# Demand Forecasting and Capacity Planning in the Semiconductor Industry *

Metin Çakanyildirim and Robin O. Roundy
*www.orie.cornell.university/∼metin/semicond*
School of Operations Research and Industrial Engineering
Cornell University

February 12, 1999

## 1 Introduction

We recently had extensive conversations with several semiconductor manufacturers to familiarize ourselves with current forecasting and capacity planning practice in the semiconductor industry. In this note, we summarize our observations. We saw a lot of variety in the approaches used for capacity planning. There seems to be substantial room for improvement. We describe clear tradeoffs in data accuracy, model accuracy and model response time. We also give a brief overview of our research agenda in this area.

The semiconductor industry has boomed to become one of the leading industries in US economy. However, astronomical fab costs (an average of $2.5 B, see [3]) combined with ever reducing chip prices are cutting into the profit margins significantly. It is reported (see [1]) that, in 1996, the industry reinvested 23% of the total revenue in capital expenses, mostly (60-70%) for tool purchases. However, tool capacity planning is not trivial. Tool plans are based on demand forecasts of 6-12 months out into the future (the tool delivery lead time). Thus, financial success is tightly tied to utilization of tools ordered almost a year in advance. The industry faces the challenge of balancing capital investment costs against the risk of losing revenue. In the next section, we will discuss our observations pertaining to demand forecasting. That will be ensued by a discussion of current practices in capacity planning. Next, we will briefly describe our future research agenda.

## 2 Demand Forecasting Practice

Demand forecasting drives many manufacturing functions. Thus, it can not be overemphasized that accurate forecasts are crucial. Nevertheless, it is commonly believed in the industry that forecast accuracy is getting worse, both because product life cycles are shortening and because line-widths are shrinking.

Semiconductor companies have parallel ways of obtaining demand forecasts. Demand is generally calculated by summing up internal and external demands. Internal demand is obtained by talking to customers,

*Technical Report 1229, SORIE. Cornell University, NY.

in the same company, downstream in the supply chain. These, upon examining their production schedule and demand profile, schedule their requirements, which become internal demand forecast for semiconductor manufacturers. External demand forecasting is strikingly similar. Outside customers provide the forecasts, which are summed up to get the total.

Forecasts are usually generated for the coming 4-5 years. They are by the part numbers for the first few years. Further out, parts are aggregated either by line-width or part families. Forecasts are generated from scratch usually twice in a year. Monthly updates are made between these generation epochs to reflect the changes in the business environment (see Table 1).

| Forecasting Characteristic | Short Term | Long Term |
|---|---|---|
| Years Out | 1-2 | 3-5 |
| Purpose | Tactical | Strategic |
| Forecast by | Parts | Aggregated parts |
| Time bucket length | Short (month) | Long (quarter) |
| Update frequency | High (monthly) | Low (quarterly-annually) |

Table 1: Forecasting characteristics

As pointed out earlier, shortening line-widths and short product life cycles make forecasting challenging. That is because, when forecasting the demand for a given product with a given line-width, the forecast depends heavily on both the time that the new line-width will be available and the time that the product moves to the new line-width. The former is specified by process development teams. For some companies, it is relatively stable; for others it is volatile. It ranges from one to three months. The latter is an outcome of market forces, and often product development schedules. Hence it is uncertain. It is possible to see up to a six month of delay in moving a product to a new line-width. (see Table 2).

| Aspects of Accuracy | Short Term | Long Term |
|---|---|---|
| Months Out | 6 | Beyond 6 |
| Accuracy at part level | Good (25% off) | Poor |
| Month to month fluctuations | Small (<5%) | Very small |
| Accuracy at aggregate | Very good (<10-20% off) | Reasonable |

Table 2: Forecasting accuracy

Demands for new products are naturally much more difficult to forecast than demands for established products. New product forecasts are often very optimistic. Hence, forecast errors for new products can have both high variance and bias. With more frequent product introductions, the effect of inaccuracies in new product forecasts is becoming more pronounced.

It is often advantageous for customers to deliberately overestimate their capacity needs to make sure that enough fab capacity is allocated to them. Some manufacturers are aware of this and monitor the quality of the each customer's forecasts to detect any consistent overestimation. Some forecasters track or estimate their customers' inventory levels. They use these figures in evaluating the forecasts given by the

customer.

Forecasters do not always agree with the internal/external customers. Sometimes forecasters want to alter the data provided to them. Due to the risk of turning out to be wrong, they prefer not to do so. This can lead to negotiations between forecasters and customers. Most forecasters do not possess an objective tool to distinguish customers who give accurate forecasts, from the rest.

## 3 Capacity Planning Practice

A capacity plan is a period by period schedule of tool purchases. Tool purchases require an enormous amount of capital: 60-70% of initial fab expenses are made for tools. Moreover, around 15% of existing tools are replaced every year.

The steps of a typical capacity planning process are as follows: The first step is to assemble demand forecasts. Next comes a corporate-level rough-cut capacity planning process. A primary output of this process is either a budget or a production target. This information guides the detailed capacity planning processes at the fabs. When a budget is provided by the corporate offices, demand forecasts are also inputs to fab-level capacity planning processes. These processes generate specific requests for new tools. Specific tool purchase requests are then sent to the corporate offices for approval. These requests may not be approved due to marketing strategy or the tool plans of other fabs. If approved, tools are ordered from vendors in a timely fashion, considering delivery lead times. Delivery lead times vary depending on the vendor's workload. However, once the vendor quotes a due date for delivery, that due date is usually not violated. (see Figure 1).


Figure 1: Tool delivery lead times

Capacity planning models need operational data such as time per part, maintenance times, ramp schedules etc. These data change frequently (e.g. bi-weekly). It is important to update the data to calculate the current capacity accurately. Some companies feed updated capacity data on many tool groups into capacity planning models twice a month. Others furnish data on critical tools only once every three months. At some companies, human resources are used to carefully filter the data for accuracy, for relevance and to account for future changes. In summary, there is a clear trade off between data accuracy and human resources expended. (see Table 3).

| Aspects of data management | Frequency |
|---|---|
| Updated data fed to capacity model | 2 weeks vs. 3 months |
| Human resources expended for data evaluation and filtering | Very high vs. very low |

Table 3: Updating operational data and human involvement

Naturally, each company uses a different capacity planning model. Some companies use more than one model, differing in terms of detail. Model accuracy differ with product aggregation, time bucket length, and number of tool groups modeled. There is a wide disparity in the computational engines used to get numbers out of these models. In almost all models, planners intervene to some degree. The impact of human intervention tends to be lower when the model represents reality well. The level of human intervention is the dominant determinant of response time of the fab-level capacity planning model. A planning horizon of 5 years is uniform in our experience. The inputs to the model keep changing, so the planning process is run from scratch at prescheduled times (Major planning iteration). For the small changes in the input, a partial rerun may be satisfactory (Minor planning update). (see Table 4).

| Aspects of capacity model | Range |
|---|---|
| Number of tools modeled | 100 vs 1 tool |
| Planning time buckets | Short (month) for near future, longer (annual) for later |
| Computational engine | Optimization software, spreadsheets, database software |
| Human intervention | Output evaluation, corrective action, gathering and analyzing data |
| Response time | Hours vs. weeks |
| Planning horizon | 5 years |
| A major planning iteration | once every 3-6 months |
| A minor planning update | weekly or monthly |

Table 4: Capacity modeling characteristics

Clearly, tool plans are created in a volatile environment. Capacity planners are well aware of these uncertainties; however, capacity models with long response time severely limit the use of scenario analysis.

# 4 Research Agenda

Our current research will address the challenge of balancing capital investment costs against the risk of losing revenue. Specifically, it will help understand and model the demand for semiconductors over, say, the next year and to effectively utilize that information to manage tool capacity.

In the overall, we are developing an integrated model to quantify the risks associated with a given tool purchase plan. At a given point in time, a company has forecasts for future demand. In the first step of this model, we estimate the variances and covariances of the errors in these forecasts ([2]). This quantified understanding of demand uncertainty can be used, among other things, to quantify the risks and the expected revenues that are associated with a given tool purchase plan.

The second step will involve stochastic capacity planing tools. We will use our forecast error model of the first step as a basis for determining tool needs. We will develop optimization algorithms that produce plans (schedules) for bringing tools into a new or an existing fab. These plans will perform well for a wide range of the most likely demand realizations. They will strike the best possible balance among financial risk, service and tool utilization.

## 5 Conclusion

With regard to forecasting processes, semiconductor manufacturers are exposed to risks related to forecast accuracy. They can neither measure nor control these risks satisfactorily. We observed two risk reduction techniques, monitoring customer's inventory levels and negotiating. If manufacturers could participate in the customers' forecasting processes, they would develop a better understanding of uncertainty and risk.

For capacity planning, each company uses a different methodology. These methodologies differ in their data accuracy, their model accuracy and their response time. For a specific manufacturer, the issue is deciding which of these three key properties of the planning process are most critical. Only then can an appropriate methodology be chosen.

*Acknowledgment: We wish to thank to many individuals from U.S. semiconductor companies who helped in the preparation of this document.*

## References

[1] Annual Report & Directory (1998). Edited by Jeff Weir. Semiconductor Industry Association, 181 Metro Drive, Suite 450. San Jose, California.

[2] Çakanyildirim, M. and R.O. Roundy. (1999). SeDFAM: Semiconductor Demand Forecast Accuracy Model. Technical paper no: 1230, SORIE, Cornell University, NY.

[3] Technology Forecast (1997). Edited by Michael Katz. Price Waterhouse World Technology Centre, 68 Willow Road. Menlo Park, California.