ELSEVIER

Evaluating alternative capacity strategies in semiconductor manufacturing under uncertain demand and price scenarios

Yon-Chun Chou*, C.-T. Cheng, Feng-Cheng Yang, Yi-Yu Liang

Graduate Institute of Industrial Engineering, National Taiwan University, 1 Roosevelt Road, Section 4, Taipei 106, Taiwan

Abstract

Because product demand is very volatile and machine and process technologies are advanced at a rapid pace, capacity planning, and investment in the semiconductor industry is a challenging task. Many studies have addressed mid- and short-term machine portfolio planning tasks of capacity planning. However, capacity planning should be considered in a framework of strategy planning in order to address the whole problem. This paper describes a method of analysis for adapting capacity strategy to changing business environment. The paper has three major parts. In the first part, empirical data analysis of a case study company yields predictive formulas for production and capacity costs. The uncertainty of demand is also calibrated using the geometric Brownian motion process. In the second part, experiments are designed to simulate demand and profit scenarios for comparing two capacity strategies. In the third part, the resultant data of simulation are analysed to gain insights to the relative performance of the two strategies in various scenarios. The analysis method provides a framework for formulating capacity strategy and for integrating capacity planning with business planning.
© 2006 Elsevier B.V. All rights reserved.

1. Introduction

Capacity planning and investment is a challenging task in the semiconductor manufacturing industry. The industry supplies integrated circuit devices to many end-product industries such as computer, communications, and electronics, which have dynamic market demands themselves. Due to the bullwhip effect of the supply chain (Geary et al.,

2006), the demand as faced by semiconductor manufacturers is very volatile and the industry is plagued with repeating cycles of over- and under-capacity. Furthermore, machine and process technologies are advanced (and become obsolete) at a rapid pace. The cycle time of one generation of technology is no more than 3 years, yet it takes at least 9 months to incrementally add capacity to an existing plant and at least a year to equip a bare clean room of the factory to produce the first wafer. Complicating the matter of capacity investment is the capital cost of capacity. A modern semiconductor wafer fabrication plant requires a capital

*Corresponding author. +886 2 3366 9501;
fax: +886 2 2362 5856.
E-mail address: ychou@ntu.edu.tw (Y.-C. Chou).

investment of 2–3 billion US dollars and equipment depreciation accounts for approximately 70% of the total production cost. Investing in one single factory is a high financial risk for even the largest companies in the industry.

Volatile demand, risk of technology obsolescence, long lead time, and high investment cost contribute to making capacity investment an overwhelming decision-making task for which a satisfactory planning methodology has not been found. Fig. 1 shows the historical data of aggregate capacity and wafer output of a major semiconductor manufacturing enterprise over a period of 10 years. The aggregate capacity increased by 10 times. The growth rate was higher in earlier years, during which times the trend was also largely predictable. However, in the later half of the period the growth rate stabilized and began to show significant oscillation. The reasons for stabilization can be attributed to market saturation and poor return on capacity investment. The reason for oscillation is attributable to the volatile nature of demand in the industry.

Fig. 2 is the quarterly total assets and the return on equity (ROE) of the company from 1994 to 2003. While the capacity continued to increase, the ROE has a decreasing slope for the trend line. The ROE of the company is at a level no better than what can be commonly found in many matured industries. This phenomenon is not unique to any particular company in the industry; there are actually many worse performers (Liang and Chou, 2003). Figs. 1 and 2 also illustrate a pattern of life cycle similar to what is exhibited in the growth process of many high-tech industries and corporations. At the initial stage of growth, product demand increases rapidly and profit margins are high. Most companies are naturally inclined to pursuing a strategy of aggressive expansion. However, as competition intensifies or the market begins to saturate, profit margins will decrease and the ROE will settle down to certain equilibriums. A conservative strategy of expansion would increasingly be justifiable, especially when the market is volatile. In fact, a higher level of volatility might call for a more conservative strategy. If the industry enters the matured stage with slowed
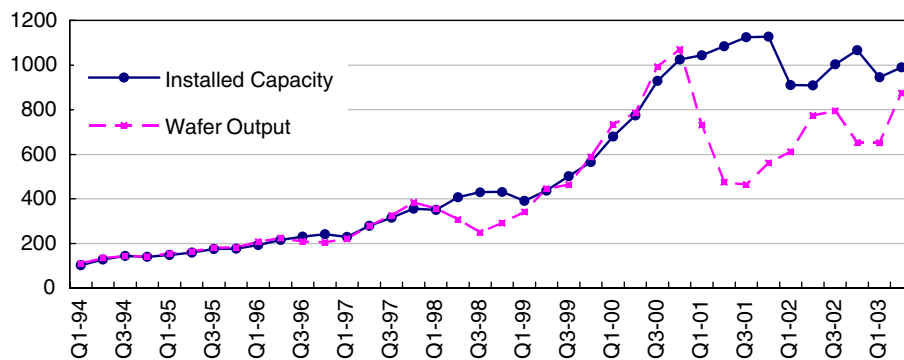

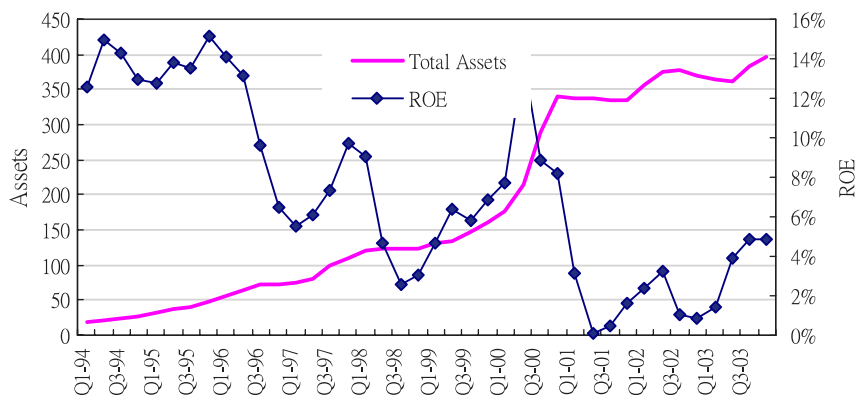Fig. 1. Growth in capacity and output of a semiconductor manufacturer.

Fig. 2. Trends of total assets and return on investment.

Table 1
Capacity planning tasks and objectives

	Objective	Tasks
Short-term	Order fulfillment	Order rescheduling Alternative routing
Mid-term	Machine portfolio optimization	Machine purchase
		Machine decommission
Long-term	Supporting business planning in technology development and product planning	

growth but the company strategy remains aggressive, aggressive expansion will put the company in great financial risk.

Capacity planning has received significant attention in the literature and in practice. Because the lead time of capacity provisioning is long, capacity planning tasks can be classified by their planning horizon (Table 1). In the short-term, the overall capacity is largely fixed, but with some room for adjustment through equipment set-up change-over (i.e., alternative routing). Therefore, capacity planning problems are mainly about capacity allocation among job orders and alternative routing planning. The granularity of capacity requirement analysis is at the machine and process step level. In the medium term, capacity can be changed by tool purchase and decommission. It should be noted that capacity is expanded in small increments, by gradually populating the factory with more machines. Capacity planning in this time frame is mainly a tool portfolio[1] configuration problem. The granularity of planning is at the critical machine and major process stage level (Nazzal et al., 2006). In the long term, the objective of capacity planning is to prepare for plant transition in anticipation of new process technology and new product and to support strategic plans of business.

A core issue of capacity planning is to configure the tool portfolio. There is a combinatorial optimization aspect to this problem. The computation of tool portfolio in semiconductor manufacturing surpasses that of other manufacturing systems in complexity. A wafer fabrication plant contains hundreds of tools which are used to process tens or hundreds of product types (called process flows).

---

[1]Machines are called tools in this industry. The cost of a single tool could be as high as 10 million US dollars. In this paper, the words machine, tool and equipment are used interchangeably.

The manufacture of a product requires several hundred steps in its process flow, with a tool visited in each step. Most of the studies in the literature address the problem of tool portfolio optimization. Swaminathan (2000) addressed the tool procurement decision over a planning horizon of multiple time periods. The uncertainty of product demand is modelled by a set of demand scenarios. A large integer programme was used to minimize the expected stock-out cost over all demand scenarios. Hood et al. (2003) followed the stochastic programming approach under demand uncertainty. The tool purchase, wafer starts, and work assignment decisions are formulated as a large mixed-integer programme. Ahmed (2002) used multi-stage stochastic programming instead of two-stage stochastic programming to better model the flexibility of tool purchase in later time periods bestowed by occurrence of new events. The paper compares the relative merit between two-stage and multi-stage stochastic programming, and finds that the merit increases with the number of stages and the number of decision branches per stage. Christie and Wu (2002) used a stochastic programming formulation for tool portfolio planning at the multiple-plant level.

A few papers addressed long-term and process issues related to capacity planning. The plant-sizing problem is addressed by Benavides et al. (1999). Product demand is assumed to follow a geometric Brownian motion (GBM) process. The timing decisions are modelled as an optimal stopping problem. The paper shows that optimal trade-off between economics of scale and flexibility can be reached by a sequential deployment of modular plants. Karabuk and Wu (2001) considered the coordination problem between production manager and marketing management within one firm, that is, to blend the two forces that drive the decisions on capacity planning. Two types of uncertainties are modelled: one is capacity estimation and the other is demand volatility. Stochastic programming is used to solve the optimal capacity plan.

There are four models of tool portfolio and capacity planning in the literature: (1) spreadsheet models, (2) multi-criteria search procedures, (3) mixed-integer mathematical programming models, and (4) stochastic programming models. The objective functions typically include terms of revenue, inventory cost, outsourcing cost, equipment cost, shortage cost, and machine utilization. Spreadsheet models are often used in practice

(Neudorff, 1999; Witte, 1996; Wu et al., 1998). They are used to compute utilization and throughout. Since flow time is also an important performance measure of tool portfolio, multi-criteria search procedures can also be used to find the optimal tool portfolio (Chou and Wu, 2002). In the literature, the most common models for capacity planning are to use mixed-integer programming or stochastic programming to optimize tool portfolio decisions. Stochastic programming is a probabilistic optimization approach for many kinds of planning problems. The uncertainty is usually modelled by a finite number of scenarios. Due to different structures of the decision process, the stochastic programs are either formulated as two-stage or multi-stage. Depending on application or focus, the decision variables range from machine to plant, and the scope may cover one or multiple time periods and one or multiple plants. In the literature, the conventional logic of capacity planning is to have sufficient capacity which will satisfy product demands. The goal is typically to maximize the profit, without considering the associated risk.

Two capacity strategies can be identified in the literature: reactive and conservative. Liang and Chou (2003) utilized the real-option theory in determining capacity level. Capacity decisions are made by taking into consideration the option of waiting for more demand information to materialize. This method amounts to a conservative strategy. An alternative strategy is to reactively adjust the capacity plan according to changes in demand forecast. If the demand is volatile, or capacity investment is irreversible, or the lead time of capacity expansion is long, this strategy could easily lead to imbalance between the capacity and demand.

The two strategies have their respective strength; their relative performance will depend on the circumstance. In practice and in the literature, the reactive strategy is the dominant strategy. Even though demand is known to be uncertain, the common objective of capacity planning is still to "satisfy" the demand. This inclination can be seen from the objective function of mathematical models for tool portfolio planning. The objective functions are primarily to maximize the revenue (or profit) subtracted by penalty for capacity shortage. This strategy is a logical consequence of the corporate structure of many enterprises. For example, in the so-called integrated design and manufacturing

companies, product plans and demand plans are developed first. A capacity plan must support their execution. Consequently, the demand forecast that is implicit in product and business plans will be used as the input to capacity planning. Although the uncertainty must have been taken into consideration in business planning, the context is likely to be lost in transferring the demand data to capacity planning.

The four planning tasks identified in Table 1 have not fully addressed the challenges posed by the uncertainties in demand, technology, and the return on investment. In industries in which manufacturing capacity is flexible and requires little capital investment, business planning of product lines, marketing, and pricing does not have to be tightly integrated with capacity planning. Capacity planning can be done sequentially after business planning is completed. However, if capacity investment requirement is large and investment is irreversible, capacity planning must be integrated with business planning; otherwise, financial well-being would be subjected to serious risk. Capacity investment in semiconductor manufacturing is huge; it has a direct impact on the subsistence of manufacturing capability. Therefore, the corporate business strategy must be translated to, supported by, and integrated with its capacity strategy. In addition, capacity strategy must be adjusted to the life cycle of product, enterprise, and even the industry. It is well known that products and processes usually have a life cycle. As revealed in Fig. 1, manufacturing enterprises also have a life cycle, although in a slightly different way. In the initial stages of rapid growth, capacity strategy should be more aggressive in order to secure market share. As the market matured, the capacity strategy should be more conservative in order to avoid technology obsolescence. However, it remains to be determined what the trigger condition is for a shift of strategy.

The focus of this paper is on the problem of adapting capacity strategy according to the characteristics of demand and business plan. The treatment of the problem will complement the four tasks of Table 1 in addressing the problem of capacity planning. The problem background—uncertainties, high investment cost, long lead time of capacity provisioning, etc.—are not unique to the semiconductor industry. Other emerging high-tech industries, such as the liquid crystal display industry, also face similar challenges. The semiconductor industry has a mass production history of

approximately 30 years. It has gone through several economic cycles and competition is fierce. Fig. 3 shows the worldwide microchip sales over the last 30 years in billions of dollars (the last data point is an estimate). The trend before 1995 appeared predictable, but the industry seemed to enter an era of great volatility after 1995. Semiconductor manufacturing as a case study offers unparallel opportunities to study modern challenges of capacity planning and investment.

In this paper, a framework of analysis and design of capacity strategies in an uncertain environment of demand and technology is described. The framework is supported by empirical data analysis in semiconductor manufacturing. The remainder of this paper comprises of four parts. In Section 2, results of empirical data analysis for the characteristics of demand, production cost, and revenue trend of semiconductor manufacturing are presented. Then, scenario design for strategy analysis is described in Section 3. In Sections 4 and 5, two strategies of capacity planning are analysed and compared by using statistical methods. Finally, Section 6 concludes the discussions.

2. Data analysis

Capacity planning usually takes into consideration market share, revenue, profit, ROE, and industry trend. In the first four subsections of this section, the financial and capacity data of a case study company are first analysed in order to estimate the trends of the fixed assets cost, material and labour costs (ML), operating expenses (OE), and average selling price (ASP). Then, in Section 2.5, worldwide demand of several market segments is analysed to characterize the variability of product demand. The analysis yields surprising results which make it possible to analyse different strategies of capacity. The data of the case study company covers a period of 10 years (from 1994 to 2003).

2.1. Capacity cost

The fixed cost of semiconductor manufacturing is primarily incurred by investment in machines and equipment. It is largely irreversible and is reflected in the fixed assets and accumulated depreciation in the financial statement. Fig. 4 is the historical
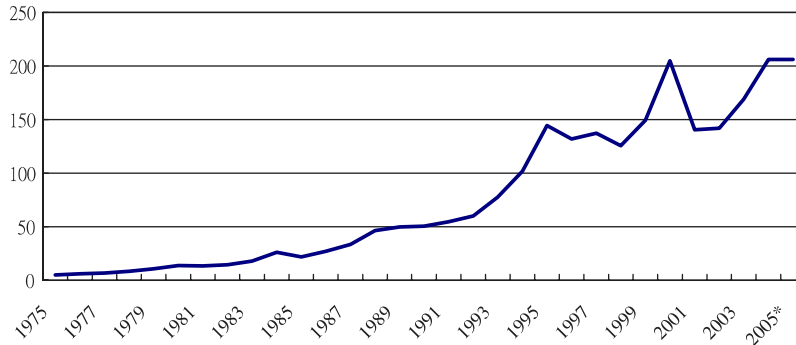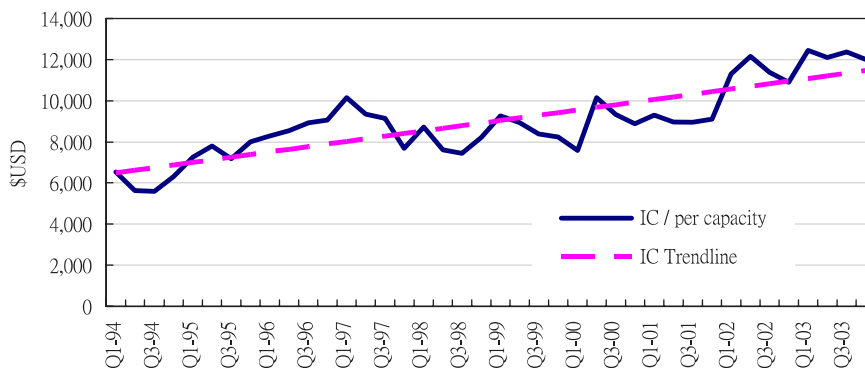

Fig. 3. Worldwide microchip sales.

Fig. 4. The fixed cost per unit capacity per quarter.

capacity cost per capacity unit (200 mm quarterly wafer equivalent). A trend of increasing cost could be easily observed. It also appears that a linear trend line is a good approximation. Using the first quarter of 1994 as the base period, Eq. (1) is the resultant regression equation for the fixed cost per capacity unit ($R^2 = 0.689$)

$$FC(t) = 6374.95 + 127.25t. \tag{1}$$

The cost increases by approximately 2% per quarter (127.25/6374.95) at the base period. Considering that prices are affected by many factors, this result that capacity cost can be reasonably predicted with a linear trend line is quite surprising; some explanation of its possible cause is warranted. The technology content of processing equipment in semiconductor manufacturing is very high. Equipment providers must continue to invest heavily in research and development (R&D). As the technology is advanced, the cost of developing new machines also increases. For each type of processing machines, there are only a few providers worldwide—and mostly in the developed countries. Because the equipment market is an oligopoly with strong competition, it is logical to expect that the profit margin and cost of R&D would maintain their stable rate of increase. Therefore, the cost of manufacturing capacity is likely to also maintain similar appreciation.

2.2. Direct materials and labour costs

By definition in cost accounting, direct materials and labour (M&L) costs are proportional to production output. Fig. 5 shows the average M&L costs per wafer that is manufactured. Similar to capacity cost, the direct M&L costs can also be reasonably modelled by a linear trend. Eq. (2) is the resultant regression equation:

$$ML = 406.28 + 3.37t \text{ (per wafer output).} \tag{2}$$

The cost increases by approximately 0.83% per quarter at the base period. Direct materials, i.e., raw wafers and labour are mostly sourced locally. Comparing Eqs. (1) and (2), the slope and the increase in percentage of the ML costs are much smaller than those of the capacity cost.

2.3. Operating expenses

OE include advertisement, marketing, and R&D costs. OE are proportional to the scale of operations, which could be measured by the capacity level or the output of production. The correlation coefficient between the OE and capacity and between OE and output turns out to be 0.94 and 0.88, respectively, indicating high level of correlation. Fig. 6 is a graph validating that it is more accurate to use capacity level than to use production output to estimate the unit OE. A second supporting argument can be gleaned from data in Fig. 1. The production output is immediately influenced by the ups and downs of the economic cycles, but OE are more resilient to economic fluctuations. In fact, some components of OE, such as R&D expenditure might be immune to short-term fluctuations of market demand. Thus, using the production output as the denominator to compute the unit OE is more likely to result in distortion.

Eq. (3) is the resultant regression equation for unit OE. It turns out that the average operating expense per unit of capacity could be reasonably treated as a constant.

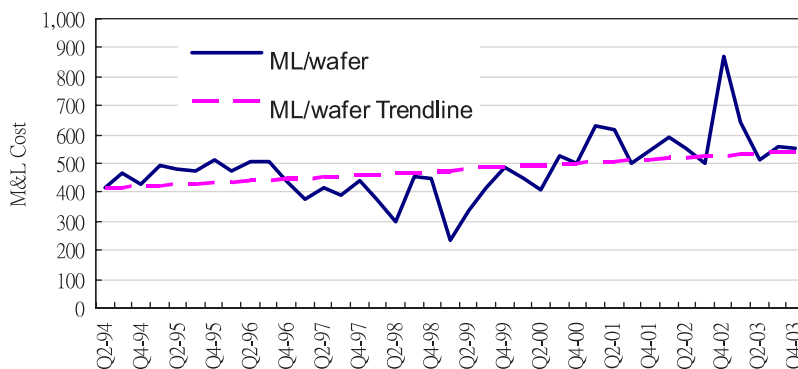$$\text{Operating expenses} = 131.44 \text{ (per unit of capacity).} \tag{3}$$

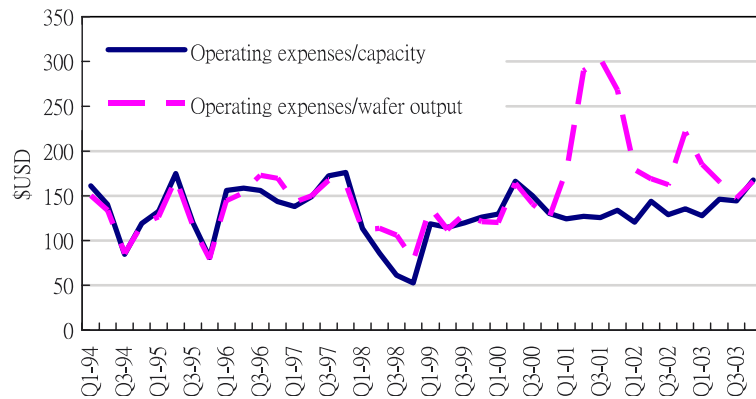Fig. 5. The direct materials and labour costs per wafer.

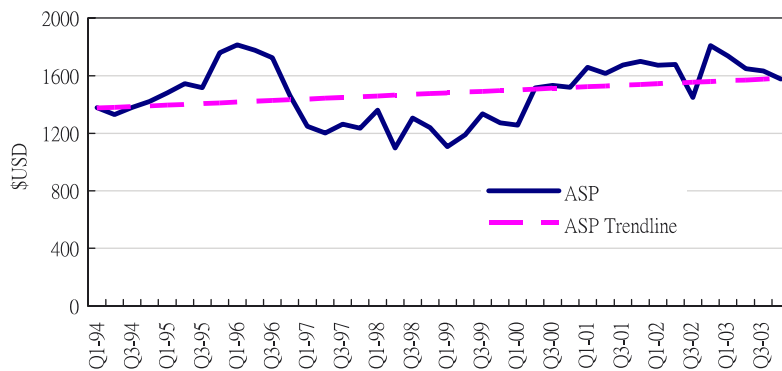Fig. 6. Operating expenses per unit of capacity and output.

Fig. 7. Average selling price and its trend.

Since one would normally expect that any cost to increase at a rate that is at least equal to the inflation rate in the long term, rather than to stay at a constant, we will offer an explanation. It is plausible that the effect of inflation is cancelled out by the increase in the scale of production. (*Note*: The average operating expense per wafer output does have a regression line of positive slope. However, the slope is mere $1.93 USD per quarter.)

Contrary to common belief, R&D cost is not significant as a percentage of the total cost. In Fig. 6, R&D expense is one of three components of the OE. The explanation is quite simple. Although R&D cost is not insignificant in absolute term. It is paled by the cost of capacity. Note that one single machine could cost 10 million US dollars and R&D work is done on production equipment.

2.4. Average selling price

As shown in the previous three subsections, the fixed costs, direct M&L costs, and OE have very

predictable trends. By using Eqs. (1)–(3), the cost of future capacity plans could be estimated with accuracy. In comparison, the ASP manifests another phenomenon. Fig. 7 is the ASP[2] of the case study company over the same 10-year duration. It is rather surprising to observe that no significant progress was made in the ASP, while the company has made significant progress in technologies and has became an industry leader in her sphere of business. Semiconductor manufacturing companies are compelled to continuously advance their process technologies. However, the ASP is under tremendous market pressure; it is not very easy to make progress in this aspect. This fact is in stark contrast to the increasing trend of capacity cost and it plays a major role in exacerbating the need to adapting capacity strategy.

Two trends can be gleaned from data in Fig. 7. First, linear regression can be used to construct a

---

[2]Computed by dividing the total revenue by the total output in wafers.

long-term trend line over the entire duration:

$$ASP(t) = 1370 + 5.28t. \tag{4}$$

Because the technology cycle time is 2–3 years on average but different companies' progress might not be in sync, a company might gain advantage in technology temporarily. But, over the long run the advantage might or might not be sustainable. Therefore, an alternative visualization of the ASP trend is possible. In addition to Eq. (4), the ASP could be alternatively modelled as a step function. The ASP remains flat over medium term. For instance, the ASP is approximately $1200 between 1997 and 2000; it then increased to $1600 after 2000. Both long-term and step function trends are relevant in capacity planning, as the life time of equipment falls somewhere in between. In the next section, both scenarios are included in the analysis of strategy. Also, we will further justify flat ASP (in the short run) by analysing the relative ASP of multiple companies.

2.5. Demand variation and modelling

That the variation of demand is very high in the semiconductor industry is well known. However, how to measure the variability and incorporate the variability factor in capacity planning remains to be studied. GBM models have frequently been used to model the movement of stock prices, a best-known example of variability. The valuation formula for stock options that were derived by Black and Scholes (1973) is based on the assumption of GBM processes. Benavides et al. (1999) also used a GBM process to model the demand of semiconductor products. The formula of the GBM model adopted in this paper is first summarized in the next paragraph, followed by calibration of its parameters using industry data.

Let q_t be the demand of time t. If q_t follows a GBM process, then the change in demand can be expressed by the equation $dq_t = \mu q_t\, dt + \sigma q_t\, dw_t$, where $dw_t = \varepsilon_t\sqrt{dt}$, and ε_t follows the standard normal distribution. The μ is called the drift parameter which can be "roughly" interpreted as a measure of the long-term growth rate and σ is the volatility parameter. Let the base period be t_0. The logarithm of the drift parameter r_t will follow a normal distribution with mean $(\mu - \sigma^2/2)t$ and variance $\sigma^2 t$ (Tsay, 2002):

$$r_t = \ln\left(\frac{q_{t_0+t}}{q_{t_0}}\right) = \ln(q_{t_0+t}) - \ln(q_{t_0}),$$

$$r_t \sim N[(\mu - \sigma^2/2)t, \sigma^2 t]. \tag{5}$$

In the Industry Economic Model (International SEMATECH, 2002), the worldwide demands of semiconductor integrated circuits (IC) are classified by process type and technology level into several categories: leading edge memory, leading edge logic, other leading edge, other IC, and other semiconductor circuits. Historical data and forecasts for the demands of the five market segments are attached in the Appendix. It can be easily observed that there is significant variation in demand. In order to calibrate the variation, we first applied the Dickey–Fuller test to the five time series. At the significant level of 0.05, the p-value is far greater than 0.05 for all five time series, indicating that they are non-stationary. Next, the original time series are converted to time series of logarithmic growth r_t. The resultant time series are stationary (left figure of Fig. 8), with a p-value of 0.0011, 0.0018, 0.0006, <0.0001, and 0.0009, respectively. Furthermore, the logarithmic growth has a bell-shaped histogram (right figure). The objective of this study is not about developing a model for demand forecast. Instead, we are interested in finding a mathematical tool that is reasonably accurate in calibrating the
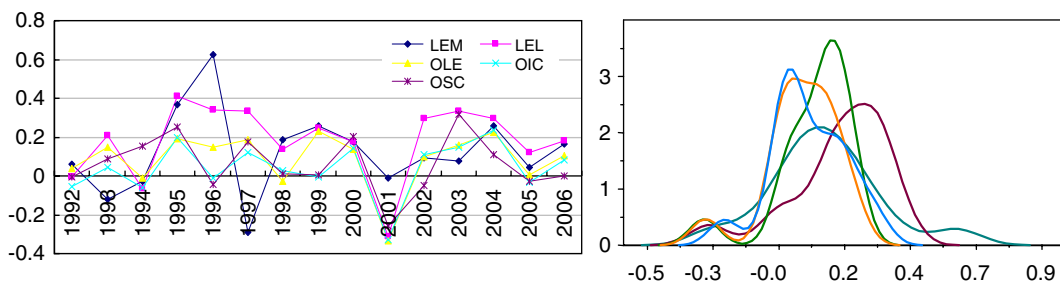

Fig. 8. The logarithmic growth of five product segments.

variability. Based on the empirical evidence of Fig. 8, we believe that the GBM process is a reasonable choice.

3. Designing scenarios for strategy analysis

In order to integrate capacity planning with business planning, we will first describe some specificity of the industry and the industrial practice in order to define the problem. In this industry, the "shell" and clean room of the factory is built first and machines are installed gradually over time. Depending on the market condition, it usually takes 2–3 years to build up a factory to its full capacity. Thus, capacity expansion is usually a "continuing" and dynamic process of small increments (as shown in Fig. 1). This is quite different from capacity expansion of most other industries, in which expansion projects have well-planned schedule and size. We believe that this continuous nature warrants some rethinking of investment analysis which is addressed in this section.

The semiconductor manufacturing industry is normally associated with high technology. However, it is rather ironic that companies with advanced process technology are not guaranteed to be the winners. Over the last 10 years, many companies in this category chose to either exit from the industry or scale down production and to outsource manufacturing to foundry manufacturers. But, if technology is not the sole sufficient condition of surviving, then what are the other conditions? Because the volatility in demand and the high cost of capacity create great financial risk, our hypothesis is that a sound strategy of capacity investment is one of them and good financial performance is the other and these two factors are closely related. Another characteristic of the industry is that companies in the industry must have the will and resources to stay in the industry for the long haul in both technology development and capacity investment. The entry and staying power barriers are high. Therefore, we argue that the normal criteria of investment analysis are not sufficient. Before we continue, we will first define the concept of perpetual manufacturing capability.

Definition (*Perpetual manufacturing capability*). Perpetual manufacturing capability is the capability of a company to continuously thrive in high-technology manufacturing for the long run. This capability will include the ability to continuously develop new manufacturing technologies and to

have sufficient financial resources to continuously invest in modern manufacturing facility. The emphasis of this definition is on perpetuality. Outsourcing management is excluded from manufacturing capability.

We next stipulate the requirements on the returns on capacity investment for perpetual manufacturing capability. Because technologies (including machines and processes) are advanced in a rapid pace in the industry, all companies have little option but to continue investing in R&D and in new technology; technology laggards will quickly be forced to make a decision to exit the industry. In most industries, investment decisions are justified by some forms of ROI analysis. A project is deemed economically feasible, if its expected ROI exceeds a certain threshold level. Because of the perpetual nature, capacity investment in this industry should not be evaluated by using the one-time ROI analysis alone; more stringent requirements need to be devised.

Investment projects can normally be funded by raising capital, through debt or equity. However, the capacity cost in this industry has another escalating cost facet, beside an increasing trend. The industry is at the juncture of migrating from older 200 mm-wafer to 300 mm-wafter fabrication plants. A 200 mm-wafer plant requires an investment of $1 billion dollars while a 300 mm wafer plant requires $2–3 billion dollars. This creates a quantum hurdle for perpetual capability. (Some companies have tried to share the cost through joint venture. But this business model has not produced success stories.)

Because of the perpetuality requirements, we argue that investment on capacity must generate sufficient net cash flow and show good performance on returns. New investment must be at least partially funded by retained cash flow. Also, the expected performance on accounting measures must be sound; otherwise, it would be very difficult to raise capital for the next generation of capacity, given the tremendous risk and poor track records of the industry.

Because capacity decision has a direct effect on total fixed assets and the depreciation of fixed assets accounts for a high percentage of the production cost, we propose that operating income (OI) and return on fixed assets (ROFA) be used as the performance measures for evaluating capacity strategies. OI is computed by subtracting ML, OE and depreciation cost (DC) from revenue (R),

as shown in the following equation. By including DC, the OI measure we used is therefore a modified measure of OI:

$$OI_t = R_t - ML_t - OE_t - DC_t$$
$$= ASP_t O_t - ML_t - OE_t - FA_t/l,$$

$$ROFA_t = OI_t/FA_t, \qquad (6)$$

where O_t is the production volume, FA_t is the fixed assets, l is the life time, and t is the time index. OI is essentially the net cash flow with some minor adjustment to cash outflow due to equipment purchase. Because the unit cost of equipment could be very significant, cash outflow normally has high level of fluctuation. In Eq. (6), equipment expense is spread over the equipment life time. We chose to use ROFA as the second performance measure, instead of the more common ROE, in order to have a more direct link between capacity decision and financial performance.

The OI and ROFA measures are complimentary. OI can be regarded as a surrogate measure for market share (since the majority of the costs are sunken), whereas ROFA is a surrogate measure for the ROE. Having a large market share does not guarantee a higher ROFA. In fact, the contrary might to be true in the problem setting that is under discussion here. We will use a numerical example to show the computation of the threshold to demonstrate how the perpetuality requirement can be translated into requirements on ROFA and OI.

Numerical Example: ROFA and OI.

The equipment life time is 5 years on average by the industry norm. We could postulate that if a company owning a 200 mm plant is to migrate to a 300 mm plant in 5 years, its equity must show an increase commensurate with the increase in the

factory cost. Otherwise, the company will not be able to survive in the industry. Therefore, to achieve the perpetual capability, and considering capacity expansion is continuous, the ROFA should be at least 2.0 in 5 years, or 3.53% per quarter:

$$(1 + ROFA)^{20} \geqslant 2.0 \Rightarrow ROFA \geqslant 3.53\%.$$

Assume a current initial capacity of one million wafers per quarter (Fig. 1). That capacity will have a remnant value of zero after 5 years of depreciation. Based on the trend of capacity cost, the unit cost of capacity will be $14,000 in 5 years. Therefore, it can be computed that a total of 14 billions of OI must be accumulated over the 5 years in order to replenish the capacity.

We will now describe the business scenarios in which competing capacity strategies are compared. The scenarios are defined by the variable factors of demand and profit. In this study, the GBM process has been used to calibrate the uncertainty of demand. As the demand is beyond the control of any single company, the drift and volatility parameters are treated as endogenous variables. It has been shown in previous section that the ASP can be modelled as a linear function of time or as a step function. Both models suggest that the ASP could be treated as constant during the life time of equipment. Fig. 9 shows the recent ASP of major competitors of the industry. This industry is an oligopoly with strong competition between firms. Microchips are used in numerous computer, communications, and consumer electronic products. But they account for a small part of the total cost of the end products. Pricing at the semiconductor manufacturing stage usually does not have a large impact on the demand of end products in the long run. Therefore, price is usually not a management lever
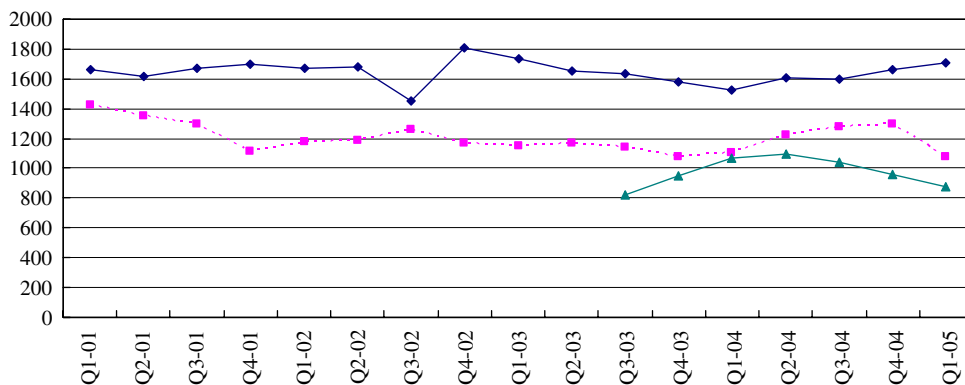

Fig. 9. Relative ASP of competing companies.

to influence demand. (An exception might be CPU products. But, CPU business is unique and is not within the scope of this paper.) Instead, the competitive factors are manufacturing technology and engineering and manufacturing services (such as design services, responsiveness, capacity plan, supply reliability, and order flexibility). The latter has become more important in the supply chain of semiconductor manufacturing. As shown in Fig. 9, although there is significant difference in prices, the relative prices are remarkably stable over the years. Therefore, in our scenario design, we will treat both demand and price as exogenous and independent factors. The difference in prices can be attributed to gaps in both technology and services, which cannot be closed in the short term. Therefore, pricing is not included as a decision variable in capacity requirement analysis.

Based on the empirical data analysis of Section 2, nine demand scenarios of Table 2 are used in the strategy analysis. The parameter μ will be alternatively referred to as the demand growth rate, and the parameter σ as the demand volatility.

The profit scenario is further defined by fixed cost, M&L costs, OE, and ASP, by using Eqs. (1)–(4),

which are summarized in Table 3. More ASP scenarios have been added in order to include other possible improvements in ASP. Eq. (4) is a linear regression for the time series of ASP. The quarterly rate of improvement is $5. However, more complex patterns could be identified by a closer examination of the time series. Because the scale of production increased by 10-folds in 10 years. The scale of production in earlier years was quite small. The company was a player in a niche market and enjoyed a higher profit margin. As the market size grew, the profit margin decreased. If the data before 1997 are disregarded, the time series could be modelled by a line or a step function of two levels. The linear line has a slope much greater than 5.0. If viewed as a step function, the ASP was kept at the level of $1200 for approximately 3 years and at $1600 for another 4 years. By hypothesizing all possible courses of ASP improvement, three levels of the step functions and three slopes of the regression line are included in Table 3. Based on Tables 2 and 3, there are nine demand scenarios and six profit scenarios. The total number of compound scenarios is 54.

4. Computation procedures for capacity strategies

In this section, the computation procedures for the reactive and conservative strategies are outlined. The GBM process is a continuous process. To derive the numerical solution, the binomial lattice approximation is frequently used in the literature (Luenberger, 1998). The approximation was first proposed by Cox et al. (1979). Jarrow and Rudd (1983) improved the binomial lattice model. In this

Table 2
Nine demand scenarios

σ	μ		
	0.23	0.28	0.33
0.25	# 1	# 2	# 3
0.33	# 4	# 5	# 6
0.41	# 7	# 8	# 9

Table 3
Factors of profit scenarios

Factors	FC		ML		OE
	By capacity		By output		By capacity
Intercept	11,465		537.71		131.44
Slope	127		3.37		0

Factors	ASP					
	Step improvement			Linear improvement		
	Step 1	Step 2	Step 3	Growth 1	Growth 2	Growth 3
Intercept	1600	1800	2000	1581.2	1581.2	1581.2
Slope	0	0	0	5	10	15

paper, the Jarrow and Rudd model is adopted. Let the demand at time t_0 be q_{t_0}, the demand in time $t_0 + \Delta t$ be $q_{t_0+\Delta t}^+$ or $q_{t_0+\Delta t}^-$ with equal probability of 0.5. Then $q_{t_0+\Delta t}^+$ and $q_{t_0+\Delta t}^-$ are given by

$$q_{t_0+\Delta t}^+ = q_{t_0}e^u, \quad u = (\mu - \sigma^2/2)\Delta t + \sigma\,\Delta t,$$

$$q_{t_0+\Delta t}^- = q_{t_0}e^v, \quad v = (\mu - \sigma^2/2)\Delta t - \sigma\,\Delta t,$$

where u and v are the up and down growth rates of the logarithm of demand, and μ and σ are the parameters of the underlying GRM process.

4.1. The reactive strategy

Let the current time be t, the capacity level be C_t, and the lead time of capacity deployment be d. The current decision is to determine the capacity level at time $t + d$. Because capacity is irreversible, the capacity decision is equivalent to determining the capacity increment Δc_t. Eq. (5) is the stochastic model for the demand. Let q_t be the demand at time t. The expected values of the drift parameter r_t and the demand q_t are

$$\bar{r}_{t+d} = (\mu - \sigma^2/2)d,$$

$$\bar{q}_{t+d} = q_t e^{\bar{r}_{t+d}}.$$

The basic concept of the reactive strategy is to compare the capacity level and expected demand. If the expected demand of a future period ($\bar{q}_{t+d}$) is higher than the then existing capacity (C_{t+d-1}), then the capacity is expanded. Therefore, the capacity increment initiated in time t is

$$\Delta c_t = \text{MAX}(0, \bar{q}_{t+d} - C_{t+d-1})$$

and the capacity trajectory is described by $C_{t+1} = C_t + \Delta c_t$. This strategy represents a now-or-never decision for each time period $(t + d)$. Without considering all possible realizations of future demands, it is also a myopic strategy.

4.2. The conservative strategy

When demands are uncertain, the value of a given capacity depends on the demands that will be realized over the life time of the capacity. That part of capacity that exceeds the demand will not contribute any revenue. The conservative strategy utilizes the real-option theory to evaluate the value of capacity in the environment of uncertain demand. The demand is represented by a binomial lattice model. Let the demand in scenario s of time t

be represented as q_t^s and the probability that scenario s will occur be p_s^t. For notation brevity, p_s^t is simplified as p_s with time t implicitly included in s. By this notation, a state s is sometimes expressed as $s(t)$ to make the time stage explicit. Under the conservative strategy, the capacity expansion decision is made by considering all possible demand realizations over the equipment life time. Let the current time be t_0. Given the binomial lattice model of demand, the formulas of determining capacity increment Δc_{t_0} have four major parts:

(1) Compute the expected revenue for a given Δc_{t_0} over its life time. The effective capacity increment (ECI) of Δc_{t_0} when evaluated against a particular demand realization q_t^s is defined as that part of capacity increment that can effectively generate revenue:

$$\text{ECI}(\Delta c_{t_0}; C_{t-1}, q_t^s) = \text{MIN}[\Delta c_{t_0}, \max(q_t^s - C_{t-1}, 0)].$$

The expected value of ECI with respect to all scenarios s of time period t is then

$$\delta_{t_0,t}(\Delta c_{t_0}; C_{t-1}, q_t^s) = \sum_{s \in S} p_s \text{ECI}(\Delta c_{t_0}; C_{t-1}, q_t^s),$$

where the set S contains all demand states of time t. The revenue that can be generated over the life time $[t0 + d, t0 + d + l]$ is

$$R_{t_0}(\Delta c_{t_0}; \text{ASP}, \text{PC}, d, l) = \sum_{t=t_0+d}^{t_0+d+l} e^{-r(t-(t_0+d))} \times (\text{ASP} - \text{PC})\delta_{t_0,t}(\Delta c_{t_0}),$$

where PC is the production cost and r is the discount rate. The $R_{t_0}(\Delta c_{t_0})$ is also called the current value of Δc_{t_0}.

(2) Optimize the capacity increment based on net cash flow. From Eq. (1), the fixed cost of Δc_t is $\text{FC}(t)\Delta c_t$. Therefore, the optimal capacity increment at time $t0$ and a certain state $s(t_0)$ is

$$\Delta c_{t_0,s(t_0)}^* = \underset{\Delta c_{t_0}}{\text{Argmax}} \, [R_{t_0}(\Delta c_{t_0}) - \text{FC}(t_0)\Delta c_{t_0}].$$

(7)

The corresponding net cash flow is

$$\pi_{t_0,s(t_0)}^* = R_{t_0}(\Delta c_{t_0}^*) - \text{FC}(t_0)\Delta c_{t_0,s(t_0)}^*.$$

(3) Steps 1 and 2 are repeated for each node (t, s) of the binomial lattice. In steps 1 and 2, the value of waiting is not included in the optimization of capacity increment and its associated net cash

flow. Let $F^*_{t_0,s(t_0)}$ be the total value of $\Delta^* c_{t_0}$, including the value of the waiting option. Then,

$$F^*_{t_0,s(t_0)} = e^{-r} \mathrm{E}_s[\max\{\pi^*_{t_0+1,s}, F^*_{t_0+1,s}\}],$$

where the expectation E is taken over all states in time $t_0 + 1$.

(4) When $\pi^*_{t_0,s(t_0)}$ is greater than $F^*_{t_0,s(t_0)}$, the capacity is expanded at time $t0$, otherwise, the expansion decision is postponed to time $t0 + 1$.

The industry practice is to use 4 years as the normal life time of process equipment and 1 additional year to account for the salvage value. That is, the life time is 5 years. In computing the total value of a capacity increment, the value beyond its life time is zero and the computation is performed backward. As the capacity cost is spread evenly over the life time (in calculating OI), the remaining useful life of capacity at the end of analysis period is taken care of.

5. Performance analysis and comparison

The findings in Section 2 that production costs are predictable provide a good basis for inferring the relative behaviour of capacity strategies in the uncertain environment of demand. In this section, the experimental results are summarized and discussed. In each experiment (i.e., compound scenario), 10 demand paths are generated using Monte Carlo simulation. The two capacity strategies are applied in order to simulate capacity expansion trajectories. The performance measures ROFA and OI are then computed for each strategy. The performance of the reactive strategy is computed using the demand paths. The performance of the conservative strategy is computed using a binomial lattice emanating from each node of the demand paths. The step improvement and linear improvement scenarios of ASP are analysed separately.

5.1. The step improvement scenario of ASP

A step growth of ASP represents the scenario of leaps in manufacturing technology. The regression formulas that correlate performance measures with endogenous variables are shown in Eqs. (8)–(11).

Expected performance of the conservative strategy:

$$\text{OI (millions)} = -10,670 - 243,336\mu + 156.552\mu p - 10.368\sigma p, \qquad (8)$$

$$\text{ROFA (\%)} = -7.39 + 0.00652p - 0.00074\sigma p. \qquad (9)$$

Expected performance of the reactive strategy:

$$\text{OI (millions)} = -3,107 - 163,581\mu + 141.742\mu p - 165,530\mu\sigma, \qquad (10)$$

$$\text{ROFA (\%)} = -10.267 + 0.00850p - 0.00363\sigma p. \qquad (11)$$

If the regression analysis is applied to the difference of performance between the two strategies, the formulas are shown in Eqs. (12) and (13). The difference is computed by subtracting the OI and ROFA of the conservative strategy from those of the reactive strategy.

$$\Delta\text{OI (millions)} = -7491 + 28,256\sigma, \qquad (12)$$

$$\Delta\text{ROFA (\%)} = 4.42\mu + 5.04\sigma - 0.00372\mu p. \qquad (13)$$

Insights can be gleaned from the formulas by comparing corresponding coefficients. It should be noted that the magnitude of the prices is in thousands. Therefore, all terms of the above formulas are equally significant in magnitude; no terms should be deleted for insignificance.

Insight (1): OI is affected by the combined effect of the growth rate and price (Eqs. (8) and (10)), not by the growth rate alone. In fact, if the trend of price is ignored in capacity planning, it is possible that a higher growth rate could lead to a lower OI. Take Eq. (8) as an example, if p is less than the ratio of 243336/156.552, OI will actually decrease with an increase in μ. This pitfall (that the growth in demand is necessarily beneficial) should be avoided in capacity planning.

Insight (2): For either strategy, ROFA is not affected by the growth rate (Eqs. (9) and (11)). Instead, it is determined by the price, subtracted by the joint effect of demand volatility and price.

Insight (3): The detrimental effect of the demand volatility is very significant. It is the deciding factor in choosing between the alternative capacity strategies (Eqs. (12) and (13)) to some extent.

Insight (4): For the conservative strategy, OI is negatively affected by the interaction between the price and volatility parameter (−10.368). In contrast, the reactive strategy is negatively affected by the interaction between growth rate and volatility parameter (−165530). The conservative strategy will necessitate a closer attention to the price trend and the reactive strategy will necessitate active monitoring of the demand volatility.

Insight (5): The volatility parameter has a negative effect on both performance measures. Its

effect is about 5 times more profound with the reactive strategy (0.00363–0.00074).

5.2. The linear growth scenario of ASP

A linear growth of ASP represents stable advancement in manufacturing technology. The regression formulas are shown in Eqs. (14)–(19). The parameter r represents the rate of increase of ASP in dollars per quarter (Table 3)

- Expected performance of the conservative strategy

$$OI \text{ (millions)} = 8,821 - 8,581\sigma \\ + 3,281\mu r - 813\sigma r, \quad (14)$$

$$ROFA \text{ (\%)} = 2.10 + 0.11212r. \quad (15)$$

- Expected performance of the reactive strategy

$$OI \text{ (millions)} = 2,173 + 59,132\mu \\ + 2,235\mu r - 161,521\mu\sigma, \quad (16)$$

$$ROFA \text{ (\%)} = 2.92 - 6.09\sigma + 0.12495r. \quad (17)$$

- Difference (the conservative strategy subtracted from the reactive strategy)

$$\Delta OI \text{ (millions)} = -25,466\mu + 103,386\mu\sigma, \quad (18)$$

$$\Delta ROFA \text{ (\%)} = -0.257 - 1.57\mu + 5.73\sigma \\ - 0.04695\mu r. \quad (19)$$

Insight (6): It should be noted that, besides the slope of price improvement (r), the level of price has an impact on OI and ROFA. The effect of the intercept of the prices is implicit in the constant term of the above formulas. Taking this observation into consideration, the two sets of formulas (8)–(13) and (14)–(19), have strikingly similar composition. Insights similar to those in Section 5.1 can be derived, including (1) OI is affected by the combined effect of demand growth and price, (2) ROFA is not affected by the growth rate of demand, (3) the volatility parameter is the crucial factor (with negative coefficients in Eqs. (14)–(16)); (4) demand volatility is more detrimental to the reactive strategy than to the conservative strategy (Eqs. (15) and (17)) and the conservative strategy outperforms the reactive strategy when σ is large (Eqs. (18) and (19)).

5.3. Adapting capacity strategy

Eqs. (8)–(19) provide useful information for adapting the capacity strategy in various scenarios of demand and ASP. Cross-over points could be determined by solving a set of linear inequalities as shown in Fig. 10. In the three-dimensional space of μ, σ and ASP (level), the feasible region of scenarios that satisfy the two perpetual capability conditions is marked with the word "Positive": (1) ROFA > 3.53% and (2) OI > 14 Billions. For instance, if the conservative strategy is adopted and the ASP is only maintained at \$1600, the objective of perpetual manufacturing capability cannot be achieved. Fig. 10 provides a useful planning tool for adapting capacity strategy in the uncertain environment of demand, while taking into consideration the ASP trend embedded in business plans.

6. Conclusions

Capacity planning and investment is a strategic issue in the semiconductor manufacturing industry.
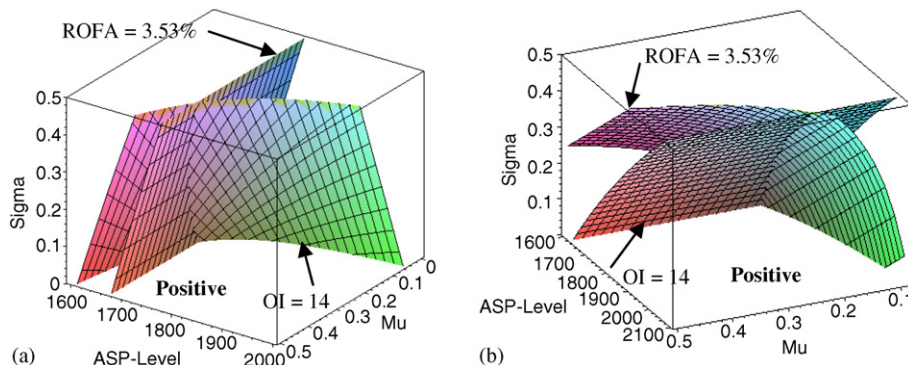

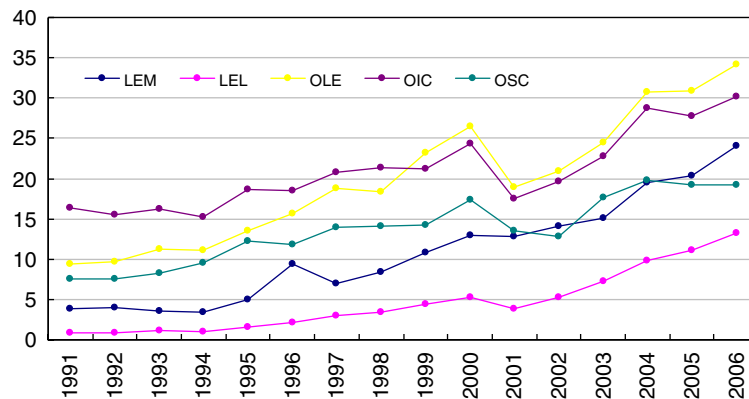Fig. 10. The feasible region of scenarios for perpetual capability.

Fig. 11. Worldwide demand of microchips by product segments.

In this paper, a new analysis method is presented for adapting capacity strategy in the environment of uncertain demand and expected growth in ASP. This method is supported by empirical data analysis of the production and capacity costs of a major semiconductor manufacturer and by calibration of the uncertainty in demand in the industry. Many insights are gained in identifying the relative strength of conservative and reactive capacity strategies in various scenarios. There are several important findings: (1) Growth in demand is not necessarily beneficial in the uncertain industry environment. Volatility in demand is one of the crucial factors in adapting capacity strategy. However, it cannot be said that volatility alone is the deciding factor. Capacity strategy should be chosen by considering the combined effect of demand growth and volatility. (2) The future trend of ASP is a more important factor than the growth in demand. (3) The corporate goals of increasing market share and increasing the return of equity are not necessarily coherent. In fact, we have observed in the industry that corporate goals do switch between them. Since the two goals have a direct impact on the choice of capacity strategy, the method described in this paper provides a framework for integrating capacity planning with business planning.

There are many other capital-intensive industries. See Luss (1982) for capacity expansion problems in heavy process industries, communication network and public services (electric power, water resources, schools, etc.) The basic capacity expansion problem consists of determining the sizes, timing, and locations of the facilities. Semiconductor manufacturing differs from those application areas in the severity of demand volatility and the new requirement of perpetual manufacturing capability. This paper identifies those differences, defines a new facet of perpetual capability, and develops a method for evaluating alternative capacity strategies.

Acknowledgement

Appendix:. Worldwide demands of semiconductor ICs

Data from 1991 to 2002 are historical data in millions of wafers. Data after 2003 are forecasts made by market research. All data are courtesy of the Industry Economic Model (Fig. 11).

References

Ahmed, S., 2002. Semiconductor tool planning via multi-stage stochastic programming. In: Proceeding of the 2002 International Conference on Modeling and Analysis of Semiconductor Manufacturing, Tempe, Arizona, USA, pp. 153–157.

Benavides, D.L., Duley, J.R., Johnson, B.E., 1999. As good as it gets: Optimal Fab design and deployment. IEEE Transactions on Semiconductor Manufacturing 12 (3), 281–287.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. The Journal of Political Economy 81 (3), 637–654.

Chou, Y.-C., Wu, C.-S., 2002. Economic analysis and optimization of tool portfolio in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 15 (4), 447–453.

Christie, M.E.R., Wu, D., 2002. Semiconductor capacity planning: Stochastic modeling and computational studies. IIE Transactions 34, 131–143.

Cox, J., Ross, S., Rubinstein, M., 1979. Option pricing: A simplified approach. Journal of Financial Economics 7, 229–263.

Geary, S., Disney, S.M., Towill, D.R., 2006. On bullwhip in supply chains—Historical review, present practice and expected future impact. International Journal of Production Economics 101 (1), 2–18.

Hood, S.J., Bermon, S., Barahona, F., 2003. Capacity planning under demand uncertainty for semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 16 (2), 273–280.

International SEMATECH, 2002. Industry Economic Model.

Jarrow, A.R., Rudd, A., 1983. Option pricing (Irwin).

Karabuk, S., Wu, S.D., 2001. Coordinating strategic capacity planning in the semiconductor industry. Technical Report 99T-11, Department of IMSE, Lehigh University.

Liang, Y.-Y., Chou, Y.-C., 2003. Option-based capacity planning for semiconductor manufacturing. In: IEEE International Symposium on Semiconductor Manufacturing, San Jose, California, USA, September 30–October 2, pp. 77–80.

Luenberger, D.G., 1998. Investment Science. Oxford University Press, Oxford (Chapter 11).

Luss, H., 1982. Operations research and capacity expansion problems: A review. Operations Research 30 (5), 907–947.

Nazzal, D., Mollaghasemi, M., Anderson, D., 2006. A simulation-based evaluation of the cost of cycle time reduction in Agere Systems wafer fabrication facility—A case study. International Journal of Production Economics 100 (1), 300–313.

Neudorff, J., 1999. Static capacity analysis using Microsoft Visual Basic. In: International Conference on Semiconductor Manufacturing Operational Modeling and Simulation, San Francisco, pp. 207–212.

Swaminathan, J.M., 2000. Tool capacity planning for semiconductor fabrication facilities under demand uncertainty. European Journal of Operations Research 120, 545–558.

Tsay, R.S., 2002. Analysis of Financial Time Series. Wiley, New York.

Witte, J.D., 1996. Using static capacity modelling techniques in semiconductor manufacturing. In: IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop, pp. 31–35.

Wu, W.-F., Yang, J.-L., Liao, J.-T., 1998. Static capacity checking system with cycle time considered. In: The Seventh International Symposium on Semiconductor Manufacturing, pp. 307–310.