A Tool Portfolio Planning Methodology for Semiconductor Wafer Fabs

Chung-En Kao and Y-C Chou
Institute of Industrial Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
ychou@ccms.ntu.edu.tw

Abstract

Tool portfolio planning is a frequent task in semiconductor wafer fabrication plants, as process, machine and product technologies evolve rapidly and the plants go through capacity expansion. As wafer fabrication plants are complex integrated factories with conspicuous phenomena of queuing effects, portfolio planning must take into consideration machine utilization, factory throughput, and total flow time simultaneously. This paper describes a tool portfolio planning methodology for wafer fabrication foundry plants. An improved static capacity model is first presented. A portfolio planning procedure based on static capacity estimation and queuing analysis is next described. A software implementation of the procedure is also used to clarify dilemmas of planning. It is shown that empirical formulas can be used to estimate the efficiency of batch machines. It is also used to show three types of portfolio adjustment action: flow time reduction, cost reduction and effectiveness improvement.

1. INTRODUCTION

Tool portfolio refers to the makeup, in quantity and type, of the set of processing tools in a fab. It is closely related to capacity planning. In most foundry fabs, tool portfolio planning is an ongoing task, as product mix, process and tool technology and factory scale usually change continuously. Tool portfolio planning is an essential task that supports business strategy to exploit market opportunities and to reduce the risk of tool obsolescence. Furthermore, because fabs manifest the behavior of complex queuing networks, their operation performance is difficult to analyze and predict. Therefore, soundness of portfolio planning is a crucial capability of competitiveness.

Capacity estimation is a major issue in wafer fab operation. Simulation, queuing models and static models are three common methods for capacity analysis, of which static models are usually used due to their relatively quick response time and ease of use. However, static models suffer from two major drawbacks, namely, inaccuracy of estimation and lack of queuing delay information. Many research studies have focused on improving the accuracy of static models. Table 1 summarizes the scope of static capacity models described in the literature, with the factors of inaccuracy listed down the table.

The first four factors are relative straightforward to include in a capacity model, but the next three factors require mathematical treatment. In calculating tool usage time, the total time is divided between operational time and non-operational time and operational time is further distinguished between up time and down time. Tool availability is defined as the ratio of tool uptime to tool operation time. Tool efficiency is the statistical mean of the ratio of actual throughput to maximum throughput. Yield is the ratio of good product output to input amount of material. Yield adversely affects workload in the form of reworks and scraps. Semiconductor manufacturing is characterized by long process flow time. Jobs released to the fab in one time period may introduce workload to other time periods. The expected arrival time of the workload

0-7803-6374-4/00/$10.00©2000 IEEE

due to a job at a tool should be calculated using lead-time and process flow information. Normally, tool availability and efficiency, process yield, and flow time data are collected on shop floor [5] or they can be assumed as premises in capacity planning.

Table 1. Scope of static models.

Factors	Witte [4]	Hsih [2]	Wu [5]	Neudorff [3]	Chou et al.
Tool availability	√	?	?	?	√
Tool efficiency	√	?	?	?	√
Yield	√	?	?	?	√
Lead time offset	?	√	√	?	√
Batching efficiency				√	√
Tool dedication		√	√		√
Tool backup		Rudi-mental		Rudi-mental	√

Notation ?: uncertain but presumed

The first four factors in Table 1 make up a basic static model [4]. To further enhance the precision of capacity models, the effect of nominal operation policy on tool efficiency can be incorporated. Tools can be classified as serial or batch tools. Serial tools have the first wafer effect errors while batch tools introduce batching efficiency errors. In [3], batching efficiency is estimated through regression analysis of the visits/starts ratio and efficiency, while the first wafer effect is estimated by analyzing the occurrence times of tool idleness based on historical data. Tool dedication and back-up are the second source of inaccuracy. Process constraints on dedication and backup can be maintained in database [5]. To forecast equipment loading, workloads are then assigned or manually shifted while observing the process capability, priority and availability of tool [2]. Tool backup has also been dealt with in a simple static manner [3]. Tool dedication and back-up practice complicates capacity estimation. It is involved with combinatorial optimization [2, 3] and is beyond the scope of static capacity modeling. There has been no satisfactory treatments in the literature.

The focus of this paper is on presenting a tool portfolio planning method. It makes use of an enhanced static capacity model and a queuing capacity model. The remainder of this paper is organized as follows. In Section 2, workload computation logic for the static capacity model is described. Batching efficiency is analyzed in Sections 3. A tool portfolio planning procedure based on trade-off between cycle time, cost and throughput is presented in Section 4. Tool backup and dedication is analyzed in Section 5. Finally, conclusions can be found in Section 6.

2. WORKLOAD COMPUTATION LOGIC

An input to portfolio planning is product demand, which are expressed as demand batches, each characterized by product type, quantity and due time. For each demand batch of product i in time t, D_{it}, unit workload can be generated for each tool. This step can be symbolically represented as:

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

where the w represents workload, the double arrow means one or more items of unit workloads are generated, $\{j\}$ represents the set of all process steps of product i, $k(i,j)$ indicates that the required tool k is specified for (i,j) in the process flow, and the term t-$l(i,j)$ indicates that the occurrence time for the unit workload is product demand time t offset forward by the lead-time $l(i,j)$. Specifically, for each demand batch D_{it} there will be J(i) unit workloads in total. Each unit

workload is identifiable by product-step pair (i,j) and its occurrence time is $t\text{-}l(i,j)$. Adjusting for yield allowance, the workload for each product-step pair of (i,j) at time t is calculated as

$$w_{i,j(i),k(i,j),t(i,j)} = \frac{D_{i,t} \cdot p_{i,j,k}}{ya_{i,j,t}} \qquad \forall i$$

where

$p_{i,j,k}$ = processing time

$ya_{i,j} = sy_{i,j} * ya_{i,j+1} \quad j = 1,\ldots,J(i)$

$ya_{i,J(i)} = sy_{i,j}$

$sy_{i,j}$ = yield of process step (i, j)

Here, yield allowance (ya) for each step is computed backward from the last step to the first step of the process flow. 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. The total workload for toolsets, $W_{k,t}$, and tool requirement, $q_{k,t}$, can be computed as

$$W_{k,t} = \sum_{i} \sum_{j(i)} w_{i,j,k,t} \qquad \forall k,t$$

$$q_{k,t} = \frac{W_{k,t}}{(availability)_k} \cdot (operation - time)_t \qquad \forall k,t$$

3. BATCHING EFFICIENCY

To reduce cycle time, batch tools may not be loaded to capacity. Loading decisions for batch tools have been thoroughly studied [1]. With local information (without arrival forecasts), the general conclusion is that the greedy loading rule is close to optimal. Lower bound formulas for average batch size were also derived for the special case of independent job streams and constant arrival rates. In practice, loading decisions are usually based on a minimum number of lots and a maximum wait time.

Batching efficiency is expressed as the statistical mean of the ratio of actual loading size to tool capacity. The effect of tool capacity (B_g^{max}), tool quantity (C_g) and tool downtime (ρ_g^{inc} %) on batching efficiency is analyzed in this study. Table 2 shows the factors of a two-level experiment matrix of simulation.

Table 2. Experiment Design Matrix.

Factor	Low (-)	High (+)
Tool Capacity	2	10
Tool quantity	1	10
Downtime fraction	0.1	0.3

Define traffic intensity (ρ_g) of a tool as the total workload over its uptime. Figure 1 shows typical characteristic curve between average batch size and traffic intensity. For the eight cases studied, the downtime ratio does not have strong effect on the curve, but tool capacity and quantity do have an effect on the location of the excursion point D.

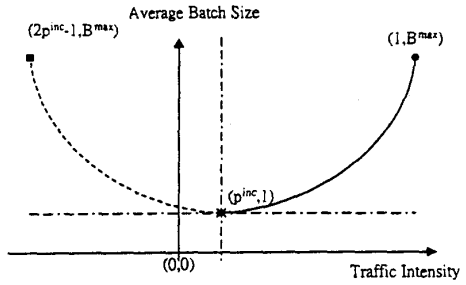

Figure 1. Characteristic curves for the average batch size.

Quadratic curve is fitted to the simulated data to obtain the following approximation formula:

$$\overline{B_g} = \frac{(B_g^{max} - 1)}{(1 - \rho_g^{inc})^2} \cdot \rho_g^2 + 0 \cdot \rho_g + 1$$

A separate fab simulator that captures the interaction between job streams is used to validate the accuracy of the fitted lines. The results for 41 batch toolsets are shown in Figure 2. The mean and standard deviation of errors are approximately 4.93 percent and 9.76. The tool requirement for batch tools is:

$$q_{k,t} = \frac{W_{k,t}}{(availability)_k} \cdot (operation-time)_t \cdot \frac{B_g^{max}}{\overline{B_g}} \qquad \forall\, k,t$$

4. PLANNING PROCEDURE

A tool portfolio can be represented as an ordered list of tool quantities, such as $s = (n_1, n_2, \ldots, n_N)$, where n_i is the number of tools for toolset i. The performance measures of interest to capacity planning are throughput, utilization and cycle time. Since static capacity models provide limited information about throughput and utilization and do not provide cycle time information, an open queuing network capacity model was used to evaluate tool portfolios in this study. This queuing model has the following premises: (1) no scrap and rework, (2) two classes of customers: work-in-process and machine breakdown, and (3) two types of tools: batch and non-batch tools.

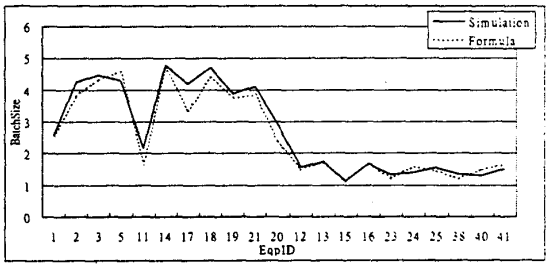

Figure 2. Performance of batch size formula.

Figure 3 illustrates the structure of portfolio solution space for portfolio adjustment. Each portfolio can be characterized by three attributes: throughput, cycle time and investment cost. For a given tool portfolio, the tool count of certain toolsets can be increased (or decreased) to reduce (or increase) cycle time while maintaining the same level of throughput. This phenomenon is shown by equi-throughput curves in the figure. Two curves, such as TP_1 and TP_2, represent relative effectiveness of portfolio. In the figure, curve TP_2 has a higher throughput than that of curve TP_1. For the same investment cost (the vertical dotted line), TP_2 has a more balanced portfolio, resulting in higher throughput and lower cycle time. In contrast, to achieve

the same cycle time (the horizontal dotted line), TP_1 will require a higher investment. Figure 3 reveals three types of portfolio adjustment actions for improving cycle time, investment cost and effectiveness as indicated by the arrows T, C and E, respectively.

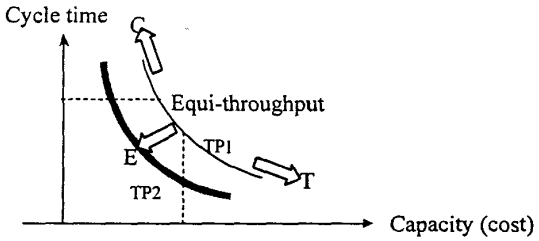

Figure 3. Portfolio adjustment strategies.

Portfolio adjustment is an iterative process based on marginal analysis. For each toolset of a current portfolio, a tool is added and subtracted to generate neighboring portfolios. The performance of all neighboring portfolios are evaluated using a queuing model and the ratio of cycle time decrement (or increment) over cost increment (or decrement) is computed. Two separate lists are maintained, one for cycle time reduction and the other for cost reduction. The two lists can be sorted to rank order action types T and C respectively. Type E actions are composed of a type T and a type C actions as shown in Figure 4. Portfolios b and c are obtained from a current portfolio (point a) by adding and subtracting a tool respectively. If the combined effect of the two actions result in a reduction of both cost and cycle time, they then constitute a type E action (point d).

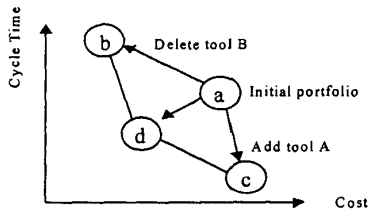

Figure 4. Type E adjustment action.

Figure 5 illustrates the T action in greater detail for a second data set of product demands. The resultant configuration and performance data can be found in Appendix B. (The investment and flow time data are business sensitive information, therefore, they have been normalized. Similarly, machine groups are identified by numbers.) Let the marginal cost of flow time be defined as the ratio of the flow time reduction to the investment cost increment that is associated with adding a machine to the portfolio. The marginal cost is computed for each machine group and the top five machine groups are listed in the second column of Appendix B. In each iteration, the top one is selected. Starting with a given initial portfolio (at the upper-left of the trade-off curve), a sequence of portfolios is then generated. This figure shows that 4% more investment (on critical machine groups) could reduce the flow time by 27%.

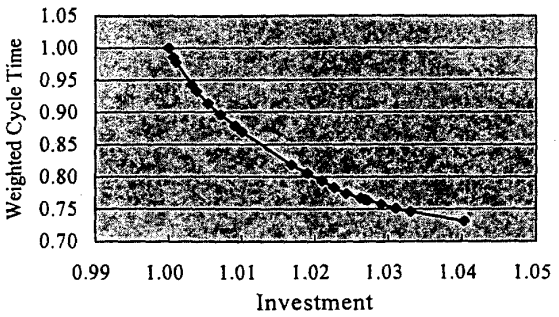

Figure 5. Tool portfolio adjustment trajectory.

This procedure utilizes both a static capacity model and a queuing capacity model. The static model is used to generate an initial tool portfolio and portfolio adjustment is based on performance data provided by the queuing model. There are a number of fine discrepancies between the premises of static and queuing capacity models. Therefore, the resultant portfolio should be subjected to human interpretation, taking into consideration the following analysis. Figure 6 shows the workload distribution for demand batches 1 and 2. The workload for time t will be B plus C in the static model. When product demands are stationary, this workload will equal A plus B. However, if the demands increase from time t to time t+1, as in the case of fab capacity expansion, C will be greater than A. (Figure 5 is based on capacity expansion.) Similarly, the queuing capacity model is applicable to steady state. Therefore, the workload computed will be for B plus A. As such there is a time lag between portfolios generated by the static model and queuing model.

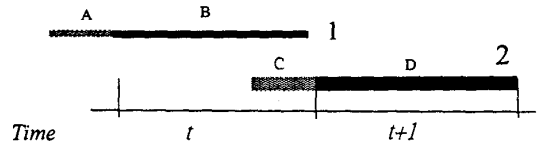

Figure 6. Workload distribution with lead time offset.

5. TOOL BACKUP AND DEDICATION

Tool backup and dedication has been implemented in a mixed integer linear program. The major purpose of tool backup is to reduce tool investment by fine tuning the workload assignment. Workload for each toolset is first calculated. These workloads are then re-assigned according to constraints imposed by dedication decisions and backup relationships. The major constraints are listed below:

Notation:

BG_k — Tool groups that can backup up tool type k
$BE_{m,k}$ — Backup efficiency of tool m w.r.t. tool k
$X_{k,t}$ — Original workload assignment for tool k at time t
$Y_{m,k,t}$ — Workload shifted to tool m from tool k at time t
$Q_{k,t}$ — Quantity of tool type k at time t
$a_{m,t}$ — Tool availability

$$\sum_k Y_{m,k,t} \le Q_{m,t} \cdot a_{m,t} \quad \forall m,t$$

$$\sum_{m \in BG_k} Y_{m,k,t} = X_{k,t} \quad \forall k,t$$

Tool backup planning is a task separate from the above portfolio planning procedure. This model is suitable for single or multi-period planning. For single period planning, the objective function is to minimize the total investment cost. For multiple period planning, the cost must be adjusted for its time value such that tools are commisioned earlier than is necessary. Figure 7 illustrates the differences between an initial and backed up portfolios.

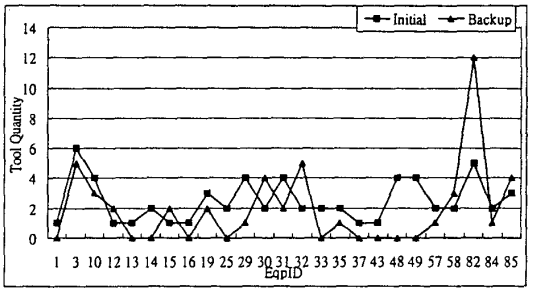

Figure 7. Effect of tool back up adjustment.

6. CONCLUSIONS

A tool portfolio planning methodology is presented in this paper. The methodology has three major components: an improved static capacity model, a queuing capacity model and a portfolio adjustment procedure. It takes into account batching efficiency, and tool dedication and backup. The methodology enables capacity planners to explore the solution space and to better evaluate the trade-off between cycle time, investment cost and throughput.

REFERENCES

[1] C. R. Glassey, F. Markgraf and H. Fromm, "Real time scheduling of batch operations," Optimization in Industry, pp. 113-137, 1993.

[2] H. W. Hsih, H. C. Wu, et al., "Equipment loading dynamic forecasting system," The Seventh International Symposium of Semiconductor Manufacturing, pp. 83-86, 1998.

[3] J. Neudorff, "Static capacity analysis using Microsoft Visual Basic," International Conference on Semiconductor Manufacturing Operational Modeling and Simulation, pp. 207-212, 1999.

[4] J. D. Witte, "Using static capacity modeling techniques in semiconductor manufacturing," IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop, pp. 31-35, 1996.

[5] W. F. Wu, J. L. Yang and J. T. Liao, "Static capacity checking system with cycle time considered," The Seventh International Symposium of Semiconductor Manufacturing, pp. 307-310, 1998.