Journal of Manufacturing Systems
Vol. 23/No. 1
2004

A Heuristic Input Control Method for a Single-Product, High-Volume Wafer Fabrication Process to Minimize the Number of Photomask Changes

Ching-Chin Chern and **Kwei-Long Huang,** Dept. of Information Management, National Taiwan University, Taipei, Taiwan

Abstract

This study proposed an input control policy, called the (k,W) rule, for a single-product, high-volume wafer fabrication process based on the special characteristics of steppers in the photolithography area, implying that the production environment possesses reentrant production flows. To add to the complexity, the setup time of each operation is time consuming. The (k,W) rule adopts the concepts of workload regulation and batch-sizing policies to release k lots of wafers for a stepper when the workload of this stepper is lower than a predefined threshold, W, for reducing setup time and raising stepper utilization. A search algorithm is proposed to determine k and W, to minimize the weighted sum of wafer waiting time, stepper idle time, and setup time. To compare the performace of the (k,W) rule with other input control policies, a simulation model is built with data collected from a DRAM wafer fabrication process in Taiwan. As observed from the simulation results, the (k,W) rule decreases setup time and increases throughput without increasing cycle time.

Keywords: *Input Control Policy, Workload Regulation, Lot Size, Photolithography, Wafer Fabrication*

1. Introduction

Due to the complexities of wafer fabrication processes and the intensities of capital and technical investment, it is important for semiconductor manufacturers to utilize their equipment efficiently. However, wafer fabrication processes are different from general production processes in that producing one wafer needs hundreds of operations and takes a few months to complete. Meanwhile, the production characteristics such as wafer reentrance, equipment downtime, random yields, and so on, make scheduling and dispatching in semiconductor wafer fabrication particularly difficult. These difficulties are elaborated in Uzsoy, Lee, and Martin-Vega (1992) as well as in Hughes and Schott (1986) and Bai and Gershwin (1990). Uzsoy, Lee, and Martin-Vega (1994) review the research related to shop floor control in the semiconductor industry. They classify the research by the techniques used in the solution and discuss the pros and cons of these approaches. A thorough description of wafer fabrication and specific modeling and analysis problems are given in Johri (1992). Leachman and Hodges (1996) conducted on-site interviews with manufacturing personnel in 16 plants to identify factors strongly correlated with productivity.

Hence, an effective operation control strategy is essential for a wafer fabrication facility. An operation control strategy is usually a combined choice of an input control policy and a dispatch rule. In the study of input control methods, Glassey and Resende (1988) propose a "starvation avoidance" input control method similar to the reorder-point concept in the traditional EOQ model. They compute the virtual inventory needed at the workcenter before the new job arrives as the reorder point. Lozinski and Glassey (1988) further develop a mechanism to determine bottleneck starvation indicators closely related to the starvation avoidance input control method. Lawton et al. (1990) propose an input control method called workload regulation that regulates wafer input by workload measurement of the bottleneck machines. That is, the wafers are input into the system when the workload is less than a threshold that is chosen to guarantee a desired level

This work was supported by the National Science Committee in Taiwan, project number NSC88-2416-H-002-056.

of throughput. Wein (1988) uses a Brownian network model to approximate a multiclass queueing network model with dynamic control capability, which emulates a wafer fabrication plant. He concludes that input control methods are more effective than dispatch rules in terms of reducing the cycle time and WIP.

In all of the above studies, simulation is used as a tool to either evaluate the algorithms or justify their arguments. Dayhoff and Atherton (1984) first utilize the discrete event simulation technique to model the process of a semiconductor wafer fabrication. Miller (1990) reviews the experience of working in IBM's General Technology Division. Spence and Welter (1987) also adopt a simulation model to evaluate the performance of working units in the photolithography area of a wafer fabrication. Instead of simulation, Chen et al. (1988) develop a mathematical model of a wafer fabricating process and derive the formula of average queue length and average waiting time by using the theory of queueing networks. However, the formula could be applied to simple cases and the error of approximation could be up to 14%. Lin and Raghavendra (1996) study the "join the shortest queue" (JSQ) policy through a queueing network model and they assume that all the queues are identical. They develop a method to approximate the performance of the system and show that the method provides accurate estimates of the mean response time.

Although many input control methods as well as dispatch algorithms have been developed and tested in the above studies, few of them are actually implemented in a wafer fabrication plant. One of the reasons is that they are too complicated to implement. This work has resulted in a simple and easy-to-implement input control policy, called the workload regulated rule by batch size or the (k,W) rule, which is constructed to effectively reduce the setup time on steppers for a wafer fabrication plant. Like the above studies, simulation will be used to evaluate this input control policy because the system is too complicated to be modeled analytically. The data needed in the simulation model are collected from a large wafer fabrication plant in Hsin-Chu, Taiwan.

2. Workload-Regulated Input Control Rule by Batch Size

A typical wafer production cycle starts with nonphotolithography processes such as diffusion, thin film, or etching, as seen in *Figure 1*. Once the wafer

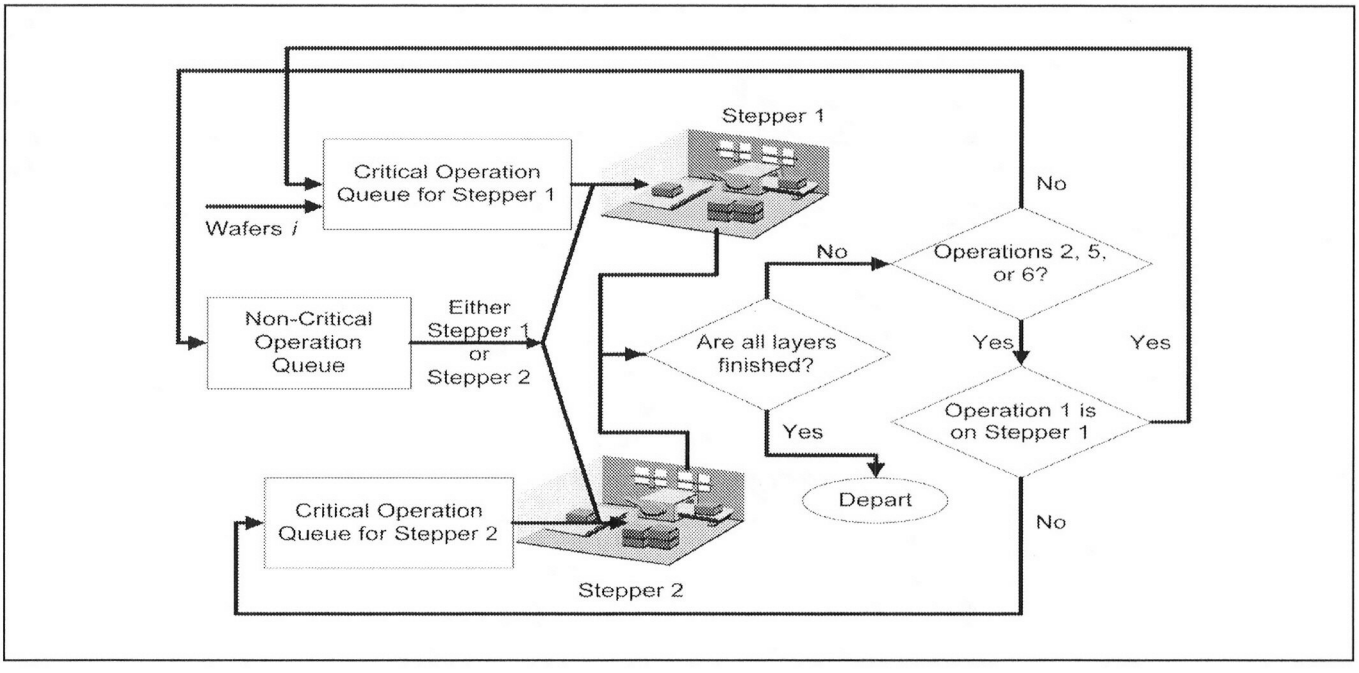

Figure 1
Critical and Noncritical Operations of Wafer *i*

enters the photolithography area, it is coated, photomasked, and developed for creating the first layer of circuitry. The wafer is then transferred to non-photolithography areas for other processes such as ion implant, diffusion, etc. It then returns to the photolithography area for the second-layer circuitry and so on. The entire cycle of wafer production finishes when all the required layers of circuitry have been properly produced. A wafer often needs more than 10 layers of circuitry, requiring it to go through hundreds of different repetitions, of which at least 10 are photomask operations processed on steppers. Two types of photomask operations are usually performed: critical and noncritical. For the same wafer, critical operations require very precise accuracy and, thus, must be performed on the same stepper; noncritical operations do not have this restriction. For instance, photomask operations 2, 5, and 6 are critical, and thus, they are required to be processed on the stepper as operation 1. Further assume that there are two steppers available. If photomask operation 1 of wafer i is processed by stepper 1, operations 2, 5, and 6 of this particular wafer must be performed on stepper 1 by requirement, but operations 3 and 4 can be processed on stepper 1 or stepper 2 depending on which one is available, as seen in *Figure 1*. Every photomask operation needs a distinguishing mask, but mask changes on steppers are troublesome and time consuming. Therefore, steppers are usually the bottleneck of the entire wafer fabrication process. Twenty-five (25) wafers are grouped and stored in a wafer boat as a "lot" in the wafer fabrication process. Therefore, the wafers are counted by numbers of lots when they are released and processed in the system. For example, 5 lots of wafers are referred to as 5 wafer boats or 125 wafers.

As mentioned in Chern, Liu, and Shieh (2003), a wafer production process can benefit from a family-based scheduling rule, which schedules the wafers needing the same operation as the current finished one ahead of other wafers, by saving photomask setup time. This study further utilizes the concept of family-based rules by grouping wafers at the releasing point. To minimize the number of mask changes, wafers are released to the steppers in a batch so that they can be processed together. The time intervals between successive photomask operations for the same wafer are usually different, which implies that wafers might wait for steppers for some operations, while steppers might be idle between some operations when no wafers are present in the queue. For example, assume that the processing times of the first and the second photomask operations are 36 and 17 minutes, respectively, and the mask changing time is 6 minutes. Furthermore, assume that the time interval between the first and the second operations is 30 minutes and only one stepper is available. New lots of wafers are released to the system only when the system is empty. If two lots of new wafers are released to the system, lot 1 finishes the first critical operation and goes to the non-photolithography areas. Lot 1 comes back to the same stepper in 30 minutes only to find that this stepper is still processing the first critical operation of lot 2 with 6 minutes to go, as seen in *Figure 2a*. Because the second critical operation of lot 1 requires it to be processed on the same stepper, it must wait in queue until lot 2 finishes its first critical operation. Lot 1 then waits 6 minutes for the stepper to change the mask and finishes its second operation 7 minutes before lot 2 returns, causing the stepper to be idle. If three lots are released, the stepper will not be idle, but wafers will wait much longer, as seen in *Figure 2b*. If only one lot is released, the stepper will be idle and the wafers need not wait, as seen in *Figure 2c*. In any of the cases, the resource time, either of stepper or of wafer, is wasted. To avoid this condition, an effective input control rule is necessary.

An input control rule tells the system when to release new wafers and how many lots of new wafers to be released. As mentioned in Lawton et al. (1990), the time to release new jobs is usually determined by the workload in the system, which is called the workload regulation rule. However, the time to release new wafers and the quantity of new wafers to be released are interrelated. This study, therefore, adopts a workload regulation rule using a batch-size releasing policy and proposes an input control method, called the (k,W) rule. In general, more than one stepper may be available. As seen in *Figure 3*, the (k,W) rule assumes that each stepper is an independent bottleneck and determines the time to release k lots of new wafers by comparing the current total workload, TW, with a pre-determined threshold, W. In other words, whenever the current total workload, TW, is less than a pre-determined threshold, W, k lots of new wafers are released to the system for this stepper. Therefore, the current total

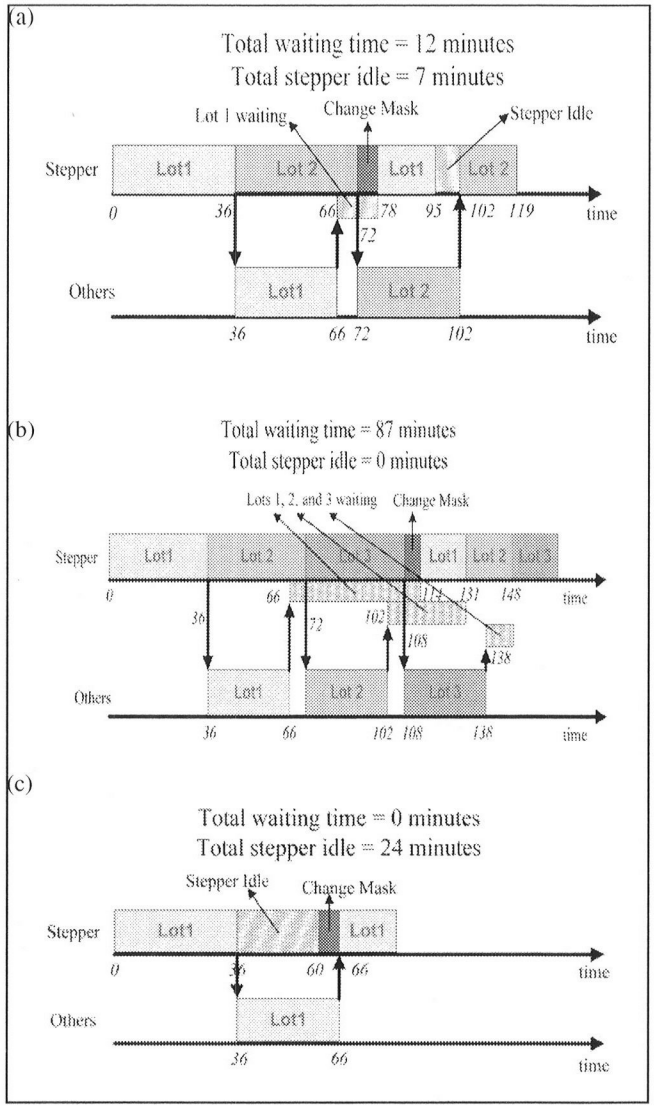

Figure 2
(a) Stepper Time Shared by Two (2) Lots
(b) Stepper Time Shared by Three (3) Lots
(c) Stepper Time Shared by One (1) Lot

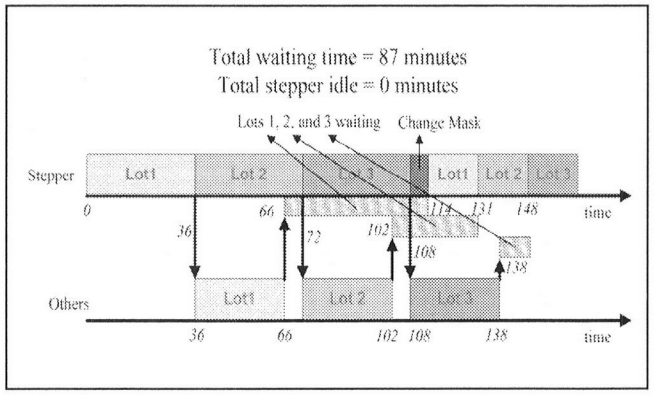

Figure 3
Workload Control Mechanism of (*k*,*W*) Rule

workload, *TW*, is updated when *k* lots of new wafers are released to the system or when an operation of a lot of wafers is finished. The (*k*,*W*) rule is based on the following rationale:

- Because steppers are the most crucial bottleneck of the entire wafer fabrication process, the (*k*,*W*) rule releases new jobs based only on the workload of steppers without considering the workloads of other resources.
- As mentioned above, a specific restriction applies to critical photomask operations but not to noncritical operations. To simplify the planning process, a wafer fabrication usually plans to perform both critical and noncritical operations of each wafer on the same steppers. That is, the same stepper is used to process all operations, critical or noncritical, for a given wafer. This policy might fail to optimize the schedule of the entire wafer production process because allowing noncritical operations to be performed on any available stepper may reduce wafer waiting time or stepper idle time. However, allowing wafers to switch to any available stepper makes production scheduling particularly difficult because no stepper can be assigned when new wafers are released and, thus, the workload of steppers cannot be determined. Restricting all operations of each wafer to be processed on the same stepper can overcome this difficulty because the processing time of new released wafers can be added to the workload of the assigned stepper. Therefore, the (*k*,*W*) rule requires the same stepper to process all operations of a given wafer and implement the policy by releasing new wafers for each stepper independently.
- Requiring $k \geq 2$ allows steppers to save photomask changing time because the (*k*,*W*) rule changes operations on steppers only when all *k* lots of wafers finish that particular operation. It is proven in Chern, Liu, and Shieh (2003) that family-based scheduling algorithm, which schedules the wafers needing the same operation as the current finished one ahead of other wafers, results in significant improvement on stepper utilization. The (*k*,*W*) rule also adopts this concept by releasing *k* lots of new wafers each time the releasing point is triggered, which

guarantees that at least k lots of wafers arrive at the same steppers consecutively. This requirement also assures that at most one mask change is performed for each k lots of wafers.

Based on the above rationale, a search algorithm is constructed to find the optimal batch size, k, and workload threshold, W, that minimize the average waiting time per lot plus total mask changing time and stepper idle time under certain constraints related to cycle time and utilization rate.

3. Problem Formulation: Determining Workload and Lot Size

The problem formulation in this section determines the decision variables such as the workload threshold, W, and the batch size of new jobs, k. Before the problem is stated, some parameters are defined as follows first.

- n total number of photomask operations for each wafer
- C_{ij} time to change photomasks from operation i to operation j
- P_i processing time of photomask operation i where $1 \le i \le n$
- A_i time interval between operations i and $i+1$ during which the lot is in the nonstepper area where $0 \le i \le n$
- m total number of batches, each consisting of k lots, in the system

Three state variables are also defined and will be used later in the formulation, as follows.

- T_{ijl} finishing processing time of lot j for operation i in batch l where $0 \le i \le n$, $1 \le j \le k$, and $1 \le l \le m$
- $T_{i'j'l'}$ finishing processing time of lot j' immediately before lot j for operation i' in batch l' where $0 \le i' \le n$, $1 \le j' \le k$, and $1 \le l' \le m$.
- N_C total number of mask changes
- TC total mask changing time

Because the (k,W) rule releases k lots of new wafers for each stepper independently, this formulation assumes only one stepper in the system. The finishing processing time of lot j in batch l for critical operation i, T_{ijl}, and the total number of mask changes, N_C, can be computed as follows.

$$\begin{aligned}
&\text{If } i \neq i',\ T_{ijl} = \max[T_{(i-1)jl} + A_{(i-1)},\ T_{i'j'l'}] + P_i + C_{ii'},\\
&TC = TC + C_{ii'},\ \text{and } N_C = N_C + 1.\\
&\text{If } i = i',\ T_{ijl} = \max[T_{(i-1)jl} + A_{(i-1)},\ T_{i'j'l'}] + P_i
\end{aligned} \tag{3.1}$$

The finishing time of lot j in batch l for operation i, T_{ijl}, is determined by choosing the maximum between the arrival time of itself, $T_{(i-1)jl} + A_{(i-1)}$, and the departure time of the previous lot, $T_{i'j'l'}$, plus lot processing time and mask changing time if the previous operation, i', is different from the required current operation, i. The total number of mask changes increases by 1 when the previous operation, i', is different from the required current operation, i.

The objective is to find the optimal k and W that minimize the average waiting time per lot of wafer plus stepper idle time and mask changing time under the requirements that the stepper utilization, ρ, is within the required maximum and minimum utilization rates, such as $\rho_u \ge \rho \ge \rho_l$, or average cycle time is less than a certain level, such as $\le T_r$. Define $f(k,W)$ as the average waiting time per lot of wafer, $g(k,W)$ as the stepper idle time, and $h(k,W)$ as the mask changing time.

Based on the above definitions, they can be calculated as follows.

$$\begin{aligned}
&f(k,W) = \left[\Sigma_{l=1}^{m}\Sigma_{i=1}^{n-1}\Sigma_{j=2}^{k}\left(T_{i'j'l'} - T_{(i-1)jl} - A_{(i-1)}\right)\right]/(m\,n\,k)\\
&\quad \text{if } \left(T_{(i-1)jl} + A_{(i-1)} - T_{i'j'l'}\right) < 0 \text{ for all } i,\ j,\ \text{and } l,\\
&g(k,W) = \Sigma_{l=1}^{m}\Sigma_{i=1}^{n-1}\Sigma_{j=2}^{k}\left(T_{(i-1)jl} + A_{(i-1)} - T_{i'j'l'}\right)\\
&\quad \text{if } \left(T_{(i-1)jl} + A_{(i-1)} - T_{i'j'l'}\right) > 0 \text{ for all } i,\ j,\ \text{and } l,\ \text{and}\\
&h(k,W) = TC
\end{aligned} \tag{3.2}$$

Let δ_0 be a weight assigned to each unit of wafer waiting time, δ_1 be a weight assigned to each unit of stepper idle time, and δ_2 be a weight assigned to each unit of mask changing time. The problem to determine the workload threshold, W, and the lot size, k, can be formulated as follows.

$$\begin{aligned}
&\min_{k,W}\left[\delta_0 f(k,W) + \delta_1 g(k,W) + \delta_2 h(k,W)\right]\\
&\text{s.t.}\quad \rho_l \le 1 - g(k,W)/T_{nkm} \le \rho_u\\
&\qquad \Sigma_{l=1}^{m}\Sigma_{j=1}^{k}\left(T_{njl} - T_{0jl}\right)/mk \le T_r
\end{aligned} \tag{3.3}$$

where T_{ijl} is computed as in (3.1) and $f(k,W)$, $g(k,W)$, and $h(k,W)$ are computed as in (3.2).

In the following section, a search algorithm for lot size, k, and workload threshold, W, is constructed and described.

4. Search Algorithm for k and W

As seen in *Figure 4a*, the waiting time increases but the stepper idle time decreases as k increases. With fixed δ_i where $0 \le i \le 2$, it is possible to find a k that minimizes the weighted sum of wafer waiting time, stepper idle time, and mask changing time. A search algorithm for (k,W) is constructed to find the optimal solution to problem (3.3) based on the following bounds on k.

- Lower bound on $k = 2$.
- Upper bound on $k = \lceil \max [A_i] / \min[P_i] \rceil$.

The reason for the lower bound on k to be 2 is based on the rationale mentioned earlier. The reason to justify the upper bound of $k = \lceil \max [A_i] / \min[P_i] \rceil$ is that the stepper will not be idle when $\lceil \max [A_i] / \min[P_i] \rceil$ lots of wafers are released to the system. Therefore, more than $\lceil \max [A_i] / \min[P_i] \rceil$ lots of new jobs released into the system only increases total wafer waiting time.

To determine the releasing times of new lots, a workload threshold, W, has to be found first. Define TW as the total current workload of the stepper. Whenever a new lot of wafers is released to the system, the total future processing time on stepper of this lot will be added to TW. Whenever a wafer finishes a stepper operation, its processing time is deducted from TW. The total current workload, TW, is then compared with the threshold, W. If $W \ge TW$, k lots of new wafers are released. Otherwise, no action is taken. An example with $k = 10$, as seen in *Figure 4b*, could demonstrate the relationship between time and workload. A search algorithm for (k,W) is constructed to find the optimal solution to problem (3.3) based on the following bounds on W.

- Lower bound on $W = 0$ unit of time or the total required processing time of 0 lots of wafers on stepper.
- Upper bound on W = the total required processing time of $\lceil \max [A_i] / \min[P_i] \rceil$ lots of wafers on stepper $\lceil \max [A_i] / \min[P_i] \rceil * \sum_{i=1}^{n} P_i$ unit of time.

The idea behind the lower bound on W is that the new wafers will be released only when the system is

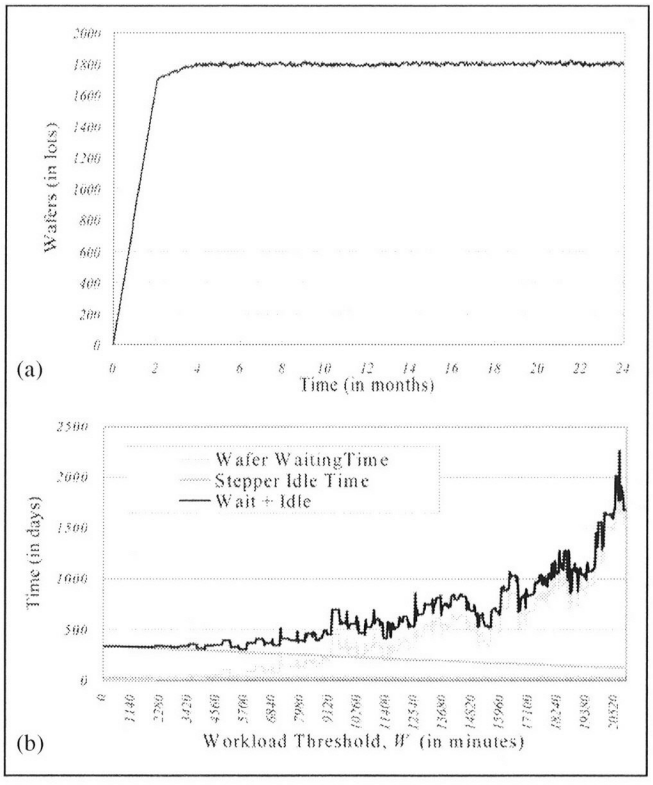

Figure 4
(a) Relationship Between k and Time (W = 2010 min.)
(b) Relationship Between W and Time (k = 10)

empty. If W is equal to the total processing time of $\lceil \max [A_i] / \min[P_i] \rceil$ lots of wafers, the system releases new lots when the first operation of the first lot is finished. At this moment, so many wafers already saturate the stepper, leaving the stepper with no idle time. Releasing more wafers only makes the wafers wait longer. As mentioned in the above rationale, the (k,W) rule changes operations on steppers only when all k lots of wafers finish that particular operation. Because the processing time of each operation is at least 1 unit of time, the system status does not change during at least k units of time or during the processing of these k lots.

For example, two lots of wafers are released to a stepper, and assume that each lot needs only two operations and the processing time of each operation is equal to 1 as well as one unit of interprocessing time. Set $W = 1$. At time 0 and time 1, $TW = 4$, implying that no action is taken. At time 2, the first operations of lots 1 and 2 are finished and still no action is taken because $TW = 2 > W = 1$. At time 4, the second operations of lots 1 and 2 are finished and two lots of wafers are re-

leased because $TW = 0 < W = 1$. The system does not check TW at time 3 because all two lots must be finished. Even if TW is checked at time 3, new lots of wafers won't be released because lot 2 must be processed, and releasing new lots at this moment worsens the system performance by increasing wafer waiting time.

Therefore, the performances of $W = 1$ and $W = 0$ are the same. If set $W = 2$, two new lots will be released at time 2, and thus, system performances will be different from the performances when $W = 1$ and $W = 0$. However, the performances of $W = 2$ and $W = 3$ are the same. It turns out that the performances stay the same on every two units of workload. When $k = 3$, the performances stay the same on every three units of workload. Because the processing time of each operation is usually larger than one unit of time and there are more than two lots of wafers in the system, the system status stays the same every k unit of time. In other words, $f(k,0) = f(k,1) = ... = f(k,k-1)$, $f(k,k) = f(k,k+1) = ... = f(k,2k-1)$, and so on. Similar relationships hold for $g(k,W)$ and $h(k,W)$. In other words, it is only necessary to search one value of W in each of the ranges of $[0,k-1]$, $[k,2k-1]$, $[2k,3k-1]$, ..., etc. This also implies that the search algorithm is more and more efficient as k gets larger and larger.

Let $M = \lceil \max [A_i] / \min[P_i] \rceil * \lceil \max [A_i] / \min[P_i] \rceil * \Sigma_{i=1}^{n} (P_i + A_i)$ units of time. That is, the algorithm stops at time M. The reason to set M equal to this large number is based on the upper bounds on k and W. The total number of batches to be released into the system, m, is bounded by the upper bounds of k and W. The largest searching point of the searching algorithm is when $k = \lceil \max [A_i] / \min[P_i] \rceil$ and $W =$ the total required processing time of $\lceil \max [A_i] / \min[P_i] \rceil$ lots of wafers on the stepper. M must be larger than the required processing time of $\lceil \max [A_i] / \min[P_i] \rceil * \lceil \max [A_i] / \min[P_i] \rceil$ lots of wafers, the combined search range of k and W.

The searching range of the algorithm based on the bounds on W computed above is very large. Based on the required stepper utilization rates, a lot of points in the searching range may be eliminated. Two heuristics are constructed to further restrict the bounds on W by applying a steady-state Poisson process. The average workload can be computed as $W = \sum_{i=1}^{n} L_i \sum_{j=i}^{n} P_j$ where L_i is the average number of wafers waiting for operation i and $\sum_{i=1}^{n} L_i \sum_{j=i}^{n} P_j$ is the remaining processing time on the stepper for wafers of operation i. Because P_j's are known and constant, the average workload can be computed if L_i's are known. To determine L_i, average arrival rate and average service rate are needed because it can be computed by $\frac{\lambda}{\mu_i - \lambda}$ if assuming a Poisson process based on Ross (1989). The wafer fabrication process is not a Poisson process if the workload regulated batch input control rule is applied. One reason to assume this process is Poisson is that the formula to compute L_i is ready and easy to apply. Another reason for the Poisson process to be assumed here is that the average arrival rate and the service rate are the only available information that can be obtained. If applying other types of stochastic processes, detail information such as the distribution or density function of the arrival process is needed, which is impossible to know at this moment. For each photomask operation i, the average service rate, μ_i, is approximated by $\frac{1}{nP_i}$ because the stepper is shared among all the photomask operations. Average arrival rate can be computed as the required utilization rate divided by the total processing time of one lot of wafers or $\lambda = \frac{\rho}{\Sigma_{i=1}^{n} P_i}$ because $\rho = \frac{\lambda}{\mu} = \frac{\lambda}{1/\left(\Sigma_{i=1}^{n} P_i\right)}$. Based on the maximum and the minimum utilization rates given in (3.3), the new bounds on W can be further restricted by the following heuristics.

- Modified lower bound on W: Let $\lambda_l = \frac{\rho_l}{\left(\Sigma_{i=1}^{n} P_i\right)}$, which means that λ_l is the arrival rate to achieve the stepper utilization requirement ρ_l. The required condition for the system to be stable is the average of $\mu_i > \lambda_l$. The average number of jobs waiting for operation i, L_i, could be approximated by $\frac{\lambda_l}{\mu_i - \lambda_l}$. Therefore, the modified lower bound of W can be approximated by $W_l = \Sigma_{i=1}^{n} \frac{\lambda_l}{\mu_i - \lambda_l} \Sigma_{j=i}^{n} P_j$.

- Modified upper bound on W: Let $\lambda_u = \frac{\rho_u}{\left(\Sigma_{i=1}^{n} P_i\right)}$. The average number of jobs waiting for operation i, L_i, could be approximated by $\frac{\lambda_u}{\mu_i - \lambda_u}$. The modified upper bound of W can be approximated by $W_u = \Sigma_{i=1}^{n} \frac{\lambda_u}{\mu_i - \lambda_u} \Sigma_{j=i}^{n} P_j$.

Combining the above argument, the search algorithm is constructed as follows.

Algorithm 1:

(0) Initialize Z^*, k^*, and W^* to be a very large number.

Let $M = \lceil \max [A_i] / \min[P_i] \rceil \lceil \max [A_i] / \min[P_i] \rceil * \left[\sum_{i=1}^{n} (P_i + A_i)\right]$.

(1) For $k = 2$ to $\lceil \max [A_{ij}] / \min[P_i] \rceil$, + + 1
(2) For $W = W_l$ to W_u, + + k
(3) Initialize TW, TNOW, $f(k,W)$, $g(k,W)$, $h(k,W)$, N_c, and m to be 0.
Empty the wafer set, S. Let $i' = 1$.
(4) While TNOW < M
(4.1) If $TW \leq W$, release another k lots of new wafers into the wafer set, S, and add the total processing time of k lots of new wafers, or $k\Sigma_{i=1}^{n} P_i$, to TW.
(4.2) Retrieve the next lot of wafer from the wafer set.
(4.3) If arrival time of this lot $(T_{(i-1)jl} + A(i\ 1))$ > TNOW,
TNOW = $T_{(i-1)jl} + A(i-1)$.
Endif.
If $i \neq i'$, TNOW = TNOW + $C_{ii'}$.
Endif.
TNOW = TNOW + P_i.
(4.4) Update T_{ijl} as in (3.1).
(4.5) Compute $f(k,W)$, $g(k,W)$, $h(k,W)$, N_C, and m as in (3.3).
(4.6) If the last operation, the lot is finished.
Else, the lot returns to the wafer set.
Let $i' = i$.
(5) Endwhile
(6) If $\rho_l \leq 1 - g(k,W) / M \leq \rho_u$ and $\Sigma_{l=1}^{m} \Sigma_{j=1}^{k} \left(T_{njl} - T_{0jl}\right) / mk \leq T_r$, compute $Z = \delta_0 f(k,W) + \delta_1 g(k,W) + \delta_2 h(k,W)$,
If $Z < Z^*$, let $Z^* = Z$, $k^* = k$, and $W^* = W$.
Endif
Endif
(7) Endfor
(8) Endfor
(9) Output Z^*, k^*, and W^*.

The search algorithm initializes the values of decision variables Z^*, k^*, and W^* as a very large number in (0) and compute the value of M. It then set the bounds on k and W to be the search ranges in (1) and (2). The current workload, TNOW, $f(k,W)$, $g(k,W)$, $h(k,W)$, N_C, and m for each pair of (k,W) are all initialized as 0, the wafer set, S, is emptied, and the first operation is set to be 1 in (3). The length of each run is set for M in (4). The algorithm determines whether to release k lots of new wafers into the wafer set, S, in (4.1). If yes, then the total processing time of these new jobs is added to the current workload. The algorithm then retrieves the next lot in the wafer set and obtains the attributes of this lot, including $T_{(i-1)jl}$ and $A_{(i-1)}$. The attributes are then used to compute TNOW in (4.3) and T_{ijl} in (4.4). The finishing time of the current job on stepper, waiting time, idle time, and mask changing time are computed in (4.5), if necessary. After each run, the algorithm checks the required conditions in (6). If they are met, the algorithm further computes the weighted objective function and compares it with the current best one. The algorithm outputs the optimal solution within the search range in (9). It should be noted that not all the points in the search ranges are needed as seen in (2). Only one in every k values of W is checked in this algorithm. That makes this algorithm very efficient because the larger k is, the fewer points of W are searched. To evaluate the (k,W) rule with the workload and lot size found in this section, a simulation model is built and other input control rules are included in the next section.

5. Simulation Model

The simulation model was based on a wafer production process of a high-volume, single-product semiconductor fab, which starts with non-photolithography processes such as cleaning and diffusion,

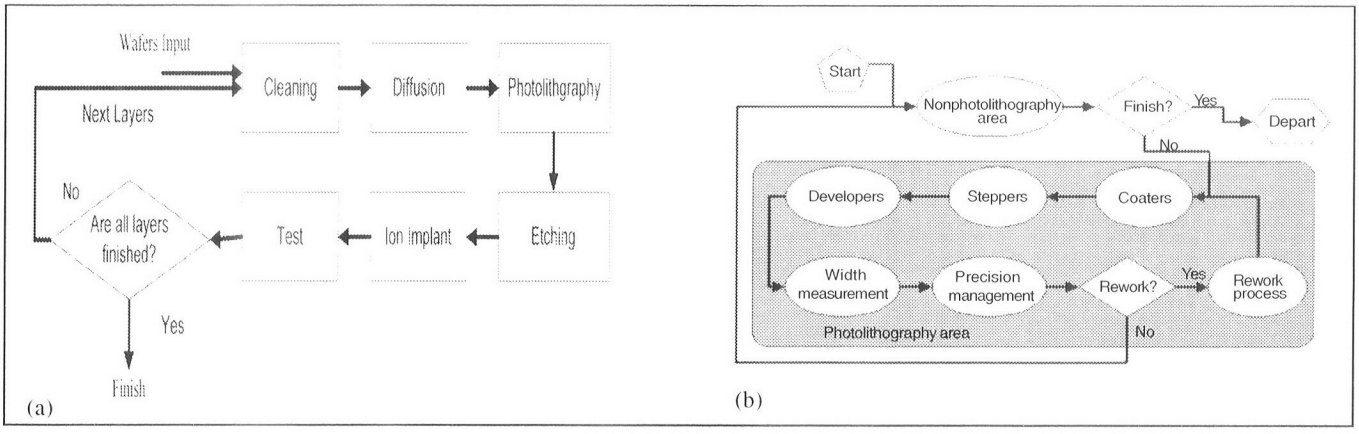

Figure 5
(a) Basic Flow of Wafer Production Process
(b) Five Operations in Photolithography Area

Table 1
Times Needed for Operations *i* on Steppers and Other Workcenters for All Lots (in minutes)

Critical Operation *i*	Coater	Stepper	Developer	Width Measurement	Precision Measurement	Nonlithography Area
0						14400
1	35	34	36	9	6	7200
2	35	31	36	9	6	7560
3	35	32	36	9	6	10080
4	35	33	36	9	6	10560
5	35	32	36	9	6	3210
6	35	35	36	9	6	5050
7	35	35	36	9	6	3890
8	35	35	36	9	6	5350
9	35	34	36	9	6	4220
10	35	34	36	9	6	14760

Table 2
Rework Rate (%)

Operation *i*	Rate (%)
1	1.0
2	1.2
3	3.2
4	1.2
5	4.0
6	2.2
7	1.1
8	1.6
9	0.6
10	2.9

as seen in *Figure 5a*. Once the wafers enter photolithography area, five operations such as coating on coaters, photolithographing on steppers, developing on developers, and width and precision measurement on measurers, are performed successively, as shown in *Figure 5b*. It is assumed that the photolithography process results in either good product or defective product that requires rework and that this outcome is an independent Bernoulli trial with rework probabilities, listed in *Table 2*. SLAM II is used to build this simulation model. The data needed in this simulation model were collected from a DRAM wafer fabrication located in Hsin-Chu, Taiwan. Each wafer has to go through similar processes, but takes different lengths of stepper time as well as other resource time, as shown by *Table 1*. The following assumptions underlie this model:

1. As mentioned earlier, each lot, or wafer boat, consists of 25 wafers.
2. Each reworking lot delays a constant time of 24 hours plus a variable time following an exponential distribution with mean equal to 36 hours.
3. The time between machine failures and repair times for each stepper and other machines are both assumed to be random variables following exponential distributions with means, listed in *Tables 3a*, *3b*, and *3c*. Each machine is also scheduled for regular maintenance, as seen in *Table 4*.
4. The time to prepare and load each input lot on a stepper is a constant of 12 minutes.

Table 3a
Machine Failures of Steppers (in hours)

Stepper	Mean Time Between Failures	Mean Repair Time
1	145	1.3
2	388	0.5
3	211	0.6
4	138	0.3
5	482	0.1
6	63	2.2
7	96	0.6

Table 3b
Machine Failures of Coaters (in hours)

Coater	Mean Time Between Failures	Mean Repair Time
1	260	1.5
2	260	1.5
3	134	1
4	197	1.6
5	210	0.9
6	329	0.7
7	211	0.1
8	480	4.6

Table 3c
Machine Failures of Developers (in hours)

Developer	Mean Time Between Failure	Mean Repair Time
1	65	0.7
2	979	0.7
3	650	0.9
4	215	0.8
5	490	0.8
6	647	0.3
7	993	0.4
8	152	0.4

Table 4
Regular Maintenance of Work Units in Photolithography Area

Work Units	Type of Regular Maintenances	Interval (days)	Repair Time (minutes)
Coater	1	60	120
Developer	1	60	120
Measurer for width	1	1	10
	2	30	60
	3	180	480
Measurer for precision	1	7	20
	2	30	60
Stepper	1	30	84
	2	90	150
	3	180	240

5. In a wafer fabrication process, special requirements for certain products with various quantities, called hot lots, arise occasionally. Hot lots are assigned the highest priorities according to the engineering principle in general wafer fabs. This rule is also adopted here. Furthermore, if more than one hot lot is in the waiting area, the process sequencing policy is first-come first-serve (FCFS). Hot lot percentages are 7% of the overall wafers released to the system.
6. Four photomasks are available for each photomask operation. Changing photomask usually takes 6 to 20 minutes. Therefore, two extreme cases of mask changing time are considered in this simulation model, one with 6 minutes and the other with 20 minutes.

To verify and validate the simulation model, suppose that the input control method is uniform input with 27.857 lots each day (in a seven-day period, 28 lots for six days, and 27 lots for one day.) The dispatch rule used for steppers is FIFO. The total simulation time is set to 720 days with a 360-day warm-up period. *Figure 6a* shows the WIP vs. time, and *Figure 6b* shows the utilization of steppers vs. time from simulation runs. As seen in both figures, after one year the system becomes quite stable. Also from the simulation, the average throughput per day is 27.88 lots and average cycle time is 64.93 days, figures that are close to 28.37 lots and 68 days, the data collected from the real system. The average WIP from simulation run is 1808.7 lots, which is equal to 1808.8 lots from Little's formula as well as close to 1766.7 lots from the real system. According to this input rate, the utilization rate of each stepper is 94.37% theoretically. The failure and maintenance rates are 1.09% and 0.28%, respectively. The theoretical total stepper occupancy rate, including utilization, repair, and maintenance, is 95.46%, which is close to 95.48% from the simulation run. Therefore, this simulation model is accurate and could be used to compare (k,W) with other input control policies.

6. Input Control Policies in the Simulation Model

In this section, four input control policies are chosen. It should be noted that 95% of steppers' utiliza-

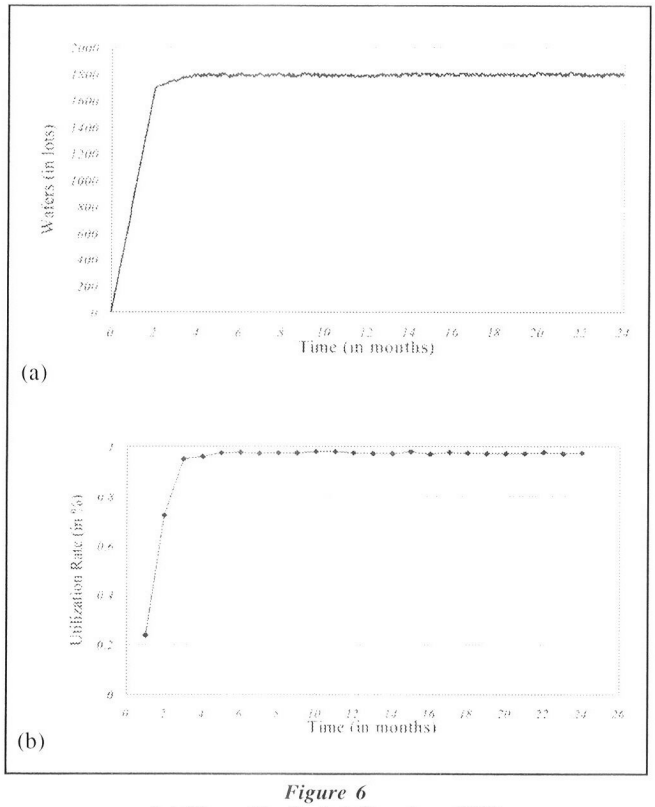

Figure 6
(a) Warm-Up Period Based on WIP
(b) Warm-Up Period Based on Utilization

tion rate is assumed. By this assumption, the parameters of these four input control policies are computed and described as follows:

- Uniform Input Control (UN): The policy releases new jobs into system at a constant rate every day. In the case of 6-minute mask changing time, the rate is 27.857 lots of wafers (in a seven-day period, 28 lots for six days, and 27 lots for one day) every day. In the case of 20-minute mask changing time, the rate is 25.857 lots of wafers (in a seven-day period, 26 lots for six days, and 25 lots for one day) every day. The interarrival time between two lots is a random variable following the Exponential distribution with mean equal to 7 minutes.
- Fix-WIP Input Control (FW): The policy keeps a constant WIP level of 1805 lots of wafers in the system when mask changing time is 6 minutes and the level of WIP is to be 1647 lots of wafers in the system when mask changing time is 20 minutes. Whenever the WIP drops below these levels, a new lot of wafers is released into system. This policy has been mentioned and compared in Wein (1988), Levitt and Abraham (1990), and Glassey, Shanthikumar, and Seshadri (1996). This policy is also similar to the starvation avoidance input control policy proposed by Glassey and Petrakian (1989), which computes the virtual inventory between new wafers and the return wafers for steppers because steppers are the only bottleneck in the whole process.
- Workload Regulation Input Control (WL): The policy keeps a constant workload of each stepper to be 327,000 minutes (or 227.08 days) for the case with 6-minute mask changing time or 298,000 minutes (or 206.94 days) for the case with 20-minute mask changing time. The system releases new wafers whenever the workload of each stepper drops below this threshold level. This policy is proposed by Wein (1988) and Lawton et al. (1990).
- (k,W) Input Control: Two sets of parameters must be defined in this rule. Based on the requirements of a wafer fabrication, the utilization rate of each stepper must be greater than 90%, and the average cycle time of each lot must be less than 65 days (or 93,600 minutes). If the average cycle time of each lot is not restricted to 65 days, the best solution chosen in this algorithm will have an average cycle time of 72 days, which is considered too long for any wafer fabrication. Finally, assume that $d_l = 1$ for all l. Based on the derivation of sections 2 and 3, the problem becomes

$$\min_{k,W} \left[f(k,W) + g(k,W) + h(k,W) \right]$$
$$\text{s.t.} \quad 90\% \le 1 - g(k,W)/T_{nkm} \le 95\%$$
$$\Sigma_{l=1}^{m} \Sigma_{j=1}^{k} \left(T_{njl} - T_{0jl} \right) / mk \le 93600.$$

where T_{ijl} is computed as (3.1) and $f(k,W)$, $g(k,W)$, and $h(k,W)$ are computed as (3.2).

The lower and upper bounds of k are 2 and 478 lots, while the lower and upper bounds of W are 17,402 minutes (or 12.08 days) and 68,125 minutes (or 47.31 days), respectively, based on the heuristics in Section 4. Using Algorithm 1 in Section 4, the optimal solution of k and W are 8 and 32550 for the case of 6-minute mask changing time. In other words, the policy releases 8 lots of new wafers into system when the workload of a stepper is less than 32,550 minutes (or 22.60 days)

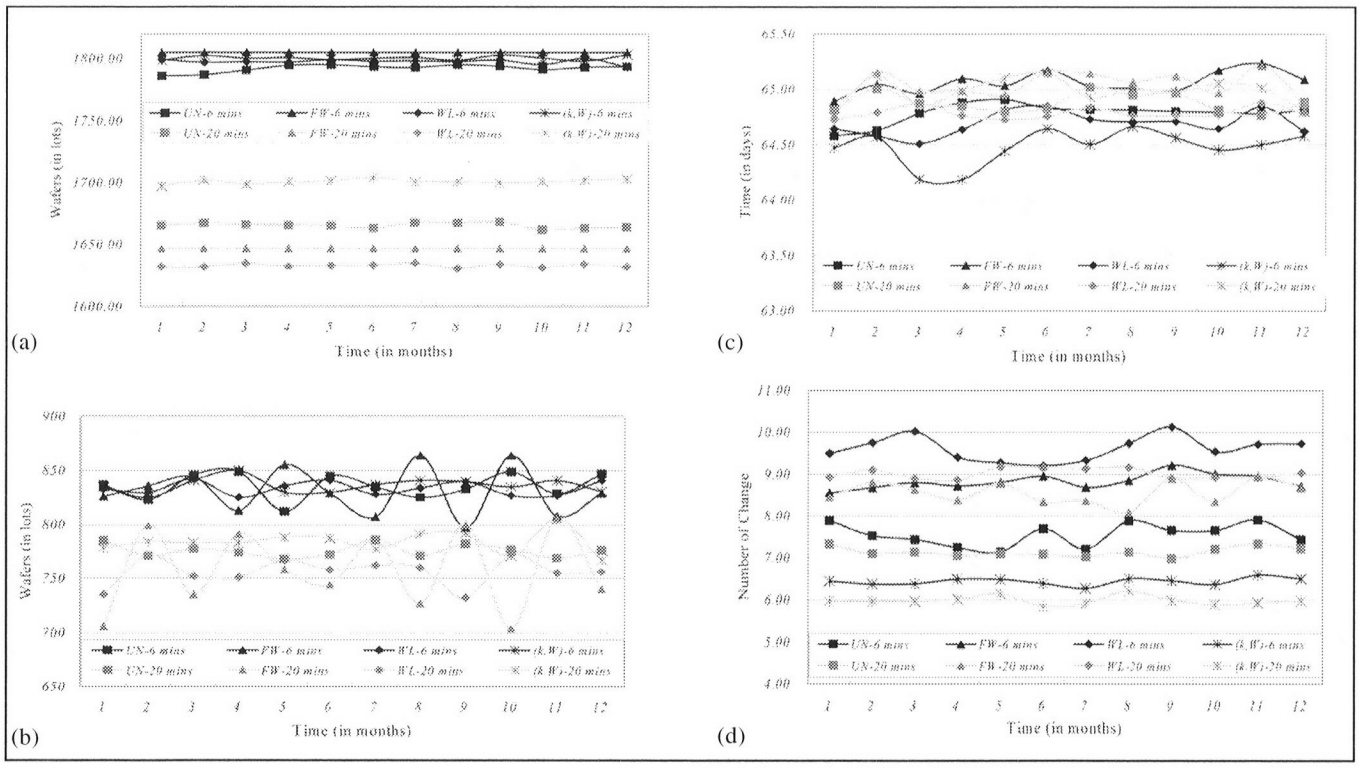

Figure 7
(a) Time Plot for WIP, (b) Time Plot for Monthly Throughput, (c) Time Plot for Cycle Time, (d) Time Plot for Number of Mask Changes

for the case with 6-minute mask changing time. For the case of 20-minute mask changing time, the policy releases 8 lots of new wafers when the workload of a stepper is less than 30,750 minutes (or 21.35 days).

After trying to match several known dispatch rules with the above four input control policies, it was found that a modified FIFO rule performs best with all four of these policies. Therefore, the modified FIFO is the only dispatch rules selected in the simulation model and its mechanism is elaborated as follows:

1. A stepper is idle.
2. Any hot lot? If yes, assign the lot to this stepper and go to step 4. Else continue.
3. Any lots in the queue waiting for this stepper? If no, go to step 6. Else continue.
4. Any lots that need the same mask? If yes, assign the lots to this stepper and go to step 6. Else continue.
5. Assign lot with earliest arrival time to this stepper.
6. Update the workload for this stepper. Assignment completes.

In the following section, the above four input control methods combined with the modified FIFO dispatch rule are compared through the simulation model built in section 5.

7. Computational Analysis

To compare these policies on the same ground, the occupancy rate of each stepper is set to be approximately 98%, which includes the utilization (95%), failure, repair, and regular maintenance. The comparison measurements used in this study to evaluate the performance of (k,W) as well as other input policies are as follows.

- Mean cycle time (in hours): the average time a wafer stays in the system.
- Average monthly throughput rate (in lots): the average number of good wafers produced each month.
- Average number of mask changes per stepper per day.
- Mean WIP (in lots): the average number of wafers in the system.

Figures 7a to *7d* show the time plots of average monthly throughput rate, average cycle time per lot of wafer, average WIP, and average number of mask

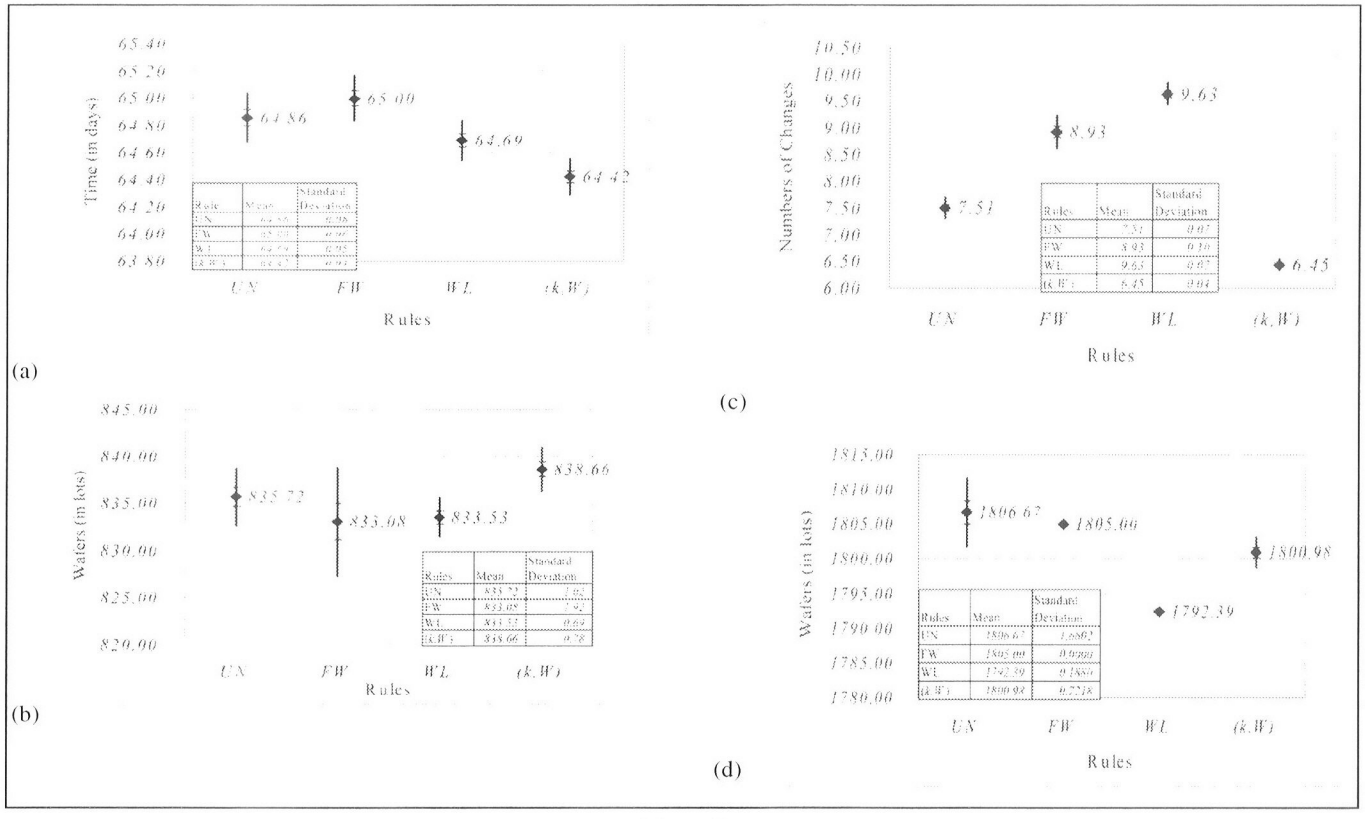

Figure 8
(a) Average Cycle Time with 6-Minute Mask Changing Time
(b) Average Monthly Throughput with 6-Minute Mask Changing Time
(c) Average Number of Mask Changes with 6-Minute Mask Changing Time
(d) Average WIP with 6-Minute Mask Changing Time

changes per day for both 6-minute and 20-minute cases from the simulation runs. The time plots show not only the levels of all these measurements from different input rules but also the variations through time. It is clear that WIP levels and cycle time are stable due to the principle adopted by these rules. However, average monthly throughput rate oscillates if applying the FW policy. That is, the FW rule considers only the number of lots of wafers in the system but not the remaining processing time needed for these wafers. Thus, the FW rule tends to release a large amount of new wafers to the system and waits a long time for the wafers to finish. The oscillating pattern of average monthly throughput rate shows in both 6-minute and 20-minute cases for the FW rule. Because the system spends more time on changing masks in the 20-minute case, the average monthly throughput rate, the average WIP, and the average number of mask changes per day are lower than those of the 6-minute case. The pattern of the average cycle time shows mixed results. The average cycle time should have been larger in the 20-minute case than that in the 6-minute case if the occupancy rates of steppers for both cases would have been kept the same. In the case of 6-minute mask changing time, all four policies can achieve the requirement of 98% stepper occupancy rate. However, only the (k,W) rule can reach 98% stepper occupancy rate by saving mask changing time in the case of 20-minute mask changing time.

The other three input control policies can only reach less than 95% of the stepper occupancy rate for the system spends more time in changing masks. Less new lots inputted into the system causes less average WIP levels, or less average cycle time, or both for the UN, FW, and WL input policies in this case (input rates per month dropping from 835 lots to 775 lots for the UN rule, from 833 lots to 759 lots for the FW rule, and from 833 lots to 755 lots for the WL rule). The (k,W) rule still maintains relatively higher input rates (dropping from 838 lots to 785 lots per month) than the other three rules, which causes lower average WIP\ level (dropping from 1800 lots to 1700 lots) and higher average cycle time

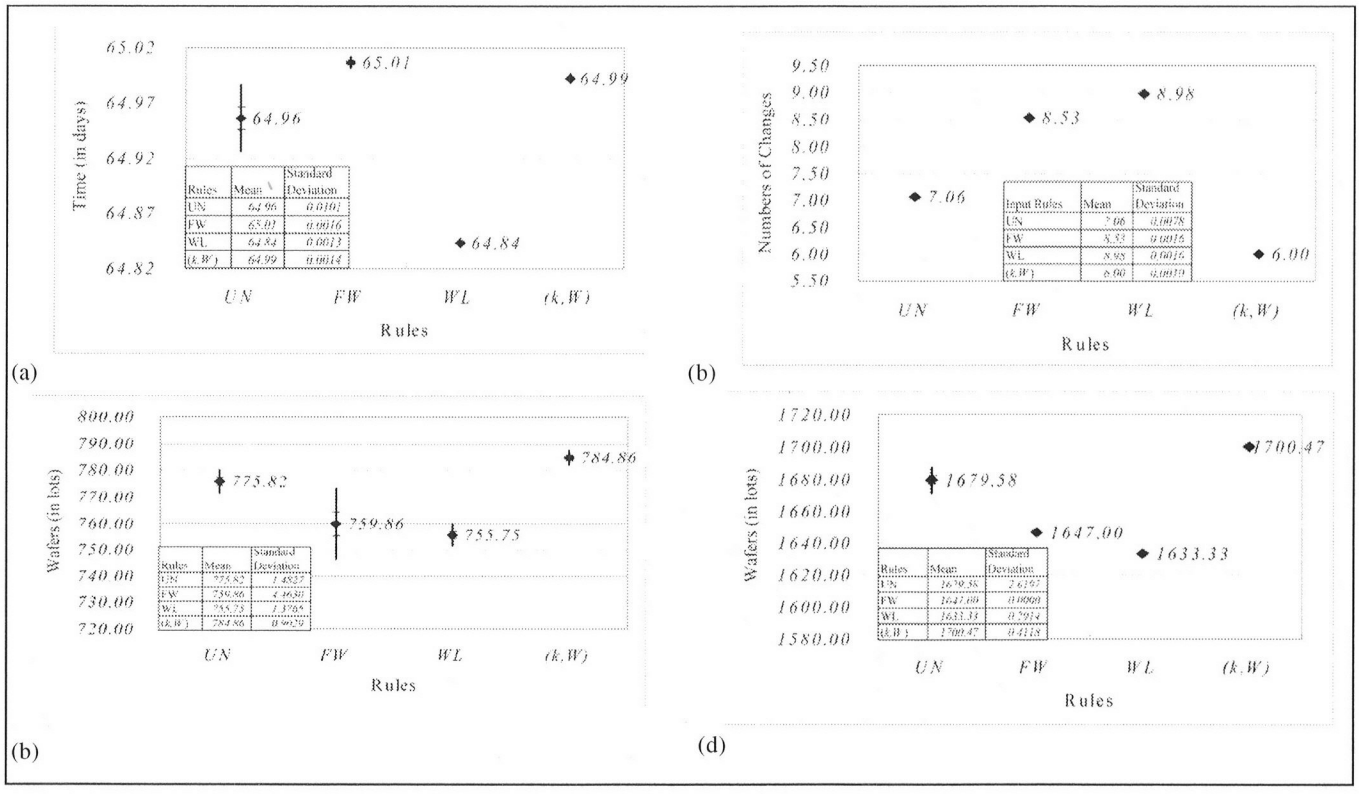

Figure 9
(a) Average Cycle Time with 20-Minute Mask Changing Time
(b) Average Monthly Throughput with 20-Minute Mask Changing Time
(c) Average Number of Mask Changes with 20-Minute Mask Changing Time
(d) Average WIP with 20-Minute Mask Changing Time

(jumping from 64.42 hours to 64.49 hours). The (k,W) rule could have reduced the input rate (dropping below 780 lots per month) to lower the average cycle time to the same level as the ones generated by the other three rules. However, increasing the idle time of steppers caused by reducing input rate is against the principle of the (k,W) rule to fully utilize the steppers by saving mask changing time. Therefore, it is suggested to keep a higher input rate for the (k,W) rule and at the same time maintain a slightly higher average cycle time, still meeting the average the average cycle time requirement of 65 days.

Ten runs of simulation are done for each input control policy with 6- and 20-minute mask changing time, respectively. Ninety-five percent (95%) confidence intervals of all the measurements for all four rules with 6-minute mask changing time are plotted in *Figures 8a* to *8d* as well as in *Figures 9a* to *9d* for 20-minute mask changing time. The diamond marks are the position of means and the little horizontal dashes indicate the positions of ± one standard deviation. In the 6-minute case, the (k,W) rule performs best with smallest average cycle time, largest average monthly throughput, and lowest number of mask changes per day among four policies, as seen in *Figures 8a* to *8c*. The (k,W) rule changes masks only six times compared with 8 to 10 times done by other policies, which implies that the (k,W) rule saves up to 24 minutes per day per stepper, or 0.5 days per month per stepper, or 1.67% utilization rate per stepper. The other policies have mixed performance results. The WL policy outperforms the (k,W) rule in WIP, as shown in *Figure 8d*, but is the second in average throughput and cycle time in the 6-minute case, which implies that lower WIP level comes from smaller throughput rate. In addition, the WL rule spend the largest amount of time in changing masks, as seen in *Figure 8c*, among all four rules. The UN and FW rules perform similarly, except that the FW rule shows most unstable pattern in average monthly throughput rate because its 95% confidence interval is larger than the ones generated by the other three rules.

In the 20-minute case, the (k,W) rule generates the lowest number of mask changes per day among

the four policies, as seen in *Figure 9c*. However, the (k,W) rule keeps higher WIP level and average monthly throughput for the reason mentioned above. The cycle time of the (k,W) rule is the same as those of other rules. This happens because (k,W) rule inputs more wafers to the system than other policies. In other words, the (k,W) rule allows steppers to spend more time on processing wafers than being idle for mask changing. Thus, the (k,W) rule changes masks for only six times comparing with seven to nine times if applying other policies, which implies that the (k,W) rule saves up to 60 minutes per day per stepper, or 1.25 days per month per stepper, or 4.17% utilization rate per stepper. Also because of this large saving in utilization rate, the (k,W) rule is the only input control rule that is able to reach 98% stepper occupancy rate in the case of 20-minute mask changing time. On the other hand, UN and WL policies outperform the (k,W) rule in WIP but are the second in average monthly throughput and almost the same in cycle time. Based on the discussion above, the (k,W) rule generates not only best but also stable results in the simulation runs and thus, could be adopted to use in a single-product, high-volume wafer fabrication plant.

8. Conclusion

This research introduces a heuristic input control policy, called the workload regulated batch-sizing rule, or the (k,W) rule, which adopts the workload regulation method with a family-based scheduling concept and can be easily implemented to help solve the new job releasing problem for steppers in the photolithography area of a single-product, high-volume wafer fabrication plant. The (k,W) policy releases k lots of new wafers into system when the total current workload associated with a stepper is less than a threshold, W. To adopt the (k,W) rule, two decision variables, batch size, k, and workload threshold, W, must be determined first. A search algorithm is constructed to find k and W within two sets of bounds. Simulation is used to evaluate the performance of the (k,W) rule as well as to compare the (k,W) rule with other known input control methods such as uniform input, fix-WIP, and workload regulation. A Taiwan DRAM wafer fabrication plant is used as a case study. As results of this work, the (k,W) rule is shown to be better than other known input control policies in terms of average monthly throughput, mean cycle time, and the number of mask changes. It is of future interest to include other work areas such as diffusion area into this study. As mentioned before, the bottleneck of the whole wafer fabrication process is the photolithography operations. However, the wafers must go through metal operations in diffusion area before being processed in photolithography area. These metal operations, including cleaning, oxidation, deposition, and metallization, must be processed by specific furnaces in large batch size (about four to six lots of wafers). The wafers must be properly cleaned before being grouped and loaded into furnaces. It generally takes half an hour to clean one lot of wafers and more than 10 hours for each batch of wafers. Although wafer fabs usually have enough number of furnaces, the wafer dispatching on the furnaces must be careful. If the furnaces are not properly assigned, the bottleneck of the whole process might switch to diffusion operations because they are the prior operations of photolithography operations as mentioned in Chern, Hsu, and Shieh (1999). Therefore, it is important to develop a combined production operation strategy for the whole wafer fabrication, including diffusion area, photolithography area, and so on, into this study.

References

Bai, X. and Gershwin, S.B. (1990). "A manufacturing scheduler's perspective on semiconductor fabrication." *VLSI Memo*. Cambridge, MA: Massachusetts Institute of Technology, pp89-518.

Chen, H.; Harrison, J.M.; Mandelbaum, A.; Ackere, A.V.; and Wein, L.M. (1988). "Empirical evaluation of a queueing network model for semiconductor wafer fabrication." *Operations Research* (v36, n2), pp202-215.

Chern, C.C.; Hsu, E.; and Shieh, C.C. (1999). "Time-performance-mask algorithm (TPMA) for photolithography and diffusion areas in a wafer manufacturing plant." *Proc.* of 5th Int'l Conf. of the Decision Science Institute (v2), pp1680-1682.

Chern, C.C.; Liu, Y.L.; and Shieh, C.C. (2003). "Family-based scheduling rules of a sequence-dependent wafer fabrication system." IEEE Trans. on Semiconductor Mfg. (v16, n1), pp15-25.

Dayhoff, J.E. and Atherton, R.W. (1984). "Simulation of VLSI manufacturing areas." *VLSI Design* (Dec. 1984), pp84-92.

Glassey, C.R. and Petrakian, R.G. (1989). "The use of bottleneck starvation avoidance with queue predictions." Berkeley, CA: Univ. of California-Berkeley.

Glassey, C.R. and Resende, M.G.C. (1988). "Closed-loop job release control for VLSI circuit manufacturing." *IEEE Trans. on Semiconductor Mfg.* (v1, n1), pp.36-46, February 1988.

Glassey, C.R.; Shanthikumar, J.G.; and Seshadri, S. (1996). "Linear control rules for production control of semiconductor fabs." *IEEE Trans. on Semiconductor Mfg.* (v9, n4), pp536-549.

Hughes, R.A. and Schott, J.D. (1986). "The future of automation for high volume wafer fabrication and ASIC manufacturing." *Proc.* of the IEEE (v74, n12, pp1775-1793.

Johri, P.K. (1992). "Practical issues in scheduling and dispatching in semiconductor wafer fabrication." *Journal of Manufacturing Systems* (v12, n6), pp474-485.

Lawton, J.W.; Drake, A.; Henderson, R.; Wein, L.M.; Whitney, R.; and Zuanich, D. (1990). "Workload regulating wafer release in a GaAs fab facility." Int'l Semiconductor Mfg. Science Symp., pp33-38.

Leachman, R.C. and Hodges, D.A. (1996). "Benchmarking semiconductor manufacturing." *IEEE Trans. on Semiconductor Mfg.* (v9, n2), pp156-169.

Levitt, M.E. and Abraham, J.A. (1990). "Just-in-time methods for semiconductor manufacturing." IEEE/SEMI Advanced Semiconductor Mfg. Conf., pp3-9.

Lin, H.C. and Raghavendra, C.S. (1996). "Approximating the mean response time of parallel queues with JSQ policy." *Computer Operations Research* (v23, n8), pp733-740.

Lozinski, C. and Glassey, C.R. (1988). "Bottleneck starvation indicator for shop floor." *IEEE Trans. on Semiconductor Mfg.* (v1, n4), pp147-153.

Miller, D.J. (1990). "Simulation of a semiconductor manufacturing line." *Communications of the ACM* (v33), pp98-108.

Ross, S.M. (1989). *Probability Models*. Boston: Academic Press.

Spence, A.M. and Welter, D.J. (1987). "Capacity planning of a photolithography work cell in a wafer manufacturing line." *Proc.* of IEEE Conf. on Robotics and Automation, pp702-708.

Uzsoy, R.; Lee, C.Y.; and Martin-Vega, L.A. (1992). "A review of production planning and scheduling models in the semiconductor industry. Part I: system characteristics, performance evaluation and production planning." *IIE Trans.* (v24, n4), pp47-60.

Uzsoy, R.; Lee, C.Y.; and Martin-Vega, L.A. (1994). "A review of production planning and scheduling models in the semiconductor industry. Part II: shop-floor control." *IIE Trans.* (v26, n5), pp44-55.

Wein, L.M. (1988). "Scheduling semiconductor wafer fabrication." *IEEE Trans. on Semiconductor Mfg.* (v1, n3), pp202-215.

Authors' Biographies

Ching-Chin Chern is an associate professor in the School of Management at National Taiwan University. She holds a bachelor's degree in business administration from National Taiwan University and a master's degree in business administration from the University of Texas at Arlington. She was employed as a systems analyst at the National Hand Tool Corp. (Farmers Branch, TX) from 1987 to 1992. She received her PhD in 1995 from the University of Texas at Dallas and joined the faculty at National Taiwan University the same year. During her research period of a doctoral student, she also worked as a research consultant for Bell North Research Lab (BNR), primarily on evaluating telecommunication systems by applying simulation models. Her teaching and research interests are currently concentrated on applying quantitative methods to solve problems related to supply chain management, especially in the semiconductor industry.

Kwei-Long Huang is currently a master's degree student in industrial engineering at Pennsylvania State University. He received his MS and BS in information management from National Taiwan University in 1999 and 1997, respectively. He has been a systems analyst in the development of advanced planning and scheduling (APS) packages at AsiaTek and a senior engineer of ERP at Compal in Taiwan. Operations research, production planning, and supply chain management are his main interests.