Available online at www.sciencedirect.com
ScienceDirect
Computers & Industrial Engineering 55 (2008) 647–662
computers & industrial engineering
www.elsevier.com/locate/caie

Capacity pricing mechanism for wafer fabrication

Shu-Hsing Chung [a], Amy H.I. Lee [b,*], Chun-Ying Huang [c], Chia-Chien Chuang [a]

[a] Department of Industrial Engineering and Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan
[b] Department of Industrial Engineering and System Management, Chung Hua University, No. 707, Sec. 2, WuFu Road, Hsinchu, Taiwan
[c] Department of Business Administration, Ching Yun University, No. 229, Chien-Hsin Road, Jung-Li, Taiwan

Received 17 October 2006; received in revised form 2 November 2007; accepted 6 February 2008
Available online 19 February 2008

Abstract

Wafer fabrication is perhaps the most capital-intensive and technology-intensive industry. Due to customer demand uncertainty, the wafer fabrication industry in Taiwan became dramatically competitive. Emergency orders demanded from customers may exist for the need of time-to-market. How to respond to emergency orders from customers, to analyze the impact to production and cost, and to design an appropriate order acceptance plan, have become an important task for enterprises in pursuing higher service quality and ultimate profit maximization. In this paper, we propose a capacity pricing mechanism to evaluate the impacts of emergency orders. The mechanism is constructed under the circumstances that master production scheduling (MPS), capacity requirement planning (CRP), and data for manufacturing costs are known. Through production planning and profit analysis, the mechanism can analyze pricing factors including the process plans of products, priority levels of products, urgency of orders, and number of layers of poly and metal, so as to reflect the length of cycle time, waiting-time savings, impact to cycle time variance of all orders, and usage amount of critical resources. The capacity price for each product type under each priority level can be determined as a result.

Keywords: Wafer fabrication; Production planning; Capacity pricing; Emergency order; Order acceptance

1. Introduction

Manufacturers or suppliers often need to determine products' selling prices in order to achieve a given profit target. If a product is overpriced, customers may be discouraged and seek other sources of supply. On the other hand, if a product is under-priced, the supplier may fail to achieve its profit target. Therefore, an appropriate pricing for each product is necessary. Simultaneous determinations of pricing, capacity and quality decisions, and buyer–supplier coordination, have been researched on, especially in service systems (Banejee, 2005).

Dynamic pricing embraces the concepts such as real time (spot) pricing, responsive pricing, state preference pricing, flexible pricing and incentive pricing (Sanghvi, 1989). The reason for developing dynamic pricing is

* Corresponding author. Tel.: +886 3 518 6582; fax: +886 3 518 6575.
E-mail address: amylee@chu.edu.tw (A.H.I. Lee).

0360-8352/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cie.2008.02.006

due to the unprecedented cost pressures, competitive challenges and a highly uncertain planning and operating environment. With dynamic pricing, an efficient market is created by providing a mechanism to let suppliers be able to decrease costs and increase revenues so as to eliminate inefficiencies, decrease overhead costs and increase inventory turns (Minga, Feng, & Li, 2003). Dynamic pricing gives advantages to both buyers and sellers in the process, and a demand sensitive model can help sellers to change prices spontaneously on the consequence of the buyers' needs (Minga et al., 2003). Therefore, even though there are difficulties and limitations in implementation, dynamic pricing methods have been applied in various industries, including retail, electric utilities, airlines, hotels and shipping companies, and a good dynamic pricing method can help firms increase their revenues and better utilize their capacity (Elmaghraby, Gulcu, & Kesinocak, 2001).

Wafer fabrication is one of the most capital-intensive and technology-intensive industries. A 300 mm (12-in) wafer fabrication factory costs approximately 2.5–3.5 billion US dollars, and the cost of equipment is approaching 70–80% of the factory capital costs. Clearly, it is essential to fully utilize existing machine capacity to meet customer demand in order to acquire the maximum profit. In general, lowering the cost per unit of wafer (a cost-driven strategy) and providing more service or flexibility to customers (a profit-driven strategy) are the most common competitive strategies used in the wafer fabrication industry.

The unit cost effectiveness of semiconductor equipment can be measured by cost of ownership (COO), a SEMI standard metric (SEMI, 1995). COO is the total cost needed to carry out a specific processing goal, and includes the costs such as acquisition cost of equipment, installation cost, applying cost, maintenance cost, etc (Ruzyllo, 2004). However, when evaluating equipment by COO, one does not consider the effect of equipment versatility, process diversity or wafer fabrication factory size explicitly (Iwata & Wood, 2002). Leachman, Plummer, and Misawa (1999) estimate wafer fabrication factory cost by considering the impact of yield ramp up and equipment efficiency. However, the effects of process diversity on wafer processing cost are not taken into account. In addition, manufacturing overhead of a product is usually calculated based on its usage of total labor hours or total machine hours under traditional accounting principle (Barfield, Raiborn, & Kinney, 2003). However, the utilization of capacity constraint resources, which reflect the true total wafer fabrication cost and determine the maximum profit a company can achieve, is not taken into consideration.

A simple equation-based model is constructed by Iwata and Wood (2002) to present the relationship between the cost and the capacity of a wafer fabrication factory running multiple processes. The model divides fixed costs into capacity-dependent and capacity-independent costs, where capacity-dependent cost is similar to the fixed component of COO. The objective of the model is to quickly and accurately estimate the effects of process diversity, wafer fabrication capacity and setup policies on fabrication costs and wafer processing costs. Even though the above-mentioned two costs can be estimated by the model, the model is highly simplified and does not consider the fact that most wafer fabrication factories process different priority levels of orders.

As integrated circuit (IC) products enter the maturity stage of product life cycle, manufacturers, in order to survive, not only need to provide diversified services to customers, but also have to invent services that differentiate from those of other competitors. The different profit rate of products and the varied importance level of clients result in different levels of orders in a wafer fabrication factory, and higher priority orders, such as hot or rush orders, oppress normal orders because of their priority for processing. In order to increase overall profit, a company needs comprehensive strategies to utilize existed capacity efficiently. The strategies must provide functions such as setting differentiated pricing for orders of different levels of urgency to increase customer satisfactory level and company competitiveness, analyzing system contribution for order acceptance, and responding quickly to the need of emergency order from customers.

Production plan, constructed by production planning department with the negotiation with sales department, may need to be updated in the shop floor due to the request of emergency orders from customers to meet time-to-market. An emergency order, however, will intrude the production logistics of existing orders in the shop floor environment. When an order with higher priority arrives to a workstation, it will surpass orders with lower priority and wait behind orders with the same priority. This will lengthen the lower-priority orders' queueing time in front of the workstation and oppress their production performance. Therefore, cost of waiting-time savings before critical resources for higher priority orders

should be calculated to compensate the loss of system production performance. Capacity pricing mechanism, making a fair judgment based on the price of machine capacity, is the solution for the decision of accepting or rejecting emergency orders. Up until now, very little research has emphasized capacity pricing model of wafer fabrication factories even though this kind of pricing analysis problem always bothers decision makers. This research, therefore, will provide a solution, which combines the cost/profit factors, the concept of modern manufacturing management and the idea of cost of waiting-time savings, to assist decision makers in capacity pricing.

A good capacity pricing for each class of orders should reflect the length of the cycle time of the order, the impact to the cycle time variance of all orders, and the usage amount of critical resources. With the proposed capacity pricing mechanism, a wafer fabrication factory can make appropriate decisions to increase the profitability and flexibility of order acceptance and to enhance customer service quality. In addition, for solving the problem of order exchange by customers due to market demand change, which is often encountered by wafer fabrication factory, capacity pricing mechanism can be applied to determine appropriate prices for dynamic customer orders. The rest of the paper is organized as follows. The next section presents the construction of capacity pricing mechanism. In Section 3, an example case is presented to verify the proposed mechanism. Section 4 concludes the paper.

2. Capacity pricing mechanism

For each product, the process flow and manufacturing difficulty for each layer are significantly different. A product can be divided into several layers based on the bottleneck resource (BN), which is often the photolithography machine (also called stepper machine). The initial layers after releasing are basic operations for all kinds of products, and several layers, including poly and metal operations, can be identified distinctly according to product specification, generation and product type. The union of processes between two neighboring BNs and excluding the activities of the front BN is called a layer.

A good capacity pricing mechanism have many factors need to be considered for translating capacity cost into product price because each layer goes through different process and passes different critical resources and because the functions made by each layer are also different. Capacity pricing related resources must be categorized to reflect their impact by different products and different priority orders in the system. According to the management principle of theory of constraint (TOC) (Goldratt & Cox, 1992), there exists at least one bottleneck (BN) in the system that prevents the system from attaining a higher performance. Due to the dependence among manufacturing events, the output of the system will be constrained by the BN. When proceeding with the capacity pricing of product, we must first consider the loading level of BN resource. Next, capacity constraint resource (CCR), defined here as a resource which has a pretty close utilization to that of the BN, must be examined since the change of product mix may cause bottleneck shifting and as a consequence, impact the throughput performance. The loading of CCR is also important to the production performance and the smoothness of the entire production system (Chung & Huang, 2003). As a result, CCR will be included in the set of critical resources in estimating capacity usage cost. Moreover, poly and metal operations are very essential for the functions of the finished products, and their manufacturing process is unique and complex. The number of poly and metal operations a product needs will affect the market price for the product. Therefore,

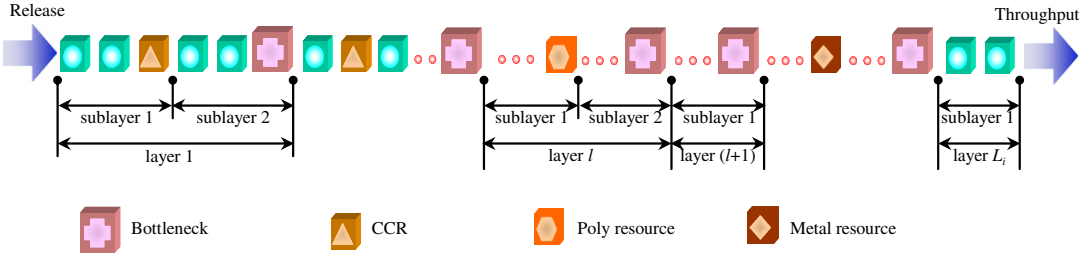

Fig. 1. Layers and sublayers in wafer fabrication.

when estimating capacity price, BN, CCR, poly and metal process resources are the core of cost consideration, and we define them as pricing related resources (PRR).

For building the capacity pricing mechanism, we also define the union of processes between two neighboring PRRs and excluding the activities of the front PRR as a sublayer. Layers and sublayers are illustrated as in Fig. 1. When calculating capacity price for each work order, the production cycle time for each sublayer/layer for each product needs to be derived so as to estimate the arrival time for an order to the PRR. Production cycle time for each layer is different due to the difference in machine units, utilization rates, and operation times of the workstation type specified in a product's process plan. The ratio of production cycle time to theoretical processing time of a product is called X-factor, which is usually used as a reference for production progress control (Kishimoto, Ozawa, Watanabe, & Martin, 2001). Therefore, based on X-factor and theoretical processing time of a layer, layer cycle time for a product can be estimated. In addition, sublayer cycle time is estimated by the same concept. With given information, the production cycle time can be estimated by statistical data, simulation model, or BBCT-MP algorithm (Chung, Pearn, Kang, Chen, & Ke, 2001).

To achieve the purpose of this research, capacity pricing mechanism applies the information such as production cycle time, order due date and utilization rates of PRRs from MPS. By adopting the concept of TOC (Corbett, 1998), this mechanism first calculates the basic capacity usage cost of each work order given its processing time on each PRR. Next, based on the degree of urgency in each layer of each work order, the mechanism can set a differentiated cost of waiting-time savings before critical resources for higher priority orders by applying the concept of bottleneck dynamics method (Morton & Pentico, 1993). Then, all cost factors will be summarized to get a capacity cost for each priority order. Based on markup method, the basic capacity price for each order can be determined at last. The framework for capacity pricing mechanism is presented in Fig. 2.

In order to simplify the complexity of the model, work order is defined as the basic unit for capacity planning. Each work order, depending on its priority, may have different release size. The following sections will explain each part of the capacity pricing mechanism in detail, and the notations are first defined as below:

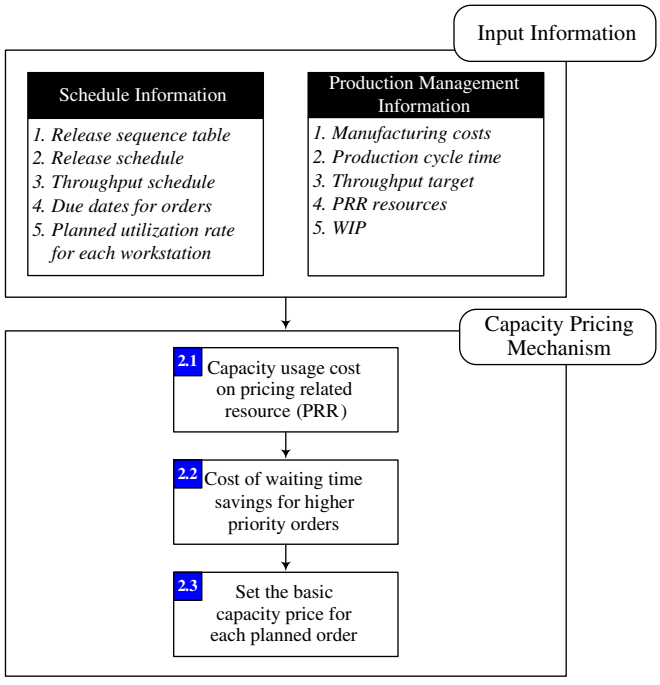

Fig. 2. Framework for capacity pricing mechanism.

Suffixes

i	Product type, i=1, 2,..., I
j	Work order, j=1, 2,..., J
l	Layer, l=1, 2, ..., L_i
x	Sublayer, x=1, 2, ..., SL_{il}
k	PRR type, $k \in$ {BN, CCR, P, M}
pri	Priority type, pri=1, 2, ..., PRI. A smaller value denotes a higher priority

Notations

d_j^{pri}	Due date of work order j with priority pri
n^{pri}	Release size of work order with priority pri
n_k	Equivalent machine units in PRR k
ρ_k	Utilization rate of PRR k
λ_{ik}^{pri}	Average hourly arrival rate of product i with priority pri to PRR k
π_i^{pri}	Mix ratio of product i with priority pri to planned target
AT_{jlx}^{pri}	Expected arrival time of work order j to sublayer x, layer l
B_{ilx}^{pri}	Processing batch size of product i in sublayer x, layer l with priority pri in PRR
C_k	Capacity usage cost per unit time for PRR k (US$/min)
C_{jlx}^{pri}	Capacity cost of work order j with priority pri in sublayer x, layer l in PRR
CC_B	Estimated conversion cost (i.e., the sum of manufacturing overhead and direct labor cost) in the planning period
CP_j^{pri}	Basic capacity price of work order j with priority pri
CT_i^{pri}	Cycle time of product i with priority pri
CT_{il}^{pri}	Layer cycle time of product i with priority pri in layer l, i.e., $\sum_l CT_{il}^{pri} = CT_i^{pri}$
CT_{ilx}^{pri}	Sublayer cycle time of product i with priority pri in sublayer x, layer l, i.e., $\sum_x CT_{ilx}^{pri} = CT_{il}^{pri}$
FT_{jlx}^{pri}	Expected completion time of work order j in sublayer x, layer l
I	Reinvestment rate of capital
$I(j)$	Product type of work order j
L_i	Number of layers for product i
LQ_{ilx}^{pri}	Average queueing length of product i with priority pri in sublayer x, *layer* l in PRR
$M(i,l,x)$	The PRR used to process product i in sublayer x, layer l
PT_{il}	Pure processing time of the last step of product i in layer l (i.e., BN processing time)
PT_{ilx}	Pure processing time of the last step of product i in sublayer x, layer l (i.e., PRR processing time)
RM_i	Raw material cost of product i
SL_{il}	Number of sublayers in layer l for product i
TPT_i	Total theoretical processing time for product i
TPT_{il}	Total theoretical processing time of layer l for product i, i.e., $\sum_l TPT_{il} = TPT_i$
TPT_{ilx}	Total theoretical processing time of sublayer x, layer l for product i, i.e., $\sum_x TPT_{ilx} = TPT_{il}$
R_j	Given planned release time for work order j
RC_{jlx}^{pri}	Cost of waiting-time savings before PRR for work order j with priority pri in sublayer x, layer l
S_{jlx}^{pri}	Slack value for work order j with priority pri in sublayer x, layer l
U_{jlx}^{pri}	Urgency factor of sublayer x, layer l for work order j with priority pri
VC_{il}	Variable cost of product i in layer l
W_k	Weight of conversion cost of PRR k, $\sum_k W_k = 1$
XF_i^{pri}	Ratio of production cycle time to theoretical processing time for product i with priority pri (i.e., X-factor)

2.1. Unit capacity usage cost of pricing related resource

In this section, planned utilization rate of each pricing related resource (PRR) k, is applied to calculate its capacity usage cost. When the utilization rate of a machine is higher, the loading of the machine is bigger, and

this specific machine plays a more important role in system performance. Therefore, the loading ratio of PRR k is applied to allocate the system conversion cost to PRR k. The steps of estimation are as follows:

Step 1: Calculate the loading ratio of PRR k by dividing the planned utilization rate of PRR k by the sum of utilization rates of all PRRs.

$$W_k = \frac{\rho_k n_k}{\sum_{k^*} \rho_{k^*} n_{k^*}}, \quad \text{for each } k \tag{1}$$

Step 2: Calculate the capacity usage cost per unit time for PRR k, C_k, by obtaining the portion of conversion cost allocated to PRR k and dividing it by the capacity available for PRR k in a planning period (28 days), unit: (US$/min).

$$C_k = \frac{CC_B \times W_k}{28 \times 24 \times 60 \times n_k}, \quad \text{for each } k \tag{2}$$

2.2. Cost of waiting-time savings for higher priority orders

As stated before, an order with a higher priority will oppress production performance of lower priority orders. Cost of waiting-time savings before critical resources for higher priority orders therefore should be calculated to compensate the loss of system production performance. Assume the processing time of higher priority order j with priority pri in sublayer x of layer l, on PRR is $PT_{I(j),lx}$. When this work order arrives at the queuing line of PRR, each order with lower priority will extend its queueing time by $PT_{I(j),lx}$. For the queueing line, if the number of lower priority orders is u, work order j must be responsible for $u \times PT_{I(j),lx}$ time cost for delaying lower-priority orders at PRR k. The concept is shown in Fig. 3.

Three stages are required to calculate the cost of waiting-time savings for higher priority orders. Stage one, estimate sublayer (and layer) cycle time by multiplying the X-factor and theoretical processing time of each sublayer (and layer), and the results are the basis for estimating the arrival time of orders to PRR. Stage two, estimate the expected arrival time and expected completion time of each sublayer (and layer) for each work order in the PRR based on the expected release time of work orders and the forward scheduling concept. The estimation is as depicted in Fig. 4. Stage three, calculate the cost of waiting-time savings of each order based on the urgency factor of the order and the specific priority level the order is in. The three stages are explained in detail as follows.

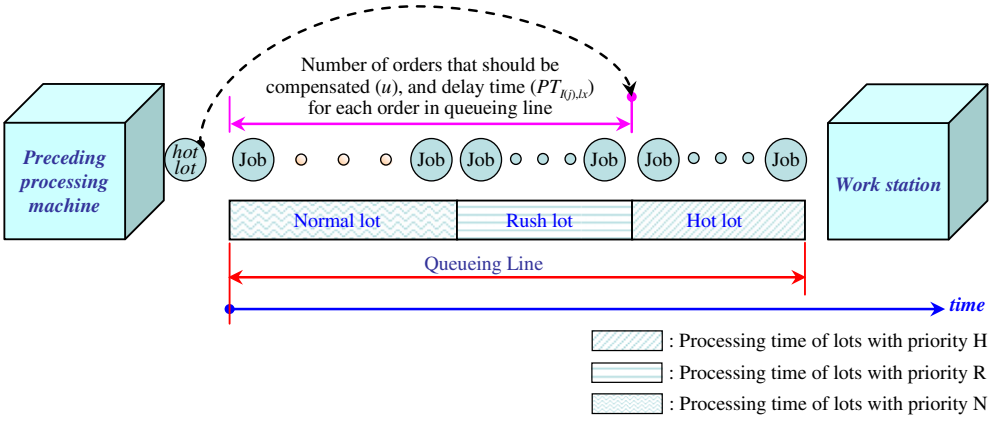

Fig. 3. Concept of estimating cost of waiting-time savings for higher priority orders.

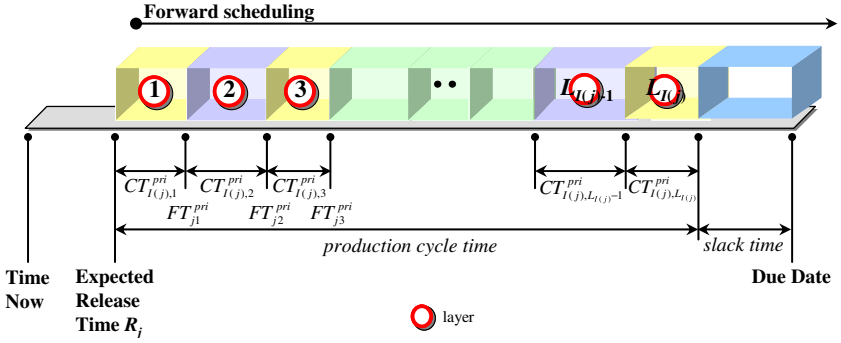

Fig. 4. Estimation of the expected completion time for each layer of an order.

2.2.1. Stage 1. Estimate cycle time of each sublayer and each layer for each product

Because the production cycle time of wafer fabrication is very long, the ratio of production cycle time to theoretical processing time, or X-factor, is practically adopted to control the production progress (see also Kishimoto et al., 2001). In other words, by using layer theoretical cycle time and X-factor of each order, we can control the processing step in each layer to be finished in a given time. Therefore, with X-factor, theoretical processing time per layer and per sublayer, layer cycle time and sublayer cycle time can be estimated.

Step 1: Set $i = 1$, $l = 1$, $x = 1$, $pri = 1$.

Step 2: Calculate X-factor of product i.

$$XF_i^{pri} = \frac{CT_i^{pri}}{TPT_i} \tag{3}$$

Step 3: Calculate the cycle time of sublayer x by multiplying theoretical processing time of sublayer x with X-factor of product i.

$$CT_{ilx}^{pri} = TPT_{ilx} \times XF_i^{pri} \tag{4}$$

Step 4: Set $x = x+ 1$. If $x \leqslant SL_{il}$, go to step 3. Else, go to step 5.

Step 5: Calculate the cycle time of layer l by multiplying theoretical processing time of layer l with X-factor of product i.

$$CT_{il}^{pri} = TPT_{il} \times XF_i^{pri} \tag{5}$$

Step 6: Set $l = l+ 1$. If $l \leqslant L_i$, set $x = 1$, and go to step 3. Else, go to step 7.

Step 7: Set $pri = pri+ 1$. If $pri \leqslant PRI$, set $l = 1$, $x = 1$, and go to step 2. Else, go to step 8.

Step 8: If $i = I$, end of the procedure. Else, set $i = i+ 1$, $l = 1$, $x = 1$, and go to step 2.

2.2.2. Stage 2. Estimate expected arrival time of each order to each sublayer and each layer of PRR

Based on the given planned release time (R_j), sublayer cycle time ($CT_{I(j),lx}^{pri}$), and layer cycle time ($CT_{I(j),l}^{pri}$) of the work orders from detail production scheduling, the expected completion time of each order on the PRR can be obtained using forward scheduling. Next, expected arrival time of each work order to the PRR can be estimated based on the pure processing time ratio.

Step 1: Set $j = 1$.
Step 2: Set $l = 1$, $x = 1$, and $temp = R_j$.
Step 3: The expected completion time of sublayer x of layer l of work order j is:

$$FT^{pri}_{jlx} = temp + CT^{pri}_{I(j),lx} \tag{6}$$

Step 4: Estimate the expected arrival time to the PRR corresponding to sublayer x.

$$AT^{pri}_{jlx} = FT^{pri}_{jlx} - XF^{pri}_{I(j)} \times PT_{I(j),lx} \tag{7}$$

Step 5: Set $x = x+ 1$, $temp = FT^{pri}_{jlx}$. If $x \leqslant SL_{I(j),l}$, go to step 3. Else, go to step 6.
Step 6: Set $l = l+ 1$. If $l \leqslant L_i$, set $x = 1$, and go to step 3. Else, go to step 7.
Step 7: If $j = J$, end of the procedure. Else, set $j = j+ 1$, and go to step 2.

2.2.3. Stage 3. Calculate cost of waiting-time savings for each order based on its priority level

With expected arrival time to each layer and each sublayer for each order, the urgency factor (U^{pri}_{jlx}) and the saved queuing length (LQ^{pri}_{ilx}) can be obtained. Urgency factor of work order j represents the slack value of work order j between the current state of the order and its due date. If work order j does not have a higher priority than other orders, it may not be finished on time for delivery. Therefore, for work order j, a higher urgency factor means that the order benefits more in its higher priority in using PRR. Based on this, urgency factor can be seen as the ratio for compensating the shortening of queueing time. As a result, cost of waiting savings is calculated by the multiplication of queueing time of all lower-priority orders by urgency factor. The steps are as follows:

Step 1: Calculate the slack value, S^{pri}_{jlx}, for each job j, which is the expected waiting time for the process of sublayer x in PRR, by using due date of the job to deduct its expected arrival time to sublayer x, layer l, its processing time in the PRR and its remaining production cycle time.

$$S^{pri}_{jlx} = d^{pri}_{j} - AT^{pri}_{jlx} - PT_{I(j),lx} - \sum_{l^*=l+1}^{L_{I(j)}} \sum_{x^*=x+1}^{SL_{I(j),l}} CT^{pri}_{I(j),l^*x^*}, \quad \text{for each } j,\ l,\ x \tag{8}$$

Step 2: Calculate the urgency factor, U^{pri}_{jlx}, of sub-layer x of layer l for each job by applying the concept of bottleneck dynamics method (Morton & Pentico, 1993).

$$U^{pri}_{jlx} = \exp\{-(S^{pri}_{jlx})/(K \times PT_{I(j),lx})\}, \quad \text{for each } j,\ l,\ x \tag{9}$$

where K is a free parameter that should be experimentally fit for each problem class. Morton and Pentico (1993) suggested that a good solution is generally attainable when K is between 1.0 and 3.0. A more detailed discussion is given in Appendix A. In this paper, we set K=2.0 through a preliminary study.

Step 3: Calculate queuing line length, LQ^{pri}_{ilx}, of product i with priority pri in sublayer x, layer l in PRR, by applying the Little's Law (Hiller & Lieberman, 2005) for queuing model. The average hourly arrival rate, λ^{pri}_{ilx}, of product i with priority pri in sublayer x, layer l to PRR is estimated first. Under the concept of throughput leveling (Chung, Yang, & Cheng, 1997), the throughput of each product with each priority in each sublayer x, layer l equals to the planned throughput of each product with each priority. Therefore, λ^{pri}_{i}, λ^{pri}_{il} and λ^{pri}_{ilx} are calculated based on throughput target ($\Re$) in a planning horizon (H) and the product mix ratio of the respective product i, priority pri (π^{pri}_{i}).

$$\lambda_i^{pri} = \lambda_{il}^{pri} = \lambda_{ilx}^{pri} = \frac{\Re}{H \times 24} \times \pi_i^{pri}, \quad \text{for each } i,\ l,\ x,\ pri \tag{10}$$

$$LQ_{ilx}^{pri} = \lambda_{ilx}^{pri} \times CT_{ilx}^{pri} - n_{M(i,l,x)}, \quad \text{for each } i,\ l,\ x,\ pri \tag{11}$$

Step 4: Calculate the cost of waiting-time savings before PRR for each higher priority order, $RC_{jlx}^{y=pri}$(US$/lot), by applying the weighted concept on the urgency factor of each order.

$$RC_{jlx}^{y=pri} = \left[\sum_{y=pri+1}^{PRI} \frac{LQ_{I(j),lx}^{y}}{B_{I(j),lx}^{y}} \right] \times PT_{I(j),lx} \times C_{M(I(j),l,x)} \times \left(1 + U_{jlx}^{y=pri}\right), \quad \text{for each } j,\ l,\ x \tag{12}$$

2.3. Basic capacity price of work orders

With the relevant data obtained in the previous sections, the basic capacity price of a planned work order can be calculated, and the three major components of the price include:

(1) Capacity usage cost of PRR $\left(\left(\frac{PT_{ilx}}{B_{ilx}^{pri}} \right) \times C_{M(I(j),l,x)} \right)$.

(2) Cost of waiting-time savings for higher priority orders (RC_{jlx}^{pri}), and

(3) Raw material and variable costs ($RM_{I(j)} + \sum_l VC_{I(j),l}$).

The steps for obtaining capacity price of a work order are as follows:

Step 1: Calculate the capacity cost of job j with priority pri in sublayer x, layer l in PRR by summing up the capacity usage cost and cost of waiting-time savings for higher priority orders for processing in PRR.

$$C_{jlx}^{pri} = \left(\frac{PT_{I(j),lx}}{B_{I(j),lx}^{pri}} \right) \times C_{M(I(j),l,x)} + RC_{jlx}^{pri} \times (1+I)^z, \quad \text{for each } j,\ l,\ x,\ pri \tag{13}$$

$$z = \begin{cases} 0, & \text{if } pri = PRI \\ \frac{CT_i^{PRI}}{CT_i^{pri}}, & \text{if } pri = 1, 2, \ldots, (PRI-1) \end{cases}, \quad \text{for each } i \tag{14}$$

Step 2: Calculate the basic capacity price of job j by summing up the capacity cost, raw material and variable costs of job j in all sub-layers.

$$CP_j^{pri} = \sum_{l=1}^{L_{I(j)}} \sum_{x=1}^{SL_{I(j),l}} C_{jlx}^{pri} + \left(RM_{I(j)} + \sum_{l=1}^{L_{I(j)}} VC_{I(j),l} \right), \quad \text{for each } j,\ pri \tag{15}$$

3. Experimental verification

To verify the effectiveness of this mechanism, data taken from a wafer fabrication factory in the Science-Based Industrial Park in Taiwan is used. To simplify the complexity of the environment, this paper is based on the following assumptions and limitations:

1. Production information: There are five different product types, A, B, C, D and E, to be fabricated in this system. The products are categorized into two product families, logic I.C. and memory I.C. Product A and B are logic products, while product C, D, and E are memory products. All product types have different process routes, and each product type has only one distinct route. The workstations and processing time required in each route are known.

2. Release size policies:
   a. Hot: highest priority. Work orders are not limited by batching policy, and they can be released into shop floor and be loaded onto any batch machine with only a single lot.
   b. Rush: secondary priority. A full batch of six lots must be formed before releasing to the floor.
   c. Normal: lowest priority. Full loading policy is also required, and release batch size is six lots.
3. Workstation information: There are 83 different types of workstations (coded from W1 to W83), including serial and batch workstations. Photo stepper, W46, is the bottleneck, and furnace, W24, is the CCR. The critical poly workstation is W29, a poly photo stepper; and the critical metal workstation is W18, a metal etch workstation.
4. Down time distribution of workstations: Meantime between failures (*MTBF*) and meantime to repairs (*MTTR*) are exponentially distributed and are traced for each workstation, while meantime between preventive maintenance (*MTBPM*) and meantime to preventative maintenance (*MTTPM*) are known constants.
5. Dispatching rule: For jobs waiting before a resource, the one with a higher priority level has a preferential right. For jobs with the same priority, first-in first-out is applied.
6. Allowance for setting due date: The due date allowances for hot, rush and normal level of orders are respectively 0.10, 0.20 and 0.30 of the estimated cycle time.
7. Raw material cost: The cost of raw material does not have a very big variation in short term, and a fixed material price is assumed here.
8. Conversion costs: Since the production environment does not have major investment nor big changes, the conversion costs, which consist of direct labor cost and manufacturing overhead, is assumed to be a fixed cost in a planning period (US$ 7,150,000/ month).
9. Variable costs: The variable processing cost per layer is US$ 300.
10. Reinvestment rate: Because hot lots and rush lots have shorter waiting time and cycle time in production, cash can be received faster from customers. The amount can be further invested with a 20% profit rate.

The simulation model is built by eM-Plant (Tecnomatix, 2000). A simulation horizon is set to 168 days. The first 84 days are a warm-up period; hence, only results belonging to the next 84 days are collected. The simulation model is run 15 times to get adequate statistical results.

3.1. Verification of capacity pricing mechanism

In order to examine the applicability of capacity pricing mechanism, we design two different system environments: (1) a basic environment with only normal orders and (2) a practical environment with hot, rush and normal orders. Under different environments, basic capacity pricing are calculated and compared.

3.1.1. Case 1: Basic environment without emergency orders

In this environment, only normal orders are existed, and the parameter *pri* is fixed at 3. The number of lots processed for each product is shown in Table 1. The capacity price of each product is obtained through the methodology proposed in Section 2, and is shown in Table 2. Simulated production cycle time and waiting time of each product under each priority are shown in Tables 3 and 4, respectively.

3.1.2. Case 2: Production environment with emergency orders

In this environment, three levels of orders are existed, hot, rush and normal orders with *pri* of 1, 2 and 3, respectively. The number of lots processed for each product is shown in Table 5. Based on the methodology proposed in Section 2, basic capacity price for each product is obtained as in Table 6. Tables 7 and 8 show the simulated production cycle time and waiting time of each product under each priority.

3.1.3. Result analysis

Table 9 presents the basic capacity pricing results under Case 2. When there are three levels of priority, lower-priority orders are suppressed by upper-priority orders. The basic capacity price of a hot lot is approx-

imately 2 times of the capacity price of a rush lot, while the price of a rush lot is also two times (1.95–1.99 times) of that of a normal lot. In consequence, the capacity price of a hot lot is approximately four times (3.90–4.12 times) of that of a normal lot. By comparing the basic environment of Case 1 and the practical environment in Case 2, we can find that the capacity price of a hot lot and a rush lot in Case 2 is 4 (3.74–4.30 times) and 2 (1.90–1.96 times) times respectively of the basic price of a normal lot in Case 1. The capacity price of a normal lot in Case 2 is 0.96–0.99 times of a normal lot in Case 1 due to the process delay of the lot in Case 2, where normal lots belong to the lowest priority. The results reflect the pricing perception of different price for different product/priority orders in real practice. In other words, a higher priority order demands a shorter production cycle time, and thus a higher price is charged. A lower priority order results in a longer production cycle time due to the impact of a faster process of higher priority orders, and as a result, a discounted price must be given. Fig. 5 also shows that the waiting time of higher priority orders decreases a lot and the profit increases concurrently in comparison with results from the basic environment in Case 1. Table 10 shows that waiting time of normal orders in a multi-priority environment in comparison with that in basic environment increases tremendously (more than 26.65%). For the system as a whole, waiting time in Case 2 is 19.91% longer than that in Case 1. Since higher priority orders have lower waiting time and have an impact on the system waiting time, higher prices must be charged to cover the loss of the system. As a result, the profit of the system increases by US$ 2,914,020.

Based on the utilization rate of PRR, conversion cost of the system can be allocated quickly and capacity usage cost of orders can be estimated. Since higher priority orders are processed more quickly, they must be charged with a cost of waiting-time savings. Capacity price of a work order can then be determined. Simulation results show that the capacity pricing mechanism could reflect the length of cycle time, the impact to cycle time variance and the usage amount of critical resources in determining an appropriate price for an order.

3.2. An application of capacity pricing mechanism

As the emphasis of the competition of wafer fabrication industry changes from production cost and quality to customer service, there is a need to allow customers to exchange planned or dispatched orders to meet market demand. When customers have an emergency order or order adjustment such as in product type, quantity and due date, production planning department must determine a balance between customer satisfaction and production performance. This is because emergency order or order adjustment may change the product mix in

Table 1
Product mix in Case 1

Priority	Product type					Total	Ratio
	A	B	C	D	E		
Hot	0	0	0	0	0	0	0
Rush	0	0	0	0	0	0	0
Normal	22	22	22	22	22	110	1
Total	22	22	22	22	22	110	1
Ratio	0.2	0.2	0.2	0.2	0.2	1	

Table 2
Basic capacity price for each product under each priority in Case 1 (US$,000/lot)

Priority	Product type				
	A	B	C	D	E
Hot	–	–	–	–	–
Rush	–	–	–	–	–
Normal	14.82	16.82	13.78	16.12	15.65

Table 3
Simulated production cycle time of each product under each priority in Case 1 (h)

Priority	Product type				
	A	B	C	D	E
Hot	–	–	–	–	–
Rush	–	–	–	–	–
Normal	271.48	297.17	268.61	315.50	308.87

Table 4
Simulated waiting time of each product under each priority in Case 1 (h)

Priority	Product type				
	A	B	C	D	E
Hot	–	–	–	–	–
Rush	–	–	–	–	–
Normal	84.68	95.37	81.49	99.27	97.09

Table 5
Product mix in Case 2

Priority	Product type					Total	Ratio
	A	B	C	D	E		
Hot	1	1	1	1	1	5	0.0455
Rush	4	4	4	4	4	20	0.1818
Normal	17	17	17	17	17	85	0.7727
Total	22	22	22	22	22	110	1
Ratio	0.2	0.2	0.2	0.2	0.2	1	

Table 6
Basic capacity price for each product under each priority in Case 2 (US$,000/lot)

Priority	Product type				
	A	B	C	D	E
Hot	59.65	62.87	59.29	63.85	63.17
Rush	28.82	32.06	27.06	30.66	29.96
Normal	14.47	16.14	13.64	15.76	15.33

Table 7
Simulated production cycle time and relative percentage of each product under each priority in Case 2

Priority	A		B		C		D		E	
	Hour	(%)	Hour	(%)	Hour	(%)	Hour	(%)	Hour	(%)
Hot	208.86	69.20[a]	228.2	70.74	212.86	71.85	245.22	70.27	243.18	70.99
Rush	263.77	87.39	276.47	85.70	259.91	87.73	302.83	86.78	298.63	87.18
Normal	301.82	100.00	322.59	100.00	296.26	100.00	348.97	100.00	342.54	100.00

[a] Based on production cycle time of normal products of the product type.

the system or the processing sequence of orders, and as a result, may delay the delivery of existed orders (Chung, Lee, Lai, Kuo, & Chen, 2002; Chung & Huang, 2003). However, if only production performance

Table 8
Simulated waiting time and relative percentage of each product under each priority in Case 2

Priority	A		B		C		D		E	
	Hour	(%)	Hour	(%)	Hour	(%)	Hour	(%)	Hour	(%)
Hot	22.06	19.18[a]	26.4	21.86	25.75	23.59	28.99	21.84	31.4	24.01
Rush	76.97	66.92	74.67	61.82	72.79	66.69	86.6	65.24	86.85	66.42
Normal	115.02	100.00	120.79	100.00	109.15	100.00	132.74	100.00	130.76	100.00

[a] Based on waiting time of normal products of the product type.

Table 9
Basic capacity price result analysis

Product	Price per lot (US$,000)			Ratio			Ratio to the normal lot in Case 1		
	(1) Hot	(2) Rush	(3) Normal	(1):(2)	(1):(3)	(2):(3)	H:N(Base)	R:N(Base)	N:N(Base)
A	59.65	28.82	14.47	2.07	4.12	1.99	4.02	1.94	0.98
B	62.87	32.06	16.14	1.96	3.90	1.99	3.74	1.91	0.96
C	59.29	27.06	13.64	2.19	4.35	1.98	4.30	1.96	0.99
D	63.85	30.66	15.76	2.08	4.05	1.95	3.96	1.90	0.98
E	63.17	29.96	15.33	2.11	4.12	1.95	4.03	1.91	0.98

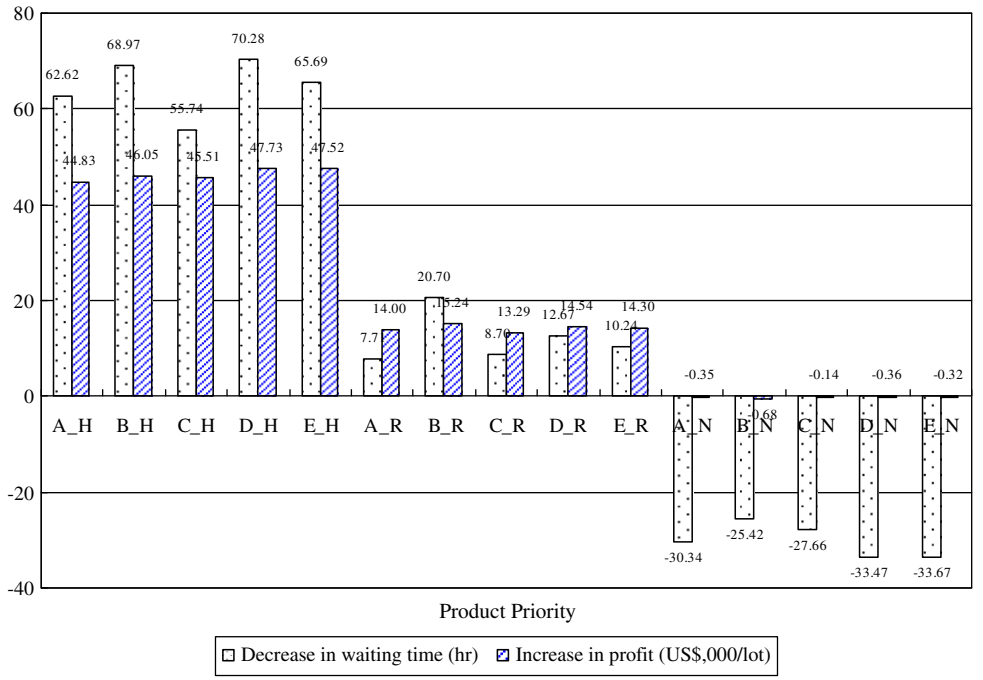

Fig. 5. Profitability of each product/priority.

is considered and the request of order adjustment by customers is ignored, this is a contradiction with customer-orientation and a loss of order acceptance flexibility. The capacity pricing mechanism proposed in this paper, by adopting dynamic pricing theory, can consider the factors such as the urgency of order delivery and capacity utilization rates, in the evaluation and determination of the price of orders. Therefore, this mechanism has its practical value.

Table 10
Overall profit analysis between Case 2 and Case 1

Product priority	(a) Throughput (lot)	(b) Decrease in waiting time (h/lot)	(c) Decrease in waiting time (%)	(d) Increase in profit (US$,000/lot)	Decrease in total waiting time (h) (e) = (a) × (b)	Increase in total profit (US$,000) (f) = (a) × (d)
A_H	6	62.62[a]	73.95[b]	44.83[c]	375.72	268.98
A_R	24	7.70	9.09	14.00	184.80	336.00
A_N	102	−30.34	−35.83	−0.35	−3094.68	−35.70
B_H	6	68.97	72.32	46.05	413.82	276.30
B_R	24	20.70	21.70	15.24	496.80	365.76
B_N	102	−25.42	−26.65	−0.68	−2592.84	−69.36
C_H	6	55.74	68.40	45.51	334.44	273.06
C_R	24	8.70	10.68	13.28	208.80	318.72
C_N	102	−27.66	−33.94	−0.14	−2821.32	−14.28
D_H	6	70.28	70.80	47.73	421.68	286.38
D_R	24	12.67	12.76	14.54	304.08	348.96
D_N	102	−33.47	−33.72	−0.36	−3413.94	−36.72
E_H	6	65.70	67.67	47.52	394.20	285.12
E_R	24	10.24	10.55	14.31	245.76	343.44
E_N	102	−33.67	−34.68	−0.32	−3434.34	−32.64
Average	–	–	−19.91[d]	–	−18.15[e]	4.42[f]
Total	660	–	–	–	−11977.02	2914.02

[a] Simulated waiting time of product A with hot priority in Case 2 minus simulated waiting time of product A in Case 1.
[b] The percentage of reduction in waiting time for product A with hot priority in Case 2 compared to the waiting time of product A in Case 1.
[c] Basic capacity price of product A with hot priority in Case 2 minus basic capacity price of product A in Case 1.
[d,e,f] Weighted average. The throughput (column (a)) is used as the weight.

4. Conclusions

Wafer fabrication is a very capital-intensive industry, and the investment in equipment can be several billion US dollars. Therefore, how to design an appropriate capacity pricing mechanism is important and urgent in practice. A good pricing mechanism not only needs to consider the static machine and equipment data, it also needs to reflect the capacity utilization of production system since the utilization rate is highly positively correlated with the economic cycle of the wafer fabrication industry. Under a high-utilization environment, customers are willing to pay at a higher unit price. On the other hand, under an economic downturn, companies need to lower product price in order to attract customers. This paper proposes a capacity pricing mechanism that is based on dynamic pricing. The mechanism analyzes pricing factors including the process plans of products, priority levels of products, urgency of orders, and number of layers of poly and metal, so as to set the capacity price for producing each product type under each priority level. The experimental results show that the proposed capacity pricing mechanism can set up an appropriate capacity price in an environment with different characteristics among products and multiple priority orders and with the consideration of both production performance and financial performance. The results can be applied in a wafer fabrication factory to determine the emergency order price when customers need to exchange orders. In conclusion, the proposed mechanism is rational and has its practical value.

The capacity pricing mechanism proposed in this paper is based on the point of view of a capacity supplier; however, wafer fabrication in recent years shows that capacity demander usually has a bargaining power in an economic downturn. Therefore, a future research direction of capacity pricing mechanism can consider the inclusion of appropriate economic indicators.

Acknowledgements

The authors gratefully acknowledge the partial support of the National Science Council in Taiwan under Grant No. NSC90-2622-E-009-001 and the UMC Corporation in Taiwan for providing the opportunity for this study.

Appendix A.

Urgency factor, U_j, is the job-time urgency of job j in using a resource in a specific time point t, and is calculated as follows:

$$U_j = \exp\left\{-\left(\frac{S_j}{K \times PT_j}\right)\right\} = \exp\left\{-\left(\frac{S_j}{PT_j}\right) \times \left(\frac{1}{K}\right)\right\}, \quad \text{for each } j$$

where

$$S_j = d_j - (PT_j + t)$$

If $U_j = 1$, the remaining time of job j from time point t to the work order due date (d_j) is exactly equal to the processing time of job j; i.e., the slack value of the work order is zero ($S_j = 0$). Thus, the job must be processed immediately in order to prevent the delay of the order. If $U_j < 1$, this means that the remaining time for processing job j is greater than the required processing time of job j; i.e., the job has sufficient time for processing. If $U_j > 1$, the remaining time for processing job j is less than the required processing time of job j; i.e., the job has been delayed.

From Table A1 and Fig. A1, we can see that when the slack value of job j is negative, the urgency factor increases exponentially when K decreases. That is, with the setting of a smaller K, the production system can focus more on on-time delivery of work orders. On the other hand, when the slack value of job j is positive, the impact of different values of K on urgency factors is not significant. Apparently, if a smaller K is set, when a job is delayed, the order delaying cost becomes very large since the urgency factor increases tremendously. Therefore, the production system will stress more on the delivery performance, and less on production flexibility. If a larger K is set, the delay cost is relatively smaller when a job is tardy. Thus, the production system aims to increase the production flexibility first, and delivery performance next. In conclusion, the setting of K value should consider the environment characteristics, and the most appropriate value can be determined through experimental design.

Table A1
The impact on urgency factor under different K

S_j/PT_j	K				
	1.0	1.5	2.0	2.5	3.0
−3.00	20.0855	7.3891	4.4817	3.3201	2.7183
−2.75	15.6426	6.2547	3.9551	3.0042	2.5009
−2.50	12.1825	5.2945	3.4903	2.7183	2.3010
−2.25	9.4877	4.4817	3.0802	2.4596	2.1170
−2.00	7.3891	3.7937	2.7183	2.2255	1.9477
−1.75	5.7546	3.2113	2.3989	2.0138	1.7920
−1.50	4.4817	2.7183	2.1170	1.8221	1.6487
−1.25	3.4903	2.3010	1.8682	1.6487	1.5169
−1.00	2.7183	1.9477	1.6487	1.4918	1.3956
−0.75	2.1170	1.6487	1.4550	1.3499	1.2840
−0.50	1.6487	1.3956	1.2840	1.2214	1.1814
−0.25	1.2840	1.1814	1.1331	1.1052	1.0869
0.00	1.0000	1.0000	1.0000	1.0000	1.0000
0.25	0.7788	0.8465	0.8825	0.9048	0.9200
0.50	0.6065	0.7165	0.7788	0.8187	0.8465
0.75	0.4724	0.6065	0.6873	0.7408	0.7788
1.00	0.3679	0.5134	0.6065	0.6703	0.7165
1.25	0.2865	0.4346	0.5353	0.6065	0.6592
1.50	0.2231	0.3679	0.4724	0.5488	0.6065
1.75	0.1738	0.3114	0.4169	0.4966	0.5580
2.00	0.1353	0.2636	0.3679	0.4493	0.5134
2.25	0.1054	0.2231	0.3247	0.4066	0.4724
2.50	0.0821	0.1889	0.2865	0.3679	0.4346
2.75	0.0639	0.1599	0.2528	0.3329	0.3998
3.00	0.0498	0.1353	0.2231	0.3012	0.3679

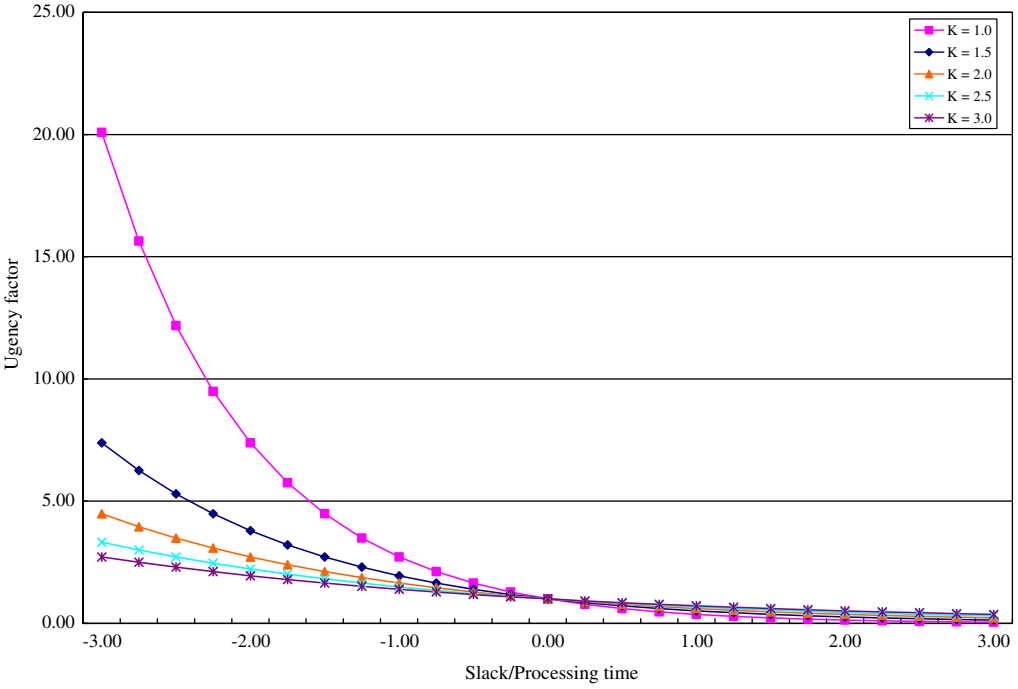

Fig. A1. The relationship between urgency factor and slackness.

References

Banejee, A. (2005). Concurrent pricing and lot sizing for make-to-order contract production. *International Journal of Production Economics, 93-94*, 189–195.

Barfield, J. T., Raiborn, C. A., & Kinney, M. R. (2003). *Cost accounting: Traditions and innovations*. South-Western: Mason.

Chung, S. H., & Huang, C. Y. (2003). The design of rapid production planning mechanism for the product mix changing in a wafer fabrication. *Journal of the Chinese Institute of Industrial Engineers, 20*(2), 153–168.

Chung, S.H., Lee, A.H.I., Lai, C.M., Kuo, N.C., & Chen, J.R. (2002). The construction of an order exchange evaluation mechanism for wafer fabs. In *2002 International conference on modeling and analysis of semiconductor manufacturing*. Tempe, Arizona.

Chung, S.H., Pearn, W.L., Kang, H.Y., Chen, C.C., & Ke, W.T. (2001). Cycle time estimation for wafer fabrication with multiple-priority orders. In *2001 Advanced simulated technologies conference*. Seattle, Washington.

Chung, S. H., Yang, M. H., & Cheng, C. M. (1997). The design of due-date assignment model and the determination of flow time control parameters for the wafer fabrication factories. *IEEE Transaction on Components, Packaging, and Manufacturing Technology, 20*(4), 278–287.

Corbett, T. (1998). *Throughput accounting: TOC's management accounting system*. Great Barrington: North River Press.

Elmaghraby, W., Gulcu, A., & Kesinocak, P. (2001). Analysis of a price markdown mechanism. In *Third International workshop on advanced issues of E-commerce and web-based information systems* (pp. 170–177). Atlanta, Georgia.

Goldratt, E. M., & Cox, J. (1992). *The Goal – A process of ongoing improvement*. Great Barrington: North River Press.

Hiller, F. S., & Lieberman, G. J. (2005). *Introduction to operations research* (8th ed.). New York: McGraw-Hill.

Iwata, Y., & Wood, S. M. (2002). Simple cost models of high-process-mix wafer fabs at different capacities. *IEEE Transactions on Semiconductor Manufacturing, 15*(2), 267–273.

Kishimoto, M., Ozawa, K., Watanabe, K., & Martin, D. (2001). Optimized operations by extended X-factor theory including unit hours concept. *IEEE Transactions on Semiconductor Manufacturing, 14*(3), 187–195.

Leachman, R.C., Plummer, J., & Misawa, N.S. (1999). Understanding fab economics. *Report CSM-47*, University of California at Berkeley.

Minga, L.M., Feng, Y.Q., & Li, Y.J. (2003). Dynamic pricing: E-commerce-oriented price setting algorithm. In *Proceedings of the 2nd international conference on machine learning and cybernetics*. Xi'an, China.

Morton, T. E., & Pentico, D. W. (1993). *Heuristic scheduling systems*. New York: Wiley.

Ruzyllo, J. (2004). *Semiconductor glossary: An introduction to semiconductor terminology* (1st ed.). State College, PA: Prosto Multimedia Publishing.

Sanghvi, A. P. (1989). Flexible strategies for load/demand management using dynamic pricing. *IEEE Transactions on Power Systems, 4*(1), 83–93.

SEMI (1995). Cost of ownership for semiconductor manufacturing equipment. *SEMI E35-95A*, http://www.semi.org, SEMI. CA: Mountain View.

Tecnomatix Technologies Ltd. (2000). *eM-Plant objects manual*. Germany: Tecnomatix Software Company.