Economic Analysis and Optimization in Tool Portfolio Planning in Semiconductor Manufacturing

Y-C Chou, C-S Wu and J-Z Lin
National Taiwan University
Taipei, Taiwan, R.O.C.
ychou@ccms.ntu.edu.tw

Abstract – The tool portfolio of a plant refers to the makeup, in quantity and type, of processing machines in the plant. Portfolio planning is a multi-criteria decision-making task involving trade-offs among investment cost, throughput, cycle time and risk. In this paper, an economic decision model is first presented for optimal configuration of portfolio and to determine optimal factory loading. If plants are closely located or have a twin-fab design, portfolio planning at multiple plants can be integrated to enhance the overall effectiveness of portfolios. A novel methodology for arbitrating capacity backup between multiple plants is described in the second part. Finally, robust configuration of portfolio in a dynamic demand environment is addressed. Industry data have been utilized to run through the developed methodologies.

1. INTRODUCTION

The tool portfolio of a plant refers to the makeup, in quantity and type, of processing machines in the plant. What portfolio a plant should have is determined by the state and future technology trends of process, machine, product and product demands. Because of the dynamic environment in the semiconductor industry, there is a high risk of under- or over-capacity and mismatch between actual demands and the right types of capacity. Tool portfolio planning is a task that has tremendous bearing on manufacturing efficiency.

There are three important issues in portfolio planning. They are configuration design, performance evaluation and risk analysis. Static modeling and queuing modeling are commonly used to evaluate capacity requirements and performance. A simple method of portfolio planning uses a static capacity model. More advanced methods might use queuing analysis of one form or another [1,2,4]. To configure a portfolio, marginal analysis of performance measures is usually applied to adjust tool quantities [1,4]. Because of the uncertainties in product demand, there is sometimes a need to address the robustness of tool portfolio under multiple demand scenarios. This problem was addressed in [3] by finding out the tool groups whose workload is sensitive to the changes in product mixes.

Portfolio planning is a multi-criteria decision task involving trade-offs among investment cost, throughput and cycle time. Not only that there are more than one portfolios that will satisfy a specified set of production goals, but also that each portfolio can be operated in a multitude of load scenarios, yielding various combinations of performance measures. The treatment of this trade-off analysis has not appeared in the literature and is the focus of this paper.

This paper addresses the optimization and economic analysis of tool portfolio. In Section 2, a procedure to generate a multitude of feasible portfolios is described. In Section 3, an economic decision model is presented for optimal configuration of portfolio and to determine optimal operation loading. In Section 4, a novel methodology for capacity sharing between plants is described. Finally, the robustness of portfolio under dynamic environment is addressed in Section 5, and conclusions can be found in Section 6.

2. GENERATION OF THE SOLUTION SPACE

Figure 1 is a flow diagram for a two-stage procedure we use to generate the solution space of portfolio. A static capacity model is first applied to generate an initial solution. In the second stage, the initial portfolio is evaluated using a queuing model to estimate its performance in throughput, flow time, and utilization. The portfolio is then modified by increasing the machine quantity of the bottleneck tool group. This improvement process continues for a number of iterations until all performance requirements are met.

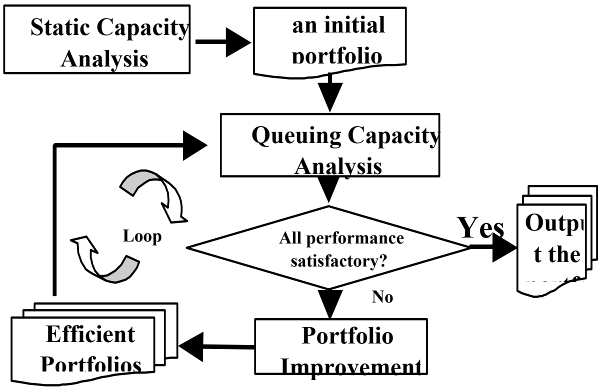

Figure 1. Generating a solution space of efficient portfolios

In the second stage, there could be multiple views of what exactly the bottleneck is. The available time of a tool can be divided between regular utilization (ρ), incapacitation utilization (ρ^{inc}), and idleness. In this study, three bottleneck indicators have been compared for effectiveness. They are utilization (regular plus incapacitation), queuing delay, and remnant capacity (*rc*), where the remnant capacity of a tool group *g* is defined as:

$$rc_g = \begin{cases} \dfrac{1-\rho_g^{inc}}{\rho_g} - 1 & \textit{for non-batch tools} \\[2ex] \dfrac{(1-\rho_g^{inc})\cdot MaxBatchSize}{\rho_g \cdot MeanBatchSize} - 1 & \textit{for batch tools} \end{cases} \qquad \dots(1)$$

A series of portfolios can be generated by using each indicator. Figure 2 shows that the queuing delay and remnant capacity indicators are more cost-effective than the utilization indicator.

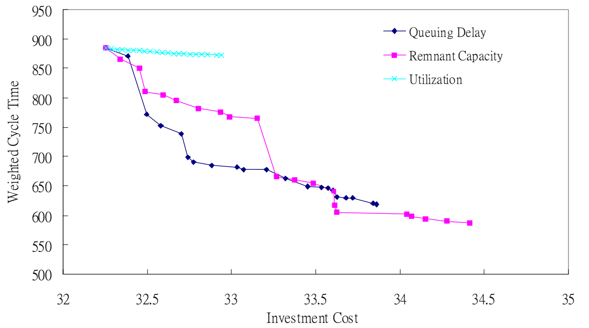

Figure 2. Effectiveness of different bottleneck indicators

Each point in Figure 2 represents a portfolio. Because a plant with a certain portfolio can be loaded differently, its operation performance actually would vary over a range. Figure 3 shows the range of performance for twenty portfolios of Figure 2. Each curve represents the operation options for one portfolio. These curves are called option curves (OC) in this paper and the space that all option curves lie in, i.e., the 2-dimentional Euclidean space depicted in Figure 3, will be called the option space.

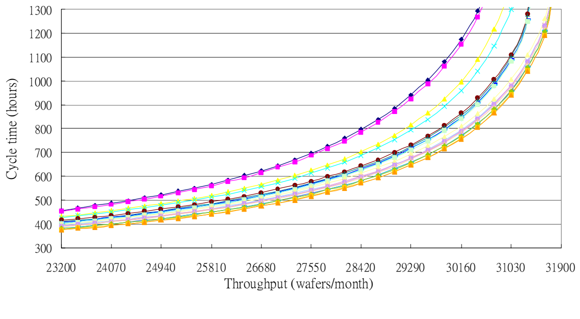

Figure 3. Option curves in the option space

3. A DECISION MODEL FOR OPTIMIZATION

In this section, we will present a methodology to determine the optimal portfolio based on performance measures of cycle time, throughput, and investment cost. While the value of throughput is relatively easy to quantify, the benefit of cycle time is very subjective and depends on business situation. In Economics, utility function is a framework of analysis that is used to express the change in perceived value that is assigned to goods as its consumption quantity increases. We use the following functional forms to model the utility of throughput and cycle time (Figure 4).

$$U(th) = f(x) = 1 - e^{-a\cdot x} \quad \textit{where} \;\; x = \frac{\textit{throughput}}{C0} \qquad \dots(2)$$

$$U(ct) = g(y) = \sin(\arccos(b\cdot y)) \quad \textit{where} \;\; y = \frac{\textit{cycle time}}{RPT} \qquad \dots(3)$$

where throughput and cycle time are normalized with respect to the nominal capacity (C_0) and the sum of raw processing times (RPT). The parameters *a* and *b* affect the curvature of the functions. Two questions have been designed to assist the planner to assign a value to a and b: What is the utility of a throughput that equals to 100% of the nominal capacity? What is the utility of a cycle time that equals to 3.5 times that of the raw processing time?

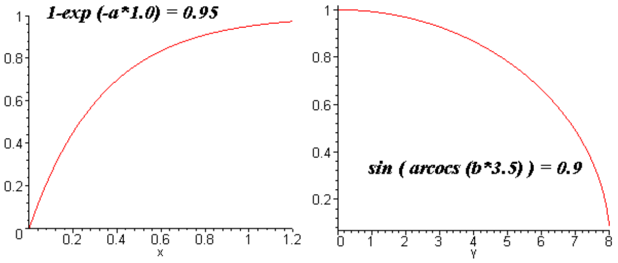

Figure 4. Utility functions of throughput and cycle Time

The total utility function, *h(x, y)*, is defined as the weighted sum of functions *f(x)* and *g*(*y*) using an assigned weight *w*.

$$h(x,y) = w\cdot f(x) + g(y) \qquad \dots(4)$$

Figure 5 is a graph that shows the total utility in the z-axis of a 3-dimentional plot for *w*=2. Each horizontal cross section of the response surface represents an indifference curve between throughput and cycle time (also shown in the right panel). That is, the total utility of all points on an indifference curve (IDC) is the same.

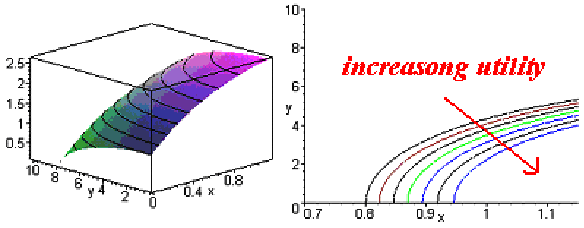

Figure 5. The total utility function and indifference curves

The optimal portfolio and its optimal operation loading can be obtained by evaluating the total utility of all points in the option space, using both Figures 3 and 5, as follows. Each option curve can be regarded as a hyper surface in the 3-dimentional space of Figure 5. The intersection between the hyper surface and the response surface of the total utility function is a hyper curve. The optimal operation loading is then the highest point of the hyper curve. This highest point can be solved mathematically by using the LaGrange multiplier method as follows. The total utility function is regarded as the objective function to be maximized and the option curve as a constraint relating throughput and cycle time. For Equation 4, the objective function *h(x, y)* and the hyper surface for the constraint OCS(x, y) are of the forms:

$$h(x,y) = w \cdot (1 - e^{-a \cdot x}) + sin(\, arccos(b \cdot y)) \qquad \ldots(5)$$

$$OCS(x,y) = \frac{1}{m - n \cdot x} - y \qquad \ldots(6)$$

Once the optimal loading is determined for each portfolio, the maximum utility that can be achieved with each portfolio is also determined. Define the investment efficiency of a portfolio as the ratio of its maximum utility to its investment cost. The portfolio with the highest investment efficiency is considered optimal.

We will now return to the issue of the weight *w*. Using the above procedure, an optimal loading level is determined using a required input of weight *w* between two utility functions. The weight is a subjective judgment of the relative utility between cycle time and throughput. As shown in Figure 6, if a weight of 2 is used, the resultant optimal point will be (x=0.940, y=3.05). The derivative of the OC at that point is equal to 14.606. Since the derivative of the OC can be interpreted as the relative utility between cycle time and throughput, a logical impasse now surfaces. That *w* equals to 2 means the utility of throughput is twice as important as that of cycle time. That the derivative equals to 14.606 means the utility of throughput is 9.28 (= $C_0 \cdot RPT^{-1} \cdot 14.606^{-1} = 29000 \cdot 214^{-1} \cdot 14.606^{-1}$) times more important than that of cycle time. In the following we will show the existence of an equilibrium weight that is inherent to the option curve of each portfolio.

We will present two numerical examples; one starts with a small value of *w* and another with a large value of *w*, to layout a framework of analysis. Each example involves a number of iterations to compute the optimal operation loading. The procedure is:

1. Iteration *i* = 1. Give an arbitrary initial weight $w = w_1$.
2. Compute the optimal loading O_i using w_i as input. Calculate the derivative of the OC at O_i. Let the derivative be D_i.
3. Set $w_{i+1} = \frac{C_0}{RPT} \cdot \frac{1}{D_i}$
4. If $|w_i - w_{i-1}| < \varepsilon$, stop. Otherwise, set *i* = *i* + 1 and go to Step 2.

The results for an initial weight of 1.0 and 8.0 are summarized in Table 1. In both examples, the weight w_i converges to a value of approximately 5.29. This convergence value is called the equilibrium weight.

Table 1. Convergence of the weight

	W1	**W2**	**W3**	**....**	**.....**	**W14**	**W15**	**W16**	**W17**
Case1	1.0	13.793	3.1562	......	....	5.2976	5.2927	5.2906	5.2916
Case2	8.0	4.2098	6.0255	....	.....	5.2922	5.2919	5.2917	5.294

4. CAPACITY SHARING AND BACKUP

In this section, we describe an application of the above decision model. Some modern plants have a twin-fab design. Two clean rooms are built side by side or stacked up one on top of another to share common utility facilities. In still some other occasions, plants are close to each other, which is the case in Taiwan. The bottleneck tools of two plants may not be the same at all times. The proximity of plants allows capacity sharing to take place. If tool capacity is shared between plants, the overall performance will be improved.

If additional capacity of the bottleneck tools is obtained from a partner plant, the option curve shifts downward to the right (Figure 6). Suppose the current operation loading is at point O. With the borrowed capacity, either the throughput could be increased from ω_1 to ω_2 (point A), or the cycle time could be reduced from τ_1 or τ_2 (point C), or any other points on the dotted curve will be achievable.

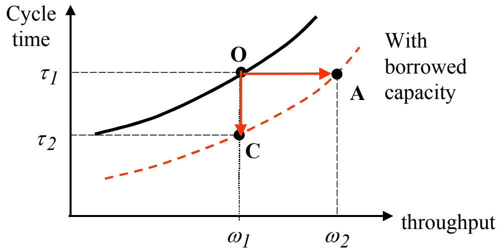

Figure 6. The effect of capacity sharing

Table 2 shows the result obtained by applying the above analysis to a set of industry data. It is shown that if 20 hours of capacity of the bottleneck tool group is borrowed from the partner plant, the throughput would be increased by 52 wafers per week, or the cycle time could be reduced by 10 hours. This methodology provides an objective arbitration for capacity sharing between plants.

Table 2. Effect of capacity sharing

	Point O	Point A	Point C
Throughput	29,000	29,052	29,000
Cycle time	883.4	883.4	873.9
WIP level	35,583	35,647	35,197

The economic benefit of increased throughput can be computed from the revenue that it brings in and the inventory cost of WIP (work-in-process). But, the benefit of cycle time improvement is a subjective matter. From the queuing theory it is known that a reduction in cycle time would affect the level of throughput and WIP. Therefore, we used an economic model to correlate the economic benefit of cycle time to that of throughput and WIP as follows. Let (ω_i, τ_i) and (ω_j, τ_j) be two points in the option space and $\omega_j > \omega_i$. The value of cycle time reduction can be computed using the following formulas with average asking price (ASP) of processed wafers, material cost (MC), production cost (PC), and an rate of return (r).

$$\text{Revenue from throughput}: \; R = \omega \cdot (ASP\text{-}PC\text{-}MC) \cdot 12 \text{ (month/year)}$$

$$\text{WIP inventory cost}: \; W\text{IC} = WIP \cdot (MC + \frac{PC}{2}) \cdot r$$

$$\text{Value of cycle time reduction} = \frac{(R_j\text{-}WIC_j) - (R_i\text{-}WIC_i)}{\tau_i - \tau_j}$$

Let r = .30, ASP = 1800, PC = 1000, and MC = 71.88, the value of cycle time reduction equals to approximately US $ 40,000 per hour. It should be noted that this figure is derived from the condition of optimal loading. This information could be used as a reference in evaluating projects of cycle time reduction.

5. EFFECT OF PROCESS FLOW GRANULARITY

There are two aspects of uncertainty in product demands. The product mixes may change over time and the bottleneck tool groups change as a result [3]. Another aspect of uncertainty has to do with products themselves. For medium- to long-term planning, products and their process flows are usually represented generically. The process flows may contain only key steps and machines. That is, the representation has a coarse granularity.

We have studied the effect of process flow granularity on the accuracy of portfolio planning. We are interested in knowing whether planning with information of only key steps of the process flows is sufficiently accurate as compared with planning with all steps. The number of tool groups is used as a measure of the granularity. A process flow with all process steps specified has the highest level of granularity. If non-critical steps are deleted from the process flow, a new process flow with a lower granularity will be derived. The queuing delay at tool groups is chosen as the surrogate measure of planing accuracy. Let $D_{g,l}$ be the queuing delay of tool group g for planning granularity level l. Figure 7 is the result of a case study. Originally, there are 101 tool groups. The horizontal axis shows the number of tool groups deleted. There are 10 granularity levels (l = 1 to 10 from the left). The vertical axis shows the average error of queuing delay between granularity levels, which is computed as

$$\text{average absolute error} = \frac{1}{|G_l|} \sum_{g \in G_l} \left| D_{g,l} - D_{g,0} \right|$$

where G_l is the set of tool groups remaining in granularity level l. The data shows that when as many as 50 tool groups are ignored in the process flows, the average error is approximately 0.1 hours, which corresponds to 5% error in cycle time. The error then increases more dramatically as the granularity level decreases further. Eventually, as the vast majority of tool groups are removed, the structure of the queuing network is destroyed. It is concluded that non-critical tool groups have very slight effect on queuing delay estimation. Ignoring non-critical tool groups will increase the errors of estimation, but the errors are very slight.

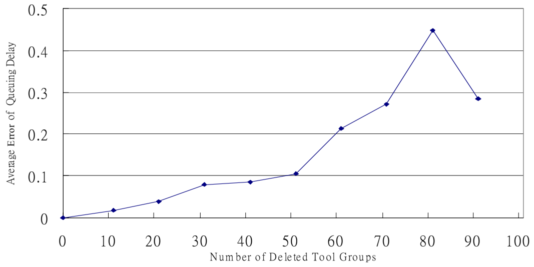

Figure 7. Effect of process flow granularity

6. DISCUSSIONS AND CONCLUSIONS

This paper presents a comprehensive methodology for tool portfolio planning, covering four advanced issues of portfolio planning: portfolio optimization, economic analysis, decision-making, and robust configuration under uncertainty. Because portfolio selection is a multi-criteria decision-making problem, we have developed a decision model for portfolio optimization and to determine the optimal operating loading. This decision model can be applied to objectively arbitrate capacity sharing between plants and to evaluate the economic value of cycle time. Finally, we presented a method to analyze the effect of process flow granularity on the accuracy of portfolio planning to address the problem caused by the uncertainty in product demands.

Acknowledgement:

This study was partially supported by Semiconductor Manufacturing Corp. under grant 201-NJ-879, and National Science Council under grants 89-2212-E-002-094 and 89-2213-E-002-117.

REFERENCES

[1] Yon-Chun Chou, and Ren-Chi You, “A Resource Portfolio Planning Methodology for Semiconductor Wafer Manufacturing,” International Journal of Advanced Manufacturing Technology, forcoming.

[2] Yon-Chun Chou, Ren-Chi You, C-R Weng, and H. Henry Wu, “A Tool Portfolio Planning Methodology for Semiconductor Wafer Fabs,” Proc. International Symposium on Semiconductor Manufacturing, October, 1999, Santa Clara, California, USA, pp. 19-22.

[3] Robert Kotcher, “Capacity Planning in the Face of Product-Mix Uncertainty,” Proc. International Symposium on Semiconductor Manufacturing, October, 1999, Santa Clara, California, U.S.A., pp. 73-76.

[4] Katsutoshi Ozawa, “Optimal Tool Planning Using the X-Factor Theory,” Proc. International Symposium on Semiconductor Manufacturing, October, 1999, Santa Clara, California, U.S.A., pp. 49-52.