Int J Adv Manuf Technol (2007) 33: 1263–1274
DOI 10.1007/s00170-006-0559-6

ORIGINAL ARTICLE

Ruey-Shan Guo · David M. Chiang · Fan-Yun Pai

A WIP-based exception-management model for integrated circuit back-end production processes

Received: 28 June 2005 / Accepted: 15 February 2006 / Published online: 6 May 2006

Abstract Meeting due dates for delivery of products is a key factor in achieving customer satisfaction in today's globally competitive semiconductor market. However, undesirable production variations are inevitable in this industry, especially for 'back-end' factories that are closer to customers. This makes it difficult for management to maintain (or improve) a factory's performance with respect to delivery on due dates. In practice, production managers ameliorate the adverse effects of manufacturing uncertainties by control of work in progress (WIP). The present study therefore proposes a WIP-exception management model to define, detect, and respond to WIP exceptions. First, a model for determining acceptable WIP deviation levels (AWDLs) is established to assist production managers in identifying WIP exceptions on monitored workstations. A correction mechanism is then proposed to adjust deviations in WIP levels ('WIP exceptions') in accordance with the projected AWDLs as soon as possible. A simulation model is constructed, and experiments are then conducted to evaluate the proposed model. The proposed model is confirmed as being able to set appropriate and effective WIP-exception conditions to trigger correction actions. The simulation experiment also demonstrates that the WIP-correction action can shorten the 'back-to-normal' duration, prolong the time between successive WIP exceptions, and improve average on-time delivery percentage. The study concludes that the proposed model does determine effective exception-triggering conditions, rectifies abnormal WIP levels promptly, and results in improved performance on due-date delivery for semiconductor 'back-end' factories.

Keywords WIP determination · WIP management technology · Integrated circuit manufacturing · On-time delivery

R.-S. Guo · D. M. Chiang · F.-Y. Pai (✉)
Graduate Institute of Business Administration,
National Taiwan University,
Taiwan, Republic of China
e-mail: d91741006@ntu.edu.tw
Tel.: +886-2-22215302

1 Introduction

Integrated circuit (IC) manufacturing is a complicated multistage process whereby silicon is transferred (in the form of thin, polished disks) into ICs. The process consists of four main stages: (1) wafer processing (or wafer fabrication, 'fab'); (2) wafer probing; (3) IC packaging; and (4) functional testing and burn-in [4]. The initial wafer fabrication is usually referred to as the 'front-end' operation, whereas the other three stages (wafer probing, IC packaging, and final testing) are referred to as the 'back-end' operation. Figure 1 shows a typical semiconductor back-end manufacturing flow.

Delivery on time of finished goods is essential for customer satisfaction, and this is a critical factor for business survival in today's highly competitive markets. Although wafer fabrication (the 'front-end' operation) is the most technologically complex and the most capital intensive of the four stages described above, the three stages of the 'back-end' operation are closer to the customer. Indeed, the overall on-time-delivery performance of the supply chain depends on the performance of the 'back-end' processes. Unfortunately, undesirable production variations—machine breakdown, material shortage, randomness of processing time, randomness of yield, reworking, and so on—inevitably occur, and these problems pose difficulties in maintaining high standards of performance in terms of due dates [21].

It is not easy to eliminate these production variations, and 'back-end' processes are more sensitive to these variations than are 'front-end fabs'. Because 'back-end' processes have shorter cycle times than 'front-end' processes, there is a smaller time buffer in which to react to variations during 'back-end' operations. These 'back-end' processes therefore require a mechanism whereby serious production variations (exceptions) can be promptly detected and corrected, thus maintaining the efficiency of the supply chain and enhancing customer satisfaction.

Factories that perform 'back-end' operations have distinctive production characteristics [11, 12, 15, 24], as summarized in Table 1. In managing these various

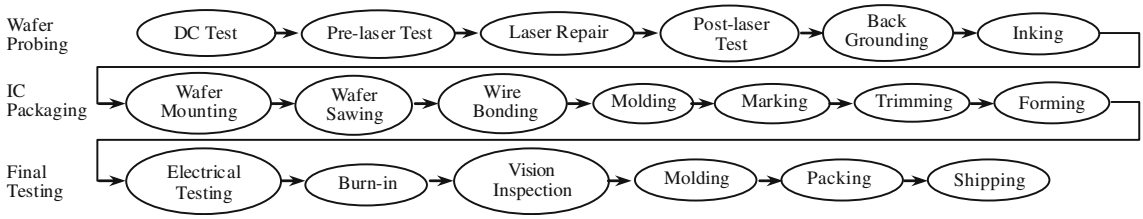

Fig. 1 Simplified semi-conductor 'back-end' flow

processes adequately, it is necessary to construct a work-in-progress (WIP) control-and-tracking mechanism to maintain the targeted throughput rate, to cushion the effect of variations in production activities, and to ensure the ultimate delivery of the ordered products to customers on time.

In practice, production managers use the WIP profile of each operational stage to reduce the effects of manufacturing uncertainties and to ensure targeted throughputs and cycle times. The WIP profile thus functions as a buffering mechanism between successive operational steps to prevent a workstation lying idle, ameliorate system disturbances, and optimize the capacity of each workstation [3, 9].

However, if the level of WIP is more than adequate, other problems can occur. A greater level of WIP prolongs cycle times, and this can diminish on-time delivery performance. In addition, too much WIP occupies space and requires additional resources for material handling and control [1, 19, 22, 26]. For these reasons, maintaining an appropriate level of WIP is critical for 'back-end' factories if they are to optimize performance while avoiding the allocation of extra resources to maintain an excessive level of WIP.

Despite the importance of these issues, few methods of shop-floor control have been proposed to establish the appropriate level of WIP in 'back-end' IC production processes. Most of the methods that have been suggested focus on the determination of levels of WIP in 'front-end' production environments [7, 8, 14, 18–20, 25]. For example, Miller [16] used simulation to determine the number of lots in a 'fab' production line under fixed WIP input policy. He pointed out that such a simulation model was applicable to a specific system, but that it was time-consuming to run a general simulation model. However, given the increasing speed and power of modern-day computers, running a general simulation model is now less time-consuming than in the past, and it is now possible to construct a simulation model to observe the behavior of real-world production systems. The main advantage of simulation is that it allows analysts to address a variety of complex issues that would otherwise defy analysis. For example, Wein [25, 26] used a simulation model to determine the total level of WIP in a production process involving four kinds of input mechanisms with different dispatching rules.

Table 1 Characteristics of various 'back-end' processes

Process	Characteristics
Wafer probe and functional test	Dynamic nature of job arrival Considerable types of products Sequence-dependent setup times Re-entrant product flows Short production cycle time Machine with different characteristics Binning phenomenon
IC packaging	Dynamic nature of job arrival Numerous types of products Different sizes of lots Machine with different setup times

Linear programming has been proposed as a method for projecting levels of WIP [10]. However, although it appears to be an appropriate way to project WIP levels at each step, it is difficult to ascertain reasonable holding costs and shortage costs using this method. This creates serious difficulties. Without knowing the correct holding costs and shortage costs, it is impossible to derive suitable levels of WIP using this method. Thus, although linear programming modeling is a popular method of handling the optimization problem, the complexity of 'back-end' production systems makes it difficult to deal with the uncertainty of such factors as equipment failure rates, repair times, sequence-dependent processing times, setup times, and re-entrant product flows. In contrast, simulation can take into account the detailed interactions among these elements in the 'back-end' manufacturing environment.

Another method that has been suggested to determine the level of WIP involved the use of a queueing network model. This approach has been shown to be useful in analyzing the performance of complex systems. For example, Burman et al. [2] developed a queueing network model for the IC manufacturing industry. They found that the performance measures, WIP levels, and cycle times that were acquired from a queueing model deviated by 7%–20% from those generated by simulation; however, they claimed that the running time of the queueing model was one-tenth of that of simulation. Lin and Lee [14] also used a queueing network model to construct an algorithm to determine the total standard level of WIP such that a fixed-WIP release control policy could apply. These models have yielded useful results with small computation capacity and short run times.

Methods based on a queueing network model are efficient in quickly estimating a wider range of manufacturing parameters (such as WIP level) than is possible with

other methods [1, 2]. However, despite their relative advantages in some respects, queueing models rely on certain broad assumptions, and they must be used with caution. In particular, queueing models are not appropriate for obtaining measures of dynamic manufacturing systems, because most queueing models assume conditions of long runs and steady states. In contrast, simulation offers greater flexibility and is closer to real-world circumstances; for these reasons, simulation is the chosen method in the present research.

The present study therefore proposes a WIP-based exception-management mechanism to detect variations in WIP levels during the production processes, thus allowing remedial action to be taken before such variations cause serious problems in on-time delivery performance. The first step in the proposed mechanism is to establish a model to determine an acceptable WIP deviation level (AWDL). Such an AWDL helps production managers to determine whether any given level of WIP is sufficient to buffer production disturbances and optimize performance. The second step in the proposed mechanism is the development of a correction mechanism to make automatic adjustments to any deviations in WIP levels (that is, any 'WIP exception') such that the WIP level is returned to the projected AWDL as soon as possible.

The present authors are confident that production managers will be able to use the proposed mechanism to determine appropriate AWDLs and correct exceptional WIP levels-thus ensuring that targeted throughputs are achieved and that due dates for delivery are satisfied.

2 AWDL determination model

The technique of simulation modeling was first developed in semiconductor manufacturing [5, 14, 16, 25], and such modeling continues to be extensively used in this field today. The reasons for using simulation modeling in this context are: (a) the intractability of detailed analysis of the semiconductor manufacturing process; (b) the uncertainties that are inherent in the manufacturing process itself; and (c) the steady improvement in computer technology that has made the building of simulation models easier and has reduced the computational expense of the resulting models [23]. The back-end manufacturing process is a typical example of a dynamic manufacturing system, whose types of product are numerous and where product mix shifts quickly from time to time, because of the short product life cycle and quick change in demand. A dynamic manufacturing system, which involves time-varying product mix, time-varying waiting time and time-varying flow time, would never enter steady state.

The present study employed a commonly used simulation package, eM-Plant, to build a simulation environment from real 'back-end' IC manufacturing processes. We interviewed several production managers who have extensive practical experience of back-end processes management and are experts on semiconductor back-end production and layouts. We asked them to verify the environment settings of our initial simulation model and corrected improper settings by implementing their suggestions for conceptual validation. Certain assumptions have been made in building the simulated model. First, a 'make-to-order' production environment (with fluctuations in the timing of arrival of jobs) was assumed. Secondly, quality issues were not considered in the model; rather, it was assumed that the yield rate is fixed at earlier stages (wafer probe, IC package, and final test stages), because significant yield-rate fluctuation rarely occurs with mature products. Thirdly, it was assumed that secondary resources (such as probe cards and handlers) will always be available. Fourthly, the model assumed a fixed capacity for the factory; order quotations and commitments were assumed to be made at the capacity-allocation planning stage to ensure that the capacity of the factory is not exceeded. Finally, the transportation time of products in the manufacturing process was not considered, because this time is insignificant compared with the time for processing and the time waiting for processing.

2.1 Performance measures

'Back-end' factories in IC production are capital-intensive and they need to attain sufficient throughput to meet customer demand on time and thus recover the initial capital investment. The performance measures of the present model thus pertain to: (a) target throughput; and (b) on-time delivery of goods. The first of these measures, throughput, determines whether an adequate quantity of goods can be delivered to customers, while the second, cycle time, determines whether ordered products can be produced and delivered on the projected dates.

In terms of WIP, both lesser levels of WIP and greater levels of WIP have the potential to diminish the delivery performance. This is because a less-than-appropriate level of WIP diminishes throughput, whereas a greater-than-appropriate level of WIP prolongs cycle time.

The present study therefore collected performance data on three measures-(1) mean cycle time; (2) average throughput rate; and (3) average on-time delivery percentage (AOTDP). These data were analyzed and compared under different levels of WIP. The performance measures are described in greater detail below.

2.2 Cycle time

Cycle time is defined as the elapsed time from the start of wafer probing until the end of final testing of the chip. It thus measures the time taken for a product to move through the manufacturing process.

In the following calculation, $WRT_{w,i}$ is the time at which wafer w of order i is released, $WCT_{w,i}$ is the time at which wafer w has completed all the required processes, $CRT_{c,w,i}$ is the time at which chip lot c (derived from wafer w of order i) is released, $CCT_{c,w,i}$ is the time at which chip lot c (derived from wafer w of order i) has completed all its

required processes, and PCT_i is the time at which all chip lots have completed all required processes in order i. Finally, n_i is the number of orders during the analyzed time slot. The mean cycle time is then given by the following equation:

$$\text{Mean cycle time} = MCT = \frac{\sum_{i=1}^{n_i} PCT_i}{n_i},$$

where

$$PCT_i = MAX\{CCT_{c,w,i} - CRT_{c,w,i} + WCT_{w,i} - WRT_{w,i}\} \forall i$$

2.3 Mean throughput rate

According to the theory of constraints, production system output is limited by the 'constraint machines' (otherwise known as the 'key machines' or the 'bottleneck machines') [6]. In the present model, the mean throughput rate is defined as the average number of chips passing through a 'bottleneck' ('constraint') workstation per calendar day. If PCL_k is the processing chip lots on the bottleneck workstation on day k ($k=1, \ldots, K$), then:

$$\text{Mean throughput rate} = MTT = \frac{\sum_{k=1}^{K} PCL_k}{K}$$

2.4 Average on-time delivery percentage

Because 'back-end' production is at the end of the supply chain, the arrival of lots at the 'back-end' stages depends on the performance of earlier stages-which introduces an element of unpredictability into the manufacturing process. AOTPD thus becomes a key performance measurement in such a complex production system. If DD_i is the due date of order i, then

$$\text{Average on-time delivery percentage } (AOTDP) = \frac{\sum_{i=1}^{n_i} l_{\{PCT_i < DD_i\}}}{n_i}$$

In this equation, $l_{\{PCTi<DDi\}}$ is an indicator function equal to 1 if the date of the completed order i is earlier than the due date. The symbol n_i represents the total number of orders.

2.5 AWDL determination

As previously noted, according to the theory of constraints, production system output is limited by the 'constraint machines' [6]. So-called 'starvation avoidance' is thus accomplished by maintaining a relatively high level of WIP at the constraint machine to ensure the availability of material in virtually any circumstance—including the occurrence of an extraordinary and unpredictable event [21].

In the 'back-end' processes of IC production, the final testing (FT) stages are considered to be the most critical stages. In the present model, the FT workstation is therefore defined as being the 'bottleneck workstation'. Also, for the most factories, die mounting and wire bonding stages are the critical stages, because there are enormous machines at these stages and those machines are under the production environment with dynamic orders, a huge number of complicated product types, and short cycle times (tight to due dates). Herein, we define the die mounting and wire bonding as the critical workstations to help production managers monitor these workstations to ensure performance. The WIP levels of the bottleneck and critical workstations are determined by the proposed mechanism.

The following steps were used to construct the proposed AWDL determination model.

Step 1.

Simulation model construction

The study modeled 533 machines in 18 'station families' with a total of 46 products in a factory located in the Hsinchu Science Park in Taiwan. Process time, yield information, and distribution arrangements were defined. Machine unavailability was also defined. All machines had respective fixed preventive maintenance (PM) schedules. PM held a higher priority than job processing. Breakdown of machines was modeled in terms of mean time between failures (MTBF), mean time to repair (MTTR), mean time between PM (MTBPM), and mean time to finish a PM (MTTPM) using appropriate distribution arrangements. Jobs were randomly fed into the wafer-probing factory as a result of the fluctuating nature of job arrival in the real-world environment. Fixed WIP was used as the releasing policy.

Step 2.

Due date assignment

To assign wafers and chip lots to the corresponding due dates, the study collected the mean cycle time, $\sigma_{c,\,w,\,i}$, and the standard deviation of cycle time, $\sigma_{c,\,w,\,i}$, for every wafer and chip group. A deterministic releasing policy, together with a 'first in–first out' (FIFO) dispatching rule were used in establishing the goal of target throughput. It was assumed that there were n_w wafer groups and n_c chip groups in the model. Every group had its own cycle time. FIFO yields longer cycle time [23, 25] and poorer delivery performance [23]. The estimated cycle time was set to ensure that AOTDP was always greater than 85%. Thus:

$$\text{Estimated cycle time of order } i = \overline{X}_{c,w,i} \times 1.05\sigma_{c,w,i}$$

$$\text{Estimated due date of order } i = DDi = CST_{c,w,i} + MAX\{\overline{X}_{c,w,i} \times 1.05\sigma_{c,w,i}\}$$

The estimated cycle time was set as the mean cycle time plus 1.05 standard deviations of the cycle time to maintain the AOTDP at greater than 85%.

Step 3.

Experiment design

Simulation experiments were designed and defined to find the optimal levels of WIP. The corresponding due dates of wafers were assigned when they were fed into the simulation model, as were those of chip lots when they were cut from their parent wafers. The total WIP level under the goal of target throughput was then collected. The total WIP in this scenario was the WIP level of the entire back-end factory. The mean and the standard deviation of the total WIP levels were then calculated.

Step 4.

AWDL determination for back-end factory

To test the performance measures, a fixed-WIP releasing policy was used. The performance measures under the total WIP level of each combination were then collected. ANOVA and multiple comparisons were used to analyze the simulation data. Using these results, the maximum and minimum total WIP levels of the simulation model were then determined. Once the total WIP level was maintained within the boundaries (between the maximum and minimum total WIP levels), mean throughput rate and AOTDP can be established.

Step 5.

AWDL determination for back-end factory

The study then collected the WIP levels of monitored workstations under the maximum total WIP level ('upper-limit AWDL') and the minimum total WIP level ('lower-limit AWDL'). The upper and lower AWDLs were then used in WIP correction actions. To facilitate a full understanding of how to use AWDLs under exception conditions, an example is shown in Fig. 2.

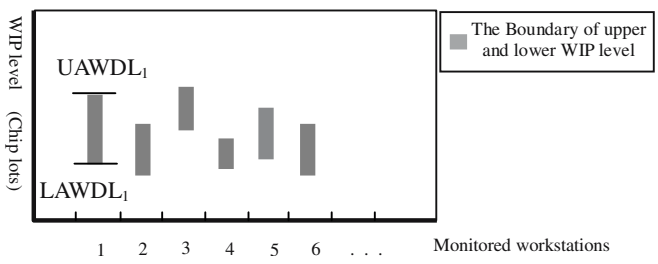

Fig. 2 An example of the application of AWDLs at monitored workstations

3 WIP correction action

After obtaining the threshold AWDL to trigger an exception, the next issue is how this can be used to rectify abnormal WIP levels. The AWDLs were set as standards to identify exceptional conditions of WIP levels. When an actual WIP level was greater or less than the corresponding upper or lower bound of AWDL at a monitored workstation, a correction action was immediately triggered to resolve this imbalance issue.

A proper and effective correction action is vital to maintain WIP levels and avoid uncertainties. Another dispatching rule to control WIP levels was triggered when an imbalance of WIP levels among workstations occurred. A minimum inventory variable scheduling (MIVS) algorithm developed by Li et al. [13] was employed as the remedial scheduling policy in the production line. The MIVS focused on the reduction of the mean and variance of cycle times while maintaining an acceptable throughput rate. Because of the trade-off between cycle time and WIP level, the MIVS was dedicated to balance WIP—that is, it minimized the difference between the WIP at any given instant and the standard WIP. Jobs were dispatched by the deviation between the actual WIP level and the standard WIP level to introduce maximum correlation between the two WIP levels. Furthermore, the MIVS not only considered the current stage, but also looked ahead to the WIP level of the next stage. The MIVS was thus designed for minimizing WIP imbalance. Although it was originally proposed to solve problems in wafer fabrication, the MIVS was selected for the present purpose after consideration of its feasibility in 'back-end' IC manufacturing environments.

Li et al. [13] took 'stage' to refer to the control phase, and assigned standard WIP levels to each stage from the historical data. The present study posits the 'workstation' as the control phase-because the main concern of the present study is with WIP control at the workstation. MIVS was invoked to dispatch queuing jobs only if the actual WIP level at a monitored workstation was beyond its AWDL boundary. MIVS then assigned job priorities according to the AWDL of each product at every monitored workstation. The upper and lower AWDL of each product were derived from the corresponding upper and lower workstation AWDLs and the product's fixed mix ratio. Priority was determined by comparing every product's actual WIP level with its AWDL at the current workstation and at the next monitored workstation. Moreover, a tie-break rule was added to the MIVS in case two jobs had the same priority. As a result, there were two phases in the dispatching rules.

Phase I:

Initial search for dispatching priority

When a workstation was idle and WIP exception occurred, the items in the queue were chosen according to their 'emergency levels' across the current workstation and that immediately following. It was assumed that there were L product types, including wafer and chip products, and M workstations (machine groups) in the 'back-end' manufacturing environment.

The MIVS ranked the dispatching sequence by nine priorities. These priorities are discussed below. Before doing so, the following notations require definition:

$UAWDL_m$ the upper AWDL in workstation m determined by AWDL determination model

$LAWDL_m$ the lower AWDL in workstation m determined by AWDL determination model

$UAWDL_{l,m}$ the WIP level of product l in workstation m and $UAWDL_{l,m}=UAWDL_m*r_l$

$LAWDL_{l,m}$ the WIP level of product l in workstation m and $LAWDL_{l,m}=LAWDL_m*r_l$

$AAWDL_{l,m}$ the actual WIP level of product l in workstation m

r_l the mix ratio of product l

The items in the queue were chosen according to the following priorities:

Priority 1:
Product l such that $AAWDL_{l,m}>UAWDL_{l,m}$ and $AAWDL_{l,m+1}<LAWDL_{l,m+1}$

Priority 2:
Product l such that $AAWDL_{l,m}>UAWDL_{l,m}$ and $LAWDL_{l,m+1}<AAWDL_{l,m+1}<UAWDL_{l,m+1}$

Priority 3:
Product l such that $AAWDL_{l,m}>UAWDL_{l,m}$ and $AAWDL_{l,m+1}<UAWDL_{l,m+1}$

Priority 4:
Product l such that $LAWDL_{l,m}<AAWDL_{l,m}<_{UAWDL_{l,m}}$ and $AAWDL_{l,m+1}<LAWDL_{l,m+1}$

Priority 5:
Product l such that $LAWDL_{l,m}<AAWDL_{l,m}<_{UAWDL_{l,m}}$ and $LAWDL_{l,m+1}<AAWDL_{l,m+1}<UAWDL_{l,m+1}$

Priority 6:
Product l such that $LAWDL_{l,m}<AAWDL_{l,m}<_{UAWDL_{l,m}}$ and $UAWDL_{l,m+1}<AAWDL_{l,m+1}$

Priority 7:
Product l such that $LAWDL_{l,m}<AAWDL_{l,m}<_{UAWDL_{l,m}}$ and $UAWDL_{l,m+1}<AAWDL_{l,m+1}$

Priority 8:
Product l such that $LAWDL_{l,m}<AAWDL_{l,m}<_{UAWDL_{l,m}}$ and $LAWDL_{l,m+1}<AAWDL_{l,m+1}<UAWDL_{l,m+1}$

Priority 9:
Product l such that $LAWDL_{l,m}<AAWDL_{l,m}<_{UAWDL_{l,m}}$ and $LAWDL_{l,m+1}<AAWDL_{l,m+1}$

Jobs ranked as priority 1 obviously had the highest priority to be processed, with those ranked as priority 9 having the lowest priority.

Intuitively, an item leaving the current workstation increases the WIP level of the workstation that follows immediately after, and decreases the WIP of the present workstation. To keep the actual WIP as close as possible to the standard WIP, any product with a greater-than-upper-boundary inventory should be given a higher priority to be selected when its WIP level at the workstation immediately following is less than the minimum requirement. The WIP status of different priorities is demonstrated in Fig. 3.

Phase II:

Searching for tie-break priority

When two jobs had the same priority, the phase II procedure was triggered. This procedure enabled a

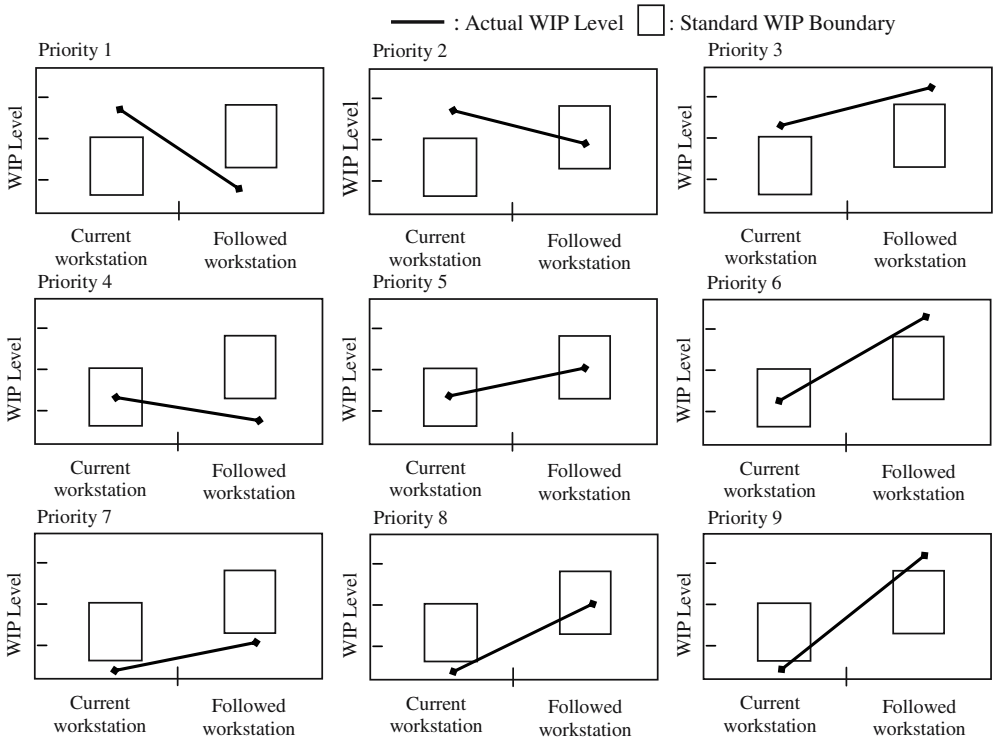

Fig. 3 WIP status of MIVS's nine dispatching priorities

decision to be made on each job's dispatching sequence under 'tie-breaking' conditions. Li et al. [23] suggested using simple static rules and Wiendahl [27] reported that, without disturbing the dispatching sequence from previous workstations, FIFO has the least variability. The present study therefore designated FIFO to dispatch jobs when two jobs had the same priority.

On the basis of the above discussion, the scheme of the correction model is presented in Fig. 4.

4 Experimentation and model performance evaluation

4.1 Experimental outline

Simulation experiments were defined to analyze the performance of a typical 'back-end' factory, located in the Hsinchu Science Park in Taiwan, under different levels of WIP. To find the appropriate AWDLs of the monitored workstations, the study first determined the appropriate total level of WIP. Initially, a deterministic lot-releasing policy was used to obtain the mean and standard deviation of the level of WIP while attaining target throughput. Seven cases were then tested under a fixed-WIP releasing policy, varying from the mean WIP level minus 3 standard deviations to the mean WIP level plus 3 standard deviations, in increments of 1 standard deviation.

There were thus seven scenarios of simulation runs. The WIP level of each scenario was posited as $\mu+3\sigma$, $\mu+2\sigma$, $\mu+\sigma$, μ, $\mu-\sigma$, $\mu-2\sigma$, and $\mu-3\sigma$ respectively. Each scenario had to run 30 replications. To reach a steady state and minimize the potential for startup bias, startup statistics were discarded for the first 90 days (taken as the 'warm-up period'). Mean cycle times, AOTPD, and mean throughput rate were then collected (after the 'warm-up period').

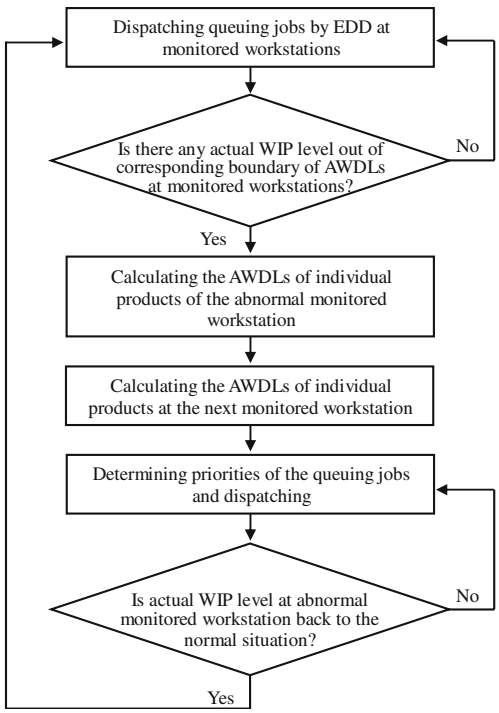

Fig. 4 Scheme of correction action

At the beginning, the utilization of bottleneck workstation, FT testing, was tracked under the minimum fixed-WIP level, $\mu-3\sigma$, to ensure that the simulation model was at a steady state. The utilization of the bottleneck workstation under a WIP level of $\mu-3\sigma$ reached 88.93% (after the 'warm-up period'). As long as reasonable and steady utilization under the minimum WIP level was maintained, it was reasonable to infer that production would remain stable under other (greater) levels of WIP.

4.2 AWDL determination by simulation results

The results of ANOVA indicated that significant differences arose at different levels of WIP with respect to mean cycle time, mean throughput, and AOTDP at $\alpha=0.05$. A Duncan's multiple test [17] was then conducted on the levels of WIP. The results of simulation runs and Duncan's multiple tests are summarized in Table 2.

The results of simulation runs and Duncan's multiple tests on mean cycle time, mean throughput rate, and AOTDP suggested that the monitored workstations should maintain their AWDLs to maintain high performance with respect to due dates. The AWDLs are shown in Table 2.

According to Table 2, when the level of WIP was at $\mu-\sigma$, the corresponding mean cycle time was at its shortest. There was no significant difference in mean cycle time between μ and $\mu-\sigma$, nor between μ and $\mu-2\sigma$, because μ and $\mu-\sigma$ were in the same Duncan group, as were μ and $\mu-2\sigma$. When the level of WIP exceeded μ, the mean cycle time increased at a faster rate; that is, a greater level of WIP led to a longer cycle time. This conforms to Little's law [7]. However, the mean cycle time began to increase at a WIP level of less than $\mu-2\sigma$. A lesser level of WIP did not guarantee a shorter cycle time because products required more time to accumulate a sufficient quantity to pass batch-processing operations (such as the burn-in stage in the final testing station), thus producing a longer cycle time.

With respect to mean throughput rate, there was no significant difference in the rate at WIP levels of $\mu+3\sigma$, $\mu+2\sigma$, and $\mu+\sigma$. There was also no significant difference at WIP levels of $\mu+\sigma$, μ, and $\mu-\sigma$. However, when the WIP levels were less than $\mu-\sigma$, the mean throughput rate was significantly less than the mean throughput rates at WIP levels of $\mu-\sigma$ and greater. Once the level of WIP decreased further, the mean throughput rates decreased more quickly at a WIP level of $\mu-\sigma$ or less.

It can also be seen from Table 2 that the AOTDP was significantly greater under WIP levels of μ and $\mu-\sigma$. This implies that these WIP levels resulted in a better performance in terms of order fulfilment than did WIP levels greater than μ. WIP levels less than $\mu-\sigma$ could not compensate for the loss of throughput rate by decreasing cycle times. The AOTDP differed when WIP levels were less than $\mu-2\sigma$.

Table 2 Duncan's multiple test results for mean cycle time, mean throughput rate and AOTDP

Total WIP Level (chip lots)	μ+3σ (4255 lots)	μ+2σ (4074 lots)	μ+σ (3893 lots)	μ (3711 lots)	μ-σ (3530 lots)	μ-2σ (3349 lots)	μ-3σ (3167 lots)
Mean cycle time (h)	313.5	304.1	290.7	281.2	278.8	283.1	293.9
Duncan grouping	A	B	C	D	E	D	C
Mean throughput rate (chip lots/day)	353.535	351.015	348.285	345.66	344.295	302.61	265.23
Duncan grouping	A	A	A B	B	B	C	D
AOTDP (AOTDP, %)	73.3%	82.4%	90.4%	95.9%	96.8%	94.7%	91.4%
Duncan grouping	E	D	C	A	A	B	C

μ=3711.2 and σ=181.3 are derived form the simulation runs under deterministic releasing policy

Considering all three performance measurements of the simulation runs, the study found that WIP levels of μ+3σ, μ+2σ, and μ+σ were associated with significantly decreased AOTDPs than those associated with other WIP levels. Moreover, the mean cycle times were significantly longer at these WIP levels (μ+3σ, μ+2σ, and μ+σ) than with other WIP levels. Because 'back-end' factories usually focus on customer satisfaction, rather than throughput rates, production managers would probably not be satisfied with these performances, even though greater mean throughput rates were achieved. However, when the WIP levels were less than μ-2σ., the mean throughput rate decreased much more rapidly, even though there was no significant difference in mean cycle time between μ-2σ and μ. Moreover, the AOTDPs were significantly less than those for WIP levels of μ or μ-σ.

In summary, in view of the main goals of 'back-end' semiconductor production with respect to mean cycle time and AOTDP, the level of WIP should be kept between μ (namely 3711) chip lots and μ-σ (namely 3530) chip lots; this will ensure greater on-time delivery percentage. Although a WIP level between μ and μ-σ cannot guarantee an optimal result, it is a safe and reliable amount to achieve a greater average on-time delivery performance with an acceptably short mean cycle time.

Focusing on managing these desirable WIP levels at the monitored workstations to improve the factory's performance, the study collected WIP profiles under different WIP levels as AWDLs for each monitored workstation. The WIP levels at μ and μ-σ were determined to be the upper and the lower ideal AWDLs for the monitored workstations (see Table 3).

4.3 Evaluation of WIP correction model

The AWDL boundaries were then set to trigger a WIP exception. If the actual WIP levels move beyond the pre-set AWDL boundaries, the MIVS is initiated to compensate for the potential effects of production variances, thus ensuring better performance in terms of due dates. When the actual WIP level returns within the AWDL boundaries, earliest job due date (EDD) is employed again.

The study first compared AOTDP and percentage improvement of the proposed WIP correction action with those of nine other popular dispatching rules to examine whether the WIP correction action (a combination of MIVS and EDD) yields better performance than other commonly used rules. Detailed descriptions of these rules and the comparison-simulation experiment are provided in Table 4.

Table 3 WIP profiles for monitored workstations

Monitored workstation	WIP profiles (chip lots)						
	μ+3σ	μ+2σ	μ+σ	μ(3711) Upper WDL	μ-σ (3530) Lower AWDL	μ-2σ	μ-3σ
Die mounting	396	379	363	346	329	309	293
Wire bonding	489	464	440	413	389	366	339
Final testing	717	688	654	629	598	552	523

Table 4 Duncan's multiple test results for AOTDP under different dispatching rules

Dispatching rule	Dispatching rule description	AOTDP	Improvement in AOTDP (%)	Duncan grouping
MIVS+EDD	WIP correction action	96.52%	16.29%	A
EDD	Earliest job due date	90.95%	9.68%	B
MIVS	Minimum inventory variability schedule	89.64%	8.10%	C
CYC	Cyclic priority	88.09%	6.22%	D
LNQ	Largest number in queue	88.03%	6.15%	D
SPT	Shortest processing time	87.02%	4.93%	D
SSPT	Shortest remaining time	87.66%	5.70%	D
RAN	Random priority	86.33%	4.09%	E
FIFO	First in first out	82.93%	-	G
SNQ	Smallest number in queue	80.42%	−3.03%	F

To evaluate the effect of the WIP correction action, three performance measures were used-(1) AOTDP; (2) average duration of an exception ('back-to-normal'); and (3) average time between successive exceptions. Table 5 shows the simulation results.

Table 4 shows that the proposed WIP correction action, EDD, and MIVS offered significantly better performance than the other rules, and that the combination of MIVS and EDD offered better performance on due dates than EDD or MIVS alone. Because EDD minimizes the maximum lateness and MIVS regulates the discrepancies between the WIP levels at any given instant and the AWDLs, it is understandable that pure EDD achieved better due-date performance than MIVS alone. For the combination of MIVS and EDD, the synergy of WIP control and lateness prevention contributed to improved AOTDP, and thus outperformed the others. For these reasons, a 16.29% improvement was achieved by changing the dispatching rules from FIFO to MIVS+EDD.

'Back-to-normal' duration and the time between successive exceptions were noted with WIP correction actions and without WIP correction actions. From Table 5, it is apparent that significant differences existed in both 'back-to-normal' duration and time between successive exceptions for all monitored workstations. Implementing WIP correction action did shorten the 'back-to-normal' duration and prolong the time between WIP exceptions.

It can be concluded that WIP correction action not only reduced the effects of production variance, but also made the production line more stable. It can thus provide improved on-time delivery performance.

4.4 Evaluation of exception detection

The WIP correction action can also be expected to reduce the occurrence of WIP exceptions. The study therefore compared WIP exception events under the proposed WIP correction action (MIVS+EDD) with those under the other two dispatching rules that occupied the highest places in Table 4: (1) pure EDD and (2) pure MIVS.

The number of WIP exceptions for each simulation run were collected under each of these dispatching rules. After averaging and sorting, the WIP exception detection results are shown in Table 6. 'False alarms' were also recorded to study the robustness of the WIP-based exception-manage-

Table 5 Test Results for the effect of WIP correction action

	With WIP correction action	Without WIP correction action	*P* value
Monitored workstation 1: FT testing			
Replications	30	30	
Average WIP exceptions	90.5	126.5	
Duration of back to normal (h)	5.17	6.83	0.030*
Duration between successive WIP exceptions (h)	24.89	13.22	0.005*
Monitored workstation 2: Die mounting			
Replications	30	30	
WIP exceptions/ per replication	77.6	99.3	
Duration of back to normal (h)	4.07	5.94	0.025*
Duration between successive WIP exceptions (h)	35.25	23.33	0.009*
Monitored workstation 3: Wire bonding			
Replications	30	30	
WIP exceptions/ per replication	84.5	122.3	
Duration of back to normal (h)	4.42	6.12	0.021*
Duration between successive WIP exceptions (h)	30.77	21.66	0.009*

*Significant ($p<0.05$)

Table 6 Results of WIP exception detection

Dispatching rule	Replications	No. of WIP exceptions	Duncan grouping	No. of false alarms	F/W*100%	Duncan grouping
MIVS+EDD	30	317.8	A	32.28	10.19%	C
EDD	30	259.3	B	43.74	16.87%	A
MIVS	30	150.6	C	18.52	12.30%	B

ment model. A 'false alarm' was defined as WIP exceptions that did not cause AOTDP to decrease by more than 1%; that is, a false alarm was a WIP exception that, if detected but not corrected, caused an insignificant decrease in due-date performance.

As can be seen in Table 6, pure MIVS led to the fewest exceptions, because it adjusts WIP levels to expected quantities. EDD led to more exceptions, because it chooses lots according to due dates, rather than discrepancy from AWDLs. A combination of MIVS and EDD (the WIP correction action) had most exceptions, but it offered much better due-date performance than MIVS alone.

Table 6 also demonstrates significant differences in false-alarm frequencies among the three dispatching rules. The WIP correction action (MIVS+EDD) caused significantly fewer false alarms than the other two. It can therefore be concluded that it not only detected more exceptions (see above), but also signalled 'real' exceptions that had the potential to diminish production performance.

Thus, taking into account AOTDP, WIP exception occurrence, and false alarm frequency, it can be concluded that the combination of MIVS and EDD outperformed the other dispatching rules in this experimental case.

4.5 AWDL determination for different workstation types

Not all WIP exceptions have the same effect on production activity. An imbalance in WIP levels occurring at a bottleneck or critical workstation has a more significant adverse influence on time-delivery performance than a similar imbalnace at a general workstation. It was therefore considered useful to attempt to establish appropriate AWDLs (triggering WIP exceptions) at various types of workstation—bottleneck, critical, and general—according to the differential effects on performance with respect to due dates. This would allow due-date performance to be improved and would free production managers to focus on WIP exceptions that really hurt a factory's performance in terms of production delivery. An experiment was therefore conducted to determine appropriate AWDL boundaries for various types of workstations.

First, WIP control-and-correction was extended to include general workstations, rather than restricting it to bottleneck workstations and critical workstations. Secondly, different exception-triggering conditions were tested for various types of workstations to determine the most appropriate and effective conditions to attain the factory's objectives.

The WIP profiles under WIP levels of μ and μ-2σ were chosen as potential upper and lower AWDLs respectively. These can be referred to as 'loose' AWDL boundaries, in contrast to the AWDL boundaries used in earlier simulation runs, which can be referred to as 'tight' AWDL boundaries. Table 7 summarizes the 'tight' and 'loose' WIP boundaries in this experiment.

WIP correction action was triggered with various combinations of 'tight' and 'loose' AWDL boundaries, and the on-time delivery percentages under each combination were then compared. Table 8 shows the simulation results.

As can be seen in Table 8, there was no significant difference between combination 1 and combination 2, or between combination 2 and combination 3. On-time delivery will therefore not be affected by choosing combination 1 over combination 2, or by choosing combination 2 over combination 3. The finding implies that identifying WIP exceptions by 'loose' WIP boundary will not lead to a significant decrease in performance with

Table 7 Tight and loose AWDL boundaries for each workstation

Workstation	CP Test 1	Laser repair	CP test 2	CP baking	Inking	Wire bonding
Tight AWDL boundary (chip lots)	228–236	96–104	202–209	153–163	95–99	387–399
Loose AWDL boundary (chip lots)	219–236	93–104	196–209	148–163	87–99	364–399
Workstation	Molding	Marking	Trimming	Forming	AS inspection	Packaging
Tight AWDL boundary (chip lots)	126–130	67–69	88–92	73–78	20–24	35–41
Loose AWDL boundary (chip lots)	123–130	61–69	85–92	69–78	17–24	30–41
Workstation	CP inspection	Tapping	Lapping	Die sawing	Die mounting	Wire bonding
Tight AWDL boundary (chip lots)	80–86	40–46	34–38	190–197	318–334	387–399
Loose AWDL boundary (chip lots)	75–86	34–46	29–38	187–197	302–334	364–399
Workstation	Final testing	Cycling	Burn-in	Laser marking	Scanning	Packaging
Tight AWDL boundary (chip lots)	581–609	277–285	205–210	174–179	99–103	35–41
Loose AWDL boundary (chip lots)	553–609	271–285	197–210	168–179	93–103	30–41

Table 8 Duncan's multiple test results for AOTDP under different AWDL boundaries

No.	Bottleneck workstation AWDL boundary	Critical workstation AWDL boundary	General workstations AWDL boundary	No. of detected WIP exceptions	AOTDP (%)	Duncan grouping
1	Tight	Tight	Tight	301.6	96.73	A
2	Tight	Tight	Loose	269.3	96.54	A B
3	Tight	Tight	-	228.7	96.26	B
4	Tight	Loose	-	199.8	95.23	C
5	Loose	Loose	-	176.5	93.82	D

respect to due dates, although it will increase WIP exceptions. The exceptions that are detected (but which will not have any significant effect on factory performance) are referred to as 'fake' exceptions. It is desirable to reduce detection of these 'fake' exceptions. Therefore, to obtain improved on-time delivery percentage and to avoid issuing 'fake' WIP exceptions, 'loose' boundaries are recommended as AWDL boundaries at general workstations. Production managers can then simply receive the notification of 'real' exceptions, thus avoiding a waste of time and resources in tracking 'fake' exceptions.

According to Table 8, there were significant differences among combinations 3, 4, and 5. AOTDP decreased when AWDL boundaries were enlarged at bottleneck and critical workstations. This suggests that replacing 'loose' AWDL boundaries with 'tight' ones will significantly improve on-time delivery percentage. It is therefore recommended that 'tight' AWDL boundaries be used at bottleneck and critical workstations to ensure that production managers are notified of abnormal WIP levels earlier, thus facilitating management of WIP exceptions.

From the results of this experiment, it is apparent that the 'tighter' the AWDL boundaries that are used, the greater the number of workstations at which corrective action can be taken, and the greater the AOTDP achieved. This supports the recommendation that AWDL boundaries be employed at all monitored workstations. However, the results also indicate that it is better to correct WIP exceptions at both general workstations and bottle-neck/critical workstations. Although this finding is not in accordance with the initial presumptions of the present study, it does confirm that WIP control at monitored workstations is the most important contributor to optimal due-date performance.

In summary, it can be concluded that monitoring WIP levels and taking WIP correction action at all workstations achieved better performance on due dates in this experiment, and that the findings clarify and support the benefits of the proposed WIP correction action.

5 Conclusion

In semiconductor manufacturing, 'back-end' manufacturing plants are close to customers and must cope with fluctuating arrivals from 'front-end' manufacturing sites. 'Back-end' processes are thus more sensitive to any unpredictable production variations. To improve customer satisfaction and avert delays, semiconductor 'back-end' factories therefore need an effective exception-management model that will enable them to detect, and cope with, unpredictable production variances. The WIP-based exception-management model that is proposed here will assist production managers in their efforts to manage WIP levels in an approporiate fashion.

The experimental study demonstrates that the AWDLs determined by the AWDL determination model are appropriate for detecting WIP exceptions and triggering WIP correction actions. The study shows that performance on due dates can be significantly improved when WIP-correction actions are triggered by such AWDLs. Production managers can also modify AWDL boundaries for 'bottleneck', 'critical', and 'general' workstations to improve due-date performance.

In addition, the experimental results show that WIP-correction action not only shortens back-to-normal duration, but also prolongs the time between successive WIP exceptions. Finally, the proposed model provides robustness in that it leads to fewer 'false alarms' than do other dispatching rules.

It is therefore concluded that the proposed model determines effective exception-triggering conditions, rectifies abnormal WIP levels promptly, and results in improved performance in terms of due-date delivery.

References

1. Askin RG, Krisht AH (1994) Optimal operation of manufacturing systems with controlled work-in-process levels. Int J Prod Econ 37(2):1637–1653
2. Burman DY, Gurrola-Gal FJ, Nozari A, Sathaye S, Sitarik JP (1986) Performance analysis techniques for IC manufacturing lines. ATT Tech J 65(4):46–57
3. Buzacott JA (1971) The role of banks in flow-line production systems. Int J Prod Res 9(4):425–436
4. Chen H, Harrison JM, Mandelbaum A, Wein LM (1998) Empirical evaluation of a queueing network model for semiconductor wafer fabrication. Oper Res 36(2) March-April
5. Chen HC, Lee CE (2003) Pull systems for control and dummy wafers. Int J Adv Manuf Technol 22:805–818
6. Goldratt EM (1990) Theory of constraints and how should it be implemented? North River Press, New York
7. Gross D, Harris CM (1974) Fundamentals of queueing theory. Wiley, New York, USA
8. Hopp WJ, Roof ML (1998) Setting WIP levels with statistical throughput control (STC) in CONWIP production lines. Int J Prod Res 36(4):867–882

9. Huang CL, Huang YH, Chang TY, Chang SH, Chung CH, Huang DT, Li RJ (1999) The construction of production performance prediction system for semiconductor manufacturing with artificial neural networks. Int J Prod Res 37 (6):1387–1402
10. Kim JS, Leachman RC (1994) Decomposition method application to a large scale linear programming WIP projection model. Eur J Oper Res 74:152–160
11. Lee YH, Lee BK, Jeong B (2000) Multi-objective production scheduling of probe process in semiconductor manufacturing. Prod Plan Control 11(7):660–669
12. Lee CY, Martin-Vega LA, Uszoy R, Hinchman J (1993) Implementation of a decision supporting for scheduling semiconductor test operations. J Electron Manuf 3:121–131
13. Li S, Tang T, Collins DW (1996) Minimum inventory dispatching in semiconductor wafer fabrication. IEEE Trans Semicond Manuf 9(1):145–149
14. Lin YH, Lee CE (2001) A total standard WIP estimation method for wafer fabrication. Eur J Oper Res 131:78–94
15. Manzione LT (1990) Plastic packaging of microelectronic devices. Technical Report, AT&T Bell Laboratories
16. Miller DJ (1990) Simulation of a semiconductor manufacturing line. Commun ACM 33(10):99–108
17. Montgomery DC (2001) Design and analysis of experiments, 5th edn. Wiley, New York
18. Robinson JK, Fowler J, Bard J (1995) The use of upstream and downstream information in scheduling semiconductor batch operations. Int J Prod Res 7:1849–1870
19. Rose O (1998) WIP evaluation of a semiconductor factory after bottleneck workcenter breakdown. Proceedings of the 1998 Winter Simulation Conference, IEEE, Piscataway, NJ, 997–1003
20. Sivakumar AI, Chong CS (2001) A simulation based analysis of cycle time and throughput in semiconductor backend manufacturing. Comput Ind 45:59–78
21. Tu et al (2005) Model to determine the backup capacity of a wafer foundry. Int J Prod Res 43(2):339–359
22. Tu YM, Li RL (1998) Constraint time buffer determination model. Int J Prod Res 36(4):1091–1130
23. Uszoy R, Church IA, Ovacik I, Hinchman J (1993) Performance evaluation of dispatching rules for semiconductor testing operations. J Electron Manuf 3:95–105
24. Uzsoy R, Lee CY, Martin-Vega LA (1992) A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning. IIE Trans 24(4):47–60
25. Wein LM (1988) Scheduling semiconductor wafer fabrication. IEEE Trans Semicond Manuf 1(3):115–130
26. Wein LM (1992) On the relationship between yield and cycle time in semiconductor wafer fabrication. IEEE Trans Semicond Manuf 5(2):156–158
27. Wiendahl H (1995) Load-oriented manufacturing control. Springer, Berlin Heidelberg New York