are confirmed, we can effectively recognize the material flow in the entire block and then catch the cycle time characteristics.

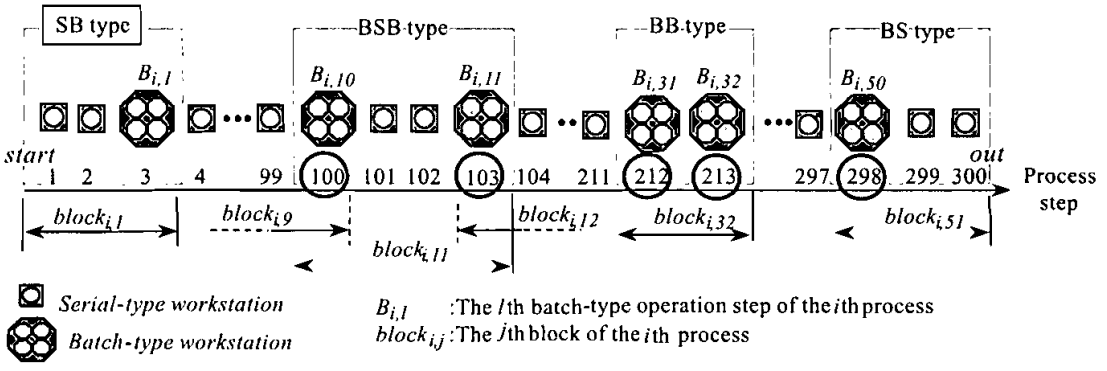

Fig. 6. Process flow in terms of blocks.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

BBCT methodology (Chung and Huang, 1999) applies the block concept and uses the output characteristics of the material flow observation points to estimate the batching-factor queue time (T'_B) and the theoretical process time (T_P). The BBCT methodology assumes that the lots in the wafer fab are all normal lots and the batch-factor queue time for each lot is estimated under the assumption that all the available capacity of each workstation can be used for producing normal lots. However, the available capacity for normal lots will decrease when an engineering lot is introduced into the fab. Therefore, it is necessary to revise the available capacity formulae for engineering lots and normal lots. The next section states how to revise the available capacity for normal lots.

3.4. *Calculation of the batching-factor queue time*

An engineering lot does not have a batching-factor queue time; it only has a loading-factor queue time, which is estimated by using the non-preemptive priority queue model as stated before.

When estimating the Batching-Factor Flow Time (BFFT) for normal lots, the c_k values in each formula for block time calculation, developed by Chung and Huang (1999), is replaced by c_k^n in order to transform the cycle time estimation algorithms that do not consider engineering lots into algorithms that do consider engineering lots. The BBCT model with a consideration of engineering lots is thus symbolized as BBCT^e.

4. Experimental design and results

In order to evaluate the performance accuracy of BBCT^e on cycle time estimation, a wafer fab simulation model is constructed. The cycle time estimation by BBCT^e is compared with a simulation result in order to understand the performance accuracy. The simulation model applies the actual production data obtained from a real wafer fab. These input data include three major factors, products, equipment and processes, as described below:

1. Product related data: There are five products, Products A, B, C, D and E in the model with a product mix of 5 : 7 : 3 : 4 : 1 in sequence, (i.e. 25, 35, 15, 20 and 5% respectively) and a weekly wafer release quantity of 4000 wafers. Product E is assumed to be an engineering lot. The release policy is applied with the fixed-WIP method. Each release quantity for Products A to D is 150 wafer pieces (or six lots) and each release quantity for Product E is 25 wafer pieces (or one lot).
2. Process related data: Products A to E corresponds to processes A to E, where the first two products are logic products and the others are memory products. The number of process steps is between 276 and 345.
3. Equipment related data: In the fab, 236 machines are separated into 83 different types of workstations, of which 37 are batch workstations. The maximum batch size for a batch workstation can be classified into six, four and two lots, each with 15, three and 19 machine types. Every workstation applies the FIFO dispatch rule, and every batch workstation applies the full-load batch rule, except for engineering lots.

The total simulation time is 240 days, of which the first 60 days is the system warm-up stage. To explain that both the BBCT^e and BBCT have a satisfactory accurate performance for cycle time estimation in a system environment with existing and non-existing engineering lots, respectively, a simulation will be carried out to show the results under these two different conditions. In addition to the experimental design, the Appendix will show a simple example to describe how to calculate the lot cycle time.

4.1. *Experiment 1: no engineering lots*

Figure 7 shows a comparison of the average lot cycle time for each product type estimated by using simulation, BBCT methodology (including T_L and T_B), $M/M/c$ queueing model (Hiller and Lieberman, 1990; Winston,

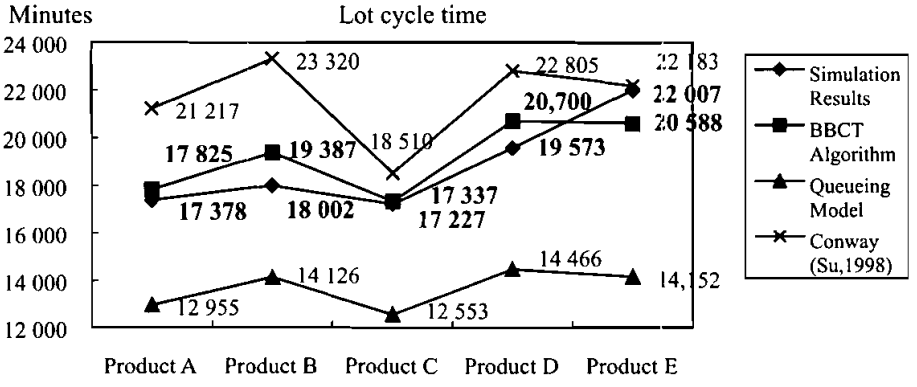

Fig. 7. The lot cycle time for each product type (after Chung and Huang (1999)).

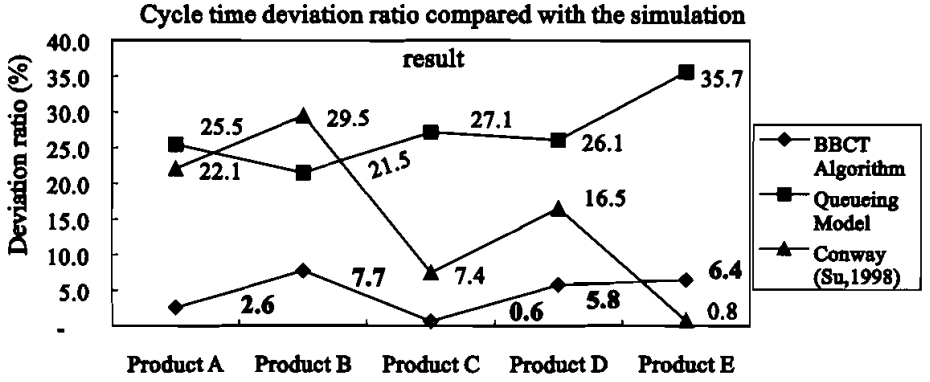

Fig. 8. A comparison of the estimation accuracy among cycle time estimation methods (after Chung and Huang (1999)).

1991) and Su's estimation formula (Su, 1998). The ratios for deviation from simulation results are shown in Fig. 8. The absolute values for the minimum deviation ratio, the maximum deviation ratio and the average deviation ratio can be read as (0.06, 7.7, 4.9%), (21.5, 35.7, 25.0%) and (0.8, 29.5, 20.3%) by using the BBCT algorithm, general $M/M/c$ queue model and Su's estimation formula respectively. The above average deviation ratio is weighted based on the product mix ratio. Apparently, the cycle time estimation performance resulting from the BBCT algorithm including the minimum deviation ratio, maximum deviation ratio and the average deviation ratio is a significant improvement over the other methods.

4.2. *Experiment 2: where product E is an engineering lot*

Here, Product E is in accordance with the dispatching principles for engineering lots. Figure 9 shows the changes in each product cycle time before and after engineering lots are placed into the system. As far as Product E is concerned, the average lot cycle time dropped from the original 22 007 minutes to 14 193 minutes, or by approximately 57.4% in cycle time length, after the lot attribute was changed to an engineering lot. On the other hand, the average lot cycle time for Products A to D increased by 4.8% to 12.5%. As the above results show, the engineering lot's cycle time will sharply decrease and the normal lot's cycle time will increase after engineering lots are placed into the system. Figure 10 shows a comparison of the cycle time estimation using BBCT^e and simulation results. Since the estimated cycle time for Product E is 13 587 minutes and its simulation result is 14 193 minutes, there is a 4.3% deviation ratio. Also, the deviation ratios between the estimated cycle time and simulation results for normal Products A to D is between 0.9 and 6.5%. Hence, BBCT^e has a very good estimation performance when engineering lots exist in the system.

It is known that there is a direct connection between the length of the loading-factor queue time and the workstation utilization. In order to ensure that the loading-factor queue time estimation is reasonable, this experiment has further demonstrated the estimation accuracy of BBCT^e in estimating each workstation's utilization rate. The results show that there are approximately 66.2, 89.2 and 98.9% of the total number of workstations whose difference in machine utilization rate between the BBCT^e estimated value and the simulation result is less than 0.01, 0.05 and 0.1 respectively. There is only one workstation, or approximately 1.2% of the total number of workstations, that has more than 0.1 as the difference between its BBCT estimate and its simulation result. The mean and standard deviation for the absolute differences in estimating machine utilization is 2 and 2.97% respec-

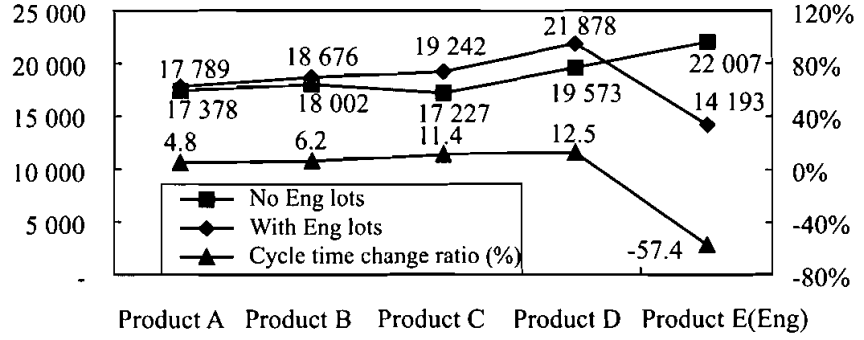

Fig. 9. The average cycle time changes before and after an engineering lot is introduced.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

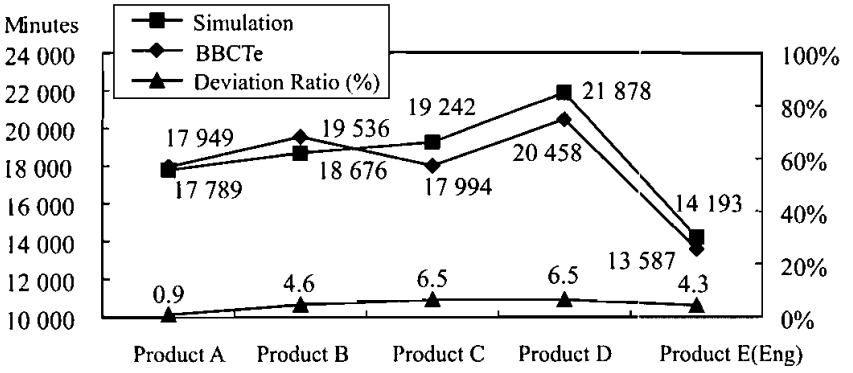

Fig. 10. The cycle time deviation between the BBCT^e and simulation results.

tively. Therefore, BBCT^e also has an outstanding performance in estimating the workstation utilization, and consequently the accuracy in estimating the loading-factor queue time is effectively confirmed. The comparison between the estimated utilization by BBCT^e and the respective simulation result is listed in Table 2 and depicted in Fig. 11.

5. Conclusions and future research

When a wafer lot is processed in a wafer fab, not only the machine load factors but also the machine batch size attributes may affect queue time formation. In this paper, the queue time formation for a wafer lot is divided into two categories: *the loading-factor*, resulting from the machine load, and *the batching-factor*, resulting from the batch size setting for the batch machine. The queue time accrued from the loading-factor is estimated by applying the non-preemptive priority $M/M/c$ queue model (Hiller and Lieberman, 1990; Winston, 1991), while the queue time accrued from the batching-factor is estimated by applying the BFFT algorithm when engineering lots exist in the system. The cycle time estimation methodology developed in this paper is symbolized as BBCT^e. The distinctive feature of the of the BBCT^e model is that all the calculations are in arithmetic form, and hence the computation time is really short.

Table 2. A comparison of the machine utilization rates

Work-station	*Simulation result*	*BBCT^e*	*Work-station*	*Simulation result*	*BBCT^e*	*Work-station*	*Simulation result*	*BBCT^e*	*Work-station*	*Simulation result*	*BBCT^e*
W1	0.30	0.30	W22	0.30	0.30	W43	0.05	0.05	W64	0.74	0.77
W2	0.45	0.46	W23	0.06	0.06	W44	0.01	0.01	W65	0.14	0.14
W3	0.02	0.02	W24	0.92	1.00	W45	0.02	0.02	W66	0.03	0.03
W4	0.07	0.07	W25	0.64	0.64	W46	0.91	0.93	W67	0.70	0.72
W5	0.37	0.38	W26	0.59	0.56	W47	0.75	0.68	W68	0.25	0.24
W6	0.02	0.01	W27	0.17	0.12	W48	0.41	0.36	W69	0.35	0.35
W7	0.56	0.57	W28	0.46	0.51	W49	0.22	0.22	W70	0.15	0.14
W8	0.41	0.42	W29	0.44	0.40	W50	0.18	0.17	W71	0.41	0.40
W9	0.50	0.51	W30	0.55	0.56	W51	0.23	0.23	W72	0.17	0.16
W10	0.32	0.32	W31	0.52	0.49	W52	0.27	0.28	W73	0.37	0.37
W11	0.46	0.48	W32	0.44	0.43	W53	0.49	0.51	W74	0.30	0.30
W12	0.58	0.65	W33	0.28	0.20	W54	0.57	0.64	W75	0.10	0.09
W13	0.64	0.73	W34	0.41	0.37	W55	0.52	0.58	W76	0.34	0.33
W14	0.37	0.40	W35	0.42	0.39	W56	0.12	0.12	W77	0.12	0.12
W15	0.38	0.36	W36	0.56	0.56	W57	0.24	0.24	W78	0.10	0.11
W16	0.29	0.29	W37	0.46	0.45	W58	0.42	0.37	W79	0.03	0.03
W17	0.15	0.14	W38	0.45	0.44	W59	0.42	0.45	W80	0.36	0.35
W18	0.57	0.64	W39	0.09	0.08	W60	0.33	0.34	W81	0.17	0.17
W19	0.24	0.24	W40	0.34	0.33	W61	0.52	0.71	W82	0.28	0.29
W20	0.36	0.36	W41	0.63	0.67	W62	0.66	0.68	W83	0.05	0.05
W21	0.63	0.63	W42	0.02	0.02	W63	0.71	0.70			

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

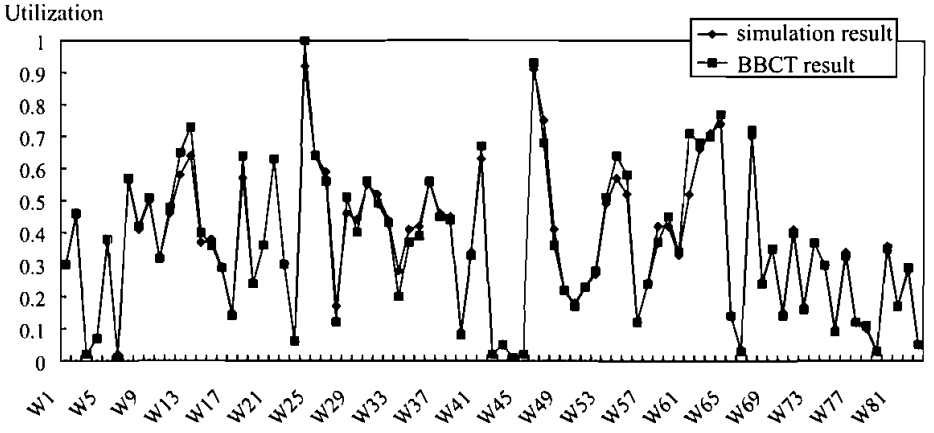

Fig. 11. A comparison between the utilization estimate for each workstation by the BBCT^e algorithm and the simulation results.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

A simulation model was built based on the production data from a real wafer fab in Taiwan. In the "no engineering lots" scenario, the lot cycle time estimated by BBCT was very close to the simulation results. The average deviation ratio was merely 4.9%, which shows that the BBCT methodology clearly performs better than the other methods in cycle time estimation. Under the "engineering lots present" scenario, the BBCT^e model showed a satisfactory performance estimation with the average deviation ratio to approximately 4.6%. The related experiment also shows that BBCT^e can effectively estimate the workstation utilization.

In this paper, the BBCT^e is used to estimate the cycle time of two-priority class lots, that is, engineering lots and normal lots. In our future research, the multiple-priority class will be considered into the model.

Acknowledgements

This paper was supported in part by the National Science Council, Taiwan, ROC, under Contract No NSC88-2213-E009-027.

References

Atherton, L.F. and Atherton, R.W. (1995) *Wafer Fabrication: Factory Performance and Analysis*, Kluwer, Massachusetts.

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.

Conway, R., Maxwell, W. and Miller, L.W. (1967) *Theory of Scheduling*, Addison-Wesley, Massachusetts.

Cunningham, J.A. (1990) The use and evaluation of yield model in integrated circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, **3**(2), 60–72.

Enns, S.T. (1995) A dynamic forecasting model for job shop flowing prediction and tardiness control. *International Journal of Production Research*, **33**(5), 1295–1312.

Etheshami, B., Petrakian, R.G. and Shabe, P.M. (1992) Trade-offs in cycle time management: hot lots. *IEEE Transactions on Semiconductor Manufacturing*, **5**(2), 101–105.

Fowler, J.W., Philips, D.T. and Hogg, G.L. (1992) Real time control of multiproduct bulk-service semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, **5**(2), 158–163.

Fronckowiak, D., Peilert, A. and Nishinohara, K. (1996) Using discrete event simulation to analyze the impact of job priorities on cycle time in semiconductor manufacturing, in *Proceedings of the 1996 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, New York, pp. 151–155.

Glassey, C.R. and Resende, M.G.C. (1988) Closed-loop job release control for VLSI circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, **1**(1), 36–46.

Glassey, C.R. and Weng, W.W. (1991) Dynamic batching heuristics for simultaneous processing. *IEEE Transactions on Semiconductor Manufacturing*, **4**(2), 77–82.

Hiller, F.S. and Lieberman, G.J. (1990) *Introduction to Operations Research*, 5th edn, McGraw-Hill, New York.

Kaplan, A.C. and Unal, A.T. (1993) A probabilistic cost-based due date assignment model for job shops. *International Journal of Production Research*, **31**(12), 2817–2834.

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

Martin, D.P. (1998) How the law of unanticipated consequences can nullify the theory of constraint: the case for balanced capacity in a semiconductor manufacturing line, *Semiconductor Fabtech*, 7th edn, ICG Publishing Ltd, pp. 29–34.

Matsuyama, A. and Atherton, R.W. (1990) Experience in simulation wafer fabs in the USA and Japan, in *Proceedings of the 1990 International Semiconductor Manufacturing Science Symposium*, Burlingame, USA, pp. 113–118.

Miller, D.J. (1989) Implementing the results of a simulation in a semiconductor line, in *Proceedings of the 1989 Winter Simulation Conference*, New York, pp. 922–929.

Narahari, Y. and Khan, L.M. (1997) Modeling the effect of hot lots in semiconductor manufacturing systems. *IEEE Transactions on Semiconductor Manufacturing*, **10**(1), 185–188.

Raddon, A. and Grigsby, B. (1997) Throughput time forecasting model, in *Proceedings of the 1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 430–433.

Su, Y.C. (1988) The construction of production planning and scheduling system for an IC foundry in ramp-up., Masters thesis, In-

dustrial Engineering and Management Department, National Chiao Tung University, Hsin-Chu, Taiwan.

Vig, M.M. and Dooley, K.J. (1991) Dynamic rules for due-date assignment. *International Journal of Production Research*, **29**(7), 1361–1377.

Wein, L.M. (1988) Scheduling semiconductor wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing*, **1**(3), 115–130.

Winston, W.L. (1991) *Operations Research: Applications and Algorithms*, PWS-Kent, Boston, MA.

Wood, S.C. (1997) Cost and cycle time performance of fabs based on intergrated single-wafer processing. *IEEE Transactions on Semiconductor Manufacturing*, **10**(1), 98–111.

Appendix

A simple model

A simple model is built in order to explain that of BBCT^e, procedure. Assume that a system has only two product types, normal lots and engineering lots, corresponding to process N and E. The product mix of N and E is 30 : 1. The process related information including process steps, workstation, and process time is shown in Table A1.

Table A2 shows the workstation related information. There are eight workstations in this system. Each workstation has a only one machine that has a 100 time unit capacity (1 time unit = 14.4 minutes) per day. Machines never break down or need maintenance. Based on the above information, c_k^e and c_k^n can be estimated by using Equations (4)–(10).

According to the definition of the block, the normal lot process is segmented into four blocks that are classified as SB, BSB, BB and BS in sequence. Based on the block style, the batching-factor flow time of each block can be effectively estimated by using the BFFT algorithm and replacing the c_k defined in Chung and Huang (1999) by c_k^n. Moreover, the batching-factor flow time of the entire process can also be estimated by using the formula for the multiple-block cycle time developed in Chung and Huang (1999).

The result is shown in Table A3. Note that in Table A3 the flow time of the normal lot was calculated as:

$$\text{Flow time of normal lot} = 27 + 42 + 52 + 46 - 14 \times \left\lceil \frac{6}{6 \times 1} \right\rceil - 5 \times \left\lceil \frac{6}{2 \times 0.97} \right\rceil - 16 \times \left\lceil \frac{6}{6 \times 0.88} \right\rceil = 101.$$

Table A1. Process-related information

Normal lot process-N				*Engineering lot process-E*			
Step	*Workstation*	*Maximum batch size*	*Process time*	*Step*	*Workstation*	*Maximum batch size*	*Process time*
1	W1	1	1	1	W1	1	5
2	W2	1	2	2	W4	1	7
3	W3	6	14	3	W6	2	3
4	W4	1	2	4	W5	1	2
5	W5	1	3	5	W7	6	12
6	W6	2	5	6	W8	1	5
7	W7	6	16				
8	W8	1	2				

Table A2. Workstation-related information

Workstation	*Workstation type*	*Maximum batch size*	*Available EQ num (c_k)*	*Occupied by engineering lots (c_k^e)*	*Reserved for normal lots (c_k^n)*
W1	S	1	1	0.05	0.95
W2	S	1	1	0.00	1.00
W3	B	6	1	0.00	1.00
W4	S	1	1	0.07	0.93
W5	S	1	1	0.02	0.98
W6	B	2	1	0.03	0.97
W7	B	6	1	0.12	0.88
W8	S	1	1	0.05	0.95

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014

Table A3. Normal lot flow time estimated by the BFFT algorithm

Block	*Type*	*Process step*	*Critical workstation*	*Block flow time*
1	SB	1, 2, 3	W3	27
2	BSB	3, 4, 5, 6	W5	42
3	BB	6, 7	W7	52
4	BS	7, 8	W7	46
Batch factor flow time of the whole process				101

Table A4. Engineering lot flow time estimated by the BFFT algorithm

Block	*Type*	*Process step*	*Critical workstation*	*Block flow time*
1	SB	1, 2, 3	W6	15
2	BSB	3, 4, 5	W7	17
3	BS	5, 6	W7	17
Batch factor flow time of the whole process				34

Table A5. The cycle time deviation ratio between the BBCT^e and the simulation results

Lot type	*Batch factor flow time (A)*	*Loading factor queueing time (B)*	*Lot cycle time estimated by BBCT^e (C)=(A)+(B)*	*Simulation cycle time (D)*	*Deviation ratio (%) ((C)−(D))/(D)*
Normal lot	101.0	105.1	206.1	192.9	6.8
Engineering lot	34.0	14.5	48.5	48.0	1.0

The calculation is based on Equation (28) in Chung and Huang (1999). Similarly, the batching-factor flow time for the engineering lot can also be estimated according to the procedure mentioned above. The result is shown in Table A4. Finally, the loading-factor queue time is estimated by using the non-preemptive priority queue model mentioned in Section 3.2.

Note that the flow time of the engineering lot was calculated as:

$$\text{Flow time of engineering lot} = 15 + 17 + 17 - 3 \times \left\lceil \frac{1}{2 \times 1} \right\rceil - 12 \times \left\lceil \frac{1}{6 \times 1} \right\rceil = 34.$$

The calculation is based on Equation (28) in Chung and Huang (1990).

To evaluate the accuracy performance of the cycle time estimation, a simulation model is constructed. Table A5 shows a comparison between the lot cycle time estimated by the BBCT^e methodology and its respective simulation result. The average lot cycle time deviation ratios of the normal and engineering lots are only 6.8 and 1.0%, respectively. This demonstrates that the BBCT^e algorithm can produce a satisfactory cycle time estimation.

Biographies

Dr. S.H. Chung is a Professor in the Department of Industrial Engineering and Management, National Chiao-Tung University, Taiwan, ROC. She received a Ph.D. degree in Industrial Engineering from Texas A&M University, College Station, TX, USA. Her research interests include production planning, scheduling, system simulation, and production planning of IC manufacturing. She has published and presented research papers in the areas of production planning, scheduling, cost analysis and IC manufacturing management.

Hung-Wen Huang is a Ph.D. candidate at the Department of Industrial Engineering and Management, National Chiao-Tung University, Taiwan, ROC. He received his M.S. degree in Industrial Engineering and Management from National Chiao-Tung University. He is also a Manager in the Department of Operation Analysis, Winbond Electronic Corp., Taiwan, ROC. His research interests include operation and performance analysis in the semiconductor industry.

Downloaded by [National Chiao Tung University] at 22:06 27 April 2014