A Methodology for Product Mix Planning in Semiconductor Foundry Manufacturing

Yon-Chun Chou and I.-Hsuan Hong

***Abstract*—Since a semiconductor foundry plant manufactures a wide range of memory and logic products using the make-to-order business model, the product mix is an important production decision. This paper first describes the characteristics of the product mix planning problem in foundry manufacturing that are attributable to the long flow time and queuing network behaviors. The issues of time bucket selection, mix optimization and bottleneck-based planning are next addressed. A decision software system based on integer linear programming techniques and a heuristic procedure has been implemented for mix planning. Data provided by a wafer plant has been used to study problems related to product mix planning. It was determined that the suitable time bucket of planning is either one week or one month and the lead-time offset factor should be included in the logic of workload calculation. This paper also presents various facets of product mix decisions and how they should be integrated with operations management.**

***Index Terms*—Bottleneck analysis, product mix planning, semiconductor foundry manufacturing.**

I. Introduction

PROCESS and machine technologies change rapidly in the semiconductor manufacturing industry. Multiple generations of technologies usually coexist in a manufacturing plant. In general, a semiconductor foundry plant has more than one hundred machine groups and three to four hundred machines in total. A foundry plant manufactures a wide range of memory and logic products using the make-to-order business model. The product types are not fixed but evolve with the time and the technology portfolio of the plant. The quantity of product types is influenced by the economics of both scale and scope and could number in the hundreds. Typically, the manufacture of a product requires several hundred processing steps and a machine group may be utilized more than once as successive circuit layers are added. This phenomenon of multiple visits to a machine group is commonly referred to as the reentry property of the routing [5]. The process routings of different products may differ significantly in the machines to be visited and in the processing times spent on the machines. A machine group is usually shared by many processing steps but, due to process requirements, some machines may be dedicated to certain process steps. The processing time requirements of the same machine group by different products may differ by as much as 100%. The large varieties of processes, machines and products lead to a conspicuous mix-planning problem in semiconductor foundry plants.

Product mix planning is a common problem in many industries. Besides strategical planning, it involves two issues at the operation level: cost accounting of capacity at the process step level [3] and the optimization of product mixes. The objective of cost accounting is to accurately estimate the manufacturing cost of each product type. This issue is important when the overhead costs need to be correctly attributed to the manufacturing activities associated with each product. Since capital investment and sunk costs account for the largest portion of the manufacturing cost in a semiconductor plant, the overhead cost accounting is not critical for the purposes of product mix planning. The second issue, the optimization of product mixes, seeks to maximize the efficiency of capacity allocation across products. Mixed integer linear programming techniques are easily applied to this problem. The manufacture of a product requires a certain amount of each type of resource. Since the resources are limited and the profits of products vary, the optimization of the product mix can be modeled as a combinatorial optimization problem. Recently, the theory of constraints was applied to product mix planning [2], [6]. However, it has been shown by numerical examples that both methods, although differing in their implementation procedure and rigorousness, are conceptually equivalent and could lead to the same solutions [7].

A semiconductor wafer plant, comprising hundreds of machines and automated material handling systems, exhibits complex queuing network behaviors [1]. The performance measures of flow time, machine utilization, work-in-process inventory, and throughput are heavily interrelated. The product mix problem in the semiconductor foundry industry has a number of unique characteristics that can be attributed to the queuing network behaviors of the plant. First, the process routing is long; the average flow time of a wafer lot is usually more than one month. If a time bucket that is smaller than the flow time is used in planning, wafer lots released to the plant in one time period will introduce workloads to several time periods. Because a wafer lot encounters significant and uncertain queuing delays as it moves through the shop floor, predicting with accuracy the workload by machine and by time would be a challenging task when a small time bucket is used. Second, production bottlenecks of the plant usually shift from one group of resources to another. If a large time bucket, such as two months, is used in planning, the effect of shifting bottlenecks will be overlooked. Table I illustrates the dilemma between using a small time bucket and a large time bucket.

Manuscript received May 5, 2000. This work was supported in part by United Integrated Circuits Corporation of United Microelectronic Corporation and the National Science Council of R.O.C. under Grants NSC 88-2212-E002 and 89-2212-E003.

Y.-C. Chou and I.-H. Hong are with the Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan 106, R.O.C. (e-mail: ychou@ccms.ntu.edu.tw).

Publisher Item Identifier S 0894-6507(00)06362-4.

TABLE I
EFFECT OF THE TIME BUCKET ON BOTTLENECK IDENTIFICATION

	Period 1	Period 2	Aggregate period (Periods 1 and 2)
Product A	260	310	570
Product B	360	270	630
Total capacity	600	600	1200

The table shows the machine time requirements of two product orders (A and B) and the capacity of a machine. If a finer time bucket (the second and third columns) is used and the workload can be predicted with accuracy, capacity violations (e.g., in Period 1) can be pinpointed, whereas the workload variation will be evened out if a coarser time bucket (the fourth column) is used. The third characteristic is concerned with the use of product mix information. Because the performance trade-off is complex in wafer plants, the product mix decision should not be treated as a rigid production schedule to be adhered to by decree. Engineers and managers on the shop floor have access to real-time information that could be used in order to dynamically enhance operation efficiency and productivity. That is to say, product mix planning as a decision task should not be separated from shop floor management. This perspective will affect how product mix planning should be done and will be elaborated on in later sections.

This paper presents a production mix planning methodology for semiconductor foundry manufacturing. The remainder of the paper is organized as follows. An analysis of the problem is presented in Section II. In Section III, the effect of time bucket granularity on planning accuracy is analyzed to determine a suitable size of time bucket for product mix planning. In Section IV, mixed integer linear programming models for multiple, heterogeneous classes of products are outlined. The integration of product mix planning with shop floor management is also presented. In Section V, a procedure based on bottleneck analysis is described to address the issue of large problem size. Discussions of numerical cases and conclusions are presented in Section VI, covering the necessity of cycle time offset, production smoothing, and tool backup.

II. PROBLEM ANALYSIS

The problem of product mix planning can be described as follows. There is a pool of customer orders to be selected for production over a certain horizon. The orders are either confirmed orders or forecasts of generic products. Although the quantity of distinct product types could number in the hundreds, some of them have similar routings. If technology and machine time requirements are used to characterize product types, the quantity could be reduced to several scores (of generic products). Because the production flow time is long, products may be represented as generic products to reserve production capacity of the later times of the planning horizon. The order pool is rolled forward when it is updated and the generic products eventually will be replaced with distinct product types.

TABLE II
PRODUCT CATEGORIES

Category	Demand type	Volume	Price
O1	Committed orders	Fixed	Fixed
O2	Fixed make-to-order	Fixed	Fixed
O3	Variable make-to-order	Range	Fixed
O4	Make-to-stock	Variable	Variable

For the purposes of planning, four categories of product demands are distinguished in our planning model: committed orders, fixed make-to-order demand, variable make-to-order demand and make-to-stock orders (Table II). Factory capacity must be reserved for committed orders as a matter of course. The unit prices of make-to-order demands are fixed but the volumes could either be fixed, according to contracted delivery terms, or be flexible to allow maneuvering room for enhancing operation efficiency. The third category (O_3) plays an important role in integrating product mix planning with shop floor operation management. Using this scheme of demand categorization, an O_3 order may become an O_2 order and then O_1 order as the order pool is rolled forward. Make-to-stock items are standard products such as memory devices. In general, memory devices require more advanced technologies. They are included in the order pool in order to fill up factory capacity and to drive the manufacturing technology. As for engineering lots, it is better to set aside machine capacity for process and product development work, as their schedule is much less predictable.

The robustness and feasibility of product mix decisions, that is, whether they can be achieved, predicate on the accuracy of workload estimation. There are several implementations of static capacity models in the literature that take into consideration the factors of lead-time offset, process yield and the efficiency of resource utilization [4], [8]–[10]. The computation logic is straightforward but there still remains a question of tradeoff between precision and computation time.

In addition to the time bucket of planning that is mentioned above, two other time buckets must be distinguished. One is associated with the work release frequency and another with the lead-time offset. In a foundry plant, jobs are released to the shop on a daily basis and the flow time is an importance performance measure that is controlled with diligence. In a well-managed plant, the flow time for each layer of circuits is less than two days with a high level of certainty. Therefore, the time bucket for lead-time offset should be one day. Candidate customer orders in the order pool are part of the input data to product mix planning. However, it may not be suitable to use one day as the time bucket for expressing the timing of those orders as the quantity of an order is usually too large to be released in the same day. On the other hand, if the fact of daily release is not modeled in one way or another in product mix planning, the feasibility of resultant mix decisions may be negatively impacted. The release frequency to be modeled in product mix planning is a parameter to be determined in later sections. This frequency will be called the *pseudo release* frequency. It may be equal to or greater than the actual release frequency.

In building a planning model, each customer order is divided into one or more work releases. Customer orders and work releases will be called order batches and release batches respectively in the remainder of this paper. Order and release batches are characterized by identifier (b), quantity (D), product type (i), and due time (t) attributes. The due time for each release batch is derived from the pseudo release frequency and the due time of the demand batch to which it belongs. A static capacity model is needed to calculate the workload primitives $(w_{b,j,k,t})$ for each unit of product that are due to the release batch b and process step j for machine group k in time t, taking into consideration the lead-time offset of each step. In calculating workload primitives, the time bucket should be one day to match the granularity of flow time data.

Because multiple steps may share the machine group (recalling the reentry property), workload primitives for the same machine group in the same time bucket of planning are summed up in the mix optimization stage of planning. These sums of workloads are called *aggregate* workloads $(W_{b,k,p})$. Here, a different subscript (p) for the time is introduced to represent the time buckets of planning. The aggregation is over all steps; therefore, the subscript j is no longer needed. It should be understood that the workloads generated by a release batch, if it is selected, would equal its aggregate workloads multiplied by its batch quantity. Let B be the set of release batches selected. The total workload $\Omega_{k,p}$ will equal to $\sum_{b\in B} D_b \cdot W_{b,k,p}$.

There are two parameters in the calculation procedure. One is the frequency of work releases and another is the time bucket of planning (p). The computation time would be less if a large time bucket of planning is used but a small time bucket will allow bottleneck machines to be better pinpointed. Similarly, if a high frequency of pseudo work releases were adopted, the prediction of workloads would be more accurate at the expense of longer computation time.

III. Precision Workload Calculation

In this section, the detailed procedure of workload calculation is presented, followed by the results of our study on the time bucket size of planning and the pseudo release frequency. Because of the reentry property, the machine subscript will be represented in a function form as $k(i,j)$ to indicate that k depends on both i and j, and that there is a many-to-one relationship between steps and machines. Similarly, the lead-time offset is represented in a function form $l(i,j)$ for the step j of product i. For each release batch of product i in time t, D_{it}, workload primitives are generated for each machine group. This step can be symbolically represented as

$$D_{it} \rightarrow\rightarrow w_{i,\{j\},k(i,j),t-\boldsymbol{l}(i,j)}$$

where the w represents a workload primitive, the double arrow implies that one or more items of workload primitives are generated, the $\{j\}$ represents the set of all process steps of product i, the $k(i,j)$ represents the required machine for (i,j), and the term $t-l(i,j)$ indicates that the occurrence time for the workload primitive is the due time t offset backward by the lead-time $l(i,j)$. Let $J(i)$ be the last step of product i, $p_{i,j,k}$ be the processing time, and $\mathrm{sy}_{i,j}$ be the yield of step (i,j). For each release batch D_{it}, there will be $J(i)$ workload primitives in total. Each workload primitive is identifiable by product-step pair (i,j) and its occurrence time is $t-l(i,j)$. The effect of the process yield is to increase the workload. Therefore, capacity allowances must be provided. Adjusting for yield allowances, the workload primitives are calculated as

$$w_{i,j(i),k(i,j),t(i,j)} = \frac{p_{i,j,k}}{\mathrm{ya}_{i,j,t}} \quad \forall i,j,k,t \tag{1}$$

where

$$\mathrm{ya}_{i,j} = \mathrm{sy}_{i,j} * ya_{i,j+1} \quad j = 1,\ldots,J(i)$$
$$\mathrm{ya}_{i,J(i)} = \mathrm{sy}_{i,j}.$$

Here, yield allowance (ya) for each step is computed backward from the last step to the first step of the process routing. The yield allowance for the last step is set equal to its step yield. The yield allowances for all other steps are iteratively accumulated backward from the last step.

To identify the suitable granularity of product mix planning, four time bucket sizes of planning and four release frequencies have been compared. Let G-1, G-2, G-4, and G-28 represent the time bucket size of four weeks, two weeks, one week and one day, respectively. And let F-1, F-2, F-4 and F-28 represent the work release frequency of quad-weekly, biweekly, weekly and daily. (The larger the number, the finer the granularity.) Together, these make up sixteen granularity schemes of planning. For each stationary product mix (i.e., a repeating set of release batches), it can be shown that the average of the total workloads $\Omega_{k,p}$ for each machine group k is the same for all granularity schemes. After all, the same set of jobs is released for production. However, the variation of the total workloads is not identical.

Using actual process routing, product and demand data provided by a foundry plant (Section 6), an empirical study has been done to evaluate the effect of the granularity level on the accuracy of workload estimation. The squared coefficient of variation (SCV) is chosen to be the measure of workload variability [(2)]. The results are summarized in Table III.

$$\mathrm{SCV}_k(\Omega_{k,p}) = \frac{E_k\{\Omega_{k,p} - E_k[\Omega_{k,p}]\}^2}{E_k^2[\Omega_{k,p}]} \quad \forall k \tag{2}$$

where E_k is the partial expectation over the time dimension p.

It can be observed from Table III that, for each time bucket size, no further accuracy on workload estimation can be gained by adopting a higher release frequency and additional noise will be introduced by adopting a lower release frequency. Thus, it is concluded that the work release frequency should be consistent with the size of time bucket.

Four granularity schemes (G1-F1, G2-F2, G4-F4 and G28-F28) on the diagonal of Table III remain to be compared. The Gregorian calendar is used to redefine the granularity schemes for its practical appeals. The four granularity levels of the new setting are one month (M), half a month (HM), one week (W) and one day (D). Because months do not have the same number of days, the work release will be less regular than the previous setting. This is not undesirable since the actual work release in the plant is usually not completely periodic.

TABLE III
WORKLOAD VARIATION FOR DIFFERENT GRANULARITY SCHEMES

	F-1	F-2	F-4	F-28
G-1	0	0	0	0
G-2	0.139	0	0	0
G-4	0.416	0.154	0	0
G-28	4.645	2.162	1.095	0

In addition, the effect of this additional source of variation on the four schemes can be evaluated in order to compare their robustness. In principle, the scheme D requires the least computation time, but is the least capable of capturing the variation of workload as daily workloads are aggregated into monthly buckets. The scheme M is favorably biased in smoothing out the workload variation. If the total workload $\Omega_{k,p}$ for each machine group k is regarded as a random variable in scheme D, the corresponding total workload in scheme M will be the sum of approximately thirty such variables. This bias, however, must be normalized. The normalization factors are shown in the third row of Table IV. The evaluation of robustness has three steps. The total workloads of the pseudo-released jobs are first calculated. The average of $\mathrm{SCV}_k(\Omega_{k,p})$ is next calculated. Finally, a magnification ratio is computed of the two quantities, with the aggregation bias of large time buckets normalized

$$magnification = \frac{\text{the average of } \mathrm{SCV}_k(\Omega_{k,p})}{\text{SCV of released work} \cdot \text{normalization}}. \quad (3)$$

A large value of *magnification* means that the variation in released work is magnified and shows up in variation of the total workloads at individual machine groups. The results are summarized in Table IV. The magnification ratios of scheme M and scheme HM are comparable, and so is that for schemes W and D. However, the schemes HM and D require higher computation load compared with schemes M and W, respectively. Therefore, it is concluded that either one week or one month should be used as the time bucket size. In the remainder of this paper, the time bucket size is set to one week, and the time subscript p will be replaced by the subscript t. Therefore, the symbol $W_{b,k,t}$ will be used, instead of $W_{b,k,p}$, to refer to the total workload contributed by batch b and the total workload $\Omega_{k,p}$ will be rewritten as $\Omega_{k,t}$.

IV. PRODUCT MIX OPTIMIZATION

Several objectives are of interest in product mix optimization: to maximize profit, to maximize wafer output, to maximize tool utilization or to maximize a hybrid model of profit and utilization. A mixed integer linear program (MILP) is used to determine the optimal product mix and mix ratio (i.e., type and volume). The decision variables, parameters, objective functions and constraints are as follows.

Decision Variables:

XF_b: 0–1 variable for $b \in O_2; X_b = 1$ if batch b is selected, $X_b = 0$ otherwise.
XR_b: Quantity of batch b, $b \in O_3$.

TABLE IV
WORKLOAD SENSITIVITY OF DIFFERENT GRANULARITY SCHEMES

Granularity Level	M	HM	W	D
Input variation	0.0009	0.0015	0.0042	0.0
Normalization factor	$\frac{\sigma^2}{30\mu^2}$	$\frac{\sigma^2}{14\mu^2}$	$\frac{\sigma^2}{7\mu^2}$	$\frac{\sigma^2}{\mu^2}$
Average $SCV_k(\Omega_{k,p})$	0.0041	0.0146	0.0144	0.0
Magnification	4.6795	4.5541	0.7954	-

XV_b: Quantity of batch b, $b \in O_4$.
R_{mkt}: Fraction of tool m to back up tool k in period t.

Parameters:

CM_i: Profit margin of product type i of O_2 and O_3.
CM_{it}: Profit margin of product type i of O_4 in period t.
Q_b: Quantity of batch b for O_2 orders.
Q_b^u: Maximal quantity of batch b for O_3 orders.
Q_b^l: Minimal quantity of batch b for O_3 orders.
E_{mk}: Backup efficiency of tool m with respect to tool k
CP_{kt}: Capacity of tool k in time period t.
S_{kt}: Residual capacity on tool k in period t.

In a foundry manufacturing environment, the objective of product mix planning is usually not fixed but changes with the business environment. Four basic objective functions can be identified:

1) ***Maximizing profit:***

$$\mathrm{Max}\, F_1 = \sum_{b\in O_2} \mathrm{CM}_{i(b)} \cdot Q_b \cdot XF_b + \sum_{b\in O_3} \mathrm{CM}_{i(b)} \cdot XR_b + \sum_{b\in O_4} \mathrm{CM}_{i(b),t(b)} \cdot XV_b.$$

2) ***Maximizing wafer output:***

$$\mathrm{Max}\, F_2 = \sum_{b\in O_2} Q_b \cdot XF_b + \sum_{b\in O_3} XR_b + \sum_{b\in O_4} XV_b.$$

3) ***Maximizing machine utilization:***

$$\mathrm{Min}\, F_3 = \sum_t \sum_k S_{kt} \quad \text{or} \quad \mathrm{Min} \sum_t \sum_k C_k \cdot S_{kt}$$

4) ***Hybrid model to maximize profit and utilization:***

$$\mathrm{Max}\, F_1 - \sum_t \sum_k P_k \cdot S_{kt}.$$

5) ***Hybrid model to maximize profit and output:***

$$\mathrm{Max}\, F_1 + C_{\mathrm{opp}} \cdot \mathrm{AM}_t \left(\sum_{b\in O_2} Q_b \cdot XF_b + \sum_{b\in O_3} XR_b + \sum_{b\in O_4} XV_b \right).$$

To maximize machine utilization is equivalent to minimizing the total residual capacity of all machines. Since machines are

not equal in cost (C_k) and criticality, a weight could be associated with each machine group in the third objective function. In the fourth objective function the penalty (P_k) for residual capacity is set equal to the weighted unit profit multiplied by machine throughput rate. The opportunity cost (C_{opp}) and the average margin (AM_t) will be discussed in detail later.

Recall that W_{bkt} is the aggregate workload contributed by one unit of batch b. The constraints for capacity, bounds on volume and the smoothness of production volume are as follows:

$$\sum_{b\in O_2} XF_b \cdot Q_b \cdot W_{bkt} + \sum_{b\in O_3} XR_b \cdot W_{bkt} + \sum_{b\in O_4} XV_b \cdot W_{bkt} + S_{kt} = \text{CP}_{kt} \quad \forall k,t \tag{4}$$

$$Q_b^l \le XR_b \le Q_b^u \quad \forall b \in O_3$$

$$(1-\delta)XR_{b'} \le XR_b \le (1+\delta)XR_{b'} \quad \text{where } t(b) = t(b')+1 \quad \forall b \in O_3$$

$$(1-\delta)XV_{b'} \le XV_b \le (1+\delta)XV_{b'} \quad \text{where } t(b) = t(b')+1 \quad \forall b \in O_4$$

where $t(b)$ is the occurrence time of batch b, and the production quantity in a period is constrained to be within the $(1 \pm \delta\%)$ range of that in the previous period.

1) Tool Backup Extension: In wafer plants, machines are, to an extent, interchangeable. The workload of a machine may be off-loaded to its backup machines. The efficiency of a backup machine may not be the same as the machine to be backed-up. A machine may be completely reassigned to backup another machine, or a fraction of its capacity is reassigned. Let $G(k)$ be the set of backup tools for tool k and $H(k)$ be the tools that are backed up by tool k. The capacity and workload reassignment constraints are

$$\sum_{b\in O_2} XF_b \cdot Q_b \cdot W_{bkt} + \sum_{b\in O_3} XR_b \cdot W_{bkt} + \sum_{b\in O_4} XV_b \cdot W_{bkt} + S_{kt} = \sum_{m\in G(k)} R_{mkt} \cdot \text{CP}_{mt} \cdot r_{mk} \quad \forall k,t$$

$$\sum_{k\in H(m)} R_{mkt} = 1 \quad \forall m,t.$$

The fraction (R_{mkt}) is a 0–1 variable if fractional backup is not allowed in a time period; otherwise it is a real number.

A planner can run the above MILP formulations at his or her discretion. Afterwards, the O_1 and O_2 type orders are fixed and the associated 0–1 variables can be regarded as constants. The degenerated formulations become Linear Programming (LP) problems. Four categories of sensitivity data are provided: 1) *shadow price* of machine capacity, 2) *marginal profit requirement* for products, 3) *unit profit allowance* and 4) *capacity allowance*. The shadow price of a machine is the marginal rate of revenue to the machine capacity. The marginal profit requirement for a product is the required increment in unit profit for a product to be selected for production. The optimal mix ratio will remain the same when the unit profit of a product fluctuates within its profit allowance. The capacity allowance of individual machines delimits the range of machine capacity within which the optimal product mix remains the same.

The sensitivity data is derived based on the duality theory. The above LP formulations can be written in a primal form

$$\text{Maximize } Z = \sum_{j=1}^{n} c_j \cdot x_j$$

$$\text{subject to } \sum_{j=1}^{n} W_{ij} \cdot x_j \le b_i \quad \forall i = 1,2,\ldots,m.$$

The c_j, W_{ij} and b_j are the parameters of the formulation. c_j is the unit profit from product j, x_j is the production quantity for product j, W_{ij} is the amount of resource j consumed by each unit of product i, and the b_i is the amount of resource i available. The dual form will be

$$\text{Minimize } y_0 = \sum_{i=1}^{m} b_i \cdot y_i$$

$$\text{subject to } \sum_{i=1}^{m} W_{ij} \cdot y_i \ge c_j \quad \forall j = 1,2,\ldots,n.$$

In solving the primal formulation, the solution x_j is obtained, but y_i is also obtained as a result of solving the problem. The economical interpretation of sensitivity data can be explained using these two forms as follows. The dual price of machine capacity is just y_i and is the contribution to profit per unit of resource i. That is, if b_i is increased by one unit, the profit will increase by an amount equal to y_i.

No matter what objective function is chosen, a common thread of the goals is to utilize the capacity as much as possible. However, the capacity constraints of the above formulations are rigid, whereas, in practice, the capacity is rather flexible. A common practice in factory management is to slightly, but temporarily, overload the factory in order to expose the bottlenecks. This is followed by focused effort to alleviate the bottlenecks. That is to say, to set the goal higher and then to achieve it. In product mix planning, this strategy can also be implemented in the following formulation.

Recall that the total workload $\Omega_{k,t}$ is the total workload at machine group k in time t. That is

$$\Omega_{k,t} = \sum_{b\in O_2} XF_b \cdot Q_b \cdot W_{bkt} + \sum_{b\in O_3} XR_b \cdot W_{bkt} + \sum_{b\in O_4} XV_b \cdot W_{bkt}. \tag{5}$$

Equation (5) can br rewritten in a succinct form

$$\Omega_{k,p} + S_{k,t} = \text{CP}_{kt} \quad \forall k,t.$$

After a product mix is generated, the workload $\Omega_{k,t}$ is analyzed to identify the bottleneck machine groups, BMG. The capacity constraints [(5)] of bottleneck machine groups are then replaced by the following constraints.

$$\Omega_{k,t} + \Omega_{k,t+1} + S_{k,t} + S_{k,t+1} = \text{CP}_{k,t} + \text{CP}_{k,t+1} \quad \forall (k,t) \in \text{BMG}. \tag{6}$$

The effect of (6) is to share the capacity of time $(t+1)$, which is under-loaded, with that of time t. Therefore, time t may become slightly overloaded. The extra workload, if not executed during time t, will be absorbed by time $t+1$. As a side effect, the queuing delay for jobs in time t might be increased but is still controllable by due diligence in dispatching and machine setup avoidance. Thus, the overloading would be temporary but judiciously focused, but offers an opportunity to drive the productivity of the plant higher.

2) Driving the Output Higher (an Advanced Application): The slack variable $S_{k,t}$ provides hints about where the bottleneck machines are (namely, the machine group k at time t). A second source of hints is the sensitivity data. The procedure for an advanced application of the product mix planning models is as follows.

1) Set a bottleneck threshold for the slack variable for two adjacent time periods t and $t+1$. For time t the threshold, U_t^l, is a lower bound and for time $t+1$ the threshold, U_{t+1}^u, is an upper bound.
2) Screen all machine-time pairs where the slack variable has a value within the specified bounds.
3) Modify the LP formulation by replacing (5) with (6) for the bottleneck machine group and solve for the product mix.

In one case study, the U_{t+1}^u was set at 0.7 and seven bottleneck machine groups (with the highest utilization) were identified. The third step was individually applied to the seven bottlenecks. Table V shows the respective increase in profit. The same procedure of temporarily overloading the plant was applied to two more data sets (of machine portfolios). Similar results were obtained. These empirical results suggest that there is a strong correlation between the dual price and the profit increment and that the correlation between the utilization and the profit increment is less certain. Therefore, it is concluded that the dual price information is more revealing and should be used to rank the bottleneck machines.

TABLE V
INCREASE IN PROFIT BY TEMPORARILY OVERLOADING THE FACTORY

EqpID (k)	Time (t)	Utilization $U_{k,t}$ (%)	Utilization $U_{k,t+1}$ (%)	Slack $S_{k,t}$	Slack $S_{k,t+1}$	Dual price	Profit increment
31	85	99.98	42.56	0.09	234.5	36944	0.063 %
68	78	99.91	67.25	0.22	126.2	36255	0.010 %
81	70	99.85	67.26	0.19	121.3	0	0
62	65	99.69	53.21	0.41	127.3	109757	0.236 %
4	65	98.14	46.61	2.54	145.3	0	0
57	65	97.20	46.73	4.02	153.0	0	0
73	70	96.21	31.09	3.96	125.7	0	0

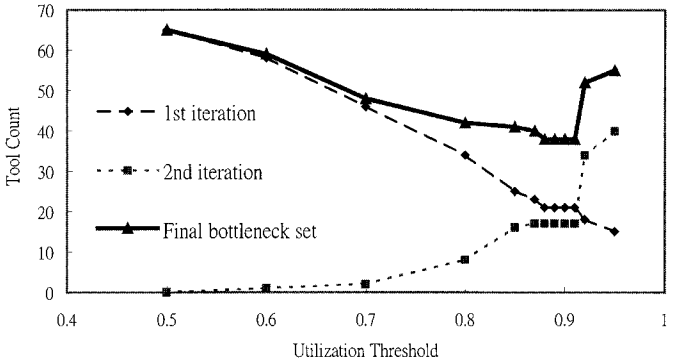

Fig. 1. Determination of the bottleneck machine set.

V. BOTTLENECK-BASED PLANNING

For MILP formulations, the computation efficiency is strongly affected by the number of variables and constraints. The above MILP model is suitable for small formulations, measured by the length of the planning horizon, the number of product types and other parameters. For large problems, computation time problems may arise. We observed that only a few constraints are binding and it is binding constraints that dominate the solution. Therefore, nonbottleneck machines could be excluded to reduce the problem size. However, this must be done in a systematic way. The bottleneck tends to shift when product mix changes from period to period.

Bottleneck determination and product mix are two convoluted decisions. An iterative procedure has been developed to identify the bottleneck machine group set. A reduced formulation covering a section of the planning horizon is first used to determine an initial product mix. The bottleneck set is next determined from the product mix. A formulation based on the bottleneck set is in turn used for the entire planning horizon to optimize the product mix. The product mix and bottleneck set are revised iteratively. The procedure is as follows.

1) Select a reduced time span, which should be longer than the average flow time, at the center of the specified planning horizon. Let $n = 0$.
2) Solve an MILP formulation for time periods in the reduced time span. Let $\mathrm{BS}_n = \phi$. Identify bottleneck machines based on a utilization threshold. These machines are new bottleneck machines $(\Delta\mathrm{BM})$ to be included in the bottleneck tool set.
3) Set $n = n+1$. Set the bottleneck set $\mathrm{BS}_n = \mathrm{BS}_{n-1} \cup \Delta\mathrm{BM}$.
4) Solve an MILP formulation on the bottleneck set for all periods.
5) Check the workload of machine groups. If there are overloaded machine groups, set $\Delta\mathrm{BM}$ to them and go to Step (4), otherwise, stop.

This procedure is an iterative one. The bottleneck tool set is monotonically increasing in its size. The setting of bottleneck threshold affects the efficiency of the procedure. If a lower value is used, more machines will be deemed the bottleneck—some erroneously. If a high value is used, initially few machines will be included in the bottleneck set. But, as more time periods are included in the formulation in later iterations, new bottleneck machines will surface. Fig. 1 shows a relation between the utilization threshold and the size of the bottleneck set in a case study. When the threshold is set at 0.5, 65 machine groups are included in the bottleneck set. No additional bottleneck tools surfaced afterward. When the threshold is set at 0.95, 15 machine groups are initially included in the set. But in a second

iteration, 40 additional groups surfaced as new bottleneck machines, bringing the total count of bottleneck machine groups to 55. Fig. 1 shows that a threshold between 0.87 to 0.92 results in the smallest bottleneck set (39 out of 116 machine groups).

VI. Discussions and Conclusions

The current practice of product mix planning uses rough granularity. A large time bucket of one month is used and the flow time offset is ignored in the logic of workload calculation. The above product mix methodology has been implemented as a software decision system and run on industry data comprising 116 machine groups, 34 backup relations and 4 representative memory and logic products of different technology generations. In this section, the computation experience related to improving the current practice is presented.

A. Capacity Allocation

The fifth objective in Section IV is of the greatest interest in practice. It merits elaboration. Although the profit is the cardinal objective for a manufacturing enterprise, in the foundry business there are usually obligations to reduce backlog of products with low profit margins. With the first objective function, the backlog might not be cleared up timely. As a result, the resultant product mix may not be sufficiently convincing in joint meetings of sales and production. Two conflicting goals are involved: to schedule high margin orders and to clear up low margin backlogs. If the latter concern is not addressed, the driving force of profit optimization will delay the production for lower margin products. Since the backlog is an obligation that must eventually be met, it will use up capacity of future time periods. When the industry-wide capacity is tight and the profit margin is in an increasing trend, the opportunity cost for postponing the fulfillment of obligations must be taken into consideration. This is an issue related to capacity allocation.

To model the fact that each customer order is to be scheduled for one (or none) of a number of time periods, a 0–1 selection variable is created for each release batch for each time period in the planning horizon $t = 1, \ldots, T$. Without loss of generality, a set of simplified symbols is used here for clarity. Each release batch b in time t is now represented by X_b^t. Let AM_t be the average margin for one unit of product at time t and the profit margin for batch b be $\mathrm{CM}_{i(b)}$. The ratio of the average margin to the profit margin of a product is used to represent the relative opportunity cost. Namely, the second term of the fifth objective function is expressed as

$$\sum_b \frac{\mathrm{AM}_t}{\mathrm{CM}_{i(b)}} \cdot \mathrm{AM}_t \cdot X_b^t \tag{7}$$

At the time that an order batch is booked, its margin should be in line with AM_t. Therefore, the term of formula (7) reduces to the average profit. If the production of a batch is postponed, the relative opportunity cost ratio will increase, thus increasing the likelihood of being chosen by the optimization code. In addition, the following inequality needs to be included in the formulation:

$$\sum_{t=1}^{T} X_b^t \le 1 \quad \forall b.$$

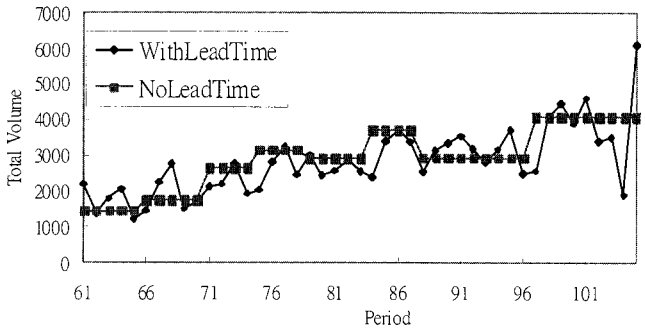

Fig. 2. Effect of lead-time offset.

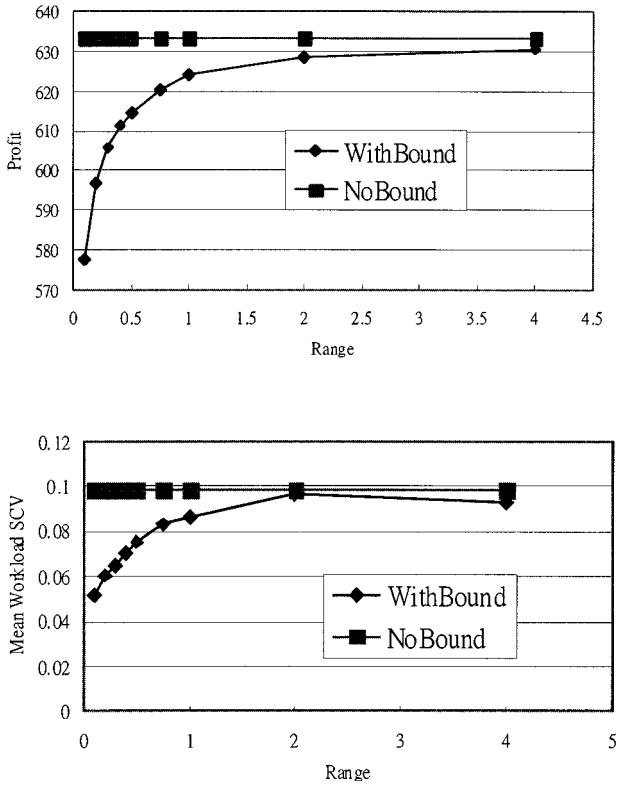

Fig. 3. Effect of production smoothing.

B. Necessity of Including Lead-Time Offset in Workload Calculation

The workload calculation logic described in Section III includes the factor of the lead-time offset. This meticulous detail of calculation is necessary. Fig. 2 shows the production volume for two product mix solutions, one with the lead-time offset and another without the lead-time offset. Although the output and profit would be higher if the lead-time factor is not included, the resultant product mix is actually infeasible as the machine capacity is violated at 122 locations of (k, t) after the total workloads are examined in the case study.

C. Production Smoothing

Although the product mix will change with time, it is desirable that the change be controlled and smoothed. Production smoothing imposes more constraints on the mix optimization problem. It will adversely affect the expected profit. On the other

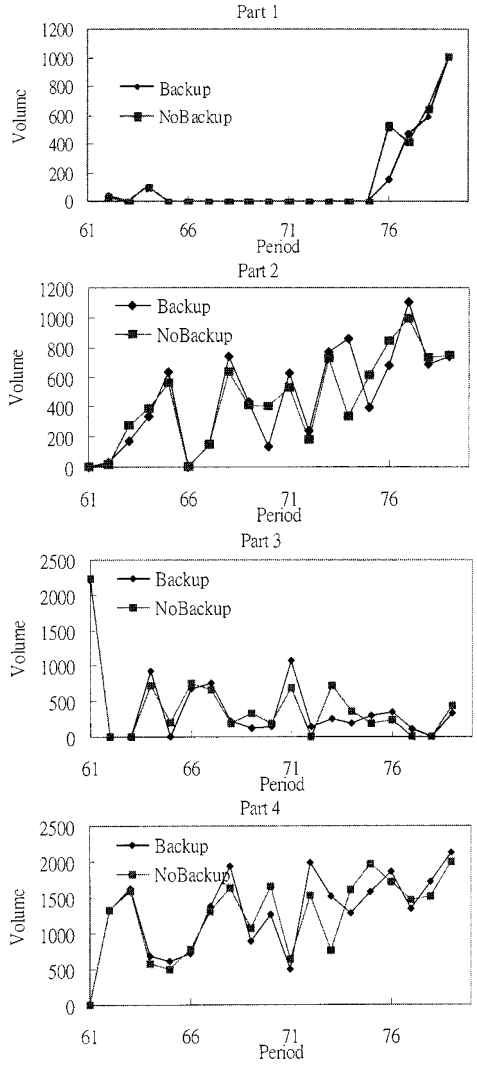

Fig. 4. Effect of machine backup.

hand, it will reduce the variation of the total workloads, thus, creating a favorable setting for enhancing productivity. Fig. 3 shows the effect of production smoothing. The range parameter is varied from 10% to 400%. A value of 100% seems to be a good demarcating point, beyond which the effect of product smoothing will diminish rapidly.

D. Machine Backup

The primary objective of machine backup is to dynamically reassign workloads in order to reduce the machine requirements and the flow time. On the operation aspect, the profit and output seem to increase as a result of exploiting machine backup. Machine backup will result in a slight change in product mixes. In the case study, the wafer output increases by 0.8% and the profit by 0.48% (Fig. 4).

This paper presents a methodology for product mix planning. It is shown that the time bucket size of one month or one week, instead of one day or two weeks, should be used for workload and product mix calculation. It is also concluded that the work release frequency should be at the same granularity level as the time bucket size. Mixed integer linear programming formulations have been developed to optimize product mix, taking into consideration the requirement of production smoothing and machine backup. A bottleneck-based procedure has been developed for problems of large size. A procedure for judiciously overloading the plant to drive the productivity higher is demonstrated. Finally, it is also shown that it is essential to include the lead-time offset factor in product mix planning.

REFERENCES

[1] D. P. Connors, G. E. Feigin, and D. D. Yao, "A queuing network model for semiconductor manufacturing," *IEEE Trans. Semiconduct. Manufact.*, vol. 9, pp. 412–427, 1996.
[2] E. M. Goldratt, *The Haystack Syndrome, Croton-on-Hudson*. Great Barrington, MA: North River , 1990.
[3] J. G. Helmkamp, "Decision making based on relevant information," in *Managerial Accounting*, 1st ed. New York: Wiley, 1987, pp. 236–237.
[4] H. W. Hsih and H. C. Wu *et al.*, "Equipment loading dynamic forecasting system," in *Proc. 7th Int. Symp. Semiconductor Manufacturing*, 1998, pp. 83–86.
[5] P. K. Johri, "Practical issues in scheduling and dispatching in semiconductor wafer fabrication," *J. Manufact. Syst.*, vol. 12, pp. 474–485, 1993.
[6] T. N. Lee and G. Plenert, "Optimizing theory of constraints when new product alternative exist," *Product. Inventory Manage. J.*, vol. 30, pp. 51–57, 1993.
[7] R. Luebbe and B. Finch, "Theory of constraints and linear programming: A comparison," *Int. J. Product. Res.*, vol. 30, pp. 1471–1478, 1992.
[8] J. Neudorff, "Static capacity analysis using microsoft visual basic," in *Proc. Int. Conf. Semiconductor Manufacturing Operational Modeling and Simulation*, 1999, pp. 207–212.
[9] J. D. Witte, "Using static capacity modeling techniques in semiconductor manufacturing," in *Proc. IEEE/SEMI Advanced Semiconductor Manufacturing Conf. Workshop*, 1996, pp. 31–35.
[10] W.-F. Wu, J.-L. Yang, and J.-T. Liao, "Static capacity checking system with cycle time considered," in *Proc. 7th Int. Symp. Semiconductor Manufacturing*, 1998, pp. 307–310.

Yon-Chun Chou received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1988.

He is currently a Professor and Head of the Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. He was an Assistant Professor and an Associate Professor at the University of Massachusetts, Amherst. His research interests are in the areas of manufacturing, production, logistics systems design, production scheduling, and information systems. His recent work includes electronics and semiconductor manufacturing, and e-business.

I-Hsuan Hong received the M.S. degree in industrial engineering from National Taiwan University (NTU), Taipei, Taiwan, R.O.C., in 1999.

He is currently with the Institute of Industrial Engineering, NTU.