A Continuous-Time Strategic Capacity Planning Model Based on the Minimum-Cut Problem

Woonghee Tim Huh
School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853
huh@orie.cornell.edu

Advisor: Robin O. Roundy
School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853

The semiconductor industry has been one of the driving forces of the "new" economy. It boasts of the exponentially growing performance of semiconductor devices, coupled with rapidly decreasing chip prices; however, it faces highly volatile demands, and copes with astronomical fab costs, most of which can be attributed to tool costs. The leadtime for purchasing tools is between 6 and 18 months, upon which tools quickly become obsolete. Thus, semiconductor companies need to recover their capital investment in the tools over a short period of time.

We develop models and algorithms for strategic capacity planning, which is to determine the sequence and timing of acquiring tools. Strategic planning decisions are made in the presence of high uncertainty. Uncertainty comes from factors such as technology, the market, and its products, and becomes amplified by long leadtimes. Although capacity planning decisions need to be made in the presence of high uncertainty, early research and even some current practices overlook the stochastic nature of planning, with the exception of simple case analyses. An extensive review of literature can be found in Çakanyildirim et al. (1999) and Roundy et al. (2000). Typical methods of stochastic optimization include stochastic-linear programming, stochastic-integer programming, and Markov decisions processes; yet, they have not been able to solve real-world capacity planning problems on the scale faced by the semiconductor industry.

This paper takes the stochastic-optimization approach, which explicitly incorporates randomness in the model. We assume nonstationary stochastic demand, with the expected demand for product families increasing over time. We also assume lost sales and no finished good inventory. As in Çakanyildirim et al. (1999), we continue to explore alternative approaches based on *continuous-time models*. The time at which a machine is purchased becomes a continuous-decision variable. These models are more compact than traditional stochastic-programming methods based on discrete-time models. It is hoped that the small dimensionality of continuous-time models will make the strategic capacity planning problem computationally tractable.

In this paper, we model multiple resource types used for multiple product families. The resulting problem is related to the continuous relaxation of the lot-sizing problem. We present an efficient divide-and-conquer algorithm that will find a locally optimal solution of this problem. A subroutine to this algorithm is the parametric minimum-cut problem.

1. Formulation

We provide a mathematical formulation of the strategic capacity planning problem. Due to the high rate of obsolescence, industries such as the semiconductor industry have low finished-goods inventory. This model assumes that negligible amounts of finished-goods inventories are held. Motivated by current industry practices, it also assumes that backorders are negligible. These assumptions imply that in the

1523-4614/03/0501/0063$05.00
1526-5498 electronic ISSN

recourse, the production quantities at a given time are functions of the capacity and demand at that time only, and not of other time instances. At Time 0, all capacity acquisition plans are made, whereas production decisions are made at each time instance after instantaneous demands have been observed. We use this model as a part of a rolling-horizon implementation.

We denote $t \in [0, T]$ as a continuous time between 0 and T, where T is the planning horizon. We use p and m to index product families $\mathcal{P}$ and tool types $\mathcal{M}$, respectively. For each tool of type $m \in \mathcal{M}$, let $\mathcal{N}_m$ be the set of tools of type m. Let n index tools in the set $\mathcal{N}_m$ in the order that purchases will be made. The ordered set $\mathcal{N}_m$ determines the sequence of tool purchases of type m. We also use $j = (m, n) \in \mathcal{J}$ to index all tools of all types that we contemplate purchasing over the planning horizon.

The price of purchasing tool j at time t is given by a decreasing convex function $P_j(t)$ of t. The instantaneous lost sales cost is c_{pt} per unit of product family p at time t. Let u_{mn} be the capacity of the nth tool of the tool type m. For any given subset $Q \subseteq \mathcal{J}$ of tools and a given tool group m, let the associated tool capacity of the type m tools be $\mu_m(Q) = \sum_{n'<n} u_{mn'}$, where $n = \min\{n' \colon (m, n') \notin Q\}$. The definition of μ_m ensures that tools of the same type should be purchased in the given order because any tool purchased out of sequence does not contribute to the tool capacity of type m. To produce one unit of product family p, we utilize $U(m, p)$ units of capacity from each tool type m.

The decision variables we are interested in are the purchase times $\tau = (\tau_j \mid j \in \mathcal{J})$ of the tools. We minimize the sum of tool purchase costs and expected lost sales costs. The tool purchase cost is $\eta^P(\tau) = \sum_{j=1}^{J} P_j(\tau_j)$. Let $\xi(Q, t)$ be the expected instantaneous lost sales cost provided that $Q \subseteq \mathcal{J}$ is the subset of tools available at time t. We denote $Q_t^\tau = \{j \colon \tau_j \le t\}$ as the set of tools available at time t given purchase times τ. We can write the expected lost sales cost η^{LS} as an integral of instantaneous lost sales cost $\eta^{LS}(\tau) = \int_{t=0}^{T} \xi(Q_t^\tau, t)\, dt$. The problem we want to solve is the following:

$$(P) \quad \min \quad \eta(\tau) = \eta^P(\tau) + \eta^{LS}(\tau)$$
$$\text{s.t.} \quad 0 \le \tau_j \le T \quad \text{for all } j \in \mathcal{J}.$$

We derive another expression for $\eta^{LS}(\tau)$ within a subset of the feasible region and develop some properties of η. Let Π be the set of all permutations on $\mathcal{J}$, or bijective maps from $\{1, \dots, |\mathcal{J}|\}$ to $\mathcal{J}$. Each $\pi \in \Pi$ corresponds to a sequence of tool purchases, and the *permutation simplex* defined by π is $PS(\pi) = \{\tau \in [0, T]^{\mathcal{J}} \mid \tau_{\pi(1)} \le \tau_{\pi(2)} \le \cdots \le \tau_{\pi(|\mathcal{J}|)}\}$, which corresponds to the set of valid τ's for that sequence.

Suppose $\tau \in PS(\pi)$ where $\pi \in \Pi$. For each $r \in \{1, \dots, |\mathcal{J}|\}$, let $\pi^-(r) = \{\pi(r') \mid r' < r\}$. Suppose that $Q_t^\tau = \pi^-(r)$ for some r. The amount of reduction in the expected instantaneous lost sales cost ξ at time t by adding the tool $\pi(r)$ to the set of available tools is denoted by $g_{\pi(r)}^{\pi^-(r)}(t)$. Formally we define, for any $Q^o \subseteq \mathcal{J}$ and $j \in \mathcal{J} \backslash Q^o$, $g_j^{Q^o}(t) = \xi(Q^o, t) - \xi(Q^o \cup \{j\}, t)$. Note that $g_j^{Q^o}(t)$ is the difference, in lost sales cost, of having the tool set Q^o and that of having $Q^o \cup \{j\}$ at time t.

Suppose $\tau \in PS(\pi)$, i.e., τ follows the sequence given by π. Then for fixed t, $Q_t^\tau = \{j \in \mathcal{J} \mid \tau_j \le t\}$ can be expressed as $\pi^-(r_o) \cup \{\pi(r_o)\}$ for some $r_o \in \{1, \dots, |\mathcal{J}|\}$. Within the permutation simplex $PS(\pi)$, the expected lost sales cost η^{LS} is continuous and separable. It is also differentiable and its partial derivative with respect to $\tau_{\pi(r)}$ is $(\partial/\partial\tau_{\pi(r)})\eta^{LS}(\tau) = g_{\pi(r)}^{\pi^-(r)}(\tau_{\pi(r)})$, in the interior of $PS(\pi)$. Furthermore, η^{LS} is continuously differentiable if each g_r^π, $r \in \{1, \dots, |\mathcal{J}|\}$, is continuous with respect to τ. Whenever $\pi(r) = j$ and $\pi^-(r) = Q$, we have $(\partial/\partial\tau_j)\eta^{LS}(\tau) = g_j^Q(\tau_j)$. This is a much stronger separability of the expected lost sales cost η^{LS} than separability in each permutation simplex. We generalize the definition of g: For any disjoint sets Q^o, $Q \subseteq \mathcal{J}$ of tools, we define $g_Q^{Q^o}(t) = \xi(Q^o, t) - \xi(Q^o \cup Q, t)$. This quantity corresponds to the marginal benefit of adding the tool set Q to the existing set Q_o at time t. It can be shown $g_{Q^1}^{Q^o}(t) + g_{Q^2}^{Q^o \cup Q^1}(t) = g_{Q^1 \cup Q^2}^{Q^o}(t)$ if Q^o, Q^1, $Q^2 \subseteq \mathcal{J}$ are disjoint. It is a strong additivity property of derivatives of η^{LS} that spans many neighboring permutation simplices.

We let $h_j(t) = (d/dt)P_j(t) \le 0$ be the rate of change in the tool cost at t. By the convexity of the tool cost P_j, $h_j(t)$ is nondecreasing. For $Q \in \mathcal{J}$, we set $h_Q(t) = \sum_{j \in Q} h_j(t)$. We remark that within the permutation simplex $PS(\pi)$ defined by π, the objective function η is separable and its partial derivative with respect to j is $(\partial/\partial\tau_j)\eta(\tau) = h(\tau_j) + g_j^{\pi^-(r)}(\tau_j)$, provided $j = \pi(r)$.

Suppose at time t, we have a partition Q_L, Q_o, and Q_U of $\mathcal{J}$, where Q_L is the set of tools we have purchased prior to t and Q_U is the set of tools we will purchase after t. Currently, we purchase tools in Q_o at t. If we split Q_o, and uniformly slide $Q \subseteq Q_o$ earlier and $Q_o \backslash Q$ later, then η changes at the rate of $-[h_Q(t) + g_Q^{Q_L}(t)] + [h_{Q_o \backslash Q}(t) + g_{Q_o \backslash Q}^{Q_L \cup Q}(t)]$. Minimizing this expression is called the *cluster-splitting* of Q_o given Q_L and Q_U.

2. Demand Modeling

As in Roudy et al. (2000), we model the random-demand vector D_t at time t as a sum of a deterministic part and a stochastic part, i.e., $D_t = b_t + \Delta_{I_t,t}\phi_{I_t,t}$, where $b_t = (b_{pt} \mid p \in \mathcal{P}) \in \mathbb{R}^{\mathcal{P}}$ is a deterministic nonnegative vector that is nondecreasing in t; I_t is a discrete random variable whose support is a finite set $\mathcal{I}_t$ such that $P[I_t = i] = w_{it}$ for each $i \in \mathcal{I}_t$; $\phi_{it} = (\phi_{ipt} \mid p \in \mathcal{P})$ is a deterministic nonnegative unit-norm directional vector in $\mathbb{R}^{\mathcal{P}}$; and Δ_{it} is a continuous nonnegative random scalar along ϕ_{it}. Intuitively, the demand D_t is determined by starting at b_t, randomly selecting a direction by observing I_t, and moving a random distance $\Delta_{I_t,t}$ in the direction $\phi_{I_t,t}$.

Currently, most models of high-dimensional random vectors are either continuous (e.g., multivariable normal) or discrete (e.g., multinomial). Our demand model is a hybrid of both: No point in $\mathbb{R}^{\mathcal{P}}$ has any nonzero probability mass. The support of D_t is a finite collection of rays emanating from b_t and has measure zero. It is shown in Roudy et al. (2000) that by a variance-reduction technique called conditioning, our demand model can approximate a continuous distribution in $\Re^{\mathcal{P}}$ more accurately than the conventional method of sampling points, provided that the number of vectors is the same as the number of points.

There is no demand shortfall if the capacity $\mu_m(Q_t^\tau)$ is sufficient to meet the demand d_t, i.e., $\sum_{p=1}^{P} U(m,p)\, d_{pt} \le \mu_t(Q_t^\tau)$ for all $m = 1, \ldots, M$. Otherwise, we are unable to meet all demands. The following section outlines a policy we use to allocate insufficient capacity to product families.

3. Shortfall Allocation

This section explains how we determine the expected value ξ of the instantaneous lost sales cost. The lost sales at time t depend on demands for product families at time t, capacities of tool types at time t, and the allocation of tool capacities to product families. Given a set Q of tools that are available at time t (which is determined by τ), tool type m's capacity is given by $\mu_m(Q)$. Given the capacity $\mu(Q) = (\mu_m(Q) \mid m \in \mathcal{M})$ of all tool types and the *realized* demand $d_t = (d_{pt} \mid p \in \mathcal{P})$ of all product families at time t, we determine both the production quantity $v_t = (v_{pt} \mid p \in \mathcal{P})$ of product family p and the allocation $x_t = (x_{mpt} \mid m \in \mathcal{M},\ p \in \mathcal{P})$ of tool type m's capacity to p. A *capacity allocation policy* is a way of selecting x_t and v_t.

As in Çakanyildirim et al. (1999) and Roundy et al. (2000), we assume no finished-goods inventory and no backorders. In other words, demand at time t can be satisfied by what is produced at time t only. Thus, in any capacity allocation policy, production should not exceed demand, i.e., $v_{pt} \le d_{pt}$ for all $p \in \mathcal{P}$. Production v and allocation x must obey the capacity limit of each tool type: $\sum_{p=1}^{P} x_{mpt} \le \mu_m(Q)$ for all $m \in \mathcal{M}$ and $t \in [0, T]$, and $U(m,p)v_{pt} \le x_{mpt}$ for all $p \in \mathcal{P}$ and $t \in [0, T]$.

We conceptually divide the demand into a deterministic portion $b_t \ge 0$ and a stochastic portion $\Delta_{I_t,t} \cdot \phi_{I_t,t} \ge 0$. We assume that there is enough capacity to meet the deterministic part b_t of the demand. We may ensure this assumption by imposing upper bounds on purchase times τ. Because $D_t \ge b_t$, our allocation policy meets the deterministic part b_t of demand before allocating resources to the stochastic part.

We use an allocation policy that determines production quantities v_t, which equalizes the instantaneous fill rates of stochastic portion of demand at time t across all products. In the recourse at time t, after the demand $d_{it} = b_t + \delta_{it}\phi_{it}$ is realized, this implies that we select production quantities $v_t = b_t + \zeta\phi_{it}$ for some $\zeta \in [0, \delta_{it}]$. Thus, v_t also lies on the ray defined by the starting point b_t and the direction ϕ_{it}. The value ζ indicates the magnitude of production along this ray. It is easy to see that the fill rate of the stochastic part for product p is $(v_{pt} - b_{it})/(d_{ipt} - b_{it}) = \zeta/\delta_{it}$, which is independent of the product family p. If $b_t = 0$, then this corresponds to the classical fill rate.

4. Divide-and-Conquer Algorithm

We outline an efficient divide-and-conquer algorithm to minimize the total cost η. This algorithm finds a solution that satisfies the first-order necessary condition for the optimality of (P)—Namely, this solution has no feasible descent direction. Our algorithm tracks and modifies clusters C that have the following properties: (1) C is a subset of the set $\mathcal{J}$ of all tools, and (2) there exists a lower bound $lb(C)$ and an upper bound $ub(C)$ such that we know there exists a solution τ^* where $lb(C) \le \tau_j^* \le ub(C)$ for all $j \in C$ such that τ^* satisfies the first-order necessary condition of (P). We note that if $lb(C) = ub(C)$, then we have found the desired purchase times τ_j^* for all $j \in C$. At the start of each iteration of the algorithm, we maintain an ordered collection $\mathcal{C}$ of sets, each of which has the above two properties. We note that $\mathcal{C}$ is a partition of the set $\mathcal{J}$ of all tools, and the intervals $[lb(C), ub(C)]$ defined for these clusters are mutually disjoint, except possibly at endpoints. If C_1 and C_2 are two members of $\mathcal{C}$ such that C_1 precedes C_2, then we have $ub(C_1) \le lb(C_2)$.

Here are the steps of the divide-and-conquer algorithm: (0) Initially, set $\mathcal{C} = \{\mathcal{J}\}$, $lb(\mathcal{J}) = 0$, and $ub(\mathcal{J}) = T$. (1) Choose some $\omega_C \in [lb(C), ub(C)]$, for each $C \in \mathcal{C}$. (2) Choose some $C \in \mathcal{C}$ such that $lb(C) < ub(C)$. Perform cluster-splitting of C at ω_C given Q_L and $\mathcal{J}\backslash(Q_L \cup C)$, and let $S \subseteq C$ be its optimal solution, i.e., let S minimize $\varphi_{\omega_C}(\cdot \mid Q_L, C, \mathcal{J}\backslash(Q_L \cup C))$, where Q_L is the union of all clusters preceding C in $\mathcal{C}$ and $S \subseteq C$. If the optimal value is nonnegative, set $lb(S) = ub(S) = \omega_C$. Otherwise, replace C with S and $\bar{S}$ in $\mathcal{C}$, where S precedes $\bar{S} = C\backslash S$. Let $lb(S) = lb(C)$, $ub(S) = \omega_C$, $lb(\bar{S}) = \omega_C(C)$, and $ub(\bar{S}) = ub(C)$. (3) Go to Step 1 unless $lb(C) = ub(C)$ for all $C \in \mathcal{C}$.

In general, finding the minimizer of φ_t may not be easy. Using explicit enumeration takes $O(2^{|Q_o|})$ computational time. Yet, under our modeling assumptions, we can minimize φ_t efficiently by constructing a minimum-flow network of $O(|\mathcal{I}||\mathcal{J}|)$ nodes and arcs. The network, similar to one found in Roundy et al. (2000), exploits the separability properties of the expected lost sales cost as well as the order in which tools become bottlenecked along each demand ray. This divide-and-conquer algorithm resembles the algorithm of Gusfield and Martel (1992), for the monotone parametric minimum-cut networks, and the algorithm of Hochbaum and Queyranne (2000), for the convex cost-closure problem.

THEOREM 1. *At each iteration of the algorithm, there exists some solution τ^* with no descent direction in (P) such that $\tau_j^* \in [lb(C), ub(C)]$ where for all $j \in C$ and $C \in \mathcal{C}$. If the algorithm terminates, we have found such a solution.*

Under some assumptions, our continuous-time model becomes a minimization of a convex function and the above algorithm find the globally optimal solutions.

References

Çakanyildirim, M., R. O. Roundy, S. C. Wood. 1999. Machine purchasing strategies under demand- and technology-driven uncertainties. Technical Report TR1250, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York.

Gusfield, D., C. Martel. 1992. A fast algorithm for the generalized parametric minimum cut problem and applications. *Algorithmica* **7** 499–519.

Hochbaum, D. S., M. Queyranne. 2000. Minimizing a convex cost closure set. M. Paterson, ed. *Algorithms—ESA 2000, Proc. 8th Annual Eur. Sympos.*, Saarbrücken, Germany, September 5–8, 2000. Volume 1879 of *Lecture Notes in Computer Science*, Springer, New York, 2000. *SIAM J. Discrete Math.* Forthcoming.

Roundy, R. O., F. Zhang, M. Çakanyildirim, W. T. Huh. 2000. Optimal capacity expansion for multi-product, multi-machine manufacturing systems with stochastic demand. Technical Report TR1271, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York.