Int J Adv Manuf Technol (2006) 31: 316–327
DOI 10.1007/s00170-005-0185-8

ORIGINAL ARTICLE

Shu-Hsing Chung · Chun-Mei Lai

Job releasing and throughput planning for wafer fabrication under demand fluctuating make-to-stock environment

Received: 26 July 2004 / Accepted: 8 June 2005 / Published online: 10 January 2006

Abstract In the semiconductor industry, dynamic changes in demand force companies changing the product mix makes the production planning challenging. This paper aims at an environment where product mix changes periodically and presents a production scheduling system to plan the wafer lot release and throughput. The proposed system is designed on the make-to-stock basis with the objective of meeting demand forecast while maintaining production smoothness. Two modules are included in the system. Preliminary analysis module analyzes throughput and cycle time distributions for different product mixes so as to determine relative parameters to be used as the inputs to the job releasing plan. In the production scheduling module, with the considerations of the attainment of demand forecast, production smoothness, and commitment of the due dates of the released job orders, the job release schedule and completion time table are prepared. A simulation model of a semiconductor fab is used as the base case to demonstrate the effectiveness and efficiency of the proposed system.

Keywords Cycle time · Lot release scheduling · Product mix · Wafer fabrication

1 Introduction

The manufacturing of wafer fabrication is a highly complex and time-consuming process. Typically, the process involves 300 to 500 process steps on a single wafer and the product cycle time is usually more than one month. A wafer makes multiple visits to a machine group as sucessive circuit layers are added in the production process, and this is the so-called reentrant flow property. The main results of this property is that wafers at different layers in their production process have to compete with each other for the same machines. Furthermore, according to the number of lots being processed simultaneously, machines are classified into serial or batch types. Batch operations would cause wafer lots additional waiting time because of batch size transformation. All these characteristics complicate the analysis of the production planning.

As the competition becomes much fiercer, semiconductor companies must quickly respond to customers' fluctuating demand in order to survive. Dynamic changes in demand force companies to make changes on the product mix. Product mix changes complicate the already-complex system. In a semiconductor fab, machines must process a huge number of different products, resulting in a heavy load sharing of precious resources, and consequently a long queue may be present. Product mix level has considerable impact on production throughput, cycle time, cycle time spread, and the capability of meeting due dates. Production throughput, cycle time, machine utilization, and work in process (WIP) inventory are highly interrelated [1–3]. Under different product mixes, the overall manufacturing performance of the system would also be different. Facing the environment with volatile demand, production planning for make-to-stock wafer fabrication is an even complicated problem compared to other manufacturing industries.

Cycle time is the time elapsed from the release of a lot into the plant until its emergence as a finished product [4]. The accuracy of cycle time estimation is important as it strongly influences the stability of a production plan. However, cycle time estimation is quite difficult because of the complexity of the wafer fabrication process. Lawrence [5], Matsuyama and Atherton [6], Glynn and O'Dea [7], and Raddon and Grigsby [8] classified cycle time algorithms into data analysis, simulation method, queueing analysis and statistical regression analysis. Chung and Huang [9], with the application of queueing theory and the observation of the characteristics of material flow, developed a production cycle time estimating formulation, the block-based cycle time (BBCT) estimation algorithm. The BBCT algorithm has distinguishable performance in es-

S.-H. Chung (✉) · C.-M. Lai
Department of Industrial Engineering and Management,
National Chiao Tung University,
Taiwan, Republic of China
e-mail: shchung@mail.nctu.edu.tw
Tel.: +886-3-5731638
Fax: +886-3-5722392

timating mean cycle time under an environment that the product mix is fixed during all the time periods. The disadvantage of the BBCT algorithm is that estimating completion time of each wafer lot would be difficult without knowing cycle time variances.

The planning tool for master production scheduling can be categorized into simulation, mathematical programming, and the combination of the fore-mentioned. By using a simulation tool, a system can be constructed easily to match with the production activity control [10]. The simulation results can offer what-if analysis and can help the control of WIP level in a certain range well. However, building and/or running a simulation model is time-consuming. Linear programming can quickly derive the best production plan based on capacity constraints, but there are too many assumptions in describing a real phenomenon [11, 12]. For this reason, some research combined simulation and mathematic algorithm for better describing the environment than by separately adopting one of the two methods [13–15].

This paper presents a production scheduling system that deals with periodical product mix changes under an environment with volatile demand. The proposed planning system is designed on the make-to-stock basis with the objective of meeting demand forecast while maintaining production smoothness. Throughput and cycle times are first analyzed for each product type under different product mixes. Wafer release schedule and completion timetable are then prepared based on the analyzed cycle time information and properly controlled WIP level. Such planning results can be valuable for preparing available-to-promise information and for decision-making in a demand management system.

The remainder of this paper is organized as follows. Section 2 describes the system environment and shows the system framework. Section 3 describes the preliminary analysis module to determine the relative parameters to be the inputs of the production scheduling module. Section 4 is devoted to derive a job release plan and throughput planning. Section 5 presents case studies and shows the effectiveness of the proposed system. Section 6 is the conclusions.

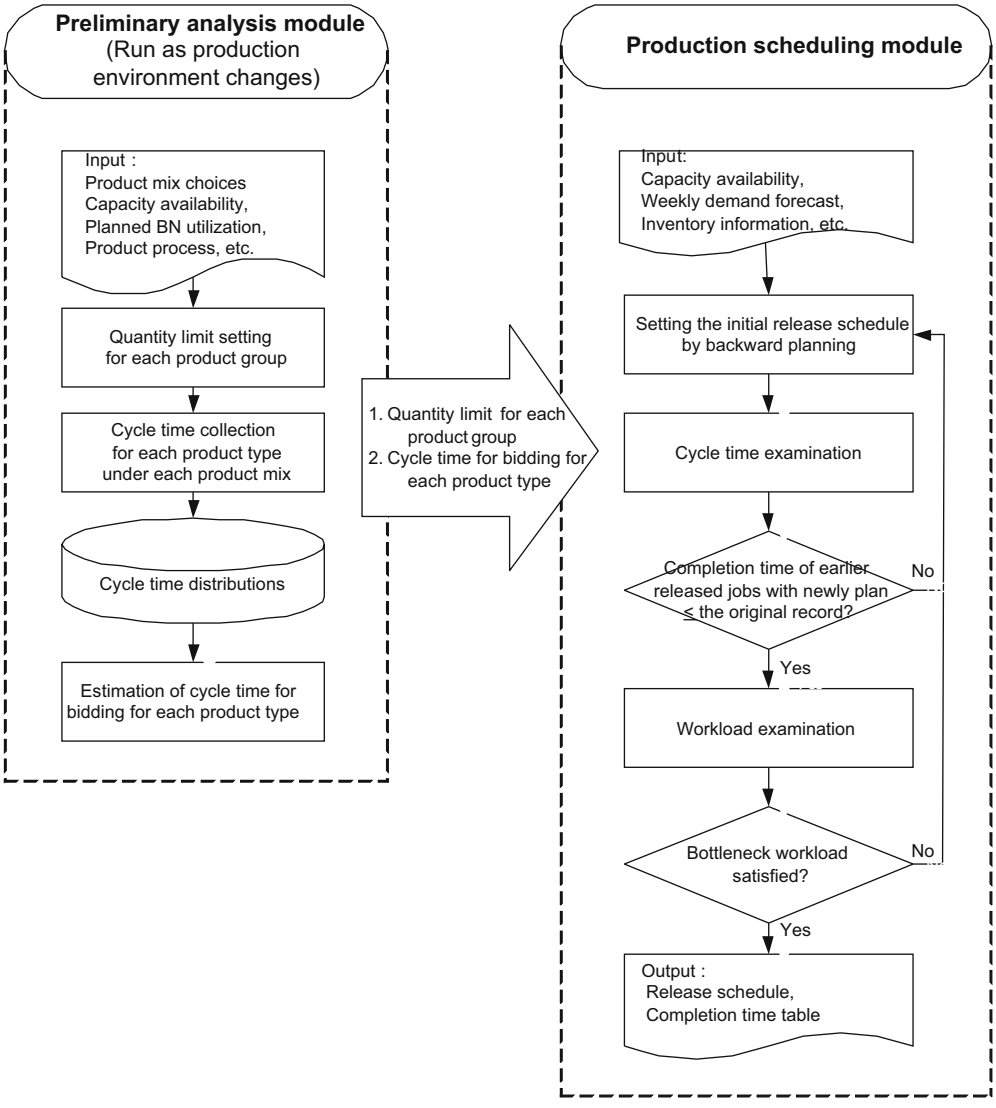


Fig. 1 The framework of the proposed system

2 Production system environment

For a make-to-stock environment, such as DRAM chip manufacturing, customer orders are met from the finished goods inventory. In such a situation, the primary goals of production planning are meeting demand forecast while maintaining production smoothness.

2.1 System configuration and notation

Modern fab requires a very high capital investment, which usually amounts to a billion dollars or more [16]. Generally, the wafer stepper machine is the most expensive machine in wafer fabrication factories. Due to the cost consideration, the capacity expansion of wafer stepper machines is carefully evaluated and is often limited. Therefore, the economic necessity of increasing return of capital investment makes the maintenance of a high utilization rate of stepper machines important. On the other hand, when the utilization rate of stepper machine is set too high, the system may tend to be unstable due to unforeseen disruptions. Therefore, in this paper, a planned utilization rate is kept in a certain range for the wafer stepper machine, which is treated as the bottleneck (BN) resource. In addition, to be consistent with the processing batch size of the thermal oxidation process so as to raise the utilization rate of equipment, the batch size of wafer release is six lots where one lot consists of 25 pieces of wafers.

The planning horizon for master production schedule is 12 weeks. The planning period is one week, and the product mix can only be changed weekly. Wafer lots are released under a CONWIP release policy. For the wafer lots already being processed in the plant, the first-in-first-out rule is used. The framework for the proposed system is shown in Fig. 1 and the notations used in the paper are defined in Table 1.

The proposed system comprises two modules: preliminary analysis module and production scheduling module. Preliminary analysis module has the objective of determining two sets of parameters that are used by the production scheduling module. These parameters are: (1) quantity limit for each product group, which is to provide the throughput upper bound for each product group in order to quickly examine the feasibility of new product mix, and (2) the cycle time for bidding for each product type, which is used as the base for setting the initial release schedule. In order to grasp the fluctuation of demand, several representative product mixes are chosen to run the simulation model for collecting cycle times. The α percentile cycle time for each product type is then determined as the cycle time for bidding. Those results will be the inputs to the production scheduling module.

In the production scheduling module, the release schedule is prepared with the considerations of meeting demand forecast, keeping production smoothness, and protecting due date commitment of the earlier released job orders. The initial release schedule is determined by backward planning based on the cycle time for bidding and demand forecast. Then, the final release schedule is set by forward examining completion date and bottleneck workload. The corresponding job order completion time will be planned in consequence.

Table 1 Mathematical notations

i	Index of product type;
l	Index of layer;
f	Index of product family;
t	Index of time period;
k	Index of workstation;
bat_k	Batch size setting for workstation k;
cap_k	Planned capacity of workstation k;
CTi	Mean cycle time for product type i;
CT_i^{BBCT}	Cycle time of product type i estimated by BBCT algorithm;
CT_i^{α}	α percentile cycle time for product type i;
$DF_{i,t}$	Demand forecast for product type i at time period t;
$DN_{i,t}$	Net demand for product type i at time period t;
$EI_{i,t}$	Ending inventory for product type i at time period t;
fpw	The first week that product completion through current release plan;
h	Planning horizon;
k_f	The workstation with the maximum difference in average theoretical process time between product family f and any other family;
L_t	System WIP level at time period t;
$L_{i,t}$	WIP level of product type i at time period t;
$L_{i,t}^l$	WIP level of layer l of product type i at time period t;
O	Weekly throughput target;
$Q_f^{\max,k}$	Maximum affordable quantity for product family f processed on workstation k;
$Q_f^{\max}$	The production quantity limit for product family f in the system;
Np_i	The number of time period needed based on planning cycle time;
$PT_f^{avg,k}$	Average theoretical process time for product family f processed on workstation k;
PT_i^k	Theoretical process time processed on workstation k for product type i;
PT_i	Theoretical process time of the whole process for product type i;
$LPT_{i,l}$	Theoretical process time for layer l of product type i;
$PR_{i,t}$	Planned release amount for product type i at time period t;
$SI_{i,t}$	Scheduled receipt for product type i at time period t;
U_k	Planned utilization rate of workstation k;
$\lambda_{i,t}$	Average system hourly arrival rate for product type i at time period t;
π_i	The product mix ratio for product type i;

2.2 Definition of circuitry layer

Since the characteristic of reentry complicates the material flow, Chung et al. [17] presents the circuit layer segmentation concept to simplify the schedule planning. As shown in Fig. 2, a wafer's process is divided into several subsets according to the positions of stepper machine operation.

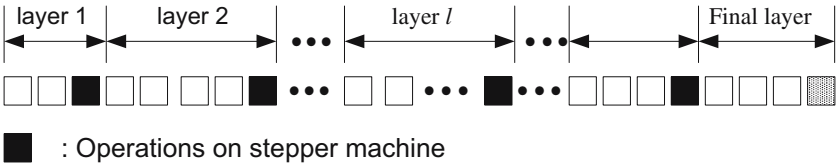


Fig. 2 The operation set of circuitry layer [17]

Except for the final layer that does not pass by the stepper machine, the last operation of all the other layers is processed on the stepper machine. Such a segmentation is very useful in exploiting the reentrant flow property; therefore, we utilize this concept in our proposed system to regard the circuitry layer as the object for WIP controlling and bottleneck workload monitoring.

2.3 Wafer release policy and WIP target setting

In order to keep production smoothness, the system WIP level requires to be properly distributed to all layers of each product type such that the fabrication of wafers will not be concentrated on some specific layers. Based on CONWIP release policy, wafer lots are released into the plant only when the WIP level is lower than the system WIP target. In the environment of periodical product mix changes, the planned throughput could also vary periodically and, consequently, the system WIP target may change with the time period. In such a situation, if we release wafer lots only according to the time when the WIP level is lower than the WIP target of the current period, the WIP level of the first layer in the beginning of each time period could increase or decrease drastically. Material flow is disturbed in consequence. For solving such a problem, two boundaries are set in the release policy: with the consideration of system WIP level, L_t, and WIP level of the first layer, L_t^1.

Under the revised CONWIP policy, once the system WIP level is lower than L_t and WIP level of the first layer is lower than L_t^1, six lots, the release batch size, of a product type which has the largest accumulated unreleased quantity is released into the plant. The calculation of "accumulated unreleased quantity" is based on the planned daily release amount. When the product type is assigned to releasing, six lots are deducted from the corresponding unreleased quantity. On the other hand, if there are remaining quantities not released to the plant, the unreleased quantities will be accumulated to the next day.

For each time period t, the suitable WIP level for each product i is determined by using Little's law [18], $L_{i,t} = \lambda_{i,t} \times CT_i$, where $\lambda_{i,t}$ is the average arrival rate and CT_i is the corresponding cycle time for the specific product type. The system WIP target, L_t, is calculated by summing up all the estimated values of $L_{i,t}$.

The distribution of the WIP level of all layers is related to the layer cycle time. That is, the amount of WIP for each layer is proportional to the length of the corresponding layer cycle time. WIP level of layer l of product i, $L_{i,t}^l$, is derived from the system WIP level multiplying by the proportion of layer cycle time to product cycle time, as shown in Eq. 2-1.

$$L_{i,t}^l = L_{i,t} \times \frac{LPT_{i,l}}{PT_i}, \text{ for each layer } l \text{ and each product type } i, \tag{2-1}$$

where $\sum_l LPT_{i,l} = PT_i$.

3 Determination of production parameter setting

The main work of the preliminary analysis module is to determine two sets of parameters that are used by the production scheduling module. Quantity limit for each product family is first determined in order to prevent bottleneck shifting and cycle time variance increasing when choosing a new product mix.

Several representative product mixes in the plant are then selected on the basis of the seasonal demand of each product type. For each product mix, cycle times of the job orders belonging to each distinct product type are collected by running a simulation model, which describes the system environment and operation behavior. Cycle time analysis is performed for each set of collected cycle times. After that, cycle time for bidding of each product type, defined as the α percentile cycle time of the corresponding product type, is determined as the input for planning lot release schedule. The α percentile cycle time is defined as the longest cycle time of $\alpha\%$ out of the wafer lots completed during a specific time period.

3.1 Setting quantity limit for product family

Product types of the same product family often have a similar manufacturing process. Releasing a great quantity of products belonging to the same product family at one time period may cause some specific workstations being overly demanded. This will result in bottleneck shifting and cycle time variation in shop floor. Therefore, the quantity limit for each product family f needs to be established before setting a new product mix.

The quantity limit setting is based on the capacity that can offer to produce the maximum quantities of the corre-

sponding product family in the system. Two kinds of capacity constraints must be considered, namely the system constraint and the family constraint. The system constraint comes from the capacity of bottleneck while the family constraint comes from the capacity of the workstation k_f, which has the maximum difference in average theoretical process time between product family f and any other families.

For each product family f, the quantity limit setting is as follows. Workstation k_f is first identified as shown in Eqs. 3-1 and 3-2. Then, the maximum affordable quantity for product family f processed on workstation k_f, $Q_f^{\max,k_f}$, is derived by dividing planned usable capacity of workstation k_f by the average process time for product family f processed on workstation k_f, PT_f^{avg,k_f}, as shown in Eq. 3-3. By using the same concept, the maximum affordable quantity for family f processed on BN, $Q_f^{\max,BN}$, can also be derived as shown in Eq. 3-4. Finally, Eq. 3-5 is to set the production quantity limit, $Q_f^{\max}$, by selecting the smaller number between $Q_f^{\max,k_f}$ and $Q_f^{\max,BN}$.

$$d\left(k,f\right) = \max\left\{\left|PT_f^{avg,k} - PT_{f'}^{avg,k}\right| \forall f' \notin f\right\} \tag{3-1}$$

$$k_f = \max_k \{d(k,f)\} \tag{3-2}$$

$$Q_f^{\max,k_f} = \left(cap_{k_f} \times U_{k_f} \times bat_{k_f}\right) \Big/ PT_f^{avg,k_f} \tag{3-3}$$

$$Q_f^{\max,BN} = \left(cap_{BN} \times U_{BN} \times bat_{BN}\right) \Big/ PT_f^{avg,BN} \tag{3-4}$$

$$Q_f^{\max} = \min\left(Q_f^{\max,k_f}, Q_f^{\max,BN}\right) \tag{3-5}$$

3.2 Simulation model description

According to the seasonal demand of each product type, several representative product mixes are selected. For each product mix scheme, cycle times are collected by running a simulation model, which is built with eM-Plant [19]. Actual production data of a wafer plant in Taiwan is used as the input data, including machine, product, and product routes information.

(1) Product-related data: Five products are classified into two product families: A and B belonging to the family of SDRAM, and C, D, and E belonging to the family of DDR. All product types have different process routes and each process contains process steps in a range of 276 to 338 operations. The required workstations and theoretical process times are known.

(2) Equipment-related data: There are 83 workstations (coded from w1 to w83) including 48 serial workstations and 35 batch workstations in which there are thirteen 6-lot workstations, three 4-lot workstations, and nineteen 2-lot workstations. The distributions of mean time between failures (MTBF), mean time to repair (MTTR), mean time between preventive maintenance (MTBPM), and mean time to preventive maintenance (MTTPM) for each workstation are known.

For each product mix scheme, simulation is run for collecting cycle times. Since the product mix is fixed all the time period for each simulation run, the throughput target of each time period would be equal. The weekly throughput target, O, is calculated based on the achievement of the planned bottleneck utilization. It is derived by dividing the bottleneck capacity by the theoretical process time on bottleneck according to the given product mix, as shown in Eq. 3-6.

$$O = \left(cap_{BN} \times U_{BN} \times bat_{BN}\right) \Big/ \sum_i \left(PT_i^{BN} \times \pi_i\right) \tag{3-6}$$

The system WIP target setting is by applying Little's law, as stated in Section 2.3. Since the CONWIP policy is adopted in the system, wafer lots are released into the plant only when equal quantity of wafers are finished and transferred out. Therefore, the arrival rate for each product type is equivalent to the corresponding throughput rate.

The block-based cycle time estimation algorithm (BBCT), developed by Chung and Huang [9], has a remarkable performance in cycle time estimation under the assumption that product mix is fixed. The BBCT algorithm estimates the cycle time by cutting the production process into several blocks based on the characteristics of material flow. In each block, the load factor waiting time and batch factor flow time are estimated. Finally, the cycle time of a product is calculated by summing up all the block cycle times related to the process of the corresponding product type. To utilize the advantages of quick response and satisfactory accuracy, the BBCT algorithm is applied to estimate the cycle time for each product type in the preliminary analysis.

For each simulation run, the system WIP target is determined as follows:

Step 1 Estimate the average hourly arrival rate, λ_i. It is derived by multiplying product mix ratio of product i, π_i, with planned hourly throughput.

$$\lambda_i = \pi_i \times \frac{O}{24 \times 7}, \text{ for each product type } i, \tag{3-7}$$

where 24 represents 24 hours per day, and 7 represents seven working days per time period (week).

Step 2 Estimate the cycle time of each product type, CT_i^{BBCT}, by applying BBCT algorithm. Input data include planned throughput, product mix, process plan, workstation related information and batch policy.

Step 3 Estimate the appropriate WIP level for product type i, L_i, by Little's law.

$$L_i = \lambda_i \times CT_i^{\mathrm{BBCT}}, \text{ for each product type } i. \tag{3-8}$$

Step 4 Setting the system WIP level, L, by summing up L_i values of all product type i.

$$L = \sum\nolimits_i L_i. \tag{3-9}$$

3.3 Cycle time collection and analysis

The simulation horizon is set to be 24 weeks, in which the first 12 weeks are the warm-up period. In order to eliminate simulation errors, 15 replications with different random seeds are run.

For each product type under each product mix, the histogram of the cycle time frequency distribution is plotted to figure out the pattern of the data. The related parameters, such as average, variance, and α percentile cycle time can be determined. The information is saved in the database and is the input to production scheduling plan.

Due to long cycle time, the job release time and job completion time may not belong to the same planning period. The cycle time of each job order thus may be affected by the product mix settings in successive periods. When estimating cycle time for bidding, we concern not only the effects of product mix composition but also the meeting of the planned on-time delivery rate under the environment that product mix changes periodically. In order to tackle the interference in periodical product mix changes, the α percentile cycle time, CT_i^{α}, is derived in order to ensure the achievement of planned on-time delivery rate.

4 Job releasing planning

The objective of the production scheduling module is to establish wafer lot release schedule and the completion timetable under the environment that product mix changes periodically. On the premise of meeting demand forecast, the initial release schedule is prepared by backward planning. In order to ensure the feasibility of the release schedule, cycle time examination of released job orders and workload examination are performed by forward planning.

4.1 Setting the initial release schedule by backward planning

The initial release amount is determined by net demands offset cycle time for bidding. The net demand of each product type in one specific time period is calculated by deducing the wafer lot scheduled receipt and the amount of beginning inventory from demand forecast. The scheduled receipt is the job orders that have been released earlier and planned to be finished in the corresponding time period according to the completion timetable.

Practically, in order to achieve a certain degree of stability in the production system under the demand fluctuating environment, the first period of the schedule is frozen, which means all but the most critical changes in this period cannot be made to the production schedule. Therefore, in the frozen period, if there is any shortage resulted from demand fluctuation or forecast error, the shortage will be backlogged.

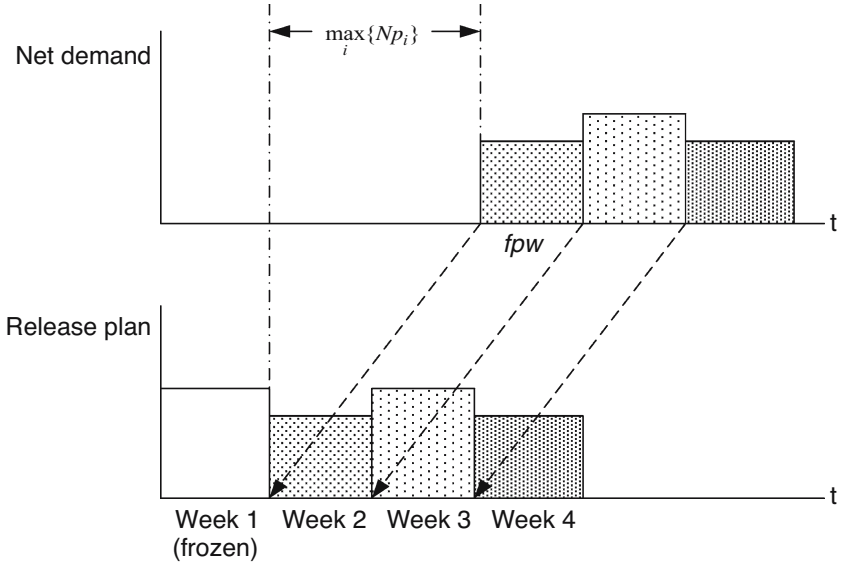


Fig. 3 An illustration of the first planning week

The 95-percentile cycle time in the proposed system is used as the cycle time for bidding. The time unit for cycle time collection in the simulation model is in hours while the planning period is in weeks; therefore, the number of time periods covered for producing product type i, Np_i, needs to be calculated first.

Since the job released to the plant at week one will be finished at the time period after the length of cycle time, the demands before this period are supplied by inventory and scheduled receipt. The first week that product completion through current release plan, *fpw*, would be in the week that the length of cycle time for bidding after the end of frozen period. Thus, we define *fpw* is $\max_i\{Np_i\}+1$, and we derive the net weekly demand in the planning horizon starting from *fpw*. Figure 3 shows an illustration of the first week that product completion through current release plan, *fpw*.

The procedures for deriving the initial release schedule are as follows:

Step 1 Calculate the number of time periods needed, Np_i, according to the length of 95-percentile cycle time, $CT_i^{95\%}$, for each product type i.

$$Np_i = \left\lceil \frac{CT_i^{95\%}}{24\times 7} \right\rceil, \text{ for each product type } i. \tag{4-1}$$

Step 2 Set the initial planning period as $t=fpw$, where $fpw = \max_i\{Np_i\}+1$.

Step 3 Calculate net demand, $DN_{i,t}$, and expected ending inventory, $EI_{i,t}$, for each product type i in the planning period t. Net demand, $DN_{i,t}$, is derived by subtracting beginning inventory, $EI_{i,t-1}$, and scheduled receipt, $SI_{i,t}$, from demand forecast, $DF_{i,t}$.

$$DN_{i,t} = \max\left[(DF_{i,t} - EI_{i,t-1} - SI_{i,t}),\ 0\right], \text{ for each product type, } i. \tag{4-2}$$

$$EI_{i,t} = \max\left[(EI_{i,t-1} + SI_{i,t} - DF_{i,t}),\ 0\right], \text{ for each product type } i. \tag{4-3}$$

$$t = t+1. \tag{4-4}$$

Step 4 Repeat step 3 until $t > fpw + h - 1$.

Step 5 Calculate the planned release amount for product type i at planning period t-Np_i, $PR_{i,t\text{-}Npi}$. It equals net demands, $DN_{i,t}$, offset by the number of time periods, Np_i.

$$PR_{i,t-Npi} = DN_{i,t}, \text{ for each product type } i \text{ and for } t = fpw \text{ to } fpw + h - 1 \tag{4-5}$$

Step 6 In each planning period, compare the planned release amount of each product family to the quantity limit for the corresponding product family, $Q_f^{\max}$. If the planned release amount is larger than $Q_f^{\max}$, we need to fine-tune the planned released quantity in the successive planning period.

4.2 Planning the wafer release time and sequence

The lot release time is related to the system WIP level and WIP level of the first layer under the revised CONWIP release mechanism. Once the system WIP level is lower than L_t and the WIP level of the first layer is lower than L_t^1, under the revised CONWIP policy, six lots of the product type which has the largest accumulated unreleased quantity is released into the plant. For each time period t, the suitable WIP level for each product i in each time period t is determined by using Little's law [18], $L_{i,t} = \lambda_{i,t} \times CT_i$, where $\lambda_{i,t}$ is obtained by dividing the weekly release amount, $PR_{i,t}$, by 24(hours/day)*7(days/week), and the cycle time for each product type, CT_i, is the mean cycle time derived in the preliminary analysis module. The system WIP target, L_t, is then calculated by summing up all the estimated values of WIP level for each product i, $L_{i,t}$. The WIP level of the first layer can also be obtained based on Eq. 2-1.

After the release sequence is determined, the release schedule can then be derived.

4.3 Cycle time examination

As product mix level affects the length of cycle time, to protect the due date commitment of earlier released job orders from the disturbance of new product mix, the cycle times of those job orders are examined to ensure that their new completion date will not be later than the date recorded in the completion table.

For each earlier released job orders, the remaining layers to be processed are identified first. By estimating the

Table 2 Product mix ratio and throughput target in the preliminary analysis

Product mix scheme	Product mix ratio (1/25)					Weekly throughput target (lot)
	Product A	Product B	Product C	Product D	Product E	
1	8	5	4	4	4	164
2	7	7	4	3	4	163
3	6	6	3	7	3	162
4	5	5	5	5	5	164
5	4	3	7	3	8	166

number of periods we still need to process the job order, we can estimate the layer cycle time based on the corresponding product mix. Then, we estimate the completion time of the job order by summing up the remaining layer cycle times.

4.4 Bottleneck workload examination

In wafer fabrication, wafer lots released to the plant at one time period will induce workloads in several time periods. To accomplish the planned bottleneck utilization and to have a leveled bottleneck workload, bottleneck workload is examined for each planning period. The circuit layer segmentation concept is adopted to estimate the bottleneck workload. Restated, the last operation of each layer is processed on the bottleneck. Based on the information of release date and layer cycle time, wafer release time and sequence, and bottleneck capacity required by each job order can be easily estimated. Finally, the daily capacity required can be estimated by summing up the required bottleneck capacity in the corresponding date.

5 Simulation investigation and verifications

This section demonstrates the process of the proposed system using actual production data of a wafer plant as input data.

5.1 System design for simulation

The simulation model is built with eM-Plant [19]. The simulation horizon is set to be 24 weeks, in which the first 12 weeks are the warm-up period. In order to eliminate simulation errors, 15 replications with different seeds are run.

5.2 The results of production parameters setting

5.2.1 Quantity limit for each product family setting

In the system, the planned bottleneck utilization rate is set as 90%. For preventing bottleneck shifting, the utilization rate of family constraint machine is set as 80%. W64 for SDRAM family and W29 for DDR family are selected as the family constraint machines based on Eqs. 3-1 and 3-2. The quantity limit of SDRAM family, $Q_{\mathrm{SDRAM}}^{\max}$, and that of DDR family, $Q_{\mathrm{DDR}}^{\max}$, are calculated based on Eqs. 3-3 to 3-5.

For SDRAM family,

$$Q_{\mathrm{SDRAM}}^{\max,\,\mathrm{W64}} = (73,627.83 \times 80\% \times 1)/420 = 140.24\,(\mathrm{lot}) \tag{5-1}$$

$$Q_{\mathrm{SDRAM}}^{\max,\,\mathrm{BN}} = (127,231.8 \times 90\% \times 1)/681 = 168.47\,(\mathrm{lot}) \tag{5-2}$$

$$Q_{\mathrm{SDRAM}}^{\max} = \min\left(Q_f^{\max,\mathrm{W64}}, Q_f^{\max,\mathrm{BN}}\right) = 140.24\,(\mathrm{lot}) \tag{5-3}$$

Table 3 Cycle time estimated by BBCT algorithm and WIP level

Product mix scheme	Estimated cycle time, CT_i^{BBCT} (hour)					WIP level	WIP level of the first layer
	Product A	Product B	Product C	Product D	Product E		
1	276.02	299.62	280.88	320.36	314.48	288	9.9
2	277.91	302.18	281.21	320.56	314.73	288	9.9
3	274.88	298.25	280.52	320.31	314.30	288	9.8
4	273.23	296.09	280.46	320.16	314.18	290	9.9
5	271.26	293.52	280.73	320.47	314.46	294	10.0

Table 4 Cycle time information of each product type from simulation (Unit: hour)

Product mix scheme		1	2	3	4	5
Product						
Product A	Mean CT	276.00	276.58	279.76	286.87	287.42
	$CT^{95\%}$	311.10	324.17	367.16	382.93	418.63
Product B	Mean CT	305.48	299.68	299.91	308.28	315.68
	$CT^{95\%}$	417.23	341.53	378.12	420.99	487.17
Product C	Mean CT	283.04	283.39	290.60	283.49	276.24
	$CT^{95\%}$	411.70	412.57	446.82	400.41	325.60
Product D	Mean CT	332.46	331.98	323.90	331.56	336.42
	$CT^{95\%}$	499.61	519.00	389.65	476.91	528.96
Product E	Mean CT	323.28	322.50	330.35	322.59	313.20
	$CT^{95\%}$	478.50	487.09	504.65	453.71	345.63

For DDR family,

$$Q_{\mathrm{DDR}}^{\max,\,\mathrm{W29}} = (51,950.48 \times 80\% \times 6)/922.67 = 270.26\,(\mathrm{lot}) \quad (5\text{-}4)$$

$$Q_{\mathrm{DDR}}^{\max,\,\mathrm{BN}} = (127,231.8 \times 90\% \times 1)/653.33 = 175.27(\mathrm{lot}) \quad (5\text{-}5)$$

$$Q_{\mathrm{DDR}}^{\max} = \min\left(Q_f^{\max,k_f}, Q_f^{\max,\mathrm{BN}}\right) = 175.27(\mathrm{lot}) \quad (5\text{-}6)$$

5.2.2 Cycle time collection

With the consideration of seasonal demand of each product type, five sets of representative product mixes in the plant are selected as the product mix schemes to collect cycle time information in this section. For each product mix scheme, the weekly throughput target can be obtained based on Eq. 3-6. The product mixes and weekly throughput target are shown in Table 2.

In order to determine the system WIP level and WIP level of the first layer, the BBCT algorithm is applied to estimate mean cycle time of each product type for each product mix scheme. System WIP level is then determined based on Eqs. 3-7 to 3-9, and WIP level of the first layer is determined based on Eq. 2-1. The mean cycle time estimated by BBCT algorithm, system WIP level, and WIP level of the first layer for each product mix scheme are shown in Table 3.

Once the related inputs are determined, the simulation model is run for collecting the information of cycle time of each product type under each product mix scheme. The mean cycle time and 95-percentile cycle time of each product mix scheme collected from simulation are shown in Table 4. The mean cycle time from simulations are compared with that from the BBCT algorithm. As shown in Table 5, the percentage in error in cycle time estimation is between-5.62 to 1.63%. Based on the analysis, BBCT algorithm has a satisfactory result in estimating mean production cycle time when the product mix is fixed all the planning periods.

As fore-mentioned, cycle time and cycle time spread are strongly influenced by product mix level. Taking cycle time of product A for example, the wide variations in cycle time can be seen in Fig. 4a– e.

By summarizing the collected cycle times, the cycle time for bidding of each product type can be obtained and number of the time period covered for producing each product type, Np_i, can then be calculated based on Eq. 4-1. The information is shown in Table 6.

5.3 The results of job release planning

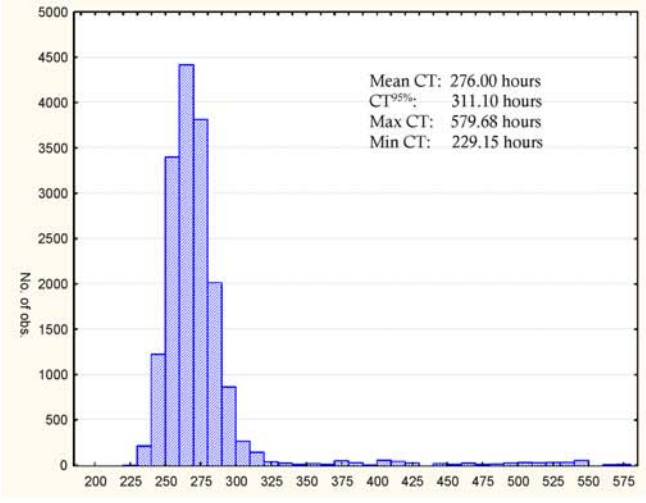
5.3.1 Initial release schedule setting

The frozen period is set as one week and $\max_i\{Np_i\}$, as shown in Table 6, equals three-week, therefore, the first planning week to calculate net demand is the fourth week. Based on Eqs. 4-2 to 4-4, the net demand and expected ending inventory for each product type for week 4 to week 16. As Np_i for each product type i is three week, the

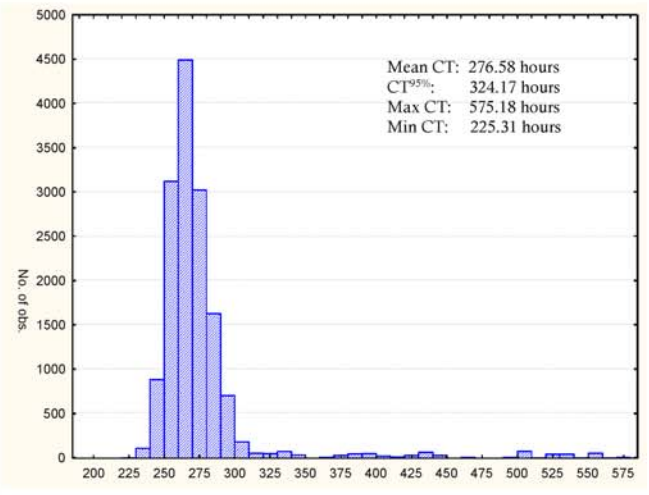
Table 5 Error between cycle time estimated from BBCT and from simulation (%)

Product mix scheme	1	2	3	4	5
Product					
Product A	0.01	0.48	–1.74	–4.75	–5.62
Product B	–0.02	0.01	–0.01	–0.04	–0.07
Product C	–0.76	–0.77	–3.47	–1.07	1.63
Product D	–3.64	–3.44	–1.11	–3.44	–4.74
Product E	–2.72	–2.41	–4.86	–2.61	0.40

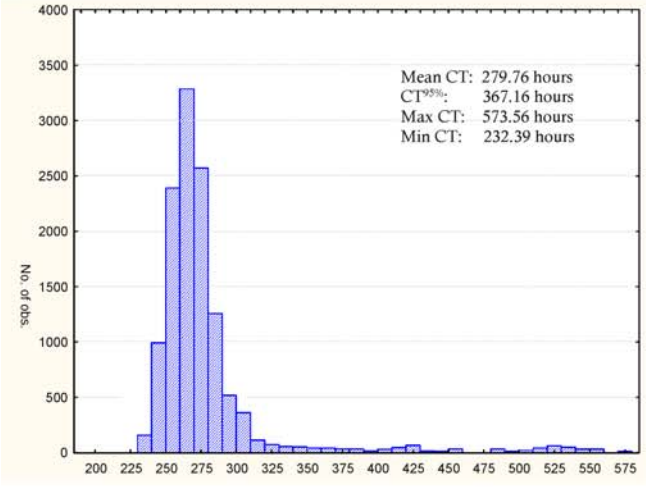
Error%=(CT from BBCT – CT from simulation)/CT from simulation 100%



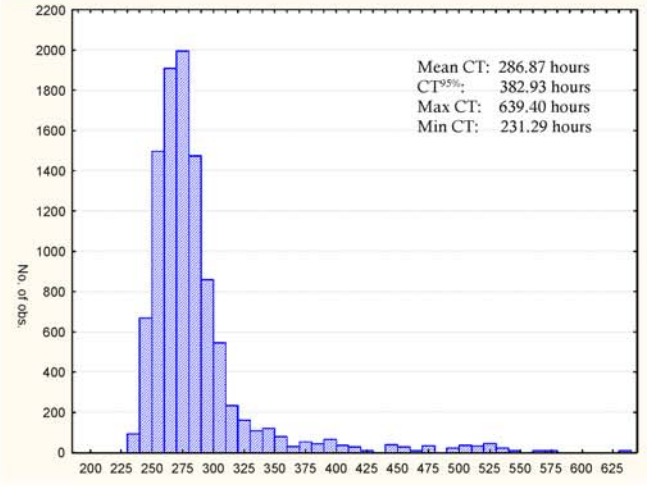
a product mix scheme 1



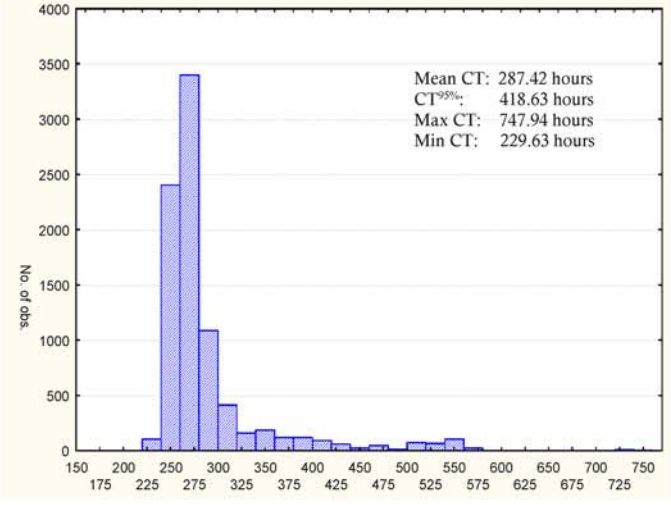
b product mix scheme 2



c product mix scheme 3



d product mix scheme 4



e product mix scheme 5

Fig. 4 The histogram of cycle time of product A under each product mix scheme

Table 6 Cycle time for bidding (hour)

Product type	Product A	Product B	Product C	Product D	Product E
Mean cycle time	280.18	304.31	282.00	329.99	320.39
Cycle time for bidding	364.58	412.58	404.09	489.22	459.82
Np_i	3	3	3	3	3

planned release amount for each product equals net demand offset by three weeks while the planned release amount of week 1 remains the same as determined by previous planned. The net demand and release plan are shown in Table 7.

All the planned release amount of each product family is smaller than the quantity limit of the corresponding product family; the next step is to determine the system WIP level of each week by using Little's law [18], $L_{i,t} = \lambda_{i,t} \times CT_i$. The average arrival rate of each product type in each week, $\lambda_{i,t}$, is obtained by dividing the weekly release amount by 24(hours/day)*7(days/week), and the mean cycle time for each product type, CT_i, is shown in Table 6.

The release plan also passes the cycle time examination and workload examination; therefore, the release schedule and completion time table is then derived.

Table 7 Net demand and release plan

Net demand				(lot)										
Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	–	–	39	46	52	46	39	33	27	33	39	46	52	46
B	–	–	39	46	33	46	39	33	20	33	39	46	33	46
C	–	–	19	26	26	26	19	33	46	33	19	26	26	26
D	–	–	45	20	26	20	45	33	20	33	45	20	26	20
E	–	–	19	26	26	26	19	33	53	33	19	26	26	26
Total	–	–	161	164	163	164	161	165	166	165	161	164	163	164

Release plan			(lot)									
Week	1	2	3	4	5	6	7	8	9	10	11	12
A	39	46	52	46	39	33	27	33	39	46	52	46
B	39	46	33	46	39	33	20	33	39	46	33	46
C	19	26	26	26	19	33	46	33	19	26	26	26
D	45	20	26	20	45	33	20	33	45	20	26	20
E	19	26	26	26	19	33	53	33	19	26	26	26
Total	161	164	163	164	161	165	166	165	161	164	163	164

Note: weeks 1 in release plan is the frozen period

Table 8 Accumulative throughput from release plan and simulation (lot)

Week		3	4	5	6	7	8	9	10	11	12
Product											
A	Release plan	468	507	553	605	651	690	723	750	783	822	868
	simulation	472	508	560	610	652	692	719	749	790	830	882
B	Release plan	468	507	553	586	632	671	704	724	757	796	842
	simulation	475	518	559	593	643	678	710	733	771	808	852
C	Release plan	228	247	273	299	325	344	377	423	456	475	501
	simulation	225	247	274	298	323	347	385	421	448	470	497
D	Release plan	540	585	605	631	651	696	729	749	782	827	847
	simulation	554	592	614	639	664	706	738	760	795	833	855
E	Release plan	228	247	273	299	325	344	377	430	463	482	508
	simulation	226	247	273	299	324	343	379	425	455	475	503
Total	Release plan	1932	2093	2257	2420	2584	2745	2910	3076	3241	3402	3566
	simulation	1952	2112	2280	2439	2606	2766	2931	3088	3259	3416	3589

5.3.2 Results of the release schedule

To verify the effectiveness of the release schedule, the simulation model is run for 15 replications with different seeds. As shown in Table 8, the release plan can not only achieve the throughput of each product in each week on time or earlier, but can keep production smoothness by well controlled bottleneck utilization. Based on the results, we can conclude that the proposed system is effective and efficient.

6 Conclusions

Quick response to customers' fluctuating demand is one of the critical issues for market competence. This paper has presented a planning system, which comprises two modules, on the make-to-stock basis under periodical product mix changes for the environment with volatile demand. With preliminary analysis, the quantity limit for product family is set to provide a fast check of the availability of a new product mix for preventing bottleneck shifting. In addition, mean cycle time and cycle time distributions are analyzed for all the possible different product mix choices so as to decide the cycle time for bidding. The production scheduling module is to plan a release schedule and completion time table with the considerations of demand variation, production smoothness and due date protection for released job orders. The output of this module includes the suitable WIP level, release amount for each product type, release schedule, and completion timetable.

The example cases showed that the proposed system can effectively solve the product mix setting problem for the make-to-stock wafer fabrication under the demand fluctuating environment. We have demonstrated that product mix changes have a great impact on cycle time distributions. In such a case, when setting completion time of wafer lots, using cycle time for bidding determined in the proposed system would be more appropriate than using mean cycle time plus safety allowance. We also present a revised CONWIP release policy, which can effectively keep production smoothness.

The simplicity and ease of implementation make the proposed system useful for demand management and production control. Future research can stress on the product mix optimization for each planning period under demand fluctuating make-to-stock environment.

Acknowledgement This paper was supported by the National Science Council, Taiwan, R.O.C., under the contract NSC92-2213-E-009-044.

References

1. Chou YC, Hong IH (2000) A methodology for product mix planning in semiconductor foundry manufacturing. IEEE Trans Semicond Manuf 13(3):278–285
2. Dümmler MA (2000) Analysis of the instationary behavior of a wafer fab during product mix changes. Proc 2000 Winter Simulation Conference, pp 1436–1442
3. Sivakumar AI, Chong CS (2001) A simulation based analysis of cycle time distribution, and throughput in semiconductor backend manufacturing. Comput Ind 45:59–78
4. Lu SCH, Ramaswamy D, Kumar PR (1994) Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. IEEE Trans Semicond Manuf 7(3):374–388
5. Lawrence SR (1995) Estimating flow times and setting due-dates in complex production systems. IIE Trans 27(5):657–668
6. Matsuyama A, Atherton RW (1990) Experience in simulation wafer fabs in the USA and Japan. 1990 International Semiconductor Manufacturing Science Symposium
7. Glynn PM, O'Dea M (1997) How to get predictable throughput times in a multiple product environment. Semiconductor Manufacturing Conference Proc, 1997 IEEE International Symposium, pp 27–30
8. Raddon A, Grigsby B (1997) Throughput time forecasting model. 1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference, pp 430–433
9. Chung SH, Huang HW (1999) The block-based cycle time estimation algorithm for wafer fabrication factories. Int J Ind Eng 6(4):307–316
10. Liu C, Thongmee S, Hepburn P (1995) A methodology for improving on-time delivery and load levelling starts. Proc IEEE/SEMI Advanced Semiconductor Manufacturing Conference, pp 95–100
11. Hackman ST, Leachman RC (1989) A general framework for modeling production. Manage Sci 35(4):478–495
12. Chu SCK (1995) A mathematical programming approach towards optimized master production scheduling. Int J Prod Econ 38:269–270
13. Burman DY, Gurrola-Gal F, Nozari JA, Sathaye S, Sitarik JP (1986) Performance analysis techniques for IC manufacturing lines. AT&T Technical J, pp 46–56
14. Thompson SD, Davis WJ (1990) An integrated approach for modeling uncertainty in aggregate production planning. IEEE Trans Syst Man Cybern 20:1000–1012
15. Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on interactive simulation and linear programming calculations. IEEE Trans Semicond Manuf 9(2):257–269
16. Hood SJ, Bermon S, Barahona F (2003) Capacity planning under demand uncertainty for semiconductor manufacturing. IEEE Trans Semicond Manuf 16(2):273–280
17. Chung SH, Yang MH, Cheng CM (1997) The design of due-date assignment model and the determination of flow time control parameter for the wafer fabrication factories. IEEE Trans Compon, Packag, Manuf Technol 20:278–287
18. Little JDC (1961) A proof for the queueing formula L=L=λW. Oper Res 9:383–387
19. Tecnomatix Technologies Ltd. (2000) eM-plant pbjects manual. Tecnomatix Software Company, Germany