A Capacity Trading and Transfer Pricing Method for Semiconductor Manufacturing

Muh-Cherng Wu †
Department of Industrial Engineering and Management
National Chiao Tung University (Taipei Tech), Hsin-Chu City, Taiwan
Tel: (+886) 3-5712-121 ext. 31913, Email: mcwu@mail.nctu.edu.tw

Jheng-Jie Peng
Department of Industrial Engineering and Management
National Chiao Tung University (Taipei Tech), Hsin-Chu City, Taiwan
Email: summit0311@hotmail.com

Chao-Wei Wang
Department of Industrial Engineering and Management
National Chiao Tung University (Taipei Tech), Hsin-Chu City, Taiwan
Email: lianfon@yahoo.com.tw

Abstract. A semiconductor foundry company usually involves several factories (fabs). Due to dynamic change of market demand, the product mix assigned to a fab may not be compatible to its originally designed tool mix. Therefore, these different fabs have to frequently support capacity to each other in a dynamic way in order to increase the capacity utilization of the company. However, these fabs are essentially competitors from the perspective of performance measurement; and may not fully support the other fabs. This research attempts to develop a capacity trading and pricing mechanism to enable these fabs to effectively cooperate with each other to increase overall capacity utilization. This research involves two decisions. The first decision is to find an optimum capacity trading portfolio among the fabs. We use discrete event simulation, artificial neural network, and genetic algorithm to make such a decision. The second decision is to set a "fair transfer price" for each unit of trading capacity. Herein, the fair transfer price denotes that these fabs that participate in the capacity trading shall "fairly" share the overall benefit so generated. This benefit sharing shall enable these fabs to have stronger intention to effectively support each other in capacity.

Keywords: Capacity Trading, Semiconductor Foundry, Transfer Pricing, Performance Measure.

1. INTRODUCTION

Semiconductor foundry is a business model in which a semiconductor foundry company is dedicated to provide manufacturing services. Namely, it only manufactures IC (integrated circuit) products for its customers (IC design houses) and does not involve in designing ICs. The investment of a leading-edge semiconductor factory (fab) is tremendously high, around 4-5 billion dollars. Therefore, IC design houses tend not to build their own semiconductor fabs and always adopt outsourcing policy in manufacturing.

Due to economy of scale, a semiconductor foundry company is usually large in scale and involves many fabs that are located near to each other. Due to dynamic change of market demand, the product mix assigned to a fab may not be quite compatible to its originally designed tool mix. Namely, a workstation in one fab may be lack of capacity while the same type of workstation may be underutilized in another fab. Therefore, these different fabs must frequently support capacity to each other in a dynamic way in order to increase the capacity utilization of the company. In practice, a foundry company would call a weekly meeting for discussing the capacity supporting plan among fabs for the coming week.

However, the weekly capacity supporting plan is usually concluded through a "negotiation" process and always not so effective. The reason is that these fabs even in the same company are essentially "competitors" from the perspective of inner performance measurement. A fab that gives capacity support to another fab may result in poor performance, while the fab that receives capacity support may leads to better performance. This internal competition inhibits the motivation of a fab to fully support capacity to other fabs. Therefore, developing a "fair" trading mechanism to motivate each fab to actively support capacity to each other in order to maximize the company's total profit is very important.

A number of prior studies on semiconductor capacity trading have been published. Some studies assume a free-will trading scenario; that is, each fab can trade capacity at its own will. Chiang et al. (2011) propose an auction system for semiconductor capacity trading within a company. Some apply game theory to obtain an equilibrium capacity trading portfolio for multiple factories (Chien, 2013; Renna and Argoneto, 2011). The free-will approach drives toward maximizing each fab's benefit rather than the whole company's benefit. Wu & Chang (2007, 2008) assume an authority-will trading scenario; that is, the corporate headquarter determine an optimum trading portfolio for two fabs and each fab must strictly follow. The authority-will approach drives toward maximizing the whole company's profit but may distort the performance of each fab.

Extending from Wu & Chang (2007 , 2008), this research attempts to develop an optimum capacity trading portfolio and a fair pricing mechanism for multiple fabs in order to enable these fabs to effectively cooperate with each other to increase overall capacity utilization. This research involves two decisions. The first decision is to find an optimum capacity trading portfolio among the fabs. We use discrete event simulation, artificial neural network, and genetic algorithm to make such a decision. The second decision is to determine a "fair transfer price" for each unit of trading capacity of a workstation. The fair transfer price denotes that these fabs that participate in the capacity trading shall "fairly" share the overall benefit generated. This benefit sharing shall enable these fabs to have stronger intention to support each other in capacity.

This paper is organized as follows. Section 2 addresses how to determine the optimum trading portfolio among fabs. Section 3 presents the fair pricing mechanism. Numerical experiments are in Section 4 and Concluding remarks are in Section 5.

2. OPTIMUM CAPACITY TRADING PORTFOLIO

2.1 Problem Description

The capacity trading scenario is introduced below. The semiconductor foundry company has n fabs, which are located near to each other so that they can mutually support capacity. Each fab has m workstations. We denote W_{ij} as workstation i in fab j, and W_{i*} (workstation i in any fab) are functionally identical and can mutually trade capacity.

Before proceeding weekly capacity trading, each fab j has the following information: (1) job input schedule of the coming week and (2) work-in-process status of each workstation. By discrete event simulation software, we can obtain the weekly throughput and (2) the utilization of workstation of each fab j at the end of the coming week. The objective is to trade capacity of workstation W_{i*} among fabs in order to maximize the total throughput of all fabs. Notation for formulating the problem is listed in Table 1.

Table 1: Notation for determining capacity trading portfolio

Symbol	**Explanation**
IS_j^0	job input schedule of fab j at beginning
WIP_{ij}^0	work-in-process of workstation i in fab j at beginning
Th_j^0	weekly throughput of fab j before capacity trading
ρ_{ij}^0	utilization of workstation W_{ij} before capacity trading
Q_{ij}	number of machines in workstation W_{ij}
$\bar{\rho}_i^0$	average utilization of workstation W_{i*} before capacity trading
Th_j	weekly throughput of fab j after capacity trading
ρ_{ij}	utilization of workstation W_{ij} after capacity trading
u	basic unit of capacity trading (say, 6 hours)
T	total working hours of a week (say, 24 hours/day *7 days)
x_{ij}	trading capacity of workstation W_{ij}
$X_j = [x_{1,j}, \dots, x_{mj}]$	a capacity trading portfolio for fab j
$\mathbf{X} = [X_1, X_2, \dots, X_n]$	a capacity trading portfolio for all fabs; $\mathbf{X} = \bar{\mathbf{0}}$ denotes no capacity trading among fabs
R_{ij}	maximum capacity unit that can be traded in workstation W_{ij}

Based on the notation, we use the following formulation to present the problem.

$$\text{Max } \sum_{j=1}^{n} Th_j \tag{1}$$

$$\left(Th_j^0, \rho_{ij}^0\right) = Evaulate(\text{no trading or } \boldsymbol{X} = \overline{\mathbf{0}\ } \mid IS_j^0, WIP_{ij}^0) \quad 1 \le j \le n \tag{2}$$

$$\left(Th_j\ , \rho_{ij}\right) = Evaluate\ (\text{trading portfolio } \boldsymbol{X}\ |IS_j^0, WIP_{ij}^0) \quad 1 \le j \le n \tag{3}$$

$$R_{ij} = Round\_down\ (\frac{1}{u}\ |\bar{\rho}_i - \rho_{ij}^0| \cdot Q_{ij} \cdot T) \quad 1 \le i \le m, 1 \le j \le n \tag{4}$$

$$\text{If } (\bar{\rho}_i - \rho_{ij}) \ge 0, \text{ then } x_{ij} \ge 0 \text{ and } x_{ij} \in \{0, 1, \dots, R_{ij}\} \tag{5}$$

$$\text{If } (\bar{\rho}_i - \rho_{ij}) \le 0, \text{ then } x_{ij} \le 0 \text{ and } x_{ij} \in \{0, -1, \dots, -R_{ij}\} \tag{6}$$

$$\sum_{j=1}^{n} x_{ij} = 0 \tag{7}$$

In the above formulation, Eq (1) denotes that the objective is to maximize the total throughput of all fabs after trading capacity. Eq (2) denotes that the throughput and workstation utilization before trading can be evaluated by simulation, where $\mathrm{WIP}_{\mathrm{ij}}^{0}$ (WIP status) is obtained from MES (manufacturing execution system) and $\mathrm{IS}_{\mathrm{j}}^{0}$ (job input schedule of the coming week) is obtained from the production plan. Likewise, Eq. (3) denotes that the throughput and workstation utilization after trading can be evaluated by simulation. Eq (4) defines the upper bound of each workstation for trading capacity; and ensure that the trade quantity is a positive integer. Eqs (5) and (6) defines the possible trading quantity for each workstation; moreover, $\mathrm{x}_{\mathrm{ij}} \ge 0$ denotes that workstation Wij can only purchase capacity while $\mathrm{x}_{\mathrm{ij}} \le 0$ denotes that workstation Wij can only sell capacity. Eq. (7) ensures that the sum of sold capacity is equal to the sum of buy-in capacity for each workstation Wi*.

Implementation of Eq. (3) is explained in more detail as follows. In carrying out a capacity trading plan, some jobs of a fab have to move to another fab for using the buy-in capacity and move back. To simulate the job moving among all fabs, we need to take all fabs as a simulation system, which becomes quite complex and computationally extensive. To reduce computation time, we simplify the simulation by scaling up/down the processing time. If a workstation W_{ij} increases in available capacity after trading, we scale down the processing times of operations that go through W_{ij}; and vice versa. By such a simplification method, each fab can be independently simulated by an individual computer; as a result, all fabs can be processed in parallel and require much less computational effort.

Even so, solving the above formulation is still computationally extensive for two reasons. First, the solution space of trading portfolio is huge. Suppose each workstation has k alternatives to trade; each fab then has m^k trading portfolio; in turn, there are $m^{k \cdot n}$ trading portfolios for all fabs. Second, for each trading portfolio $\boldsymbol{X}$, we have to carry out one simulation for each fab which sums up to n simulations for all fabs. This implies that we need to carry out $m^{k \cdot n} \cdot n$ simulations for *exhaustively* evaluating the set of all possible trading portfolios $\{\boldsymbol{X}\}$.

2.2 Neural Network and Genetic Algorithm

To resolve the computational extensive issue, we apply the techniques of neural network (Dadhoff, 1990) and genetic algorithm (Gen and Cheng, 2000). The neural network technique is designed to emulate and replace the role of simulation in order to reduce the computation time for *evaluating* a capacity trading portfolio. The genetic algorithm (GA) is designed to effectively search the solution space rather than by applying an exhaustive search.

The neural network technique is explained below. Consider each fab *j* with capacity trading portfolio X_j as a complex system. Given a set of sampled input vectors, we can obtain a set of corresponding output vectors by simulation. Then, the set of input/output vectors (obtained by simulation) can be used to establish a neural network by a training process. The trained neural network can be used as an efficient performance evaluator. Given a new input vector, the trained neural network can be used to quickly compute its corresponding output vector. That is, Eqs (2) and (3) in the evaluated formulation can be evaluated by the neural network other than by simulation. Carrying out the valuations in Eqs (2) and (3) by neural network is much faster than by simulation. Prior research (Wu and Cheng, 2007) notices that one performance evaluation that takes a few minutes by simulation may requires only one thousandth second by its corresponding neural network, about 10,000 times faster.

The genetic algorithm (GA) is a solution space search technique, which has been widely applied in various disciplines. The basic idea of GA involves three major steps. First, we model a solution by a sequence of digits, which is called a chromosome ($\boldsymbol{X}$), herein a chromosome is a trading portfolio; and a population of chromosomes is randomly created. Second, we randomly pick a few chromosomes from the population; use crossover and mutation operators to generate new chromosomes; evaluate the performance of these new chromosomes; and the chromosome population is updated by including better quality chromosomes. Third, the chromosome population updating process is iteratively performed until a satisfactory solution is obtained. The solution ultimately obtained from GA is denoted by $\boldsymbol{X}^*$, which is the best solution that can be found from GA or called a near optimum solution.

3. PRICING DECISION FOR CAPACITY TRADING

This section presents a model for determining the transfer price of trading capacity among semiconductor fabs. In the pricing model, we assume that the corporate headquarter has determined the optimum trading portfolio X* that can maximize the total profit of all fabs; and the throughput of each fab after capacity trading has been obtained by simulation. Notations of the pricing model are partially listed in Table 2 and the others have been shown in Table 1.

Table 2: Notation for the pricing model

Symbol	**Explanation**
h	selling price per throughput wafer
x_{ij}	amount of trade capacity of workstation W_{ij}
c_i	weekly depreciation cost of workstation W_{i*}
p_i	unit transfer price of workstation W_{i*}
δ_j	increased profit of fab *j* due to trading capacity
r_j	percentage of increased profit of fab *j* after trading

The pricing model is formulated as follows.

$$\text{Minimize} \quad (r_{max} - r_{min}) \tag{8}$$

s.t.

$$\delta_j = (Th_j - Th_j^0) \cdot h + \sum_{i=1}^{m}(p_i \cdot x_{ij}), \quad j = 1, \ldots, n, \tag{9}$$

$$r_j = \frac{\delta_j}{Th_j^0 \cdot h} \qquad j = 1, \ldots, n \tag{10}$$

$$c_i \le p_i, \qquad i = 1, \ldots, m \tag{11}$$

$$r_{min} \le r_j, \qquad i = 1, \ldots, m \tag{12}$$

$$r_j \le r_{max}, \qquad i = 1, \ldots, m \tag{13}$$

$$r_j \ge 0, \qquad i = 1, \ldots, m \tag{14}$$

In the above formulation, Eq (8) is the objective function which attempts to make each fab share the total benefit after trading as fair as possible. That is, the increase percentage of profit (r_j) of each fab after trading must be as equal as possible. Eq (9) describes the increased profit of fab j after trading, which involves two parts. The first part is due to the increase of throughput which receives more revenue from outside customers. The second part is due to transfer pricing, the selling or buy-in capacity from other fabs. Eq (10) describes the increased percentage of each fab. Eq (11) ensures that the transfer price of a workstation shall be higher than its depreciation cost. Eqs (12) and (13) attempts model r_{max} and r_{min} to compute objective function. Eq. (14) ensures that each fab must get more profit after trading.

4. NUMERICAL EXPERIMENTS

Numerical experiments of the proposed method have been carried out. The experiment assumptions are stated below. A semiconductor foundry company has three fabs which trade capacity weekly. Table 3 shows the production information of each fab, which is a simplified case of a real-world fab provided by a semiconductor company. A basic unit of capacity trading is 20 machine hours.

By the proposed method, the optimum capacity trading portfolio is shown in Table 4. The table indicates that 6 workstations need to trade capacity. For workstation WS6, Fab 1 has to sell 10 units, Fab 2 has to buy 4 units, and Fab 3 has to buy 6 units. Table 5 shows the profit of selling throughput wafers before and after trading capacity. See the table, after trading capacity, Fab 1 increases the throughput of 4P1M products, Fab 2 remains the same in throughput, and Fab3 increases the throughputs of both 1P5M and 1P6M products. Trading capacity among the three fabs indeed increases the overall throughput of the three fabs.

Now we discuss the transfer price obtained by the proposed method. Table 6 shows the transfer price of each workstation, and compares the transfer price against the depreciation cost of capacity. Of the 6 workstations, the transfer prices of WS5 and WS2 are most expensive, respectively 3.8 and 3.4 times of their depreciation costs. In contrast, the transfer prices of the other workstations are about equal to their depreciation costs. Namely, workstation WS5 and WS2 are two most precious workstations in the coming week, which shall be effectively used and fairly charged when a demand for supporting capacity is issued.

Table 7 compares the total profit of each fab; which includes the profit of selling throughput wafers and the profit of selling capacity to other fabs. The table shows that each fab increases profit by about 16.0% after trading capacity if we adopt the proposed transfer price system. This finding supports that the transfer prices of capacity obtained by the proposed method is quite “fair” a fair; each fab fairly shares the overall benefit generated by trading capacity. In contrast, if there is no charge when we implement the trading capacity plan (i.e., transfer price is zero), then Fab 3 receives most favor; its profit increases 42.2%; yet Fab1 and Fab receives only 7.7% and 0.7%. Apparently, such a substantial difference in profit may discourage the fab that supports capacity to other fabs. As a result, capacity support by negotiation can never be effectively implemented if a “fair” transfer pricing mechanism is not established.

Table 3: Production Information of Each Fab

FAB	Number of	Number of	Product	Total processing times	Total number of
Fab 1	60	270	4P1M	402	358
			1P7M	443	412
Fab 2	60	207	1P3M	315	276
			1P8M	477	446
Fab 3	60	210	1P5M	381	344
			1P6M	411	378

Table 4: Optimum capacity trading portfolio (1 unit = 20 machine hours)

FAB	WS 1	WS 2	WS 3	WS 4	WS 5	WS 6
Fab 1	− 2	1	2	0	4	− 10
Fab 2	2	− 1	− 11	1	9	4
Fab 3	0	0	9	− 1	− 13	6

Table 5: Wafer throughput profit before and after capacity trading

Fab	Fab 1		Fab 2		Fab 3	
Product Types	4P1M	1P7M	1P3M	1P8M	1P5M	1P6M
Profit per wafer	$ 65	$ 80	$ 50	$ 100	$ 90	$ 85
Throughput (Before Trading)	2,025	1,250	1,925	875	750	1,350
Throughput (After Trading)	2,300	1,250	1,950	875	1,250	1,725
Total Profit (Before Trading)	$ 231,625		$ 183,750		$ 182,250	
Total Profit (After Trading)	$ 249,500		$ 185,000		$ 259,125	

Table 6: Ratio of transfer price to depreciation cost

	WS 1	WS 2	WS 3	WS 4	WS 5	WS 6
Transfer Price (p_j)	$ 600	$ 2702	$ 1,200	$ 300	$ 4,608	$ 300
Depreciation Cost (c_j)	$ 600	$ 800	$ 1,200	$ 300	$ 1,200	$ 300
$\tilde{p}_j = p_j/c_j$	1	≅ 3.4	1	1	≅ 3.8	1

Table 7: Total profits before and after trading

	Fab 1	Fab 2	Fab 3	All Fabs
Before Trading	$ 231,625	$ 183,750	$ 182,250	$ 597,625
After Trading: No Transfer Price	$ 249,500	$ 185,000	$ 259,125	$ 249,500
Increased %: No Transfer Price	7.7%	0.7%	42.2%	16.0%
After Trading: With Transfer Price	$ 268,834	$ 213,270	$ 211,521	$ 693,625
Increased %: With Transfer Price	16.0%	16.0%	16.0%	16.0%

5. CONCLUSION

This research addresses an internal capacity trading problem for a semiconductor company that involves many fabs which need to trade capacity every week in order to effectively utilize the machines. The weekly capacity trading problem involves two decisions. First, the corporate headquarter has to determine an optimum trading portfolio that each fab must strictly follow for the coming week. Second, it also needs to determine appropriate transfer prices of the coming week to fairly reward the capacity support activity.

To solve the problem of finding an optimum trading portfolio, we proposed a method that uses simulation, neural network, and genetic algorithm techniques. To solve the problem of determining a fair transfer price system, we proposed a linear program. The capacity transfer price of a workstation shall be changed every week, depending upon the its demand. The higher is the demand, the higher the transfer price.

The contribution of this research is two-fold. First, we extend the work of prior studies from a two fab trading problem to a multiple fab trading problem, in the context of finding an optimum trading portfolio. Second, this research is a pioneer study in determining a fair transfer price system among semiconductor fabs. With a fair transfer price system, it shall greatly help motivate fab managers to actively support capacity to each other in order to increase the overall benefit of the company.

ACKNOWLEDGEMENTS

This research is financially supported by National Science Council, under a research contract NSC102-2221-E-009-148-MY2.

REFERENCES

Chien, C. F. and Kuo, R. T. (2013) Beyond make-or-buy: cross-company short-term capacity backup in semiconductor industry ecosystem. *Flexible Services and Manufacturing Journal*, **25** (3), 310-342.

Chiang, A. H., Chou, Y. C. and Chen, W. H. (2011) A capacity evaluation and trading model for semiconductor foundry manufacturing. *International Journal of Advanced Manufacturing Technology*, **54** (1), 1-10.

Dayhoff, J. E. (1990) *Neural network architectures: an introduction*, Van Nostrand Reinhold Co. New York.

Gen, M. and Cheng, R. (2000) *Genetic algorithms and engineering optimization*, John Wiley & Sons.

Renna, P. and Argoneto, P. (2011) Capacity sharing in a network of independent factories: A cooperative game theory approach. *Robotics and Computer-Integrated Manufacturing*, **27** (2), 405–417.

Wu, M. C. and Chang, W. J. (2007) A short-term capacity trading method for semiconductor fabs with partnership. *Expert Systems with Applications*, **33** (2), 476–483.

Wu, M. C. and Chang, W. J. (2008) A multiple criteria decision for trading capacity between two semiconductor fabs. *Expert Systems with Applications*, **35** (3), 938–945.