This article was downloaded by:[National Taiwan University]
On: 1 August 2008
Access Details: [subscription number 788856085]
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

International Journal of Production Research

Publication details, including instructions for authors and subscription information:
<http://www.informaworld.com/smpp/title~content=t713696255>

Improved customer satisfaction with a hybrid dispatching rule in semiconductor back-end factories

David M. Chiang ^a; Ruey-Shan Guo ^a; Fan-Yun Pai ^a

^a Graduate School of Business Administration, National Taiwan University, Taipei, Taiwan, R.O.C.

Online Publication Date: 01 September 2008

To cite this Article: Chiang, David M., Guo, Ruey-Shan and Pai, Fan-Yun (2008) 'Improved customer satisfaction with a hybrid dispatching rule in semiconductor back-end factories', International Journal of Production Research, 46:17, 4903 — 4923

To link to this article: DOI: 10.1080/00207540701324168

URL: <http://dx.doi.org/10.1080/00207540701324168>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Improved customer satisfaction with a hybrid dispatching rule in semiconductor back-end factories

DAVID M. CHIANG, RUEY-SHAN GUO and FAN-YUN PAI*

Graduate School of Business Administration, National Taiwan University,
Taipei, Taiwan, R.O.C.

(Revision received February 2007)

On-time delivery is a vital factor for customer satisfaction in the competitive semiconductor manufacturing industry, and to optimize on-time delivery manufacturers must continuously improve their management of work-in-progress (WIP). However, in undertaking to optimize WIP, managers are also concerned with short cycle times, high throughput, and high utilization. In an attempt to find the most satisfactory solution to these potentially conflicting requirements with regard to WIP, the present study employs fuzzy analytic hierarchy process (AHP) to determine an appropriate set of acceptable WIP deviation levels (AWDLs). These AWDLs are then used in a proposed hybrid dispatching rule to determine the operational priorities of jobs. A simulation model using real-world data is then constructed to examine the proposed mechanism for improving customer satisfaction. The findings of the study confirm that the proposed mechanism is capable of simultaneous consideration of various goals and the achievement of enhanced performances with respect to due-date delivery from semiconductor back-end processes.

Keywords: Semiconductor manufacturing; WIP balance; Dispatching; Fuzzy analytic hierarchy process

1. Introduction

The semiconductor business environment is characterized by rapid changes in business needs and increasing customer demands. In these circumstances, manufacturers must satisfy their customers if they are to stay ahead of the competition. A proven successful strategy to improve service responsiveness is enhanced management of work-in-progress (WIP) (Qiu 2005).

In undertaking such management of WIP, the goal that manufacturing enterprises are trying to achieve is the maintenance of desirable WIP subject to throughput requirements. An appropriate level of WIP functions as a buffer between successive operational steps to: (i) prevent 'starvation' at any given workstation; (ii) absorb system disturbance; and (iii) maximize the utilization of workstation capacity (Buzacott 1971, Huang *et al.* 1999). Efficient WIP ultimately brings many

*Corresponding author. Email: rubyasa@seed.net.tw

Figure 1. Simplified one-stop shopping turnkey service.

benefits to a firm—including a reduction in working capital requirements, a reduction in storage requirements (and lower associated costs), an improvement in product quality, an improvement in customer service, and the maintenance of flexibility (Wein 1992, Askin and Krisht 1994, Rose 1998, Tu and Li 1998, Wang and Prabhu 2006).

The overall semiconductor manufacturing process involves four main steps: (i) wafer processing or wafer fabrication (fab); (ii) wafer probing; (iii) integrated circuit (IC) packaging; and (iv) functional testing and burn-in (Chen *et al.* 1998, Tu *et al.* 2005). Wafer fabrication is generally referred to as the ‘front-end’ operation, whereas the remaining three stages are referred to as the ‘back end’ in the turnkey service. Figure 1 shows a simplified turnkey service.

As yield efficiencies in the front-end operations have increased, increasing attention has been directed to improving the ability of back-end factories to handle unexpected production variations. The objective of the management of back-end operations is improvement of customer service through better on-time delivery while maintaining low cycle times, high throughput, and high utilization. WIP levels must therefore be maintained at a level that leads to acceptable performance with respect to *all four* parameters (cycle time, throughput rate, due-date accuracy, and utilization), rather than at a level that performs well with respect to one objective and poorly with respect to others. The present paper addresses this important issue of improving customer satisfaction with respect to the whole semiconductor supply chain. A dispatching rule is proposed and acceptable WIP deviation levels (AWDLs) for the dispatching rule are determined by integrating the various goals of back-end processes.

The remainder of the paper is organized as follows. Following this introduction, the paper presents a literature review of decision-making models and their role in determining WIP levels. This is followed in section 3 by the presentation of a brief overview of back-end semiconductor manufacturing processes. Section 4 discusses performance measures in the back-end environment. Section 5 presents the proposed hybrid dispatching rule. This is followed by a simulated application of the rule in section 6. Finally, section 7 presents a summary of the main findings and conclusions of the study.

2. Literature review

2.1 'Fuzzy' decision-making models

To integrate the multiple objectives noted above (cycle time, throughput rate, due-date accuracy, and utilization), the present study applied fuzzy analytic hierarchy process (AHP) to determine the weight of each performance measurement. AHP (Saaty 1980) has become one of the most extensively used multiple-criteria decision-making (MCDM) methods because it is relatively easy to use and can effectively handle both qualitative and quantitative data. These data are compared in a pair-wise fashion with respect to the relative importance of (or preference for) each criterion (Saaty 1980, Zahedi 1986, Kuo *et al.* 2002).

A good decision-making model must be able to accommodate vagueness or ambiguity because such 'fuzziness' is a common characteristic of many decision-making problems (Yu 2002). In these circumstances, the transformation of qualitative preferences to precise numerical values is usually inappropriate, and decision-makers often therefore provide uncertain estimates rather than precise answers. Because uncertainty needs to be considered in some or all of the problems tackled by decision-makers, the pair-wise comparisons of traditional AHP is not always appropriate (Yu 2002), and the use of fuzzy numbers and linguistic terms might be more suitable. The utilization of so-called 'fuzzy theory' in AHP is thus often more appropriate and effective than traditional AHP in an uncertain pair-wise comparison environment, and several researchers have integrated fuzzy theory with AHP to deal with uncertainty (Boender *et al.* 1989, Cheng and Mon 1994, Cheng 1996, Cheng *et al.* 1999, Murtaza 2003).

2.2 Determining WIP levels

Several previous studies have proposed methodologies for setting suitable WIP levels for the semiconductor back-end environment. However, most of these studies have focused on the determination of WIP levels in the wafer-fabrication (front end) stage (Gross and Harris 1974, Robinson *et al.* 1990, Wein 1992, Kuroda and Kawada 1995, Rose 1998, Lin and Lee 2001, Nye *et al.* 2001, Sivakumar and Chong 2001).

Burman *et al.* (1986) developed a *queuing network model* for the IC manufacturing industry. Lin and Lee (2001) constructed a queuing network-based algorithm to determine the total WIP level. The queuing network-based method is efficient for estimating a wide range of manufacturing parameters and it can also answer questions quickly under some conditions (Burman *et al.* 1986, Askin and Krisht 1994, Kuroda and Kawada 1995). However, queuing models are not appropriate for obtaining measures in dynamic manufacturing systems because most queuing models assume long-run and steady-state conditions.

Choobineh and Sowrirajan (1996) developed a *hybrid model* to determine the total WIP level. The total WIP was specified, but the WIP levels for each product could vary dynamically within heuristic limits specified according to the product mix and the work content of each product. Luh *et al.* (2000) also proposed a hybrid method to determine the total WIP level to minimize tardiness and earliness penalties.

Simulation models can also provide detailed estimation of the WIP level. Bonvik *et al.* (1997) and Miller (1990) used simulation models to identify the minimum

WIP level, subject to a target throughput. Duenyas and Patananake (1998) used simulation to set a WIP level that achieved minimum cost. Wein's studies (1988, 1992) also utilized simulation models.

Linear programming has also been used to project WIP levels (Kim and Leachman 1994, Kuroda and Kawada 1995, Nye *et al.* 2001, Ryan and Choobineh 2003). This is an appropriate methodology for projecting WIP levels at each step of the process, but it is difficult to ascertain reasonably accurate input data (such as holding costs and shortage costs). Without such data, it is not possible to derive suitable WIP levels using this methodology.

According to Domaschke *et al.* (1998), practitioners are increasingly turning to simulation models because such models enable analysts to address a wide variety of complex issues. The 'National Technology Roadmap' has also identified simulation as a critical tool for factory integration (Semiconductor Industry Association 1997). Using this methodology, a simulated back-end factory is established, and simulation results are observed and analysed (by ANOVA and multiple comparison methodology) to derive the optimal total WIP level for the whole factory and the WIP profiles of each workstation.

After determining appropriate WIP levels, these levels can be set as control limits in a proposed dispatching rule. The rule is designed to reduce tardy jobs, enhance on-time delivery, and rectify WIP deviation. If manufacturers simply pursue due-date performance without consideration of bottleneck starvation and wasted time (changeover and waiting) at batch-processing workstations, WIP levels will increase and cycle time will be prolonged. Holding WIP or inventory in the back end is more expensive than in front-end fabs because the functions of ICs are determined in the back-end processes. Maintaining WIP in balance is therefore an important factor in ensuring short cycle time, sufficient throughput, and on-time delivery.

The present study therefore proposes and tests a hybrid dispatching rule that can be used to assign priority among various jobs.

3. Back-end semiconductor manufacturing processes

The focus of the present research is restricted to factories that have dedicated lines for particular products. As noted above, a typical back-end manufacturing system includes three main IC manufacturing processes.

- In *wafer probing*, the individual circuits (of which there can be hundreds on each wafer) are tested electronically by means of thin probes.
- In *IC packaging*, electrical leads are connected to the individual dies, which are then encapsulated in plastic or ceramic shells.
- In the *testing process*, automated test equipment is used to interrogate each chip and determine whether it is operating at the required specifications to ensure that customers receive a defect-free product (Uzsoy *et al.* 1992).

A typical back-end manufacturing system consists of 20–40 processes (Qiu 2005). Materials are advanced through some or all of these processes before being turned out as finished goods. Because products differ in terms of dimensions, consumables, and process specifications, the process flows differ from product to product. A simplified back-end process flow is depicted in figure 2.

Figure 2. A simplified back-end process flow.

The semiconductor back-end process has certain production characteristics that distinguish it from other shop-floor production processes (Mamziane 1990, Uzsoy *et al.* 1992, Lee *et al.* 1993, 2000). For example, in the semiconductor back-end process, jobs arrive over time, and the timing of these arrivals depends on the performance of the previous stages. If an order is delayed, the production schedules will be changed accordingly. Moreover, because process flows differ from product to product, various products pass through different equipment, and the size of lots depends on the size of the equipment container and/or magazine. The processing time depends on the lot size. Furthermore, a given job might return to a workstation more than once in the wafer-probe stage or in the process of final IC testing. For example, if a lot has to be tested at three temperatures, all three tests are carried out at the same test workstation. This makes production control and planning more difficult because most jobs are required to pass a series of tests at various temperatures on the same test system, resulting in re-entrant product flows.

With respect to equipment characteristics, the setting-up time is sequence-dependent for some equipment, and the changeover time varies in accordance with the sequence of processing products. In addition, the machine used for burn-in oven production belongs to batch-processing equipment. Scheduling batch processing is a crucial issue for back-end manufacturing. Undesirable production variations—such as machine breakdown, material shortage, randomness of processing time, randomness of yield, and reworking—always exist, thus increasing the challenge of meeting due-date performance (Gupta and Sivakumar 2005, Tu *et al.* 2005).

4. Performance measurements of back-end environment

4.1 Traditional performance measurements

The primary goal of back-end operations is *on-time delivery*—that is, delivering finished goods in sufficient volume on the promised date. To achieve this objective,

throughput and *cycle time* are the most important parameters of the back-end processes. Managers are also concerned with *utilization* of capital-intensive equipment, especially at bottleneck workstations. The performance measures for these various parameters are described below.

4.1.1 Average on-time delivery percentage. ‘Average on-time delivery percentage’ (AOTDP) measures the average proportion of on-time deliveries for all orders shipped.

Let DD_i be the due date of order i , and let PCT_i be the time that all chips lots are finished all required processes in order i ; then average on-time delivery percentage is defined as:

$$AOTDP = \frac{\sum_{i=1}^{n_i} I_{\{PCT_i < DD_i\}}}{n_i} \times 100\% = \frac{\sum_{i=1}^{n_i} (no_of_orders_early + no_of_orders_on-time)}{n_i}. \quad (1)$$

$I_{\{PCT_i < DD_i\}}$ is an indicator function equal to one if the complete date of delivery of order i is earlier than the due date, and n_i is the sum of orders.

4.1.2 Cycle time. Cycle time is defined as the time that elapses from the start of wafer probing until the end of the final testing of chips. It thus measures the rate at which products move through the manufacturing process. The product type changes from wafers to chips when a single wafer is cut into several chips in the die-sawing stage.

Let $WRT_{w,i}$ be the time at which wafer w of order i is released, and $WCT_{w,i}$ be the time at which wafer w has completed all the required processes. Moreover, let $CRT_{c,w,i}$ be the time at which chip lot c (as derived from wafer w of order i) is released, $CCT_{c,w,i}$ be the time at which chip lot c (as derived from wafer w of order i) finishes all its required processes, and ECT_i be the expected cycle time of order i . Then, when the cycle time is compared with the expected cycle time of order i , the ‘relative cycle time performance’ (RCTP) of order i is defined as:

$$CTP_i = \begin{cases} 1 + \left[\frac{ECT_i - PCT_i}{ECT_i} \right] \times 100\% & \text{if } ECT_i - PCT_i > 0 \\ \left[\frac{ECT_i - (PCT_i - ECT_i)}{ECT_i} \right] \times 100\% & \text{if } ECT_i - PCT_i < 0 \end{cases}, \quad (2)$$

where $PCT_i = \max\{CCT_{c,w,i} - CRT_{c,w,i} + WCT_{w,i} - WRT_{w,i}\} \forall i$, and then relative cycle time performance is:

$$RCTP = \frac{\sum_{i=1}^{n_i} CTP_i}{n_i}. \quad (3)$$

4.1.3 Throughput rate. The mean throughput rate is defined as the average number of chips performed at the bottleneck per calendar day (under conditions in which production is restricted by the constraint stage). The ‘relative throughput rate performance’ (RTRP) can then be calculated.

Let PCL_k be the processing chip lots on the bottleneck workstation on day k ($k=1, \dots, K$). Then the mean throughput rate is defined as:

$$MTR = \frac{\sum_{k=1}^K PCL_k}{K}, \quad (4)$$

and the relative throughput rate performance is defined as:

$$RTRP = \frac{MTR}{DTR} \times 100\%, \quad (5)$$

where DTR is the expected daily throughput rate.

4.1.4 Utilization. Utilization is defined as the average running (up) time of the bottleneck workstation per calendar day. The ‘average utilization’ (AU) can thus be calculated.

Let OPR_k be the operational time on day k ($k=1, \dots, K$) and OPO_k be the factory operation time. In Taiwan, back-end factories always operate for 24 hours per day. Thus, OPO_k is equal to 24 hours. Then:

$$AU = \frac{\sum_{k=1}^K (OPR_k/OPO)}{K}. \quad (6)$$

4.2 Integrated performance measurements

To attain good performance (in terms of cycle time, throughput, on-time delivery, and utilization for production efficiency), the present study combined these four objectives into an integrated performance index to determine the AWDLs for WIP flow control. Fuzzy AHP was employed.

The procedures of AHP typically involve several steps—from defining the unstructured problem and stating the objectives, to determining the relative weights of the decision elements, to obtaining an overall rating for the alternatives (Saaty 1980, Zahedi 1986). In this study, the weights of cycle time, throughput rate, and on-time delivery were determined by the following steps:

Step 1. Construct the hierarchical structure: The criteria for achieving the overall goal of customer satisfaction (on-time delivery, cycle time, throughput rate, and utilization) were deemed to be in the second level and must be analysed by experts.

Step 2. Formulate a questionnaire: Decision elements were compared pair-wise and assigned relative scales by decision-makers or experts. They were asked to compare each of the paired elements in the matrices through questionnaires. Triangular membership functions can be defined to represent linguistic terms for facilitating judgment and integrating different experts’ opinions. The membership function is shown in table 1.

With fuzzy numbers $\tilde{1}$, $\tilde{3}$, $\tilde{5}$, $\tilde{7}$, $\tilde{9}$ used to present the relative contribution of each element on the objectives, a fuzzy judgment vector can be constructed. These scales represent the options of ‘equally important’, ‘moderately important’,

Table 1. Triangular fuzzy member function.

Fuzzy number	Characteristic (membership) function*
$\tilde{1}$	(1, 1, 3)
$\tilde{3}$	(1, 3, 5)
$\tilde{5}$	(3, 5, 7)
$\tilde{7}$	(5, 7, 9)
$\tilde{9}$	(7, 9, 9)

*The member function is represented by (L, M, U) , where M is the strongest grade of membership and L and U are the lower and upper bounds.

‘strongly important’, ‘very strongly important’, and ‘extremely important’ (corresponding to pair-wise weights from $\tilde{1}$ to $\tilde{9}$ respectively).

Step 3. Determine the fuzzy weight matrix: The weight vector represents the relative importance of each criterion. This can be obtained by applying pair-wise comparisons of the AHP or it can be judged immediately by a decision-maker. Whichever method is adopted, it must not deviate from the experts’ (or decision-makers’) subjective judgment.

Different experts or decision-makers might define different weight vectors; if so, this usually causes imprecise evaluation and serious difficulties during the decision-making process. For this reason, the present study proposed a group decision based on AHP with triangular fuzzy numbers to improve original pair-wise comparisons.

Let each expert k individually carry out pair-wise comparisons using Saaty’s scale of 1–9 for all criteria. For S experts, the relative importance level of criterion i and criterion j can be expressed as

$$D_k = \begin{matrix} & \begin{matrix} C_1 & C_2 & \cdots & C_m \end{matrix} \\ \begin{matrix} C_1 \\ C_2 \\ \vdots \\ C_m \end{matrix} & \begin{bmatrix} b_{11k} & b_{12k} & \cdots & b_{1mk} \\ b_{21k} & b_{22k} & \cdots & b_{2mk} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1k} & b_{m2k} & \cdots & b_{mmk} \end{bmatrix} \end{matrix}, \quad k = 1, 2, \dots, S \quad (7)$$

$$b_{jik} = \begin{cases} b_{ijk} & \text{if } j \neq i \\ 1 & \text{if } j = i \end{cases} \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m, \quad (8)$$

where a score (b_{ijk}) represents the measurement of expert k on the relative importance of criterion i and criterion j .

A comprehensive pair-wise comparison matrix (D) can then be constructed by integrating all experts’ measurement through equations (9)–(13) as follows:

$$L_{ij} = \min(b_{ijk}), \quad k = 1, 2, \dots, S, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m \quad (9)$$

$$M_{ij} = \frac{\sum_{k=1}^S b_{ijk}}{S}, \quad k = 1, 2, \dots, S, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m \quad (10)$$

$$U_{ij} = \max(b_{ijk}), \quad k = 1, 2, \dots, S, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m \quad (11)$$

$$\tilde{b}_{ij} = (L_{ij}, M_{ij}, U_{ij}), \quad k = 1, 2, \dots, S, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m, \quad (12)$$

where a comprehensive score ($\tilde{b}_{ij}$) represents the relative importance of each criterion using triangular fuzzy numbers. The following is then obtained:

$$D = \begin{matrix} & \begin{matrix} C_1 & C_2 & \cdots & C_m \end{matrix} \\ \begin{matrix} C_1 \\ C_2 \\ \vdots \\ C_m \end{matrix} & \begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} & \cdots & \tilde{b}_{1m} \\ \tilde{b}_{21} & \tilde{b}_{22} & \cdots & \tilde{b}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{b}_{m1} & \tilde{b}_{m2} & \cdots & \tilde{b}_{mm} \end{bmatrix} \end{matrix}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m. \quad (13)$$

Each criterion has a different degree of importance. To derive a weight ($\tilde{w}_i$) that corresponds to a specific criterion (c_i), the following equation can be used to calculate relative weights of all criteria:

$$\tilde{w}_i = \frac{\sum_{j=1}^m \tilde{b}_{ij}}{\sum_{i=1}^m \sum_{j=1}^m \tilde{b}_{ij}}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, m. \quad (14)$$

The fuzzy weight vector $\mathbf{W}$ can be obtained:

$$\mathbf{W} = \begin{bmatrix} \tilde{w}_1 \\ \tilde{w}_2 \\ \vdots \\ \tilde{w}_m \end{bmatrix} \quad (15)$$

Step 4. Determine the integrated performance index: After weights of cycle time (w_{CT}), throughput rate (w_{TR}), on-time delivery (w_D), and utilization (w_u) have been derived, they can be used to construct the ‘integrated performance index’ (IPI). The IPI is defined as:

$$IPI = w_{CT} \times RCTP / \mu_{RCTP} + w_{TR} \times RTRP / \mu_{RTRP} + w_D \times AOTDP / \mu_{APTDP} + w_u \times AU / \mu_{AU}. \quad (16)$$

After determining the integrated performance index, the optimal WIP levels can be determined experimentally by: (i) collecting the performance indices; (ii) calculating the integrated performance under different WIP levels; and (iii) conducting ANOVA and multiple comparisons to determine the optimal total WIP of the entire back-end factory, together with the upper and lower WIP boundaries for the whole factory. If the total WIP level is maintained within the upper and lower WIP boundaries, optimal results will be attained for the parameters being examined here (cycle time, throughput rate, average on-time delivery percentage, and utilization).

Moreover, within the upper and lower boundaries for the total WIP level of the whole factory, corresponding WIP profiles for each workstation can be collected.

The profiles are defined as upper and lower acceptable WIP deviation levels (AWDLs). They are then used in the proposed dispatching rule to determine the priorities of processing lots.

5. Hybrid dispatching rule

According to the theory of constraints, production system output is restricted by the constraint machines (bottleneck machines) (Goldratt 1998, Wu *et al.* 2006). Starvation avoidance is accomplished by maintaining a relatively high WIP level at the constraint machine to ensure the availability of material in virtually any circumstance (including extraordinary and unpredictable circumstances) (Tu *et al.* 2005). To prevent machine starvation, the dispatching rule must therefore control the WIP at the bottleneck at a level that is no lower than a predefined lower AWDL.

For general ('non-bottleneck') workstations, the dispatching rule must control the WIP at a level that: (i) does not exceed a pre-defined upper AWDL (to maintain performance on cycle time and throughput rate); and (ii) is no lower than the pre-defined lower AWDL (to maintain high utilization).

Generally speaking, the final testing (FT) stage is usually considered to be the bottleneck stage, whereas the die-mounting and wire-bonding stages are usually considered to be high-loading stages. These stations were therefore designated as 'monitored workstations' for the purposes of this study. All the other stages are referred to as 'general workstations' for the purposes of this study.

Critical ratio (CR) is used for its good on-time delivery performance (Dabbas and Fowler 1999, Dabbas *et al.* 2001). The CR is defined as the ratio between the remaining time to the customer's due date and work remaining to complete the product. Values between 0 and 1 indicate that the lot is late, whereas values greater than 1 indicate that the lot is ahead of schedule. The reciprocal of CR is adopted in the dispatching rule. To achieve WIP balance, lots should be dispatched by WIP deviation. Thus, the priority lots are determined by the time slackness and WIP deviation. The priorities will be raised when time slackness is tightening or WIP deviation is increasing. The priority is calculated as follows:

(1) *Lot priority determination of bottleneck and high-loading workstations*

$$P_{c,i,m} = \begin{cases} \alpha \left[\frac{RMT_{c,i}}{DD_i - NT} \right] & \text{if } LAWDL_m < WL_m \\ \alpha \left[\frac{RMT_{c,i}}{DD_i - NT} \right] + \beta \left[\frac{LAWDL_m - WL_m}{LAWDL_m} \right], & \text{if } WL_m \leq LAWDL_m, \end{cases} \quad (17)$$

where:

- $RMT_{c,i}$ is estimated remaining processing time for chip lot c of order i ;
- DD_i is the due date of order i ;
- NT presents the current time;
- $P_{c,i,m}$ is the priority value of chip lot c at workstation m ;
- α is the weight of reciprocal of CR ratio;
- β is the weight of reciprocal of WIP deviation;
- WL_m is the actual WIP level of workstation m ; and
- $LAWDL_m$ is the lower AWIP level of workstation m .

(2) Lot priority of upstream workstation of general workstations

$$P_{c,i,m} = \begin{cases} \alpha \left[\frac{RMT_{c,i}}{DD_i - NT} \right] - \beta \left[\frac{WL_m - UAWDL_m}{UAWDL_m} \right] & \text{if } WL_m \geq UAWDL_m \\ \alpha \left[\frac{RMT_{c,i}}{DD_i - NT} \right] & \text{if } LAWDL_m < WL_m < UAWDL_m \\ \alpha \left[\frac{RMT_{c,i}}{DD_i - NT} \right] + \beta \left[\frac{LAWDL_m - WL_m}{LAWDL_m} \right] & \text{if } WL_m < LAWDL_m \end{cases} \quad (18)$$

where:

- $RMT_{c,i}$ is estimated remaining processing time for chip lot c of order i ;
- DD_i is the due date of order i ;
- NT presents the current time;
- $P_{c,i,m}$ is the priority value of chip lot c at workstation m ;
- α is the weight of reciprocal of CR ratio;
- β is the weight of reciprocal of WIP deviation;
- WL_m is the actual WIP level of workstation m ;
- $LAWDL_m$ is the lower AWIP level of workstation m ; and
- $UAWDL_m$ is the upper AWIP level of workstation m .

Of course, jobs with a higher priority value have a higher priority in being processed. Intuitively, an item coming from the current workstation increases the WIP level of the succeeding workstation and decreases its own WIP. To maintain WIP between the upper and lower AWDLs, the lots with greater discrepancy from actual WIP and lower WIP should be given a higher priority in being selected; conversely, the lots with a greater discrepancy between actual WIP and upper WIP should have a lower priority.

(3) Searching for tie-break priority

When more than two jobs have the same priority, it is necessary to decide each job's dispatching sequence under tie-breaking. Li *et al.* (1996) suggested using simple static rules and Wiendahl (1995) reported that, without disturbing the dispatching sequence from previous workstations, FIFO has the least variability. Thus, FIFO is designated to dispatch jobs when two lots have the same priority.

6. Simulation and experimentation

6.1 Development of simulation-based back-end environment

A typical back-end factory, located in the Hsinchu Science Park in Taiwan, was analysed using simulation experiments of performance under different WIP levels. A commonly used simulation package, eM-Plant, was used to construct a simulation environment from real back-end IC manufacturing processes. This research modelled 463 machines in 22 station families with a total of 25 products. All products had an identical process route, but their processing times differed. Machine unavailability was defined. All machines had respective fixed preventive maintenance

Figure 3. Simulation model validation based on correlated inspection.

(PM) schedules. PM held a higher priority than job processing. Breakdown was modelled with 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 arguments. Jobs were randomly fed into the wafer-probing factory to simulate the dynamic nature of job arrival in the real-world environment, and a commonly used releasing policy, Fixed-WIP, was used as the releasing policy.

Before constructing the simulation model, interviews were conducted with several production managers who had extensive experience of back-end process management and who were experts in semiconductor back-end production and layouts. These experts were asked to assess the environmental settings of the initial simulation model, and any settings that were deemed inappropriate were then corrected.

The simulation-based virtual back-end factory was also verified and validated against an actual factory performance through simulation experiments over a suitable period. The simulation outputs of the virtual model achieved 90% reliability (in terms of cycle time, WIP level, and equipment utilization) as compared with the actual factory performance measures.

These tests imply that the generated simulation was a reliable representation of a real-world back-end process under similar pre-defined production assumptions. The procedure for validation checking is depicted in figure 3.

6.2 Experimental design

To identify the appropriate AWDLs of monitored workstations, the appropriate total WIP level was first determined. Initially, a deterministic lot-releasing policy was used to establish the mean WIP level while attaining target throughput. Thirteen cases were then examined under a fixed-WIP releasing policy—varying from 0.7 times the mean WIP level to 1.3 times the mean WIP level (see table 2). To reach a steady state and to minimize potential start-up bias, statistics for the first 180 days (designated the ‘warm-up’ period) were discarded from consideration. Data on cycle time performance, average on-time delivery percentage, throughput rate, and

Table 2. Simulation scenarios.

Run #	Replicates	WIP level
1	30	1.3 μ
2	30	1.25 μ
3	30	1.2 μ
4	30	1.15 μ
5	30	1.1 μ
6	30	1.05 μ
7	30	1.0 μ
8	30	0.95 μ
9	30	0.9 μ
10	30	0.85 μ
11	30	0.8 μ
12	30	0.75 μ
13	30	0.7 μ

utilization were then collected for the first 180 days after the warm-up period to represent the back-end process performance and enable the calculation of integrated performance. From the beginning, the utilization of the bottleneck ('FT') workstation was tracked under the minimum fixed WIP level (0.7 μ) to ensure that the simulation model was at a steady state. The utilization of the bottleneck workstation under a WIP level of 0.7 μ reached 82.87% after the warm-up period. As long as reasonable and steady utilization was maintained under the minimum WIP level, it was considered reasonable to infer that production would remain stable under other (higher) WIP levels.

6.3 Integrated performance index determination

Senior managers and experts contributed their professional experience to determine the relative importance of the four individual performance measures (average on-time delivery percentage, cycle-time performance, throughput rate performance, and utilization). The Delphi method (consisting of a series of reiterated interrogations through questionnaires) was used to obtain a consensus among the people involved, and thus to construct the pair-wise matrices. The group's opinions were synchronized by applying equations (7)–(15) (above), and a fuzzy weight vector $\tilde{W}$ was thus obtained:

$$\tilde{W} = \begin{bmatrix} \tilde{w}_D \\ \tilde{w}_{CT} \\ \tilde{w}_{TR} \\ \tilde{w}_u \end{bmatrix} = \begin{bmatrix} 0.363 \\ 0.343 \\ 0.149 \\ 0.112 \end{bmatrix}$$

$$IPI = 0.396 \times \frac{AOTDP}{\mu_{AOTDP}} + 0.343 \times \frac{RCTP}{\mu_{RCTP}} + 0.149 \times \frac{RTRP}{\mu_{RTRP}} + 0.112 \times \frac{AU}{\mu_{AU}}.$$

In the opinion of the group of experts, on-time delivery (with a weight of 0.396) was the major factor in determining the overall performance of the back-end process.

Short cycle time (0.343) ranked second. Throughput (0.149) and utilization (0.112) ranked third and fourth respectively.

6.4 AWDL determination

After determining the integrated performance index (IPI), 30 simulation runs were conducted under each of the thirteen different scenarios to collect data on average on-time delivery percentage (AOTDP), relative cycle-time performance (RCTP), relative throughput rate performance (RTRP), and average utilization (AU) for further analysis. Results of an ANOVA indicated that significant differences arose at different WIP levels with respect to IPI (deriving from AOTDP, RCTP, RTRP, and AU) at $\alpha=0.05$. Duncan's multiple tests (Montgomery 2001) were then conducted on IPI under different WIP levels. The results of simulation runs and Duncan's multiple tests are summarized in table 3.

According to table 3, when the WIP level was at 0.95μ , the corresponding mean cycle time was at its shortest. When the WIP level exceeded μ , the mean cycle time increased rapidly; that is, a higher WIP level led to a longer cycle time. This result conformed to Little's Law (Gross and Harris 1974). However, as can also be seen in table 3, the mean cycle time began to increase at a WIP level of less than 0.9μ ; that is, a lower WIP level did not guarantee a shorter cycle time—primarily because products require more time to accumulate a sufficient quantity to pass each batch-processing operation (such as the burn-in stage at the final testing station), thereby producing longer cycle time (Lin and Lee 2001).

With respect to mean throughput rate, when the WIP levels were below 0.9μ , the mean throughput rate was significantly smaller than the mean throughput rates

Table 3. Performance index under different total WIP levels.

Total WIP level (chip lots)		Mean cycle time (h)	RCTP (%)	AOTDP (%)	MTR (lots/day)	AU (%)	Integrated performance index	Duncan grouping
1.3 μ	5047 lots	314.5	71.5	67.89	357.2	98.21	0.81	G
1.25 μ	4853 lots	300.3	78.1	74.68	356.1	97.99	0.86	F
1.2 μ	4659 lots	291.5	88.9	79.56	353.5	97.76	0.92	E
1.15 μ	4465 lots	283.6	99.6	84.75	352.6	97.71	0.97	D
1.1 μ	4271 lots	278.1	112.9	89.22	350.9	97.32	1.04	C
1.05 μ	4077 lots	272.3	126.7	93.34	346.5	96.91	1.09	B
1.0 μ	3882 lots	268.2	136.6	95.92	343.7	96.27	1.14	A
0.95 μ	3688 lots	266.7	141.2	96.61	337.1	95.90	1.15	A
0.9 μ	3494 lots	267.6	135.8	95.88	328.3	95.10	1.12	A
0.85 μ	3300 lots	271.9	126.4	93.82	312.7	94.30	1.08	B
0.8 μ	3106 lots	276.4	110.1	89.81	298.2	92.70	1.00	C
0.75 μ	2912 lots	284.6	96.2	86.64	280.4	88.20	0.93	E
0.7 μ	2718 lots	292.4	79.1	84.12	252.6	82.87	0.85	F

$\mu = 3882.7$ is derived from the simulation runs under fixed WIP releasing policy.

Integrated performance index: $IPI = 0.396 \times AOTDP/\mu_{AOTDP} + 0.343 \times RCTP/\mu_{RCTP} + 0.149 \times RTRP/\mu_{RTRP} + 0.112 \times AU/\mu_{AU}$.

at WIP levels of 0.9μ and above. In accordance with Little's Law, the relationship between WIP and throughput time (Hopp and Spearman 1996) is:

$$\text{Throughput time} = \frac{\text{WIP}}{\text{Throughput rate}}$$

Thus, reducing WIP levels without making any other changes will also reduce throughput. This is consistent with the present findings—that once the level of WIP decreased further, the mean throughput rates decreased more rapidly at a WIP level of 0.9μ or less.

With respect to utilization of the bottleneck workstation, higher WIP levels were positively associated with greater utilization (as might be expected because higher WIP levels mean that more lots are fed into the bottleneck workstation).

The AOTDP depends on the mean cycle time and mean throughput rate. From table 3, it is apparent that AOTDP remained at 95% (or above) at WIP levels of 0.9μ , 0.95μ , and μ ; moreover, it remained at 90% (or above) at WIP levels of 0.85μ to 1.05μ . Once the level of WIP decreased or increased beyond these levels, the AOTDP decreased dramatically. It is apparent that these WIP levels produced a better performance than WIP levels higher than 1.05μ . It is also apparent that WIP levels below 0.8μ could not compensate for the loss of throughput rate by decreasing the cycle time—thus producing lower AOTDP.

Considering the IPI of simulation runs, it was found that WIP levels of 0.9μ , 0.95μ , and μ led to an IPI that was significantly higher than those under other WIP levels. It can be assumed that production managers would not be entirely satisfied with these performances—because higher mean throughput rates and equipment utilization could be achieved at higher WIP levels. However, in view of the fact that the main goal of back-end semiconductor processes is to improve performance on due-date performance, back-end factories usually focus on parameters that enhance customer satisfaction more than parameters such as throughput rate and equipment utilization. Thus, the total WIP level should be kept between 1.05μ (4077 chip lots) and 0.9μ (3494 chip lots) to ensure greater customer satisfaction—that is, these WIP levels will achieve a greater average on-time delivery performance with acceptable results for the other parameters (short mean cycle time, acceptable throughput, and acceptable utilization).

The present study also collected WIP profiles for every workstation under different WIP levels (as AWDLs for the proposed hybrid dispatching rule). As noted above, the final testing (FT) station was designated as the 'bottleneck workstation', the die-mounting and wire-bonding stations were designated as the 'monitored workstations', and the other workstations were designated as 'general workstations'. WIP levels of 0.9μ and μ were the upper ideal AWDLs (UAWDLs) and lower ideal AWDLs (LAWDLs) for general workstations; WIP levels of 0.9μ were set to the lower AWDLs (LAWDLs) for the bottleneck and monitored workstations (see table 4).

6.5 Evaluation of hybrid dispatching rule

After determining the AWDLs, α and β were set at 0.7 and 0.3 respectively in equations (17) and (18). The priorities of jobs were thus determined (and dispatched) by applying the proposed hybrid dispatching rule.

Table 4. WIP profiles for workstations.

WIP level Workstation	WIP profiles (chip lots)												
	1.3 μ (5040)	1.25 μ (4853)	1.2 μ (4659)	1.15 μ (4465)	1.1 μ (4271)	1.05 μ (4077)	1.0 μ (3882)	0.95 μ (3688)	0.9 μ (3494)	0.85 μ (3300)	0.8 μ (3106)	0.75 μ (2912)	0.7 μ (2718)
CP_Test_1	310	302	290	280	268	259	246	237	223	212	201	188	176
Laser_Repair	147	142	136	129	121	115	108	101	94	89	85	79	71
CP_Test_2	273	264	255	244	237	223	218	208	198	188	179	170	161
CP_Baking	214	206	198	189	180	175	170	157	150	142	135	127	118
Inking	142	138	133	127	124	115	103	98	93	86	79	75	71
CP_Inspection	116	111	106	102	97	94	90	84	78	73	68	64	59
Tapping	77	72	66	63	58	52	48	44	39	35	31	26	21
Lapping	62	58	54	51	47	43	40	36	33	30	26	22	18
Die_Sawing	256	248	240	232	223	215	205	196	186	178	171	163	155
Die_Mounting (high-loading)	454	436	419	401	384	366	348	330	312	294	275	257	238
Wire_Bonding (high-loading)	559	536	513	489	465	440	415	397	379	356	332	304	276
Molding	173	167	161	155	149	142	135	129	123	118	112	106	99
Marking	96	92	88	85	81	76	72	68	65	61	57	53	50
Trimming	124	119	115	110	106	101	96	91	86	82	78	73	68
Forming	108	103	98	95	91	86	81	76	72	67	63	59	55
AS_Inspection	38	36	33	32	31	28	25	22	20	18	16	15	15
Final_Testing (Bottleneck)	821	791	761	726	691	662	634	602	569	537	504	479	453
Cycling	357	347	337	326	315	306	297	284	271	259	247	236	224
Burn-in	268	260	253	245	237	228	219	210	201	190	180	170	160
Laser_Marking	229	223	216	208	199	193	186	178	171	162	153	146	139
Scanning	139	133	127	122	117	112	107	102	97	92	86	80	73
Packaging	77	69	57	55	52	47	41	37	33	30	27	22	19

To examine whether WIP balancing action (the proposed hybrid dispatching rule) yielded better performance than other commonly used rules, tardy jobs and AOTDP according to the proposed rule were compared with those of another 11 popular dispatching rules. Detailed descriptions of these rules and comparison results are presented in table 5.

As shown in the table, significant differences in tardiness frequencies and AOTDP were observed among 12 dispatching rules. Significantly better performance (on both numbers of tardy jobs and AOTDP) was achieved by the hybrid dispatching rule, CR, and EDD than was achieved with all other rules. Furthermore, the dispatching rule offered better performance on due dates than that offered by pure CR or pure EDD. EDD minimizes the maximum lateness and CR identifies the quality of dispatching from the point of view of on-time delivery; hence, it is understandable that pure CR achieved better due-date performance than did pure EDD alone. The proposed dispatching rule is designed to prevent lateness and, simultaneously, to reduce the discrepancies between WIP levels and AWDLs at any given time. The proposed rule thus improved the AOTDP and outperformed the other rules. Indeed, a 14.38% improvement was achieved by changing the dispatching rules from FIFO to the hybrid dispatching rule. With respect to the tardy jobs, the hybrid rule led to the least tardiness; once tardy jobs are under control, on-time delivery performance will be improved.

6.6 Parameter analysis

To illustrate the characteristics of the proposed hybrid dispatching rule, the impact of the important parameters, α and β , on the due-date performance was explored. A comparison was made of the performance under combinations of different values of α and β by employing the experimental design method, ANOVA, and multiple comparison analysis. Because the sum of α and β is limited to 1, nine combinations (α , β), were tested, with the value of α varying from 0.1 to 0.9 (in increments of 0.1) and the corresponding value of β varying from 0.9 to 0.1 (in decrements of 0.1). In all, 30 simulation runs were conducted under each combination to obtain the AOTDP in each case. The results are summarized in table 6.

As can be seen in table 6, combination 7 resulted in the least tardiness and highest AOTDP. However, there were no significant differences among combinations 6, 7, and 8 with respect to AOTDP. It is thus apparent that on-time delivery would not be affected by choosing combination 6 over combination 7, nor by choosing combination 6 over combination 8, nor by choosing combination 7 over combination 8. To maintain high on-time delivery performance, production managers can thus choose any one of these combinations when employing the hybrid dispatching rule.

The weight of the CR ratio changed with the value of α , and the weight of WIP deviation regulation depended on the value of β . When the value of α decreased, the AOTDP decreased; to improve customer satisfaction, it is thus recommended that the value of α be kept higher than 0.5. On the other hand, when the value of α was too high ($\alpha=0.9$), due-date performance diminished. Because WIP balance is a critical factor in preventing bottlenecks and smoothing the production line, it is essential to control the discrepancies between the actual WIP levels and pre-set AWDLs. In determining the appropriate parameter values for the whole

Table 5. Test results for AOTDP under different dispatching rules.

Dispatching rule	Dispatching rule description	No. of tardy jobs	Duncan grouping	AOTDP (%)	Improvement in AOTDP (%)	Duncan grouping
Hybrid rule	Proposed rule in this study	8.87	H	93.71	14.38	A
CR	Critical ratio	18.68	G	89.28	9.95	B
EDD	Earliest job due date	23.71	G	88.69	8.84	B
CYC	Cyclic priority	40.54	F	86.66	5.77	C
FLNQ	Fewest lots at the next queue	57.06	F	85.39	4.22	C
LS	Least slack	83.52	E	84.87	3.59	CD
LNQ	Largest number in queue	110.29	D	83.64	2.08	D
SPT	Shortest processing time	145.96	C	83.31	2.04	D
SRPT	Shortest remaining time	154.66	C	83.13	1.46	D
RAN	Random priority	183.38	B	81.52	0.72	E
FIFO	First in first out	193.07	B	80.93	—	E
SNQ	Smallest number in queue	262.04	A	77.41	-4.35	F

Table 6. Test results for AOTDP under different combinations of α and β values.

Combination no.	α	β	Maximum tardiness	No. of tardy jobs	AOTDP (%)	Duncan grouping
1	0.1	0.9	227.39	148.62	82.29	F
2	0.2	0.8	85.15	52.89	85.74	E
3	0.3	0.7	68.22	39.21	87.67	D
4	0.4	0.6	30.30	18.38	89.11	C
5	0.5	0.5	24.19	12.67	90.19	B
6	0.6	0.4	22.32	10.63	93.36	A
7	0.7	0.3	18.75	8.87	93.71	A
8	0.8	0.2	14.44	7.22	93.88	A
9	0.9	0.1	32.19	17.69	89.71	BC

manufacturing process, it is therefore recommended that production managers should consider a trade-off between a focus on due-date performance and the pursuit of WIP balance.

7. Conclusion

Because semiconductor back-end processes are at the end of the whole semiconductor supply chain they are sensitive to any unpredictable production variations. To improve customer satisfaction and avert delays, semiconductor back-end factories need to set proper WIP levels to protect on-time delivery from unpredictable production variance. This study has proposed a dispatching rule that attempts to consolidate on-time delivery and WIP balance simultaneously. It is designed to achieve higher customer satisfaction while maintaining acceptable WIP levels, throughput, and equipment utilization. Moreover, because back-end factories usually have multiple production objectives, a fuzzy-AHP model has been employed to determine the weights of various objectives under the primary goal of customer satisfaction. An integrated performance index has thus been derived and used to determine the AWDLs, which are then applied in the hybrid dispatching rule to determine priorities for the processing of jobs.

A simulated back-end process system has been constructed to examine the performance of the proposed dispatching rule. The experimental study demonstrated that, compared with 11 other commonly used dispatching rules, the hybrid dispatching rule produced a significant improvement in average on-time delivery percentage (AOTDP) and a reduction in tardy jobs. The AWDLs, as determined by considering multiple objectives, were shown to be appropriate when these AWDLs were applied in the hybrid dispatching rule and the due-date performance was subsequently improved.

The study has also demonstrated that parameters in the dispatching rule should be decided with care. To determine the appropriate parameter values for the whole manufacturing process, it is recommended that production managers should consider a trade-off between focusing on due-date performance and pursuing WIP balance.

References

- Askin, R.G. and Krisht, A.H., Optimal operation of manufacturing systems with controlled work-in-process levels. *Int. J. Prod. Econ.*, 1994, **37**, 1637–1653.
- Boender, C.G.E., de Graan, J.G. and Lootsma, F.A., Multi-criteria decision analysis with fuzzy pairwise comparisons. *Fuzzy Set Syst.*, 1989, **29**, 133–143.
- Bonvik, A.M., Couch, C.E. and Gershwin, S.B., A comparison of production-line control mechanisms. *Int. J. Prod. Res.*, 1997, **25**, 789–804.
- Burman, D.Y., Gurrola-Gal, F.J., Nozari, A., Sathaye, S. and Sitarik, J.P., Performance analysis techniques for IC manufacturing lines. *AT&T Tech. J.*, 1986, **65**, 46–57.
- Buzacott, J.A., The role of banks in flow-line production systems. *Int. J. Prod. Res.*, 1971, **9**, 425–436.
- Chen, H., Harrison, J.M., Mandelbaum, A. and Wein, L.M., Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Oper. Res.*, 1998, **36**, 202–215.
- Cheng, C.H. and Mon, D.L., Evaluating weapon system by AHP based on fuzzy scale. *Fuzzy Set Syst.*, 1994, **63**, 1–10.
- Cheng, C.H., Evaluating naval tactical missile systems by fuzzy AHP based on the grade value of membership function. *Eur. J. Oper. Res.*, 1996, **96**, 343–350.
- Cheng, C.H., Yang, K.L. and Hwang, C.L., Evaluating attack helicopters by AHP based on linguistic variable weight. *Eur. J. Oper. Res.*, 1999, **116**, 423–435.
- Chooobineh, F. and Sowrirajan, S., Capacitated – Constant Work in Process (C-CONWIP): a Job Shop Control System. Technical report, Industrial and Management Systems Engineering, University of Nebraska, Lincoln, Nebraska, 1996.
- Dabbas, R.M. and Fowler, J., A new scheduling approach using combined dispatching criteria in wafer fab operations. Technical report IE-ORPS-99-001, Department of Industrial Engineering, Arizona State University, Tempe, AZ, 1999.
- Dabbas, R.M., Chen, H.N., Fowler, J.W. and Shunk, D., A Combined dispatching criteria approach to scheduling semiconductor manufacturing systems. *Comput. Ind. Eng.*, 2001, **39**, 307–324.
- Duenyas, I. and Patananake, P., Base-stock control for single product tandem make-to-stock systems. *IIE Transact.*, 1998, **30**, 31–39.
- Domaschke, J., Brown, S., Robinson, J. and Leibl, F., Effective implementation of cycle time reduction strategies for semiconductor back-end manufacturing, in *Proceedings of the 1998 Simulation Conference*, 1998, 985–992.
- Goldratt, E.M., *Theory of constraints and how should it be implemented?* 1998 (North River Press: New York, NY).
- Gross, D. and Harris, C.M., *Fundamentals of Queueing Theory*, 1974 (John Wiley & Sons: New York, NY).
- Gupta, A.K. and Sivakumar, A.I., Conjunctive simulated scheduling. *Int. J. Adv. Manuf. Tech.*, 2005, **26**, 1409–1413.
- Hopp, W.J. and Spearman, M.L., *Factory Physics-Foundations of Manufacturing Management*, 1996 (Irwin: Chicago, IL).
- Huang, C.L., Huang, Y.H., Chang, T.Y., Chang, S.H., Chung, C.H., Huang, D.T. and Li, R.J., The construction of production performance prediction system for semiconductor manufacturing with artificial neural networks. *Int. J. Prod. Res.*, 1999, **37**, 1387–1402.
- Kim, J.S. and Leachman, R.C., Decomposition method application to a large scale linear programming WIP projection model. *Eur. J. Oper. Res.*, 1994, **74**, 152–160.
- Kuo, R.J., Chi, S.C. and Kao, S.S., A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. *Comput. Ind.*, 2002, **47**, 199–214.
- Kuroda, M. and Kawada, A., Adaptive input control for job-shop type production systems with varying demands using inverse queueing network analysis. *Int. J. Prod. Econ.*, 1995, **41**, 217–225.

- Lee, C.Y., Martin-Vega, L.A., Uzsoy, R. and Hinchman, J., Implementation of a decision supporting for scheduling semiconductor test operations. *J. Electron. Manuf.*, 1993, **3**, 121–131.
- Lee, Y.H., Lee, B.K. and Jeong, B., Multi-objective production scheduling of probe process in semiconductor manufacturing. *Prod. Plan. & Cont.*, 2000, **11**, 660–669.
- Li, S., Tang, T. and Collins, D.W., Minimum inventory dispatching in semiconductor wafer fabrication. *IEEE Trans. Semiconduct. M.*, 1996, **9**, 145–149.
- Lin, Y.H. and Lee, C.E., A total standard wip estimation method for wafer fabrication. *Eur. J. Oper. Res.*, 2001, **131**, 78–94.
- Luh, P.B., Zhou, X. and Tomastik, R.N., An effective method to reduce inventory in job shops. *IEEE Trans. Robot. Autom.*, 2000, **16**, 420–424.
- Mamziane, L.T., Plastic Packaging of Microelectronic Devices. Technical Report, AT&T Bell Laboratories, 1990 [Also *Plastic Packaging of Microelectronic Devices*, 1990 (Van Nostrand Reinhold: New York)].
- Miller, D.J., Simulation of a semiconductor manufacturing line. *Commun. ACM*, 1990, **33**, 99–108.
- Montgomery, D.C., *Design and Analysis of Experiments*, 5th ed., 2001 (Wiley: New York, NY).
- Murtaza, M.B., Fuzzy-AHP application to country risk assessment. *Am. Bus. Rev.*, 2003, **21**, 109–116.
- Nye, T.J., Jewkes, E.M. and Dilts, D.M., Optimal investment in setup reduction in manufacturing systems with WIP inventories. *Eur. J. Oper. Res.*, 2001, **135**, 128–141.
- Qiu, R.G., Virtual production line based WIP control for semiconductor manufacturing systems. *Int. J. Prod. Econ.*, 2005, **95**, 165–178.
- Robinson, J.K., Fowler, J. and Bard, J., The use of upstream and downstream information in scheduling semiconductor batch operations. *Int. J. Prod. Res.*, 1990, **7**, 1849–1870.
- Rose, O., WIP Evaluation of a Semiconductor Factory after Bottleneck Work center Breakdown, in *Proceedings of the 1998 Winter Simulation Conference*, 1998, pp. 997–1003.
- Ryan, S.M. and Choobineh, F.F., Total WIP and WIP mix for a CONWIP controlled job shop. *IIE Transact.*, 2003, **35**, 405–418.
- Saaty, T.L., *The Analytic Hierarchy Process*, 1980 (McGraw-Hill: New York, NY).
- Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*, pp. 127–129, 1997 (Semiconductor Industry Association: San. Jose, CA).
- Sivakumar, A.I. and Chong, C.S., A simulation based analysis of cycle time distribution and throughput in semiconductor backend manufacturing. *Comput. Ind.*, 2001, **45**, 59–78.
- Tu, Y.M. and Li, R.L., Constraint time buffer determination model. *Int. J. Prod. Res.*, 1998, **36**, 1091–1130.
- Tu, Y.M., Chao, Y.H., Chang, S.H. and You, H.C., Model to determine the backup capacity of a wafer foundry. *Int. J. Prod. Res.*, 2005, **43**, 339–359.
- Uzsoy, R., Lee, C.Y. and Martin-Vega, L.A., A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning. *IIE Transact.*, 1992, **24**, 47–60.
- Wang, L. and Prabhu, V., Parallel algorithm for setting WIP levels for multi-product CONWIP system. *Int. J. Prod. Res.*, 2006, **44**, 4681–4793.
- Wein, L.M., Scheduling semiconductor wafer fabrication. *IEEE Trans. Semiconduct. Manuf.*, 1988, **1**, 115–130.
- Wein, L.M., The relationship between yield and cycle time in semiconductor wafer fabrication. *IEEE Trans. Semiconduct. Manuf.*, 1992, **5**, 156–158.
- Wiendahl, H., *Load-Oriented Manufacturing Control*, 1995 (Springer-Verlag: Berlin).
- Wu, M.C., Huang, Y.L. and Yang, K.F., Dispatching in semiconductor fabs with machine-dedication features. *Int. J. Adv. Manuf. Tech.*, 2006, **28**, 978–984.
- Yu, C.S., A GP-AHP method for solving group decision-making fuzzy AHP problems. *Comput. Oper. Res.*, 2002, **29**, 1969–2001.
- Zahedi, F., The analytic hierarchy process- a survey of the method and its applications. *Interface*, 1986, **16**, 96–108.