Tool capacity planning for semiconductor fabrication facilities under demand uncertainty

Jayashankar M. Swaminathan *

Walter A. Haas School of Business, University of California, Berkeley, CA 94720, USA

Received 1 November 1997; accepted 1 September 1998

Abstract

This research is motivated by issues faced by a large manufacturer of semiconductor devices. Semiconductor manufacturing companies allocate millions of dollars every year for new types of machine tools for their facilities. Typically these are special purpose machine tools which are made to order. The rate of change in products and technology makes it difficult for manufacturers to have a good estimate of future tool requirements. Further, manufacturers experience a long lead time while procuring these tools. In this paper, we model the tool capacity planning problem under uncertainty in demand. The number of tools required in a facility is sufficiently large (nearly hundred or more tools) to make it nearly impossible to obtain efficient exact algorithms. We provide heuristics to find efficient tool procurement plans and test their quality using lower bounds on the formulation. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Demand scenarios; Semiconductor manufacturing; Strategic planning; Optimization; Heuristics

1. Introduction

The semiconductor manufacturing process can be divided into four basic steps: wafer fabrication, wafer probe, assembly and final testing. Wafer fabrication is the most technologically complex and capital intensive of all four phases (Uzsoy et al., 1992). It involves the processing of wafers of silicon or gallium arsenide to build different layers and patterns of metal and wafer material to produce the required circuitry. The facility in which the fabrication takes place is called a wafer fab. The initial investment for building a wafer fab is close to two billion dollars (Benson, 1997). In addition, every year tool and equipment procurement could cost several million dollars per facility. Wafer fabs are complex manufacturing environment which support a wide variety of products and processes. Products undergo operations (often more than once) on different tools before completion. Production planning is difficult due to complex product flows, random yields, diverse equipment characteristics, downtimes, shared facilities and data maintenance (Leachman, 1993;
Hughes and Shott, 1986). Fordyce and Sullivan (1994) identify the procurement and allocation of tools based on a demand forecast as one of the most important issues faced by managers of wafer fabs.

In this paper, we address the procurement of tools for a wafer fab based on demand forecasts. In recent years, tool procurement planning has become more challenging at wafer fabs that produce application specific integrated circuits (ASICs). Some of the main reasons are as follows.

Changes in technology and products: Technology in this area has changed rapidly, which has meant that new products are being introduced into a facility all the time. This has led to additional requirements for new tools and equipment. Moreover, the life cycle of products is becoming shorter, making forecasting future demand for wafers even more difficult.

Lead time for procurement: Most of the tools that are procured are highly customized and made-to-order. The lead time for procuring a tool could range anywhere from three months to a year. As a result, planners have to decide on tool procurement based on forecasts for demands for the next year. In a rapidly evolving technology environment these forecasts could be highly inaccurate.

Cost of tools: The cost of tools has been on the rise as indicated in Fordyce and Sullivan (1994). Our interactions with industry indicate that a medium sized wafer fab could spend a few million dollars every quarter on procurement of new tools. Thus, even a small improvement in the procurement decisions could have a large impact on the manufacturer’s performance.

Demand characteristics: Manufacturers of ASIC devices face very stringent customer requirements. The demands are unpredictable and are lost if the manufacturer does not have enough capacity during a period of high demand. In some cases, the manufacturer can outsource production, but this may not be cost efficient during periods of high demand.

The wafer fab that motivated this research procured tools based on a single forecast of demands (we call this a coordinated plan). Our industrial contacts indicate that this is a standard practice in industry. A coordinated plan is developed by a group of experienced planners and contains the most likely forecast for demands for different wafers. Tool procurement is planned so that the tools have a high utilization while meeting the demand projections. Such a planning approach has often led to either lower utilization of tools (if the actual demand realized is less than projected) or have led to shortages (if actual demands are higher or if the mix changes). In order to overcome the above problems, the manufacturer was interested in developing a tool procurement policy that hedges against the uncertainty in future demands for various products.

Our aim in this paper is to provide an analytical model for tool procurement that incorporates the uncertainty in demand forecasts and provides methods to operationally hedge against it. Our model enables a manufacturer to plan for a set of possible demand scenarios (as opposed to a single coordinated plan) and procure an efficient set of tools. We model the problem as a discrete single period capacitated problem where one period represents a quarter since that is the planning period used in industry. Extension of this model to multiple periods is discussed in Section 6.1. The tool procurement problem consists of two parts – (1) orders for tools have to be placed before the actual demand is realized; (2) once the demand is known, different wafer types (products) need to be assigned to different tools in order to minimize shortages. We formulate this problem as a stochastic integer program with recourse. The first stage integer variables (that are decided before demand occurs) are the number of tools procured and the second stage variables (that are decided after the demand occurs) determine the allocation of different wafer types to the different tools in each demand scenario. We utilize the objective of minimizing the expected stock-out costs due to lost sales across all demand scenarios. Two variants which incorporate tool utilization are discussed in Sections 6.2 and 6.3.

2. Related literature and our contributions

There is significant amount of literature in the area of semiconductor manufacturing. Uzsoy et al.
(1992) provide an exhaustive review of production planning and scheduling models. They classify research into three broad areas – performance evaluation, production planning and shop-floor control. Uzsoy et al. (1994) provide an extensive literature review of research related to shop-floor control. Golovin (1986) discusses various issues related to production planning and scheduling and advocates a hierarchical approach. Leachman (1993) presents the initial work on production planning that utilizes linear programming and integrates decisions at different levels within the hierarchy. Leachman and Carmon (1992) present different ways to aggregate constraints in order to make the solution procedure more efficient.

There has not been much research in the area of tool procurement in production planning. Fordyce and Sullivan (1994) present an initial model for allocation of wafers to different tools which enables what if's on questions that arise during planning and management of tools. Their goal programming approach has been incorporated in ROSE; a decision support system utilized by IBM at their wafer fabs. However, they restrict their attention to a coordinated demand forecast and plan in order to maximize the utilization of tools. Other research models for capacity planning in semiconductor industry have also limited themselves to planning for a known forecast (Leachman and Carmon, 1992). A reason for this is the complexities associated with solving a stochastic program. However, increase in computational power in recent years has made it possible to address some of those concerns. As a result, scenario based planning is starting to be used for industry size problems (see Swaminathan and Tayur, 1995; Swaminathan, 1996).

In this paper, we present a model (based on stochastic program with recourse) for tool procurement planning that utilizes the information in scenario based forecast estimates and generates an efficient procurement plan. This is one of the first models that explicitly captures uncertainty in demand while generating a tool procurement plan. We model the problem as a large scale mixed integer program and develop lower bounds using Lagrangean relaxation and selectively relaxing the integrality constraints on tools. Since the industry sized problem is very large (nearly hundred or more tools) it is impossible to find the optimal procurement plan within a reasonable time. We provide two heuristics, one based on the data related to cost of tools and another based on a greedy approach to procuring tools. We also derive conditions under which optimal solution of the coordinated plan is strictly dominated by the optimal solution from a scenario based plan. In our computational section, we compare the performance of our heuristic to the solutions from a coordinated plan. Our results indicate that the heuristics provide effective solutions even for large problems and their performance is superior to the solutions of the coordinated planning approach on average.

The rest of the paper is as follows. In Section 3, we introduce our scenario based model and the assumptions associated with it. In Section 4, we discuss our lower bounds and the heuristics. In Section 5, we provide details of our computational study. In Section 6, we discuss a multi-period extension and other variants and we conclude in Section 7.

3. Model formulation

In this section we present a single period problem where the manufacturer plans procurement of tools for the next period (typically the first quarter of next year) based on the current demand forecast for the different wafer types. A multi-period model where procurement decisions are taken for all the four quarters of the next year is described in Section 6.1. The manufacturer faces long procurement lead times (three months to a year) for tools. As a result, the manufacturer has to utilize inaccurate demand forecasts for products. We model the future demand by a set of L demand scenarios where each scenario (assigned a probability of occurrence) reflects a possible set of demands for the different wafer types at the beginning of the next year. Based on our interactions with industry, we find that the management has the ability to generate about ten such scenarios with reasonable effort. For each demand scenario, wafer produc-
tion is assigned to the different tools in order to meet the demand as closely as possible. Though tool procurement decisions have to be made long before the actual demand is realized, management needs to incorporate information on how the tools would be utilized when the actual demand is realized while deciding on the procurement plan. The performance of the tool procurement plan is evaluated by the expected stock-out cost incurred due to lost sales in the next period (the first quarter of next year). Traditionally, research has considered either utilization or delivery performance as performance measures (Uzsoy et al., 1992). Delivery performance measures are becoming more and more appropriate in an environment where customer lead times are reducing and customer satisfaction is becoming more important. Since we are interested in fabrication facilities that manufacture ASIC devices, a delivery performance is more appropriate. As a result, we chose stock-out cost as the performance measure. Variants of the model that consider utilization are discussed in Sections 6.2 and 6.3.

We model the wafer production process under the following assumptions. Each wafer type $i$ ($1, \ldots, N$) has to undergo a number of operations and we assume that it takes $t_{ij}$ amount of time for processing a unit of wafer type $i$ on tool $j$ ($1, \ldots, n$). We indirectly model yield losses by inflating the process times $t_{ij}$ as suggested in Fordyce and Sullivan (1994). Traditional production planning models have considered a finer representation where each operation could be alternatively performed on different tools and the model decides not only the amount of production of wafer $i$ on tool $j$ but also which operations are performed on tool $j$. Such a variant is discussed in Section 6.4. In each demand scenario $l$ ($1, \ldots, L$), $s_l$ is the amount of wafer $i$ which is produced in scenario $l$, $\pi_l$ is the stock-out cost per unit of wafer $i$ (incurred as a result of lost sales during the quarter) and $p_l$ the probability of scenario $l$. $C$ is the total time during which the wafer fab is in operation during the period and we assume that all tools are available during this time. In addition, we assume that new and old tools have the same availability thereby neglecting some of the issues related to ramp-up of new equipments. However, the above changes can be easily incorporated into the model by adding tool specific coefficients for available capacity. $d_{il}$ is the demand forecast for wafer type $i$ in scenario $l$. $c_j$ is the cost of procurement for tool $j$. The number of existing tools of type $j$ is represented by $x_{j0}^I$, and $x^I_j$ represents the vector $(x^I_1, x^I_2, \ldots, x^I_n)$. Finally, the manufacturer has a budget $B$ which restricts the number of tools that can be bought for the next period.

The sequence of events in the model are as follows: (1) tools are ordered for the next period (before demand is known); (2) newly procured and existing tools are utilized to satisfy the demand for the wafers in the next period (after demand is realized). The problem can be formulated as a two stage stochastic program with recourse. The first stage variables representing the number of tools to be procured $X_l$ are decided before the demands are known. The second stage continuous variables representing production of wafers in each scenario ($s_l$) are the best recourse actions taken so that the stock-out cost in that scenario is minimized. The first stage variables are so chosen so that the expected cost incurred (expectation with respect to demand) in the second stage is minimized. This problem can be reformulated as a large scale mixed integer program by explicitly modeling the random demand for different wafers using a set of demand scenarios.

\[
\text{WSP:} \quad \min_{X_l, s_l} \sum_{i=1}^L p_l \sum_{i=1}^N \pi_l (d_{il} - s_{il})
\]

s.t.
\[
\sum_{i=1}^N t_{ij} \cdot s_{il} \leq (x^I_0 + x^I_j)C \quad \forall j, l, \tag{1}
\]
\[
s_{il} \leq d_{il} \quad \forall i, l, \tag{2}
\]
\[
\sum_{j=1}^n c_j x^I_j \leq B, \tag{3}
\]
\[
s_{il} \in R^+, \tag{4}
\]
\[
x^I_j \in Z^+. \tag{5}
\]

In the above formulation the expected cost incurred is computed by weighting the costs by the probability of occurrence of each scenario. Constraints (1) indicate that the total number of wafers
that can be produced on a tool in each scenario is limited by the time availability of that tool. Constraints (2) indicate that the amount of wafers produced is less than or equal to the demand in each scenario. We assume that one does not produce more than required because it may not be a very good idea to produce for inventory unless it is known that there is going to be a huge demand in the near future. Since the products are ASIC they cannot be easily substituted for other wafers. Constraint (3) makes sure that the procurement of other wafers does not exceed the budget for the period. Constraints (4) and (5) denote that the tools have to be procured in integral quantity whereas production can be denoted by non-integral values. It is to be noted that having productions as continuous variables is an approximation which is made in most planning models because the problem could become extremely difficult to handle otherwise. In addition, the production quantities are typically large values in which case such an assumption is reasonable.

3.1. Coordinated vs. scenario based planning

In this section, we introduce coordinated planning for tool procurement and present conditions under which the optimal solution of the coordinated plan will be dominated by a scenario based plan.

A coordinated plan is developed by a group of experienced planners and contains forecasts \( d_i^C \) of future demand for the different wafer types \( i \) produced at the wafer fab. As a result, planners consider only one demand scenario and try to procure tools in order to minimize the expected costs. The scenario chosen is the most likely scenario and often the estimates of demands are inflated a bit to avoid stock-outs later. The tool procurement problem for a coordinated plan can be formulated as given below.

\[
\text{WSPC: } \min_{X_i^C} \sum_{i=1}^{N} \pi_i (d_i^C - s_i^C) \\
\text{s.t. } \sum_{i=1}^{N} (t_{ij} \cdot s_i^C) \leq (x_0^C + x_i^C)C \quad \forall j, \\
\sum_{j=1}^{n} c_j x_j^i \leq B, \\
s_i^C \in R^+, \\
x_i^C \in Z^+. 
\]

As evident, the objective is to minimize the stock-out cost incurred and tools are procured based on a single demand forecast.

Remark 1. WSP and WSPC are equivalent if there exists \( l' \quad (1 \leq l' \leq L) \) such that \( p_{l'} = 1 \) and \( d_{l'} = d_i^C, \forall i \).

Definition 1. Given a constrained minimization problem \( P \), a problem \( Q \) is a quasi-relaxation of \( P \) if for every feasible solution of \( P \) with objective function value \( v \), there is a feasible solution of \( Q \) having an objective function value less than or equal to \( v \).

Remark 2. Optimal value of a quasi-relaxation \( Q \) is a lower bound on the optimal value of \( P \).

It is to be noted that linear programming relaxation is a special case of quasi-relaxation.

Definition 2. A demand scenario is termed most optimistic demand scenario \( d_i^o \) if \( d_i^o \geq d_i \forall i \) where \( i \) represents the different wafers and \( l \) represents scenarios under consideration.

One possible way to generate a most optimistic demand scenario is by setting the optimistic forecasts as follows:

\[ d_i^o = \max \{d_{i1}, d_{i2}, \ldots, d_{il} \} \quad \forall i. \]

Proposition 1. WSP is a quasi-relaxation of WSPC if the most optimistic demand scenario is used in WSPC.

Proof. Let \( l' \) denote the most optimistic demand scenario which is utilized in WSPC. Then the following three conditions are valid: (i) \( d_i^C = d_{i'} \forall i \); (ii) \( d_i^o \geq d_i \forall i \neq l' \); (iii) \( p_{l'} > 0 \). Let there be a feasible solution for WSPC given by \( X_i^C, s_i^C \), respectively. Now, \( X_i^C \) is a feasible tool procurement plan for WSP (since the budget constraint is
not violated). Construct a feasible solution for WSP in the following manner. $X_1 = X_1^C$ and $s_{jl} = \min(s^C, d_{jl}) \forall i, l$. $0 \leq (d_{jl} - s_{jl}) \leq (d_{jl} - s^C) \forall i, l$ because if $s_{jl} = d_{jl}$ then $d_{jl} - s_{jl} = 0$ and $d_{jl} \geq s^C$ from inequality (2) in WSPC; on the other hand, if $s_{jl} = s^C$ then $(d_{jl} - s_{jl}) \leq (d_{jl} - s^C)$ since $d_{jl} \geq d_{jl} \forall i, l$.

Thus, the expected cost incurred in WSP is lower than the cost incurred in WSPC. □

The above result indicates that with the same budget the optimal tool procurement for WSP incurs a lower stock-out cost than the optimal tool procurement for WSPC. In other words, with a lower budget (than $B$) in WSP it is possible to incur the same stock-out cost. Since stock-out cost is a surrogate for service, it is possible to provide same service with a lower budget when a scenario based plan is utilized for tool procurement. It implies that planning for a set of demand scenarios is always more efficient than planning for a coordinated demand forecast based on the most optimistic estimate of product demands. An important managerial insight arises from the above result. Typically, the manufacturing group analyzes the demand forecast and service requirements before requesting budget for tool procurement from corporate. The above result indicates that the manufacturing group is likely to ask for a lower budget (for a given service) from the corporate office when they plan according to a scenario based approach as compared to the case when they plan for a single most optimistic coordinated scenario. It is to be noted that the above insight holds good only if the scenarios on hand are correct, or in other words, when the demand gets realized from one of the scenarios considered in scenario based planning. However, in practice, this condition may be difficult to achieve.

Note that any tool procurement solution $X_1^C$ for WSPC under all demand conditions is a feasible tool procurement for WSP. Therefore, the expected cost incurred by the optimal tool procurement for WSPC will always be greater than the expected cost incurred by the optimal tool procurement for WSP when their performance is evaluated under stochastic demand (given by the set of demand scenarios used in WSP). In our computational study, we compare the performance of our heuristics with the performance of the coordinated plan under different probabilities for the most probable scenario.

4. Solution methodology

WSP enables modeling the uncertain environment encountered during tool procurement planning. However, it is an extremely large problem and in addition has many integer variables which makes it impossible to solve optimally within a reasonable time. In this section, we introduce lower bounds for the problem and subsequently describe our heuristics.

4.1. Lower bounds

Linear relaxation of WSP provides a lower bound on the objective value of the solution. However, our experience showed that the lower bound is very weak. We generate additional lower bounds based on Lagrangean relaxation and selectively relaxing integer variables.

4.1.1. Lagrangean relaxation

Lagrangean relaxation is generated by relaxing a set of constraints and provides a lower bound at least as good as the linear relaxation. In our problem, we relax the constraints (1) in order to obtain the relaxed problem WSPL.

$$\text{WSPL}(\lambda): \min_{X_1, s} \sum_{l=1}^{L} p_l \sum_{i=1}^{N} \pi_i (d_{jl} - s_{jl})$$
$$+ \sum_{l=1}^{L} \sum_{j=1}^{n} \lambda_{jl} \left( \sum_{i=1}^{N} (t_{ij} \cdot s_{jl}) - (x^j_0 + x^j_1)C \right)$$
subject to:
$$s_{jl} \leq d_{jl} \ \forall i, l, \quad (2)$$
$$\sum_{j=1}^{n} c_j x^j_i \leq B, \quad (3)$$
$$s_{jl} \in R^+, \quad (4)$$
$$x^j_i \in Z^+. \quad (5)$$
WSPL can be decomposed into two problems WSPLa and WSPLb as follows.

\[
\text{WSPLa}(\lambda): \quad P1(\lambda) = \min_{x_1} \sum_{j=1}^{L} \sum_{i=1}^{n} - \lambda_{ij}(x_0^i + x_1^i)C \\
\text{s.t.} \quad \sum_{j=1}^{n} c_j x_1^i \leq B, \quad x_1^i \in \mathbb{Z}^+.
\]

(1)

\[
\text{WSPLb}(\lambda): \quad P2(\lambda) = \min_{s_1} \sum_{j=1}^{L} \sum_{i=1}^{n} p_i(d_{il} - s_{il}) + \sum_{i=1}^{n} \lambda_{ij} \sum_{j=1}^{n} t_{ij} \cdot s_{il} \\
\text{s.t.} \quad s_{il} \leq d_{il} \quad \forall i, l, \quad s_{il} \in \mathbb{R}^+.
\]

(2)

The above decomposition leads to two problems that are easier (computationally) to solve as compared to the original problem. WSPLa is a knapsack problem (containing all the integer variables) and WSPLb is a linear program. WSPLb can be solved using a one pass algorithm where \( s_{il} = d_{il} \) if \( p_i, p_i \geq \sum_{j=1}^{n} (t_{ij} \cdot \lambda_{ij}) \), else \( s_{il} = 0 \). Thus, it can be solved in linear time. We solve WSPLa using CPLEX integer programming routines. The Lagrangean lower bound (LB1) can be obtained by solving the following optimization problem:

\[
\text{LB}_1 = \max_{\lambda \geq 0} P1(\lambda) + P2(\lambda).
\]

The above optimization LB1 is solved by using a sub-gradient based algorithm to find the optimal values of \( \lambda_{ij} \) (see Nemhauser and Wolsey, 1988). We start with an initial value \( \lambda_{ij}^0 \) and solve the above problems. Subsequently, the \( \lambda_{ij} \) are updated as follows.

\[
\lambda_{ij}^{k+1} = \lambda_{ij}^k + \theta_k \delta_{ij}^k \quad \text{where} \quad \delta_{ij}^k = \sum_{i=1}^{N} t_{ij} \cdot s_{il} - (x_0^i + x_1^i)C, \\
\theta_k = \text{step}_k \times \frac{(\text{UB} - \text{WSPL}(\lambda_k))}{\left(\sum_{j=1}^{n} \sum_{l=1}^{L} (\sum_{i=1}^{N} (t_{ij} \cdot s_{il}) - (x_0^j + x_1^j)C)^2\right)^{1/2}}.
\]

The step size \( \text{step}_k \) is chosen in an appropriate manner so that procedure converges. For example, when \( \text{step}_k \to 0 \) and \( \sum_{j=1}^{\infty} \text{step}_j \to \infty \); see Bazaara and Shetty (1979). UB is an upper bound found with the heuristic (described in Section 4.2). The above method converges to the optimal values of \( \lambda \) and provides the lower bound LB1.

### 4.1.2. Selective relaxation

A lower bound for the problem WSP can be obtained by considering only \( n' \) tools in WSP as integers where \( n' \leq n \). Consider the following formulation SWSP where the first \( n' \) tools are considered to be solved as integer and the rest are solved as continuous variables.

\[
\text{SWSP} : \quad \min_{x_1, x_1} \sum_{j=1}^{L} \sum_{i=1}^{n} p_i(d_{il} - s_{il}) \\
\text{s.t.} \quad \sum_{j=1}^{n} (t_{ij} \cdot s_{il}) \leq (x_0^i + x_1^i)C \quad \forall j, l, \\
\sum_{i=1}^{n} c_j x_1^i \leq B, \quad s_{il} \leq d_{il} \quad \forall i, l, \\
\sum_{i=1}^{n} c_j x_1^i \leq B, \quad s_{il} \in \mathbb{R}^+, \quad x_1^i \in \mathbb{Z}^+ \quad \forall j \leq n', \\
\sum_{i=1}^{n} c_j x_1^i \leq B, \quad s_{il} \in \mathbb{R}^+, \quad x_1^i \in \mathbb{R}^+ \quad \forall j > n'.
\]

Since only a partial set of tools are considered to be procured in an integral manner the solution to SWSP is a lower bound on the original problem WSP. We will call this lower bound LB2. We chose \( n' \) most expensive tools as integers since they have the maximum influence on the budget constraint and leave the rest as continuous variables.

### 4.2. Upper bounds

In this section we provide two heuristics to generate upper bounds for the problem.

#### 4.2.1. Slack based heuristic – SB

This heuristic utilizes the slack variables obtained from the selective relaxation SWSP. We
• Round all values \( x_j \) where \( j > n' \) to the nearest integer less than or equal to \( x_j \).

• Compute the additional budget and call it \( S \).

• Calculate the average slack value corresponding to constraints (1) across all the scenarios for \( j > n' \) call it \( \text{slack}_j \).

• Sort the variables in ascending order in terms of \( \text{slack}_j \cdot c_j \).

• While \( S > 0 \) do
  
  - Pick the first tool \( j \) such that \( c_j \leq S \).
  
  - Increase \( x_j \) by 1. Update \( S = S - c_j \)

  - Goto next \( j \)

• Output Solution.

Fig. 1. Heuristic – SB.

generate the upper bound by rounding all the variables \( x_j \) where \( j > n' \) to the nearest integer less than or equal to \( x_j \). Then we utilize the additional budget to procure some more tools. Fig. 1 shows the steps in the heuristics. The solution obtained from the heuristic, called UB1, provides an upper bound to the problem WSP. Typically, a heuristic based on selective relaxation of variables is not widely used due to the fact that it is difficult to determine effectively as to which variables to relax. Based on the discussions with our industrial contacts we find that the 80–20 rule holds good for the cost of tools. So, 20% of tools are very expensive and take up more than 80% of the budget and, the remaining 80% are relatively inexpensive and roughly total 20% of the budget. The expensive tools typically are the lithography photo clusters and ion implantation chamber tools while the inexpensive ones include inspection microscopes and bake ovens. In this heuristic we consider 30% of the tools to be integers and round the value of the remaining 70% based on their slack. We test this heuristic in our computational section.

4.2.2. Greedy heuristic – GB

In this heuristic we procure tools one at a time in a greedy manner till we run out of budget. This heuristic can be invoked anytime with a remaining budget of \( S \) and the existing tools \( x_0 \). Let us assume that before the heuristic begins somehow it is determined that \( y_j \) units of each tool have to be purchased and the remaining budget is \( S \). Then the budget constraint (3) is removed from WSP. The resulting linear program (called LINWSP) is solved with the updated number of existing tools. Dual variables from constraint (1) are utilized to find the tool that provides the maximum reduction in costs per dollar spent in procurement. This can be found by summing the dual value across all scenarios corresponding to constraint (1) for each tool and dividing that by \( c_j \). One unit of the most promising tool is procured and the number of existing tools is updated and the linear program is solved again. This goes on till we run out of budget (refer Fig. 2).

The solution obtained from the greedy heuristic called UB2 provides a second upper bound to the problem WSP. In our computational section, we use this heuristic after procuring an initial set of tools based on the coordinated demand forecast. We choose the initial procurement \( y_j \) to be one less than that required to fully satisfy the demand for the coordinated plan. We appropriately reduce the available budget \( B \) in order to obtain \( S \) and update the existing tools based on the initial procurement. This starting allocation is feasible because in all our computational study, the budget \( B \) that we choose in WSP is the amount required
so that tools can be procured to fully satisfy the demand in the coordinated plan. This way of setting the budget is consistent with the real situation where planners and finance people iterate a number of times adjusting both the demand forecasts and budget till they converge on a consistent coordinated demand forecast and a budget figure which allows procurement of tools for that forecast.

5. Computational study

In the computational study, we mainly focus on two issues. Firstly, we test the quality of our heuristic relative to the optimal solution either by a direct comparison or by comparing the gap between the bounds. Secondly, we compare the performance of a procurement plan based on coordinated planning to our heuristic solution under varying degrees of demand uncertainty.

In our computational study, the existing number of tools are generated such that there is 90% tool availability for the most probable demand scenario and we generated the processing time \( t_{ij} \) based on the aggregate information provided to us by our industrial contact. We compute the budget in the following manner. First we solve for the requirements for the different tools based on the coordinated demand forecast by solving a relaxed version of WSP where the budget constraint is removed and all the tool procurements \( x_j \) are treated as continuous variables. Using the optimal result, we round all the tool procurement variables to obtain an integer solution. We find the cost of procuring those tools and treat that as the budget. Determining budget in the above manner enables us to have a fair comparison between coordinated and scenario based planning. We base all our computations on problem sets with 25 wafer types which is representative of the product line being manufactured in the wafer fab. It is to be noted that as the number of wafer types increases the accuracy of demand forecasts typically goes down, as a result, one might have to consider more number of scenarios. Further, increasing the number of wafer types increases the number of constraints (one for each scenario) and variables (one for each scenario) in the formulation. However, the computational complexity of the problem

- Initial Procurement: \( y^j \).
- Increase \( x_j \) by \( y^j \).
- Set \( S = B - \sum_{j=1}^{n} c_j y^j \).
- Remove the budget constraint (3), fix \( x_1 \) to zero and create a linear program LINWSP.
- While \( S > 0 \) do
  - Solve LINWSP.
  - Calculate the sum of dual values corresponding to constraints (1) across all the scenarios for \( j \) and call it \( dual_j \).
  - Sort the variables in increasing order in terms of \( dual_j/c_j \).
  - Pick the first tool \( j \) such that \( c_j \leq S \).
  - Increase \( x_1 \) by 1. Update \( S = S - c_j \).
- Output Solution.

Fig. 2. Heuristic – GB.
does not change as much since no additional integer variables are added.

5.1. Goodness of heuristic

In this section, we report our findings about the goodness of the heuristics proposed in the paper. We first compare the solution to the optimal value for small problems with 30 tools, 10 demand scenarios and 25 wafer types. The most probable demand scenario is assigned a probability of 0.5 and the remaining scenarios are assigned equal probability. The demands for the different wafers are generated from a multi-variate normal distribution. We tested 120 random problems whose results are summarized in Tables 1 and 2. We solved for the selective relaxation by relaxing 21 least expensive tools. Table 1 shows the average and best case performance of the two heuristics

\[
\text{(Heuristic-Optimal) \times 100 \over \text{Optimal}}
\]

Table 2 shows the number of problems where the deviation from the optimal solution was less than 1%, between 1% and 5%, between 5% and 10% and greater than 10%.

Results for both heuristics indicate that the heuristics perform well on this sample set of problems. The slack based heuristic provides solutions on average 2.38% away from the optimal solution while providing solutions within 1% of the optimal solution in 57 problems out of 120. The greedy heuristic provides solutions on average within 5.7% and provided solutions to 6 problems within 1% of the optimal solution. The best solutions of SB and GB are 0.025% and 0.29% away from the optimal solution, respectively.

Subsequently, we generated 80 problems with 100 tools, 10 demand scenarios and 25 wafer types which closely models a real planning problem in size. We solve the selective relaxation problem by relaxing the 70 least expensive tools. Our results are summarized in Tables 3 and 4 where the gap is computed as

\[
\text{(Heuristic-LB) \times 100 \over \text{LB}}
\]

The lower bound LB is taken to be the better of Lagrangean bound and selective relaxation. Both heuristics perform reasonably well although SB outperforms GB in a relative sense. The gaps between the lower bound and the heuristic solutions have an average of 5.9% and 9.9% for SB and GB, respectively (Table 3). Part of the reason for the large gap is that our lower bounds are not very tight. The distribution of the gap between the heuristics and the lower bound is given in Table 4. Though the heuristics seem to provide reasonable solutions (gap less than 10%) still a few percent difference could translate into thousands of dollars since the actual cost incurred is very large. In order to understand the effectiveness of our heuristics, we decided to conduct another set of study to compare the performance of a coordinated planning environment with the performance of our heuristics.

5.2. Comparison of coordinated and scenario based planning

In this section, we compare the performance of the heuristic solution to the expected cost incurred with the coordinated tool procurement. We con-
Consider the coordinated demand forecast as one of the \( L \) demand scenarios in WSP. We vary the probability of occurrence of the scenario corresponding to the coordinated demand forecast as a measure for the accuracy of the forecast. The rest of the \( L - 1 \) scenarios are assumed to be equally likely. We generated 40 problems each with 100 tools, 10 scenarios and 25 wafer types and varied the probability of occurrence of the most probable scenario \( p_C \) (on which the coordinated planning is based) to 0.50, 0.65, 0.80 and 0.95. When \( p_C = 0.95 \) it implies that the coordinated forecast is extremely accurate and those forecasts are likely to occur with probability 0.95. The rest of the 9 scenarios occur each with a probability 0.05/9. The demand for wafers are generated from a multi-variate normal distribution. Table 5 shows the mean percentage difference between the solutions obtained for the scenario based planning using heuristics SB and GB and coordinated planning

\[
\left( \frac{(\text{Coordinated-Heuristic}) \times 100}{\text{Coordinated}} \right).
\]

A positive value indicates that the solution generated by the heuristic incurs a lower cost and is better by the percentage indicated.

Table 3
Performance of heuristics in terms of gap between the heuristic and the best lower bound

<table>
<thead>
<tr>
<th></th>
<th>Best case(%)</th>
<th>Average case(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slack based heuristic (SB)</td>
<td>0.01</td>
<td>5.91</td>
</tr>
<tr>
<td>Greedy heuristic (GB)</td>
<td>0.03</td>
<td>9.96</td>
</tr>
</tbody>
</table>

Table 4
Distribution of the gap between the heuristic and the best lower bound out of 80 cases

<table>
<thead>
<tr>
<th></th>
<th>( X &gt; 10% )</th>
<th>( 10% \leq X &lt; 5% )</th>
<th>( 5% \leq x &lt; 1% )</th>
<th>( x \leq 1% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slack based heuristic (SB)</td>
<td>16</td>
<td>24</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Greedy heuristic (GB)</td>
<td>39</td>
<td>34</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5
Percentage difference between coordinated planning and the heuristics under different accuracy of forecast

<table>
<thead>
<tr>
<th></th>
<th>( p_C = 0.50(%) )</th>
<th>( p_C = 0.65(%) )</th>
<th>( p_C = 0.80(%) )</th>
<th>( p_C = 0.95(%) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slack based heuristic (SB)</td>
<td>23.26</td>
<td>14.33</td>
<td>-0.81</td>
<td>-25.4</td>
</tr>
<tr>
<td>Greedy heuristic (GB)</td>
<td>16.72</td>
<td>11.15</td>
<td>4.28</td>
<td>1.37</td>
</tr>
<tr>
<td>Best = Min (GB, SB)</td>
<td>23.63</td>
<td>15.63</td>
<td>5.15</td>
<td>1.38</td>
</tr>
</tbody>
</table>

Table 5 shows that as the accuracy of the forecast (on which the coordinated plan is generated) becomes better the performance of the heuristics that generate the procurement plan for a scenario based planning worsens. The SB heuristic provides solutions are on an average 23.2% better when the probability of the occurrence of the coordinated scenario \( p_C \) is equal to 0.5. However, the performance worsens with increase in \( p_C \) and it provides solutions which are 25.4% worse than the coordinated plan when \( p_C = 0.95 \). An intuitive reason for SB worsening in performance as \( p_C \) increases could be that when \( p_C \) has a higher value it becomes more important to procure tools for the coordinated scenario which may be quite different from the most expensive tools that SB focuses on. As a result, the performance of SB is much worse. When \( p_C \) has a lower value SB performs much better because it focuses on expensive tools which are most likely to constrain the budget. On the other hand, heuristic GB outperforms the coordinated plan under all conditions and provides solutions which are at least 1.37% better than the coordinated plan on average. One of the reasons for the better performance of the heuristic GB as compared to the coordinated plan is that it first starts with a procurement plan which has one less tool of each type than suggested by the coordinated plan and uses the remaining budget to hedge against the chances for demand occurring from other scenarios.

Our results indicate that if the coordinated forecast has a low probability of occurrence then it is better to utilize the slack based heuristic SB. On
the other hand, if the probability is high then the greedy heuristic GB should be used. Both these heuristics can be solved quickly even for large problems (maximum time is close to 30 minutes to solve both heuristics on a SUN ultra-1 machine).

On comparison of the best solution from the two heuristics with coordinated tool planning, we find that the best of the two heuristics consistently outperforms coordinated planning under all conditions (Table 5). The improvement in performance decreases with increase in $p_C$ and ranges from 23.47% to 1.38%.

6. Multi-period extension and other variants

In this section we present a multi-period model for the problem and three variants of the basic model.

6.1. Multi-period model

Tool procurement decisions are sometimes made every quarter while considering the demand scenarios for all the four quarters in the next year. The additional demand information could be useful when product life cycles are towards the beginning or the end since those could be incorporated in the tool procurement. For example, if there is a tool that is used explicitly for a particular wafer whose product life cycle is expected to end after the first period, then it may be better to incur some stock-out cost in the first period as opposed to buying that tool and keeping it idle for the remaining three periods. In this section we introduce a multi-period model (having 4 periods). Here are some additional notations. (i) $x_i^t$: number of tool $j$ procured for the period $t(1 \ldots 4)$; (ii) $B_t$: total budget for period $t$; (iii) $d_{il}^t$: demand for product $i$ in scenario $l$ during period $t$; (iv) $s_{il}^t$: amount of product $i$ produced in quarter $t$ in scenario $l$; (v) $C_t$: time available per tool during quarter $t$.

\[ \text{MWSP : } \min_{x_1, x_2, x_3, x_4, s_{il1}, s_{il2}, s_{il3}, s_{il4}} \sum_{i=1}^{4} \sum_{t=1}^{L} p_i \sum_{l=1}^{N} x_{il} \times (d_{il}^t - s_{il}^t) \]

\[ \text{s.t.} \]

\[ \sum_{i=1}^{N} t_{ij} \cdot s_{il}^t \leq \sum_{p=0}^{t} x_{p}^t C_t \quad \forall j, l, t, \]

\[ s_{il}^t \leq d_{il}^t \quad \forall i, l, t, \]

\[ \sum_{j=1}^{a} c_j x_{il}^t \leq B_t \quad \forall t, \]

\[ s_{il}^t \in R^+, \]

\[ x_{il}^t \in Z^+ \quad \forall t, j. \]

The above problem is computationally very difficult because it has four times the number of integer variables in the formulation as compared to a single period problem and many more constraints. An approximate solution can be found by relaxing all the variables $x_{il}^t$ in the above formulation to be continuous for $t > 1$. Though the problem size remains the same the number of integer variables is reduced considerably. On this relaxed problem we can obtain the lower and upper bounds the same way that we did for the single period problem. This problem can be solved on a rolling basis in order to obtain approximate solutions for the multi-period problem MWSP.

6.2. Utilization measure

A scenario based approach can be utilized for procuring tools for a utilization based measure as well. We introduce some additional variables as follows: (i) $w_{il}$: amount of idle time for tool $j$ in scenario $l$; (ii) $a_{il}$: cost of one unit of idle time for tool $j$ in scenario $l$. The problem can be formulated as follows.

\[ \text{UWSP : } \min \sum_{i=1}^{L} p_i \sum_{j=1}^{N} a_{il} w_{il} \]

\[ \text{s.t.} \]

\[ \sum_{i=1}^{N} \sum_{k=1}^{K} t_{ij} \cdot s_{ij} = (x_{il}^1 + x_{il}^2)C \quad \forall j, l, \]

\[ s_{il} \leq d_{il} \quad \forall i, l, \]

\[ \sum_{j=1}^{a} c_j x_{il}^t \leq B_t, \]
6.3. Hybrid model: Utilization and stock-out costs

Utilization and delivery performance are two types of measures that have been used in the semiconductor industry. The coordinated planning approach plans for a single coordinated demand forecast and procures tools so that the stock-out cost is minimized. Since the plan is for only one scenario, optimal tool procurement is such that their utilization is high. On the other hand, a scenario based approach generates solutions which incurs the minimum cost across all scenarios and yet may not provide a high utilization of tools for any given single scenario. Since management in many cases is concerned about utilization as much as about reduction in expected cost we provide a hybrid model that considers both utilization and stock-out costs. The model determines the most efficient tool procurement plan (in terms of stock-out costs) but in addition makes sure in each demand scenario the tool \( j \) has a utilization greater than \( q_{ij} \) (0 \( \leq \rho_j \leq 1 \)). The problem can be formulated by adding an additional set of constraints to WSP in order to guarantee a minimum level of utilization.

\[
\sum_{i=1}^{N} t_{ij} \cdot s_{ij} \geq (x_{ij}^d + x_{ij}^r)C \rho_j \quad \forall j, l.
\]

6.4. Alternative operations allocation

In the basic model we assume that we are only concerned with the production of wafers in each scenario and that the processing times and routing for these wafers is given. In some cases there may be an opportunity to make use of alternative routing for the wafers in which case we need a detailed model. Let each wafer have \( K \) \((k = 1, \ldots, K)\) operations that have to be completed. Each operation needs to be done at most once on a wafer. \( t_{ijk} \) is the amount of time it takes to perform the \( k \)th operation on wafer \( i \) on machine \( j \). \( t_{ijk} = 0 \) if operation \( k \) need not be performed on wafer \( i \) and \( t_{ijk} = \infty \) if operation \( k \) needs to be performed on wafer \( i \) but can not be performed on tool \( j \). Let \( q_{ijkl} \) be the amount of wafer \( i \)'s, \( k \)th operation that is allocated on tool \( j \) in scenario \( l \). The problem can be formulated as follows.

\[
\text{WSPA} : \quad \min_{x_{1,j},q_{ijkl}} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{l=1}^{L} p_i \sum_{j=1}^{M} \pi_l (d_{ij} - s_{ij})
\]

s.t.
\[
\sum_{i=1}^{N} \sum_{k=1}^{K} t_{ijk} \cdot q_{ijkl} \leq (x_{ij}^d + x_{ij}^r)C \quad \forall j, l, \quad (1)
\]
\[
\sum_{j=1}^{M} q_{ijkl} \geq s_{ij} \quad \forall i, k, l, \quad (2)
\]
\[
s_{ij} \leq d_{ij} \quad \forall i, l, \quad (3)
\]
\[
\sum_{j=1}^{M} c_j x_{ij}^l \leq B, \quad (4)
\]
\[
s_{ij}, q_{ijkl} \in R^+, \quad (5)
\]
\[
x_{ij}^l \in Z^+ \quad \forall j. \quad (6)
\]

The above problem can be solved using heuristics outlined in Section 4; however, the outlined Lagrangean relaxation procedure may not be suitable for this formulation.

7. Conclusions

Rapid changes in products and technology combined with long procurement lead times for tools have made it extremely difficult to procure tools in an efficient manner. Motivated by a problem faced by a semiconductor manufacturer, in this paper we present a model (based on stochastic program with recourse) for tool procurement planning that utilizes the information in scenario based forecast estimates and generates an efficient procurement plan. Our model is one of the first models that explicitly captures uncertainty in demand while generating a tool procurement plan. In this paper, we model the problem as a large scale mixed integer program. We develop lower bounds using Lagrangean relaxation and selectively relaxing the integrality constraints on tools.
In addition, we provide two heuristics, one based on the data related to cost of tools and another based on a greedy approach to procuring tools. We also derive conditions under which optimal solution of the coordinated plan would be strictly dominated by the optimal solution from a scenario based plan. In our computational section, we study the quality of our heuristics and compare their performance to the solutions from a coordinated plan. The methodology and results in this paper have demonstrated the drawbacks of a coordinated plan as well as the benefits of scenario based planning.

Acknowledgements

The author wishes to thank the referee and the editor for their comments which have improved the readability of the paper to a great extent. I also wish to thank Dr. Sarah Hood and other seminar participants at IBM T.J. Watson Research Center for their valuable inputs to this paper. This research was partially supported by the Junior Faculty Research Grant at The University of California, Berkeley.

References


