SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-3801

TECHNICAL REPORT NO. 1265

August 2000

**EVALUATION OF CAPACITY
PLANNING PRACTICES FOR THE
SEMICONDUCTOR INDUSTRY**

by

Metin Cakanyildirim and Robin O. Roundy

Evaluation of Capacity Planning Practices for the Semiconductor Industry *

Metin Çakanyıldırım and Robin O. Roundy

www.orie.cornell.university/~metin/semicond

School of Operations Research and Industrial Engineering, Cornell University

August 5, 2000

Abstract

We evaluate several features of capacity planning approaches used in practice and make concrete suggestions for practitioners. We run four experiments to determine the effects of large vs. small time buckets, the cost implications of using heuristics, frequent vs. less frequent forecasting and planning, and joint optimization of tool and floor-shell expansions vs. sequential optimization. We use industrial demand and capacity data for realistic conclusions.

*Technical Report 1265, SORIE. Cornell University 14853, NY.

1 Introduction

The semiconductor industry has become one of the leading industries in US economy. However, astronomical fab costs (an average of \$2.5 B [23]) combined with ever reducing chip prices are cutting into the profit margins significantly. [2] reports that, in 1996, the industry reinvested 23% of total revenue in capital expenses, mostly (60-70%) for tool purchases. Based on [10] and [20], [19] reports that 75% of new fab expenditure is for tool purchases. He also argues that expenditures for new tools are siphoning off manufacturers' profits. Consequently tool capacity planning is very important, but it is far from trivial. Currently, most tool plans are based on demand forecasts which are at least one-year-out (tool delivery lead times) into the future. Thus, financial success is tightly tied to utilization of tools ordered at least a year in advance. The industry faces the challenge of balancing capital investment costs against the risk of losing customer demand.

We refer the reader to [9] for a detailed survey of capacity expansion models. In summary it is still a largely unmet challenge to construct capacity expansion models that capture important aspects of the semiconductor industry. Capacity expansion models are generally at the detail of planning for a single resource (mostly obtained by crude aggregation). Thus a computationally tractable model that plans each resource explicitly is desirable. A common approach, usually at the expense of losing tractability for problems of realistic size, is stochastic programming ([25], [6], [14], [15], [11] and [24]). Some other researchers focus on the structure of expansion policies ([12], [13] and [18]). There have been attempts to apply inventory theory [1], Markov decision processes [4] and real options theory [5] to capacity expansion problems as well.

As more and more tools are installed into a fab, resources other than tools can become limiting factors on production volume. Two common limiting factors are floor space and shell space. Floor space generally refers to clean room manufacturing area while shell space is the total indoor manufacturing area. If demand is expected to grow rapidly companies may take advantage of the strong economies of scale in shell space expansion by building a large shell and adding floor space in increments. This idea is also in [5] which concludes that "Sequentially deployable large fabs are ... attractive since they offer the economies of scale of larger fabs but require a smaller initial capital outlay".

In this paper we evaluate several aspects of capacity planning approaches used in practice. We use industrial capacity data, and demand data from two semiconductor manufacturers. We run four experiments to determine the effects of large vs. small time buckets, the cost implications of using heuristics, frequent vs. less frequent forecasting and planning, and joint optimization of tool and floor-shell expansions vs. sequential optimization. Using these experiments, we evaluate capacity planning practices and make concrete suggestions for practitioners.

In the next two subsections, we will briefly introduce the capacity planning technique FIFEX (details are in [9]) and define our base case. In Section 2 we will do four experiments, using FIFEX. With the availability-time rounding experiment (the first), we will study the effects of time bucket length on total cost. [22] examines the effects

of granularity of time on cost for three cases: certain demand without shortage, certain demand with shortage, and uncertain demand. It is noted that the uncertain demand case is more sensitive to time granularity. The study advises using nonuniform granularity i.e., decisions are made more frequently in the short run than in the long run.

The heuristics experiment (the second) evaluates practical tool capacity planning ideas with three heuristics: the percentile, loss and bottleneck heuristics. The first two heuristics are straightforward approaches based on Type I and II service levels [21], and the third is inspired by the “theory of constraints” [17]. With the forecasting and planning frequency (the third) experiment, we will illustrate the cost savings achieved by feeding capacity expansion models with up-to-date and accurate demand forecasts. This experiment will show the value of frequent and accurate forecasting. Although evaluation of rolling decision horizons is a common theme both in the third experiment and in [3], only the former uses up-to-date forecasts. Besides the latter deals with inventory control.

The hierarchical decomposition experiment (the fourth) compares two sequential floor-shell and tool expansion heuristics (top-down and bottom-up) with the joint optimization technique FIFEX. In particular, we will study the cost savings with joint optimization of floor-shell and tool expansion times. Similarly [5] studies times and sizes of fab expansions in a single product environment with uncertain demand. It also handles construction lead times and economies of scale in fab construction. However [5] does not consider tool purchases, and tool expenses constitute the majority of new fab expenditures. FIFEX overcomes this shortcoming by capturing both fab construction and expansion costs, and tool expenses. We will summarize the results of our experiments and make concrete suggestions for practitioners in the last section.

1.1 Overview of FIFEX

With FIFEX [9], we model three kinds of costs, tool costs, lost sales costs and floor-shell costs. Tool costs include the cost of financing the purchase, the installation and the maintenance of machines. The lost sales cost is the expected value of the lost sales incurred during the planning horizon. Floor-shell costs are infrastructure costs for expanding floor space and shell spaces.

FIFEX considers a single semiconductor product family, such as ASICS or Memory. Let D_t be the product demand at time t so that it is a stochastic process over the planning horizon $[0, T]$. Note that FIFEX is a continuous time model and t is a continuous variable. Then with available fab capacity of K_t at time t the lost sales is $E(D_t - K_t)^+$.

[9] argues that for certain service measures, including lost sales cost, there is an optimal capacity expansion plan where only bottleneck tools are installed. Note that the bottleneck can shift from a tool type to another type when enough tools of the former type are purchased. Because of the optimality of installing only bottleneck tools, the optimal sequence of tool purchases (Bottleneck Purchasing Policy or BPP order) can be constructed regardless of the demand process. First order tool purchases according to BPP. Let t_n and a_n be the availability time of the n th (according to BPP) purchase and the fab capacity after the n th purchase, clearly. Because of BPP ordering,

$t_{n-1} \leq t_n$ and $a_{n-1} \leq a_n$. For a given series of floor-shell expansions, let $f_n(t_n)$ be the sum of tool costs, floor-shell costs and expected lost sales cost associated with machine n installed at t_n . For financial and/or practical reasons, suppose that N is the number of machines that can possibly be purchased over the planning horizon. Then the machine capacity expansion problem becomes

$$\min\left\{\sum_{n=1}^N f_n(t_n) : 0 \leq t_1 \leq \dots \leq t_N \leq T\right\}.$$

This problem can be solved efficiently to optimality if D_t is growing. By solving several problems of this type, FIFEX finds the optimal sequence of floor and shell expansions, as well as machine installation times. FIFEX can also model positive tool delivery lead times and economies of scale in floor-shell construction costs.

1.2 Base Case

In this subsection we define our base case problem and use FIFEX to solve it. FIFEX needs tool related data and product demand data as inputs. Our tool related data is based on a SemaTech report about 130nm equipment performance (<http://www.sematech.org>). After pre-processing the data, we ended up with 42 different tool types. We have 3 tool types for Lithography (a Litho Stepper/Tracker, a 1-Line Expose, a 1-Line Litho Track), 9 for Etching, 5 for Ion Implantation, 3 for Wet Cleaning, 7 for Test, 3 for Metal Deposition, 5 for Chemical Vapor Deposition, 3 for Chemical Mechanical Polish, 2 for Furnace, 1 Gate Capacitor Cluster and 1 Physical Vapor Deposition. We assume that each of these tool types is needed to manufacture the semiconductor products discussed below. Each tool type has 4 attributes. *Price* is the purchase price of a tool. The Litho Stepper/Tracker costs \$ 10.1 M whereas some Test equipment cost \$0.65 M. *Capacity* is the number of wafers produced per week by a tool. Tool capacities range from 500 to 8000 wafers per week. *Lead Time* is the time between the release of a tool purchase order and the availability time for that tool. The availability time for a tool is the time its capacity hits a high percentage of its maximum achievable capacity. Lead times range from 12 months to 24 months. *Area* is the floor space consumed by a tool. That ranges from 1.6 m^2 to 33.9 m^2 . We will not consider floor-shell expansions in any of the experiments except for the hierarchical decomposition experiment where the focus is the optimization of floor-shell expansions. Thus, the discussion of floor-shell data is postponed to the Hierarchical Decomposition experiment.

We derive demand processes from historical demand forecasts. Two different semiconductor manufacturers gave us historical forecasts and demands for one product family each. We will use the term product instead of product family in the remainder of this paper. To constitute a base for our numerical experiments, we take Jan 1994 as month 1 and study capacity expansion for the next 60 months, supposing that we are now in Jan 1993. Also we suppose that 71 tools are available in Dec 1993 and that nothing is currently scheduled to be delivered afterwards. The initial capacity provided by the 71 tools is large enough that lead times do not constrain the optimal purchase plan. Because of the minimum lead time of 12 months, the period from now to Jan 1994 is irrelevant for planning.

That is why we consider [Jan 1994, Dec 1998] as the decision horizon, comprising months 1 to 60. Capacity costs are discounted with an interest rate of 5 % per annum. Lost sales costs are taken as \$3700 per wafer.

Inputting the forecasts into the SeDFAM procedure [8] we compute the expected value and the variability of the each demand process as seen from Jan 1993. The first product’s demand is forecasted to double over the decision horizon [Jan 1994,Dec 1998] (see Figure 2). The mean of the second product is set equal to the mean of the first product so that differences in the test results are only due to differences in variability. As Figure 1 shows, the relative magnitude of their coefficients of variation is remarkably different. For obvious reasons we name product demands “Light Tail” and “Heavy Tail”.

— Figure 1 —

Given the common mean and the variability for each product demand, we represent demands as a unimodal distribution which is a mixture of four trapezoidal densities. The Heavy Tail density is skewed whereas the Light Tail density is more symmetric. Mixtures of trapezoids are very versatile for demand representation. They are intuitively appealing because, unlike other commonly used densities (such as Normal), their support is bounded and can be taken nonnegative. All our figures will have two columns, Light Tail on the left and Heavy Tail on the right. A legend in a row applies to all the graphs in that row.

— Figure 2 —

Figure 2 depicts the monthly and average shortfalls incurred by the FIFEX capacity expansion plan. We define shortfall as the fraction of demand lost due to inadequate capacity. Specifically $shortfall(t)$ is defined as

$$shortfall(t) = \frac{E(D_t - K_t)^+}{E(D_t)}. \quad (1)$$

Average shortfalls are smaller for Light Tail demand because it has smaller variance. When the Heavy Tail variance is comparable to the Light Tail variance (the first 7 months), monthly shortfalls look similar. After that, Heavy Tail variance continues to grow much faster than Light Tail variance until the 13th month. This pattern is visible in the relative growth of the Heavy Tail shortfall against the Light Tail shortfall from month 7 to month 13. After month 13, both Heavy Tail and Light Tail coefficients of variation stay constant and monthly shortfalls appear to be stabilizing around their averages.

Figure 2 depicts mean product demands, the 98% and 85% percentiles (quantiles) of demands, and optimal capacities. It shows that tool availability times in the case of Light and Heavy Tail are quite similar; capacities start and end the decision horizon at the same levels. However, availability times of Heavy Tail are uniformly smaller than corresponding ones of Light Tail, resulting in higher tool purchase costs. In spite of this, the lost sales costs are also higher with the Heavy Tail demand, as the higher shortfall indicates. In both cases FIFEX achieves an optimal balance between tool costs and lost sales costs.

2 Experiments

2.1 Rounding Continuous Availability Times

Many capacity planning models treat time as a discrete quantity. Also in practice decisions are executed at certain points in time, say at the beginning of every month. On the other hand continuous time decision models, such as FIFEX, have the flexibility of executing a decision at any time over the decision horizon. With the Rounding experiment we will measure the impact of discrete decision variables on the cost.

The availability time of a tool is when it is available for production, one delivery lead time after after it is purchased. Note that initially FIFEX has 60 months in the decision horizon and tool availability times are continuous. In our rounding experiment, we will fix the number of *availability epochs* and uniformly distribute them over the decision horizon. We will round optimal availability times to the nearest availability epoch. For example, with 60 availability epochs we round the continuous FIFEX solution to 60 integers which stand for the first day of the each month in the decision horizon. In that case the time between availability epochs is 1 month. We will increase the time between epochs from 1 month to 2, 3, 6, 9, 12, 15, 18, 24 and 30 months and investigate the impact on the sum of tool and lost sales costs. This closely approximates the optimal discrete-time production plan.

— Figure 3 —

For the Light Tail and Heavy Tail demands, Figure 3 shows how the total costs go up as the time between epochs goes up. When the time between availability epochs is 0, we have the continuous solution. Note that when the time between epochs is 9 or fewer months, the cost of the rounded solution differs from the cost of the continuous solution by about \$0.9 M in both Heavy and Light Tail case. When the time between epochs is 18 months the cost difference is \$7.6 (\$10.2) M or 2.3% (2.8%) for the Light (Heavy) Tail demand. [22] runs a similar rounding experiment for a (water treatment) facilities expansion model with a decision horizon of 40 years. When the number of availability epochs decreases from 40 to 4, the costs go up by 5.5%-8.7%, see Table 5 of [22]. These numbers are comparable in magnitude to those shown by Figure 3. As a result, we conclude that within reasonable limits rounding does not increase the total cost much and the total cost function is fairly flat around its minimizer. Solution quality does not appear to be sensitive to the number of months between availability epochs, as long as that number does not exceed 9 months. If demand were growing more rapidly than the 20% per annum in this experiment, 9 months might be too long.

2.2 Tool Capacity Planning Heuristics

FIFEX computes the optimal solution and is very fast. It solves an instance of the tool capacity expansion problem to optimality in around 45 seconds (on a Sun Ultra Sparc 1) per trapezoid used in the demand density modeling. In

our case the running time is 4x45 seconds. However it is worthwhile to compare FIFEX against heuristics for two reasons. The first reason is of a practical nature; we want to see the cost savings practitioners can achieve using FIFEX instead of simpler heuristics. The second reason is rather theoretical. Intuition drawn from heuristics can be used to attack more complex problems, such as a multi-product version. We define three heuristics: Percentile, Loss and Bottleneck.

Percentile Heuristic: The percentile heuristic purchases tools as late as possible, while filling a minimal percentage of the demand at all times. For the demand stochastic process $\{D_t : 0 \leq t \leq T\}$, we let the p th-percentile of the demand process be the series of numbers d_t^p such that $P(D_t \leq d_t^p) = p$ for all t . Note that for each p , d_t^p is a nondecreasing deterministic sequence because D_t is growing in t . Thus d_t^p can meaningfully be compared with the tool capacity levels. Let t_n^p be the availability time of the n th tool derived from the p -th percentile. Then $t_n^p = t$ where t is the smallest time satisfying $d_t^p = a_n$. The cost of the percentile heuristic is calculated as

$$\min\left\{\sum_{n=1}^N f_n(t_n^p) : p \in \{0, 0.02, 0.04, \dots, 1\}\right\}. \quad (2)$$

The solution of the percentile heuristic is the percentile p and the corresponding set of t_n^p yielding the minimum cost in (2). The percentile heuristic sets t_n^p to control the probability of meeting the demand -Type I service [21].

We run heuristics for six lost sales costs per wafer (\$2.2 K, \$2.7 K, \$3.2 K, \$3.7 K, \$4.2 K and \$4.7 K; see Figure 4), and for the Light Tail and Heavy Tail demands, for a total of 12 runs. A single percentile heuristic run takes about 3 minutes: the heuristic spends most of this time computing d_t^p . Figure 4 shows the performance of the percentile heuristic in terms of lost-sales and tool costs. Figure 5 has the ratio of total FIFEX costs to the total percentile heuristic costs, and the average shortfalls. The percentile heuristic cost is within 10% of, and on average 7% worse than, the optimal FIFEX cost. Average shortfalls follow a pattern similar to that of FIFEX. Since 7% of the total costs are about \$25 M, substantial savings can be achieved using FIFEX instead of the percentile heuristic.

For both Heavy and Light Tail demand, the percentile heuristic and FIFEX exhibit similar behavior. However, the percentile heuristic becomes more conservative as lost sales cost per wafer increase. In comparison with FIFEX, it dictates installing tools earlier, incurring higher tool costs, lower lost sales costs and lower shortfalls.

— Figure 4 —

— Figure 5 —

The percentile heuristic has a shortcoming. Finding the value of the percentile p minimizing costs in (2) requires many functional evaluations of $\sum_{n=1}^N f_n(t_n^p)$. Although these p values increase with lost sales cost per wafer, the relationship between the two is complex. Thus, we are bound to search for a p for every different value of lost sales cost per wafer. The percentile heuristic requires as much time as FIFEX does, so it is not interesting as an alternative

for FIFEX. However, versions of the percentile heuristic are used in practice with p chosen in an ad hoc manner. When p is not chosen optimally, the resulting costs can be very high.

Loss Heuristic: The loss heuristic is also a service level based heuristic and is similar to the percentile heuristic. The loss heuristic controls the shortfall -Type II service [21]. It purchases tools to keep the shortfall at a target level denoted by ρ . It calculates the availability time t_n^ρ of the n th tool from

$$t_n^\rho = \min\{t : \frac{E(D_t - a_{n-1})^+}{E(D_t)} \geq \rho\}. \quad (3)$$

Since the value of ρ is not known initially, it will be chosen to minimize the total costs;

$$\min\{\sum_{n=1}^N f_n(t_n^\rho) : \rho \in \{0, 0.01, 0.02, \dots, 0.50\}\}. \quad (4)$$

Note that we do not consider $\rho > 0.50$ because such poor service levels are well beyond industry norms. The solution of the loss heuristic is the shortfall level ρ and the corresponding t_n^ρ minimizing the cost in (4).

Figures 4 and 5 show the performance of the loss heuristic. The loss heuristic performs almost identically to the percentile heuristic and takes about the same time, so our comments regarding the performance of the percentile heuristic are also valid for the loss heuristic.

Bottleneck Heuristic: The percentile and loss heuristics are based on oversimplified tool purchase triggering mechanisms. Another way of achieving simplification is using FIFEX for a single bottleneck tool type. We choose the bottleneck tool type as the one with the lowest capacity to price ratio. We apply FIFEX only to the bottleneck tool type and purchase non-bottleneck tools based on the bottleneck tool capacity.

FIFEX balances lost sales costs and tool purchase costs. When we ignore all non-bottleneck tools, we underestimate tool purchase costs and this disrupts the balance. To compensate for this anomaly, we will inflate the price of the bottleneck tool. We sum the tool prices of all of the 71 tools initially available and divide this number by the sum of the tool prices of the bottleneck tools. The ratio is about 8 so we multiply the cost of a single bottleneck tool by 8. After this price adjustment, we run FIFEX on the single bottleneck tool.

After FIFEX is applied, it results in fab capacity K_t . Then non-bottleneck tools are purchased so that the bottleneck tool always remains the bottleneck, i.e. availability time for a non-bottleneck machine n is computed from

$$t_n^b = \min\{t : a_{n-1} < K_t\}. \quad (5)$$

Since we use FIFEX for the bottleneck and an oversimplified tool purchase triggering mechanism (5) for the non-bottleneck tools, the bottleneck heuristic is a hybrid of FIFEX and ideas from the percentile and loss heuristics. The idea of focusing on bottleneck machines during the planning process is further discussed in [17].

Figures 4 and 5 show the performance of the bottleneck heuristic in terms of lost-sales, tool costs and shortfall. Clearly this heuristic fails to deliver a good solution. The bottleneck heuristic purchases the bottleneck machines

later than they would be with FIFEX. This results in an average savings of \$55 M (28%) in tool purchase costs, but it increases lost sales costs by an average of \$215 M (150%). If the bottleneck tool price is not multiplied by 8, the bottleneck heuristic incurs \$290 M (150%) in excess of tool costs and saves \$105 M (75%) in lost sales costs.

During our interviews with semiconductor companies ([7]), we have seen a company using the bottleneck heuristic, and that is why we include this heuristic here. We do not have any tool data from this company so we cannot comment on how far they are from the optimal solution. However, the bottleneck heuristic will not work well unless bottleneck tool prices are much larger relative to the sum of all tool prices. In the SemaTech data set we used, the bottleneck prices are 12.5% of all tool prices. In summary, we do not recommend the bottleneck heuristic. If FIFEX can not be used either the percentile or the loss heuristic can be used because their performances are similar, and better than the bottleneck heuristic.

2.3 Forecasting and Planning Frequency

In practice capacity plans are made for decision horizons of 4-5 years and they are updated monthly, quarterly or annually [7]. In this subsection we will investigate the cost savings that can be achieved by updating plans frequently. When FIFEX makes a decision in a month, it uses the information available in that month to estimate demand distributions. Available information expands with time, so demand forecasts change with time. Demand distributions are modeled and computed with SeDFAM [8]. We let *dup* denote the **demand update period**. For example if *dup* = 2 demand is updated every two months. Clearly frequent demand updating (forecasting) provides accurate inputs to FIFEX. However frequent forecasting is costly. Towards deciding on the forecast frequency, the first step is studying the sensitivity of costs to forecast frequency. Therein lies the principal motivation for our forecasting and planning frequency experiment.

We explain the experiment for *dup* = 2 months. In Jan 93, we first generate the demand process for the decision horizon [Jan 94, Dec 98] using the coefficient of variation data for these 60 months. Then we apply FIFEX to get a solution, but only the first two months of that solution are implemented and the rest is discarded. Two months later, in Mar 93, new and improved demand forecasts are generated using SeDFAM. This improved accuracy of the new forecasts leads to a different FIFEX solution in Mar 93 than the one in Jan 93. Again, the partial solution for the first two months after Mar 94 is implemented, and the rest is discarded. We continue rolling decision horizons forward from [Jan 94, Dec 98] to [Nov 97, Oct 02] and record the solution and costs incurred in [Jan 94, Dec 97]. For values of *dup*=4,8 and 16 (different than 2), the time between FIFEX runs (with updated demand forecast) is 4,8 and 16 months and the partial solutions for the first 4,8 and 16 months are implemented.

In Figure 6 we show the costs of the implemented solution over [Jan 94, Dec 97]. All costs are increasing with *dup*. Infrequent forecasting (large *dup*) means that decisions are made with older data. Thus, there is more uncertainty in what the demand will be. When FIFEX solves a problem with larger demand uncertainty, it purchases more tools

to keep the lost sales reasonable, but the lost sales costs are still higher. Consequently, we have increasing tool and lost sales costs in Figure 6.

— Figure 6 —

Figure 6 shows that the sensitivity of costs to forecast frequency depends on uncertainty in forecasts. Recall that the rates coefficient of variations increase in the Light and Heavy Tail demands over 8 months into the future, are close to each other (see Figure 1). That is, by reducing forecast frequency (increasing dup from 2 to 8 months), similar amounts of uncertainty is added to demands. Thus, the Light and Heavy Tail costs increase in a similar manner. However, this similarity does not persist when dup becomes 16 months. From 8 to 16 months out, there is substantial amount of uncertainty added to the Heavy Tail demand (see Figure 1). On the contrary, the Light Tail demand has constant coefficient of variation over this range. Therefore, the Heavy Tail demand costs go up faster than the Light Tail demand costs. Namely costs are more sensitive to forecast frequency when the rate coefficient of variations change is large.

Although it is hard to recommend a dup value for all the practitioners, based on our data we believe that it should be within 2-4 months range. In our experiments $dup = 4$ has a cost increase of \$7.9 M or 2% (\$15.4 M or 4%) when compared to $dup = 2$ for Light (Heavy) Tail. When comparing $dup = 8$ to $dup = 4$, the corresponding increases are \$15.6 M or 4% (\$26.4 M or 7%). These numbers can be compared with forecasting and planning costs. If duplicating this experiment is not feasible, practitioners can gain some insight from a figure analogous to Figure 1. If the standard deviation goes up more rapidly in the interval $[dup/2, dup]$, then the benefit of halving dup will be larger.

2.4 Hierarchical Decomposition

When practitioners are required to make plans for complex activities, they decompose the activities. This decomposition is often hierarchical; first plans are made for a subset of activities and then these plans are used as inputs to plan for the remaining activities. One of the methods to decompose a plan is by grouping the activities that take place at about the same frequency (see Chapter 10 of [16]). In this subsection, we will compare three capacity planning schemes: planning based on overall integrated optimization and two planning schemes based on hierarchical decomposition.

Our previous experiments assume that initial floor and shell spaces are sufficient to install all tools considered for purchase. However, we now assume that the initial floor and shell spaces are 2700 and 3300 m^2 (including corridor and storage areas). Potentially the floor space can be expanded to 3300, 3900 and 4200 m^2 ; these numbers are 3300 and 4200 for the shell space. Floor and shell space expansion costs are taken as a sum of a fixed cost and a linear (proportional to expansion size) cost. We have obtained cost figures from an industrial source. We use floor and

shell expansion lead times of 20 and 30 months respectively. With these additions to base case data, we use FIFEX to solve for optimal floor space, shell space and tool capacity expansions. According to the optimal plan floor space is expanded to 3300 m^2 in month 13 and both floor and shell spaces are expanded to 3900 m^2 in month 49.

— Figure 7 —

We consider two hierarchical heuristics depending on whether shell and floor space expansions are planned before or after tool purchases.

Top-Down Heuristic: We remove all tool capacity considerations and optimize floor and shell space expansions. This optimization is equivalent to using FIFEX with zero tool prices. Then we take the optimal floor and shell expansion sizes and times as fixed, and plan tool capacity expansions around them.

Bottom-Up Heuristic: We remove all floor and shell space considerations and optimize tool purchases. Indeed this is the base case we discussed in Section 1. We then take the optimal tool purchase times as fixed, and plan for floor and shell expansions to accommodate these tools.

For both heuristics we vary floor and shell expansion costs by multiplying them with a floor-shell cost multiplier which takes values in $\{0.6, 0.8, 1.0, 1.2, 1.4\}$. We do not alter the purchase price of tools. Floor-shell expansion costs, lost sales and tool costs are shown in Figure 7. For the bottom-up heuristic availability times are independent of floor-shell expansion costs, and the tool and lost sales costs stay constant. The floor-shell costs increase linearly as the floor-shell cost multiplier goes up. The tool and lost sales costs of the top-down heuristic roughly move in the same direction as those of FIFEX. However, both top-down and bottom-up floor-shell costs grow while FIFEX floor-shell costs stay approximately constant as the floor-shell cost multiplier goes up. Roughly speaking the heuristics keep the floor-shell expansions unaltered even if they become more expensive. On the contrary, FIFEX keeps the floor-shell expansion costs down and incurs more lost sales costs. This behavior is exemplified when the floor-cost multiplier increases from 1 to 1.2. It is also disappointing that each heuristic cost figure (floor-shell, tool or lost sales) takes a very different value than the corresponding FIFEX cost.

— Figure 8 —

We graph the ratios of total optimal (FIFEX) cost to total heuristic costs in Figure 8. Although each heuristic cost component is different from the corresponding FIFEX cost, the differences are smaller when cost components are summed. Comparing the heuristics with each other one draws the obvious conclusion: the top-down heuristic is better when floor-shell costs are large. The top-down (bottom-up) heuristic is within 13% (19%) of the optimal cost in the worst case and within 9% (11%) on average. Noting that 10% of the optimal costs are around \$50 M, the performance of the hierarchical decomposition heuristics are not satisfactory.

— Figure 9 —

In an effort to improve the heuristics by balancing the floor-shell expansion costs and the lost sales costs, we will inflate the floor-shell expansion (tool) costs in top-down (bottom-up) heuristic, as we did with the bottleneck heuristic. Heuristics are applied with inflated costs but their solutions are evaluated with the original costs. We suppose that tool purchases are 70% and floor-shell expansions are 30% of fab expenditure. Thus we multiply floor-shell expansion (tool) costs by 10/3 (10/7).

— Figure 10 —

Figures 9 and 10 are analogous to Figures 7 and 8, but they show the results for the heuristics with inflated costs. Figure 9 shows that inflating costs helps each top-down heuristic cost (floor-shell, tool or lost sales) take similar values to the corresponding FIFEX costs. Inflated costs also improves the performance of heuristics in terms of total costs as shown in Figure 10. Except for the floor-shell multiplier of 0.8 for Light Tail demand, the top-down heuristic dominates the bottom-up heuristic. The top-down (bottom-up) heuristic is within 3% (9%) of the optimal cost in the worst case and within 2% (5%) on average. 2% of optimal costs are around \$10 M, so the heuristic performance, in spite of improving with the inflated costs, is still not satisfactory. We suggest that an integrated model (i.e. FIFEX) be used. If that is not possible, the top-down heuristic with inflated costs can be used.

3 Conclusion

With this paper, we have numerically studied several capacity planning practices used in the semiconductor industry: decision making using discrete time buckets, using heuristic tool plans, frequent forecasting-planning and integrated vs. decomposed planning for tools and floor-shell space. We are concerned with the effectiveness of these practices and are interested in bringing out their weak and strong points. We use real life demand and capacity data. We duplicate each experiment for two demand processes: Heavy Tail and Light Tail demand.

One of the aims of this paper is to investigate how discrete time buckets affect optimal costs in practice. We conclude that time bucket length does not impact the costs significantly, up to 9 months.

Being aware of service-level approaches to tool capacity planning that are already used and can potentially be used, we turn these ideas into heuristics and test them by comparing their costs against optimal costs. The percentile and loss heuristics are consistently 5%-10% worse than an optimal solution -a substantial gap. The bottleneck heuristic also performs poorly and we suggest that it be avoided.

Another interesting question is how often one should forecast. Motivated by this question, we devise a forecasting-planning frequency experiment and measure the tool and lost sales cost savings achieved by inputting frequent forecasts into a planning module. Based on cost figures, one can successfully trade off the extra cost of frequent forecasting and planning against the tool and lost sales cost savings. We observe that when frequent forecasting

resolves greater amounts of demand uncertainty, it reduces costs further. Roughly speaking, we suggest forecasts be regenerated and capacity plans be rerun about every quarter.

Our final experiment is measuring the cost savings achieved by optimizing floor-shell and tool expansions jointly with an integrated model. We compare total costs given by two hierarchical decomposition heuristics (using sequential planning of tools and fabs) with those of the integrated model (FIFEX). Although the best heuristic is within 2% of optimal, considering the absolute magnitude of the costs involved we suggest integrated models be used instead of sequential planning.

Acknowledgment : This work was supported by a grant from Semiconductor Research Cooperation under the task “Modeling Random Processes”, and by the National Science Foundation under the grant DMI-9713549.

References

- [1] A. Angelus, E.L. Porteus and S.C. Wood. *Optimal sizing and timing of capacity expansions with implications for modular semiconductor wafer fabs*. Research Paper No.1479. Graduate School of Business, Stanford University, 1997.
- [2] Annual Report & Directory. Edited by Jeff Weir. Semiconductor Industry Association, 181 Metro Drive, Suite 450. San Jose, California, 1998.
- [3] K.R. Baker and D.W. Peterson. *An analytic framework for evaluating rolling schedules*. Management Science, Vol.25, No.4, pp.341-351, 1979.
- [4] S. Bhatnagar, E. Fernandez-Gaucherand, M.C. Fu, Y. He and S.I. Marcus. *A Markov decision process model for capacity expansion and allocation*. Technical report, ISR, University of Maryland, MD, 1999.
- [5] D.L. Benavides, J.R. Duley and B.E. Johnson. *As good as it gets: optimal fab design and deployment*. IEEE Transactions on Semiconductor Manufacturing, Vol.12, No.3, pp.281-287, 1999.
- [6] J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Verlag, New York, 1997.
- [7] M. Çakanyıldırım and R.O. Roundy. *Demand forecasting and capacity planning in the semiconductor industry*. Technical Paper no: 1229, SORIE, Cornell University, NY, 1999.
- [8] M. Çakanyıldırım and R.O. Roundy. *SeDFAM: Semiconductor demand forecast accuracy model*. Technical paper no: 1230, SORIE, Cornell University, NY, 1999.

- [9] M. Çakanyıldırım, R. O. Roundy and S. C. Wood. *Optimal capacity expansion strategies under demand uncertainty*. Technical Report 1250, SORIE. Cornell University, NY, 1999.
- [10] P.K. Chatterjee and G.B. Larrabee. *Gigabit age microelectronics and their manufacture*. IEEE Trans. VLSI Syst. Vol.1, No.1, pp.7-21, 1993.
- [11] Z.-L. Chen, S. Li and D. Tirupati. *A scenario based stochastic programming approach for technology and capacity planning*. Working paper 98-04, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA, 1998.
- [12] A. Dixit. *Investment and employment dynamics in the short run and the long run*. Oxford Economic Papers, Vol.49, No.1, pp.1-20, 1997.
- [13] J.C. Eberly and J.A. Van Mieghem. *Multi-factor dynamic investment under uncertainty*. Journal of Economic Theory Vol.75, pp.345-387, 1997.
- [14] G.D. Eppen, R.K. Martin and L. Schrage. *A scenario approach to capacity planning*. Operations Research, Vol.37, No.4, pp.517-527, 1989.
- [15] L.F. Escudero, P.V. Kamesam, A.J. King and R.J.B. Wets. *Production planning with scenario modeling*. Annals of Operations Research, Vol.43, pp.311-335, 1993.
- [16] S.B. Gershwin. *Manufacturing Systems Engineering*. Prentice-Hall, 1994.
- [17] E.M. Goldratt. *Theory of Constraints*. North River Press, 1990.
- [18] J.M. Harrison and J.A. Van Mieghem. *Multi-resource investment strategies: operational hedging under demand uncertainty*. European Journal of Operational Research, Vol.113, No.1, pp.17-29, 1999.
- [19] D.A. Hicks. *Evolving complexity and cost dynamics in the semiconductor industry*. IEEE Trans. on Semiconductor Manufacturing, Vol.9, No.3, pp.294-302, 1996.
- [20] G.B. Larrabee. *Flexibla intelligent manufacturing: Its role in microelectronic device manufacture*. Manuf. Rev. pp.212-226, 1994.
- [21] S. Nahmias. *Production and Operations Analysis*. The McGraw-Hill Companies, 1997.
- [22] S.L. Ong and B.J. Adams. *The effect of discretizing the planning period on the optimal cost of capacity expansion systems*. European Journal of Operational Research, Vol.43, No.3, pp.292-304, 1989.
- [23] Technology Forecast. Edited by Michael Katz. Price Waterhouse World Technology Centre, 68 Willow Road. Menlo Park, California, 1997.

- [24] S. Takriti, J.R. Birge and E. Long. *A stochastic model for the unit commitment problem*. IEEE Trans. on Power Systems, Vol.11, Iss.3, pp.1497-1506, 1996.
- [25] R.J.B. Wets. *Stochastic programming*. Optimization: Handbooks in OR & MS ed. G.L. Nemhauser et al. Elsevier Science Publishers B.V. North-Holland, 1989, Chapter VIII.

Figure 1: Coefficients of variation and standard deviations for the Light and Heavy Tail demands.

Figure 2: Demand (percentiles and mean) vs. capacity, and shortfalls for the Light and Heavy Tail demands.

Figure 3: Increase in total (lost sales plus tool) costs as the time between availability epochs increase.

Figure 4: FIFEX and Heuristic (Hrste) tool and lost sales costs as lost sales cost per wafer varies.

Figure 5: Ratio of total FIFEX cost to the percentile, loss and bottleneck heuristic costs, and shortfalls (for FIFEX and heuristics) as lost sales cost per wafer varies.

Figure 6: Increase in costs as months between forecasts (*dup*) goes up.

Figure 7: Comparison of Lost Sales, Tool, Floor-Shell expansion costs of the solutions given by FIFEX, Top-Down and Bottom-Up Heuristic.

Figure 8: Total cost ratios for FIFEX vs Top-Down and FIFEX vs. Bottom-up Heuristic.

Figure 9: With inflated costs, comparison of Lost Sales, Tool, Floor-Shell expansion costs of the solutions given by FIFEX, Top-Down and Bottom-Up Heuristic.

Figure 10: With inflated costs, total cost ratios for FIFEX vs Top-Down and FIFEX vs. Bottom-up Heuristic.