INCENTIVE SCHEMES FOR SEMICONDUCTOR CAPACITY-ALLOCATION: A GAME THEORETIC ANALYSIS

Suleyman Karabuk
Department of Industrial Engineering, College of Engineering
University of Oklahoma, Norman, OK 73019-0631
karabuk@ou.edu
S. David Wu
Department of Industrial and Systems Engineering, P. C. Rossin College of Engineering
Lehigh University, Bethlehem, PA 18015
david.wu@lehigh.edu

ABSTRACT

We study incentive issues that arise in semiconductor capacity planning and allocation. Motivated by our experience at a major U.S. semiconductor manufacturer, we model the capacity-allocation problem in a game-theoretic setting as follows: each product manager (PM) is responsible for a certain product line, while privately owning demand information through regular interaction with the customers. Capacity-allocation is carried out by the corporate headquarters (HQ), which allocates manufacturing capacity to product lines based on demand information reported by the PMs. We show that PMs have an incentive to manipulate demand information to increase their expected allocation, and that a carefully designed coordination mechanism is essential for HQ to implement the optimal allocation. To this end, we design an incentive scheme through bonus payments and participation charges that elicits private demand information from the PMs. We show that the mechanism achieves budget-balance and voluntary-participation requirements simultaneously. The results provide important insights into the treatment of misaligned incentives in the context of semiconductor capacity-allocation.

Keywords: semiconductor manufacturing, capacity planning, supply chain coordination, game theory, capacity-allocation game

1. Introduction

Semiconductor manufacturing operations consist of two main stages: the ‘front-end’ operation of wafer fabrication, and the ‘back-end’ operation of assembly and testing. The front-end operation is typically the bottleneck as the process involves a 6-12 week manufacturing lead-time, while the back-end requires 2-4 days. Moreover, the wafer fabs are extremely capital intensive (some require more than $2 billion) while requiring significant lead-time to build (12-18 months). Demand in high-tech industry is known to be volatile and particularly sensitive to economic cycles. Managing wafer fab capacity is among the most crucial activities for semiconductor firms. In this paper, we explore an aspect of semiconductor capacity management inspired by our experience at a major U.S. semiconductor-manufacturing firm.

In the semiconductor industry, *strategic capacity planning* decisions are typically made at the beginning of a fiscal year to determine how to configure and allocate wafer fab capacity with respect to aggregate microelectronics technologies required by different products. Key players in the planning process are *business units* directly responsible for customer demands, *manufacturing units* responsible for fab resources, and capacity planners at the corporate headquarters (HQ) who reconcile the goals of the business and manufacturing units while trying to realize strategic targets set by senior executives. Capacity-allocation resulting from strategic planning provides a basis for each business unit to strategize its negotiation posture with the customers while at the same time committing its financial accountability to the HQ. As such, the *strategic capacity planning* decisions are static by nature; they rarely change unless the market conditions change drastically. In between strategic planning updates, *tactical planning* takes place, where business units communicate to the wafer fabs their demand outlooks updated from new forecasts and confirmed customer orders, which results in an operational plan. In this context, capacity-allocation from strategic planning serves as a basis that specifies the level of wafer releases (capacity units) for which a business unit is entitled. In case the total demand exceeds the available capacity, the capacity is rationed proportionally to the initial capacity-allocation (known in the industry as “sharing-the-pain” policy). Strategic capacity-allocation must reflect a reasonably accurate match of supply and demand during its intended fiscal period. Otherwise, short-term fluctuations in the market may lead to great inefficiencies.

An important aspect of decision-making in the industry is the decentralized nature of demand management. Custom semiconductor manufacturers typically market application-specific chips to a wide variety of industry sectors such as telecommunications, consumer electronics, and computing equipment. As the customers and the demand characteristics in these industries are drastically different, firms often organize their business units according to industry sectors. Within each business unit, demand management authority is delegated to product managers (PMs) who are responsible for a certain line of products. PMs, as a result, have the most accurate information on the demands that they manage, as they interact with the customers on a regular basis. As an attempt to coordinate the PMs' efforts with company-wide performance, the PMs receive bonuses that are directly proportional to the total profits for their product lines. This incentive scheme is quite common in semiconductor firms. Unfortunately, the *strategic capacity planning* and the incentive structure described above lead to behaviors that have highly undesirable consequences. First, the PMs perceive the strategic capacity-allocation as capacity *guaranteed* for their products throughout the fiscal year. They would utilize this capacity as much as possible, sometimes regardless of the actual demands. Second, a PM may have incentive to inflate (or delay) his demand signal during tactical planning as a way to protect his allocated capacity. The HQ manager who oversees the capacity assignment activities describes the current allocation method and its shortcoming as follows:

> *…Our current capacity-allocation method is to assign a certain number of wafer starts to each business unit by technology groups based on some reference demand view, typically a demand view that is linked to a specific financial commitment. This type of allocation creates a sense of wafer-starts ownership, and has a tendency to cause business segments to hold on to their share of wafers until the last moment when they don't really need to make the starts, or they tend to build inventory. From a global asset utilization point of view, these allocations drive under-utilization by trapping pocket of capacity to segments with a low-swing of demand, where at the same time there are segments short of supply because of a high-swing of demand. Because it is necessary to have some finite lead-time on the high-swings of demand, wafers that are relinquished at the point of execution are*

> *sometimes too late to capture the upswing*. - Director of Integrated Circuit Business Planning & Rationalization

To address the above issues, the firm is interested in an incentive structure that would motivate PMs to reveal the true demand they observe during *tactical planning* such that the fabs may (re)allocate capacity dynamically to changing market conditions. On the other hand, due to the budgetary and fiduciary accountability demanded by corporate governance, the firm must retain the structure of *strategic capacity planning*, i.e., each PM *will* continue to receive an upfront capacity-allocations for planning purposes, knowing that the allocation may change later. While making perfect managerial sense, the above setting creates a complication from the modeling perspective since initial capacity-allocation imposes constraints that may hamper the efficiency for actual allocations that take place as the demand unfolds. This is a research issue that will be addressed in this paper.

The capacity planning and allocation dilemma described above can be observed in many industries, however, its impact is exacerbated in the semiconductor industry for the following reasons: *(i)* the lead-time required for capacity expansion is long (12-18 months), which makes strategic capacity planning essential not only for manufacturing planning, but also for capital budgeting and financial planning purposes; (*ii*) manufacturing lead-time is long while the product life-cycle is short, which severely limits the options for corrective planning; (*iii*) demand is highly volatile, and it is therefore difficult to hold a PM accountable for his demand forecasts; (*iv*) inventory carries a high risk of becoming obsolete, which diminishes its value to buffer for forecast inaccuracies; (*v*) the capacity cost (of wafer fabs) is extremely high, which means that small adjustments in capacity-allocation could have great impact on profitability.

In this paper, we propose a two-pronged approach to address the above incentive problems in semiconductor capacity-allocation; the goal is to improve efficiency in overall capacity-allocation while at the same time addressing the decision makers' (PMs) incentives. We propose a game- theoretic model for the capacity-allocation problem. We develop an incentive scheme that elicits private information from the PMs while implementing the optimal capacity-allocation that maximizes the firm's expected profits. The incentive scheme could be implemented using the executive bonus system commonly used in semiconductor firms.

The rest of the paper is organized as follows: in the next section, we review the related literature; in Section 2, we define the capacity-allocation problem and establish the solution that will optimize the firm's profit. In Section 3, we propose a capacity-allocation game that satisfies the budget-balance and voluntary-participation properties; this is followed by Section 4, where analyzes the mechanism and draws conclusions from its characteristics under various conditions. In Section 5, we describe a case study and numerical analysis based on real scenarios in the semiconductor environment, which is followed by the conclusions in Section 6.

1.1. Related Literature

Semiconductor capacity-allocation problems are extremely challenging due to the long manufacturing lead-time and the high demand uncertainty for semiconductor products. In this section, we first provide a broad overview of literature that addresses different aspects of semiconductor capacity planning, then point to a few papers that are closely related to our proposed capacity-allocation game.

The semiconductor capacity-allocation literature can be grouped into *optimization-based* methods that focus on large-scale mathematical programming models to find precise solutions, and *game-theoretic* models that capture the behavioral aspects of decision making, both provide valuable insights for the problem environment. In the former group, significant efforts have been put forward to analyze the strategic and operational trade-offs in capacity planning and capacity-allocation. Recent examples that represent this group of research include (Swaminathan 2001, 2002, Barahona et al. 2001), who develop mixed-integer stochastic programming models for tool acquisition and capacity management decisions at wafer fabs. Cakanyildirim and Roundy (2002) evaluate capacity-planning procedures used in practice, and extend these procedures while trying to maintain their simplicity. Karabuk and Wu (2002, 2003) develop stochastic programming models that address strategic capacity and operational capacity-allocation problems, respectively. Their model addresses potential conflicts regarding capacity planning between the productions and marketing functions in a semiconductor-manufacturing firm. Mechanisms are proposed to coordinate the capacity planning process. Christie and Wu (2002) propose a multi-stage stochastic programming model using scenario trees; they study different ways to characterize the dependencies among demand scenarios, and analyze the impacts on solution quality. In all of these studies an inherent assumption is that all of the input parameters of the decision models are available for the modeler. However, this assumption fails when participating decision makers have private information and are motivated to take advantage of it for local gains. This leads to the second area of literature that consider the players' incentives in a game-theoretic setting.

Several different approaches can be found in the literature that addresses the players' incentives in a capacity-allocation environment. For instance, Celikbas et al. (1999) devise penalty schemes to coordinate forecasting and production planning. Mallik and Harker (1998) develop a bonus rewarding function that provides marketing managers the incentive to improve their forecast accuracy. Porteus and Whang (1991) develop a transfer pricing scheme to coordinate inter-divisional capacity planning and allocation. Kouvelis and Lariviere (2000) generalize the approach of Porteus and Whang (1991) to a class of internal market-coordination mechanisms. Groves and Loeb (1979) challenge the effectiveness of pricing mechanisms as a means for coordinating divisional managers and propose instead performance evaluation measures based on divisional profits less the impact of bundling decisions on the profits of other divisions. On the other hand, Harris and Kriebel, (1982) design optimal transfer-pricing schemes for allocating resources under a more restrictive setting. The capacity-allocation problem has also been analyzed in the context of supply chain coordination (c.f., Cachon and Lariviere, 1999a, 1999b, Schneeweiss and Zimmer, 2004). Balasubramanian and Bhardwaj (2004) look into marketing-manufacturing coordination issues in a single company, which competes in a duopoly competition setting.

This paper contributes to the capacity coordination literature by introducing a game where a strategic capacity-allocation is first announced while actual (re)allocation takes place over time as demand unfolds at various tactical planning points. This setting captures an important characteristic of the semiconductor planning environment described earlier; the manufacturer must generate some form of long-term strategic capacity plan and make it known to external and internal constituents. Externally, the plan signals the customers the firm's future manufacturing capabilities pertaining to their products, and discloses to the investors and share-holders the firm's capacity positioning; internally, the plan serves as a reference that helps to set corporate revenue targets for different product categories while assigning priorities and financial accountabilities among business units. Although the mathematical model that we use takes a similar form to that of Cachon and Lariviere (1999b), we show that the initial allocation distorts the incentives of the players and makes coordination significantly more difficult.

To analyze the capacity-allocation game we turn to the mechanism design literature in microeconomics. There is a line of research that focuses on coordinating trades among independent agents to maximize the surplus generated by the trade, subject to the requirements of voluntary participation and budget balance. In our problem the initial capacity-allocation creates an ownership, and reallocation can be viewed as a trade among the participants. Similarly, in our case the HQ acts as an intermediary and wants to maximize the surplus generated by the reallocation of capacity while maintaining voluntary participation and recovering any additional bonuses paid to facilitate reallocation. Our analysis follows the general mechanism design framework established by Myerson (1982).

Several important findings in the literature are directly relevant to the capacity-allocation game we set out to study. Wu (2004) studies different type of supply chain intermediaries and their roles in subsiding the effects of adverse selection created by asymmetric information. Gibbard (1973), Green and Laffont (1977), and Myerson (1979) introduce the *revelation principle* for Bayesian games; it states that regardless of the actual mechanism constructed by the intermediary, given the Bayesian-Nash equilibrium outcome of the mechanism there exists an equivalent direct mechanism where the buyer and the supplier reveal their respective valuation to the intermediary, and the intermediary determines if the trade is to take place. This allows the study of a large class of asymmetric information games without the need to specify each mechanism in detail. Using the revelation principle, Myerson (1981) proposes an optimal mechanism design problem for auctions where the bidder chooses, among all possible mechanisms, one that would maximize her expected net revenue. Myerson and Satterthwaite (1983) prove that it is impossible to coordinate a buyer and a seller with an indivisible object if the supports of their belief distributions overlap. Cramton et al., (1987) develop a bidding mechanism that assigns a jointly owned asset to the partner who values it the most. Since the valuations of agents are linear functions of the quantity owned, the outcome represents the optimal reallocation of the asset. The mechanism only works if the ownership distribution satisfies certain conditions. McAfee, (1991) studies a case where the quantity of the goods to be sold is privately known by the seller. He proposes a method that would elicit private information while maximizing the total valuations. Makowski and Mezzetti (1993) analyze a trading problem for an indivisible object with two buyers and one seller. They characterize the conditions under which an implementable solution exists. Makowski and Mezzetti (1994) and Williams (1999) generalize the theory and characterize a broader set of conditions under which a mechanism with implementable properties exists.

We contribute to the literature by introducing a capacity (re)allocation game given *reference allocation* made by strategic capacity planning, and we develop a Bayesian incentive compatible mechanism that implements the firm's optimal capacity-allocation; we characterize the conditions when an individual rational and budget balanced incentive structure can be attained. We also extend the analysis of reallocation of a whole asset in continuous quantities and show that the results of Cramton et al. (1987) do not hold when the valuation function is nonlinear. More importantly, in the context of strategic capacity-allocation, we show that distributing the entire capacity among business units in an environment where there is private information may make it impossible to reallocate it efficiently. We also show that, if the HQ postpones a fraction of capacity distribution for later reallocation points, then system-wide optimal reallocation can be achieved. To our knowledge, this problem has not been studied before in the context of capacity coordination.

2. The Capacity-allocation Problem

2.1. The Product Manager's Decision Problem

We consider the decision problem of a PM during a tactical planning phase after the *reference allocation* has been made from strategic capacity planning. There can be several tactical planning points in one fiscal period. Reference allocations are typically revised on a yearly basis. We assume that the tactical planning points are infrequent enough (such as quarterly) that the decision point can be treated as independent. Therefore, we focus on a single decision point and describe a PM's decision problem at this point.

The enterprise planning system keeps track of customer orders for each product line and allows manual entry of forecast orders. Each PM prepares a priority-ordered demand list using the planning system. The ordering represents the relative importance of the entries in the list; importance is assigned based on the certainty or profitability of an order. At this point, PMs take into account their capacity share and may add or delete entries in the list to match or exceed their reference share. This list is then communicated to the wafer fab to be scheduled for production. The scheduler at the fab level releases wafers to satisfy the orders in the list; starting from the most important entries and continuing until the capacity share of the PM is filled. Capacity-allocation has a decreasing marginal profit for a business unit, because each additional unit of capacity will be used to fill a relatively less important order in the order list. The demand is highly volatile and even the actual customer orders placed at the time of planning are subject to significant changes throughout the manufacturing lead-time. We consider an aggregation of the order list that each PM submits and describe their model based on aggregate demand.

The demand analysis at the HQ level consists of historical data and input from the PMs. The analysis consists of identifying a nominal demand that is constant over the planning period. The variance on top of the nominal demand is very high and can be accurately identified by historical data. However, the nominal demand is not easy to predict and the PMs have more accurate demand information since they interact with the customers frequently and they can anticipate a shift in the nominal demand.

We describe the PMs' decision problem by a newsvendor model. Let ξ_i represent the random variable for the demand for PM *i*'s products and $F_i(\xi_i, \theta_i)$ represent its distribution with parameter θ_i. We assume that θ_i is the mean of the distribution, which represents the nominal demand for business unit *i*, and that the respective PM observes it privately.

We express the capacity as the number of wafers that can be manufactured per tactical planning period. Let r_i be the average profit from one wafer allocated to PM *i*. We normalize the unit production cost to zero without any loss of generality. In the highly volatile semiconductor industry, carrying inventory is undesirable. For custom-made products, a customer may stop ordering a particular version of a chip without any contractual liability. Other products may face a decline in demand or may even be phased out during the manufacturing lead-time, causing the inventory to be worthless to the manufacturer. In a striking example, CISCO Systems wrote off $2.25 billion in inventory during the economic downturn in early 2000's. Their customers suddenly cancelled their orders and the company did not see any possibility of selling that inventory.

> … On Tuesday, the Chief Financial Officer said the company [CISCO] plans to *scrap and destroy* the majority of the inventory because most of it *can't be sold because it was custom-built …*
>
> Cnet.com news report May 9, 2001

Therefore, we assume that the expected resale value of inventory at the end of the period is less than the production cost. Let v_i represent the potential loss from one unit of leftover wafer at the end of the period (production cost minus salvage value) in PM *i*'s newsvendor model. We can describe total profit function for PM *i*'s business unit, under a demand realization of ξ_i and a capacity-allocation of y_i, as follows:

$$\pi_i(y_i,\xi_i) = r_i \min(y_i,\xi_i) - v_i \max(y_i - \xi_i, 0) \qquad (1)$$

It is a common practice in the semiconductor industry that the PMs are rewarded bonuses based on the profits they realize after sales are finalized. Typically, the bonus is directly proportional to the total realized profits for the division the PM represents. To capture this basic bonus structure we assume that the bonus function implemented by HQ is a strictly increasing function of realized profits tallied at the end of the fiscal period, and that all decision makers are risk neutral unity maximizes. Thus, each PM's utility function reduces to maximizing the total expected profits of her respective business units since this will at the same time maximizes the total bonus she receives. Thus, the PM's utility is the total expected divisional profit functions as follows:

$$E[\pi(y_i,\xi_i)] = \Pi_i(y_i,\theta_i) = r_i y_i - (r_i + v_i)\int_0^{y_i} F_i(\xi_i,\theta_i)d\xi_i \qquad (2)$$

The decision problem for a PM is to maximize the total expected division profits subject to their reference capacity-allocation constraint, $y_i \le x_i$, where x_i is the reference allocation generated previous by HQ during strategic capacity planning. The optimal solution to the PM's decision problem is the newsvendor solution, as follows:

$$y_i^*(\theta_i) = \min\left\{F_i^{-1}\left(\frac{r_i}{r_i + v_i}, \theta_i\right),\ x_i\right\} \tag{3}$$

This clearly shows that a PM will want to utilize her reference allocation (her initial capacity share) as long as her newsvendor optimal solution is above her reference share.

2.2. The Integrated Solution for the Capacity-Allocation Problem

We consider a corporate environment where n PMs compete for scarce capacity and the HQ, acting as a central coordinator, wants to maximize the total expected profits for the corporation. We assume the net profits less total bonuses paid to PMs are always increasing in profits. Therefore, the HQ's problem is equivalent to finding the capacity-allocation that maximizes total expected profits across n business units. We formally define the coordination problem as follows:

Problem CA

$$\begin{aligned} \underset{\text{y}}{\text{Max}}\ z(\theta) &= \sum_{i=1}^{n} \Pi_i(y_i, \theta_i) \\ &s.t. \\ &\sum_{i=1}^{n} y_i \le b \end{aligned} \tag{4}$$

where b represents the total available capacity.

For notational convenience, variables without a subscript represent a vector, where applicable. Let $y^*(\theta)$ be the optimal capacity-allocation that solves problem CA for a given θ vector. We also define $z^*(\theta) = \sum_{i=1}^{n} \Pi_i(y_i^*, \theta_i)$ as the optimal solution value for problem CA, for a given θ vector.

We need the following assumptions to facilitate further analysis in the next section. We assume that the mean of demand distribution θ_i is independent across all PMs. We also assume that for every possible realization of the θ vector, all of the available capacity is allocated at the optimal solution (i.e. $\sum_{i=1}^{n} y_i^*(\theta_i) = b, \forall\theta$.) This assumption avoids trivial cases that do not require additional effort for coordination.

We make the following assumptions regarding the newsvendor problem of a PM: (*i*) it is concave in capacity-allocation quantity y_i , so y_i^* is unique and defined by the first order conditions; (*ii*) at any capacity-allocation $y(\theta)$, a marginal increase in θ_i increases the system-wide expected total profits. These assumptions are satisfied by most standard demand distribution functions such as Normal, Exponential and Weibull. A detailed analysis of suitable distribution functions in this regard is provided in Lariviere (1999).

With the above assumptions we have $\frac{\partial y_i^*(\theta)}{\partial \theta_i} > 0$, and $\frac{\partial y_i^*(\theta)}{\partial \theta_j} < 0, \forall i, \forall j \in -i$, where the notation –*i* indicates all indices other than *i* (e.g. $\theta_{-i} = \{\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_{n-1}, \theta_n\}$), and $\frac{\partial z^*(\theta)}{\partial \theta_i} > 0, \forall i$. The first two terms in the previous statement imply that a higher θ_i value will receive larger capacity-allocation at the expense of other divisions.

Finding the optimal allocation depends on private information from the respective PMs. Without acquiring the mean of the demand distribution information from the PMs, the HQ cannot implement the optimal capacity-allocation. On the other hand, the PMs are clearly motivated to exaggerate the demand mean because their allocation increases with the demand mean they report to the HQ. Therefore, it is not possible to implement the optimal allocation without an additional incentive structure. In the next section, we set up a capacity-allocation game, which applies a necessary incentive structure to implement the optimal capacity-allocation.

3. The Capacity-Allocation Game

The timing of events in the capacity-allocation game we want to design is as follows. At the start of a fiscal period strategic capacity-allocation decisions are made and an initial share of x_i, expressed as a percentage of the total capacity, is assigned to PMs. We assume that all available capacity is assigned initially. Operational decisions are based on reference allocation until the next tactical planning point. At that point the HQ solicits θ_i from the PMs and announces the allocation rule $y_i^*(\theta)$ and the bonus function $t_i(\theta)$ as a function of communicated θ_i values. The PMs observe their true demand (θ_i) privately and simultaneously announce a demand mean $\theta_i^{'}$. Capacity-allocation is implemented as $y_i^*(\theta^{'})$ effective immediately and bonuses are paid after total profits are realized. We assume that all of the parameters used in the payment and allocation functions are publicly known, except for the private information θ_i.

Our objective is to devise a bonus payment function in the capacity-allocation game so that the PMs reveal their true demand knowing that the company-wide optimal allocation will be implemented. Note that the bonus function $t_i(\theta)$ is above and beyond the divisional-profit-based bonus commonly used in the semiconductor industry (Section 2.1). In other words, we are proposing an additional bonus that is not tied to the divisional profit but is a function of the information reported by the PM.

The range of θ_i values that a PM can announce (the message space) is restricted by the *prior beliefs* of the other PMs and the HQ, which we assume to be shared by every participant. The prior belief function represents the information that participants have about each other's private information and is described by a distribution function. We will refer to the private information of the PMs as their type. Let $\Phi_i(\theta_i)$ be the distribution function that represents the prior beliefs of participants about the type of PM *i,* with support $\theta_i \in [\overline{\theta}_i, \underline{\theta}_i]$. Furthermore, let θ_i^* and $\theta_i^{'}$ denote the actual type of PM *i* as it is privately known to her and the announced type by PM *i,* respectively.

3.1 The surrogate profit function for PMs.

In order to relate the bonus function *t*$_i$*(.)* to the existing bonus system, we define the following surrogate profit function. Let $\overline{\pi}_i(\ )$ be the surrogate profit function for PM *i*, to which the bonus function *g(.)* is applied to determine the total bonuses to be paid. Let *t*$_i$*(.)* be the function that represents a side payment to PM *i*. Then we have:

$$\overline{\pi_i}(\theta_i^{'},\theta_{-i}^{'},\xi_i)=\pi_i(y_i(\theta_i^{'},\theta_{-i}^{'}),\xi_i(\theta_i^{*}))+t_i(\theta_i^{'},\theta_{-i}^{'}) \qquad for\ \ i=1..n\,. \tag{5}$$

The local problem of PM *i* reduces to maximizing the expected surrogate profit function for any given $\theta_{-i}^{'}$, and it is described as follows.

Problem PM

$$\underset{\theta_i^{'}}{Max}\ \ \overline{\Pi}_i(\theta_i^{'},\theta_{-i}^{'},\theta_i^{*})=\mathrm{E}_{\xi_i}[\overline{\pi_i}(\theta_i^{'},\theta_{-i}^{'},\xi_i)] \tag{6}$$

The surrogate profit function pays the PMs the profits from their own business units and an additional side payment. This motivates the PMs not only to consider a coordinated solution, which is induced by the side payment, but also to keep the profits of their business units high. This incentive structure should serve as a means of coordinating the PMs in such a way that the bonuses paid are increased (together with the expected profits) and this increase is shared by the PMs appropriately.

3.2. Desired properties of the capacity-allocation game

Truth telling is the Bayesian Nash equilibrium policy for the PMs. That is, for each participating PM, announcing her true type maximizes her surrogate profit function in expectation with regards to others' types. This is referred to as *Bayesian incentive compatibility* and expressed by:

$$E_{\theta_{-i}^{'}}[\overline{\Pi}_i(\theta_i^{*},\theta_{-i}^{'},\theta_i^{*})]\geq E_{\theta_{-i}^{'}}[\overline{\Pi}_i(\theta_i^{'},\theta_{-i}^{'},\theta_i^{*})] \qquad \forall i\ \forall\theta_i^{'}\ \forall\theta_i^{*} \tag{7}$$

where, $E_{\theta_{-i}^{'}}$ is the expectation operator with respect to the prior belief function $\Phi_{-i}()$.

We require that the incentive payment be at most based on the total realized profits, including the side payments. We call this a *budget-balanced scheme* with respect to the existing bonus system. This requirement is also needed to satisfy the assumption we made while describing problem

CA. The total expected profit after bonus payments to the PMs is maximized with $y^*(\theta)$ if the side payments constitute a *budget-balanced scheme*. Otherwise, depending on the bonus system, additional profits may be offset by the additional bonuses to be paid to the PMs. The coordination should ideally allocate the capacity optimally, and the increase in total profits should increase the bonuses that the PMs expect to receive. This corresponds to the following *budget- balance* constraint on the side payment function.

$$\sum_{i=1}^{n} t_i(\overset{'}{\theta}) \le 0 \qquad \forall \theta' \qquad (8)$$

Notice that the budget constraint implies that the total side payments may amount to a negative value, which indicates that the total bonus payments may be based on a value that is less than the total profits. This can be justified by the PMs by a significant increase in total profits from coordination that would otherwise not be possible. On an individual basis, a negative side payment could be acceptable by a PM only if the expected allocation, and hence the expected divisional profits, are larger than what could have been achieved with the reference allocation. In this case the PM should be willing to pay a participation fee for a higher expected divisional profit. This is ensured by the following property.

In order to ensure voluntary participation of the PMs, the bonus system should pay off at least as much as a PM would get if she chose not to participate. We consider this constraint in expectation with regard to the belief function of the participants. This is called the *interim individual rationality* constraint, which is ensured if the following relation holds:

$$E_{\theta'_{-i}}[\overline{\Pi}_i(\theta_i^*, \theta'_{-i}, \theta_i^*)] \ge \Pi_i(\mathrm{b}x_{\mathrm{i}}, \theta_i^*) \qquad \forall \theta_i^* \qquad (9)$$

where x_i is the initial capacity share of PM *i* as a percentage of total capacity *b*. The right-hand-side of the equation is the expected profit that PM *i* could have made if she did not participate in the capacity-allocation game and chose to stick with her initial capacity share.

Violation of individual rationality in our context implies that by participating in the coordination scheme a PM has to sacrifice her personal benefits, in terms of bonus payments, for the good of others. Even in an intra-company environment, this is highly undesirable, as a disadvantaged PM may show less effort to increase divisional profits, which then leads to decreased system-wide profits. In extreme cases, she may even seek employment elsewhere, as this effectively reduces her total financial compensation.

3.3. A Coordination Mechanism for Capacity-allocation

Consider the following side payment function as part of the PMs' surrogate profit function:

$$\begin{aligned} t_i(\theta_i',\theta_{-i}') &= \pi_{-i}(y^*_{-i}(\theta_i',\theta_{-i}'),\xi_{-i}) - C_i \\ &\quad + \frac{1}{2}[B + B_i(\theta_i') - B_{-i}(\theta_{-i}') - [\pi_i(y_i^*(\theta_i',\theta_{-i}'),\xi_i) + \pi_{-i}(y^*_{-i}(\theta_i',\theta_{-i}'),\xi_{-i})]] + \min\{0,q\} \qquad \forall i \end{aligned}$$

where, (10)

$$B = \int_{\underline{\theta}}^{\bar{\theta}} z^*(\theta)\,d\Phi(\theta)\,.$$

$$B_i(\theta_i) = \int_{\underline{\theta}_{-i}}^{\bar{\theta}_{-i}} z^*(\theta_i,\theta_{-i})\,d\Phi(\theta_{-i})$$

$$C_i = \min_{\theta_i}\{B_i(\theta_i) - \Pi(x_i,\theta_i)\}$$

$$q = \frac{\sum_{i=1}^{n} C_i - (n-1)B}{n}$$

When the side payment function described above is plugged into the surrogate profit function for PM *i*, the terms in the second line cancel each other in expectation. This results in PM *i* receiving system-wide total expected profits less a lump-sum participation charge C_i, in expectation to other PMs' types. The term *B* is the system-wide total expected profits over the possible types of all participating PMs. Similarly, $B(\theta_i)$ is *B* as a function of PM *i*'s type θ_i. The term C_i is the minimum of the expected system-wide profits less the expected divisional profits over all possible types of PM *i*. The last term, *q*, is the total participation charges less the total expected side payments divided by *n*, which is the surplus as distributed equally to PMs.

Theorem 1

The proposed side payment function

(a) implements the optimal capacity-allocation,
(b) is Bayesian Incentive Compatible,
(c) supports a budget-balanced bonus structure for problem CA and satisfies interim individual rationality if and only if $q \geq 0$.

Proof: (See Appendix for proof).

The surrogate profit function with the side payment function essentially pays the realized total profits to each PM and charges a lump-sum participation fee of C_i. The payment of the system-wide total profits aligns the incentives of the PMs with those of the corporation, hence induces truth telling as equilibrium strategy. The charges preserve this incentive structure because they are in a lump sum. C_i is the maximum that can be charged without knowing the type of PM *i* while still preserving voluntary participation. According to the definition of C_i, PM *i* expects to make at least that much over all the possible realizations of her type in expectation with respect to the types of other PMs. Any charge above C_i may violate the individual rationality for PM *i* for certain values of her type.

The expression *q* measures the difference between the total expected payments and the lump sum charges to the PMs. If *q* is positive, then the charges offset the payments, budget balance is achieved and individual rationality is ensured. However, if *q* is negative, then either budget balance is violated or individual rationality is jeopardized at the expense of achieving budget balance. In the former case, *q* represents the cost of private information in terms of additional bonuses to be paid to the PMs.

Mechanism CA is based on the existing bonus structure that is defined by the bonus function *g()*.With the proposed mechanism, the PMs are still paid based on their divisional profits too; consequently they are motivated to expend effort to close more sales deals and to increase the mean of their divisional demand.

4. Analysis of the Capacity-allocation game

We want to know whether we can identify cases in which we can be assured of the possibility of achieving budget balance and individual rationality simultaneously in the capacity-allocation game. In this section, we characterize the C_i values as a function of the initial capacity shares and the characteristics of private information about the PMs' types and draw some insights into their interaction.

Lemma 1. Let $\theta_i^{\min} = \{\theta_i \in [\bar{\theta}_i, \underline{\theta}_i] \mid \min\{B_i(\theta_i) - \Pi_i(x_i, \theta_i)\}\}$. There exists $\theta_{-i}^{"} \in [\bar{\theta}_{-i}, \underline{\theta}_{-i}]$ such that the following results hold for every PM *i*. The value of $\theta_{-i}^{"}$ depends on the belief function and the expected profit function for PM*s*.

Case	Condition	$\theta_i^{\min} =$	$C_i =$
(a) $x_i > 0$	(1) $y_{\text{i}}^*(\theta_i, \theta_{-i}^{"}) > x_{\text{i}} \quad \forall \theta_i \in [\bar{\theta}_i, \underline{\theta}_i]$	$\underline{\theta}_i$	$B_i(\underline{\theta}_i) - \Pi_i(x_i, \underline{\theta}_i)$
	(2) $y_{\text{i}}^*(\theta_i, \theta_{-i}^{"}) < x_{\text{i}} \quad \forall \theta_i \in [\bar{\theta}_i, \underline{\theta}_i]$	$\bar{\theta}_i$	$B_i(\bar{\theta}_i) - \Pi_i(x_i, \bar{\theta}_i)$
	(3) else	$\{\theta_i \mid y_{\text{i}}^*(\theta_i, \theta_{-i}^{"}) = x_{\text{i}}\}$	$\int_{\underline{\theta}_{-i}}^{\bar{\theta}_{-i}} \Pi_{-i}(y_{-i}^*(\theta_i^{\min}, \theta_{-i}), \theta_{-i}) d\Phi(\theta_{-i})$
(b) $x_i = 0$		$\underline{\theta}_i$	$B_i(\underline{\theta}_i)$

Table 4.1. Characterization of participation charges from the PMs.

First, we focus on the no-initial-share policy, case (b) in lemma 1. We have the following strong conclusion stated by the next theorem.

Theorem 2. Individual rationality and budget balance always hold with zero initial share: $x_i = 0 \ \forall i$. Moreover, $q > z^*(\underline{\theta})$.

Proof (See Appendix).

This result proves that discontinuing the initial capacity-assignment policy overcomes the inefficiency that may be caused by private information under any business environment that can be described by the model. This is an important conclusion in that it provides a tradeoff between disadvantaging the PMs by putting them in uncertainty during their dealings with customers and implementing an incentive-compatible, individually rational optimal capacity-allocation. However, in the business environment of semiconductor manufacturing discussed earlier, this may not be a viable policy.

Unfortunately, while the positive-initial-share case better captures the essence of semiconductor capacity planning, it is also more complex to analyze. The three regions for the initial capacity share, as shown in Table 1, have important practical implications. If x_i satisfies (a.1), then PM *i* will always have, in expectation with regard to the types of the other PMs, more capacity after the capacity-allocation game is played. That is, she expects to be a capacity buyer from the other PMs, because we assume that initial shares sum up to the total capacity. Similarly, if x_i satisfies (a.2), then PM *i* will be a capacity seller in expectation. Any initial capacity-allocation that leads one of the PMs to be an expected buyer or an expected seller is undesirable at the corporate level because of the delicate politics among the PMs, which we cannot capture with our game-theoretic model. Semiconductor manufacturers engage in internal capacity trading often report unfair practices influenced by favor trading, seniority, and power play. Therefore, we will focus on the initial capacity-allocations that fall into case (a.3) for all PMs. The next proposition restricts the feasible solution space for the initial allocation even more.

Lemma 2.
Assume case (a.3) in Table 1.
(a) $\theta_i^{\min} = \overline{\theta}_i,\ \forall i$, if and only if $\sum_{i=1}^{n} x_i > b$.
(b) $\theta_i^{\min} = \underline{\theta}_i,\ \forall i$, if and only if $\sum_{i=1}^{n} x_i < b$.

The possibility of achieving both individual rationality and budget balance is not conclusive in the positive-initial-share case. It depends on the belief functions and the cost structure of the local newsvendor problem as well as how the initial shares are distributed. However, if we relax the requirement that the total capacity is distributed as initial shares, then we find a compromise that ensures that all of the desirable properties hold.

Theorem 3.

Individual rationality and budget balance always hold if the initial shares are set as

$$x_{\mathrm{i}} = y_{\mathrm{i}}^{*}(\underline{\theta}_i, \theta''_{-i}) \qquad \forall \mathrm{i}, \qquad (11)$$

where, θ''_{-i} is defined as in Lemma 1. This means that x_i must satisfy the following expression:

$$-(r_i + v_i)\int_0^{x_i} F_i(\xi_i, \underline{\theta}_i)d\xi_i = \int_{\underline{\theta}_{-i}}^{\bar{\theta}_{-i}} \left(-(r_i + v_i) \int_0^{y_i^{*}(\underline{\theta}_i, \theta_{-i})} F_i(\xi_i, \underline{\theta}_i)d\xi_i \right) d\Phi(\theta_{-i}) \qquad \forall i\,. \qquad (12)$$

Proof (See Appendix).

The theorem states that if the initial shares are set in such a way that all of the PMs end up with the same capacity share (in expectation) with regards to the other PMs' type (at their lowest type), then individual rationality and budget balance holds simultaneously. However, by Lemma 2, we know that this can be achieved if and only if the initial shares are less than the total capacity. The initial shares described by Theorem 3 represent the least acceptable quantity, from the PM's perspective, as an initial capacity assignment. The unassigned capacity will be in the ownership of the HQ and it will be totally distributed with the mechanism CA. Therefore, this policy does not interfere with the politics between the PMs. Deferring the assignment of part of the capacity can be viewed as a bargaining tool for the HQ against the private information of the PMs. Our discussions with semiconductor manufacturers lead us to believe that the above scheme would be very easy to implement and makes intuitive sense to capacity planners, product managers and alike.

5. Case Study

In this section, we describe a case study constructed based on real-world scenarios observed in the semiconductor-manufacturing environment. The case is constructed to illustrate some of the key results developed in the previous sections.

We consider a semiconductor Integrated Circuit (IC) manufacturer who produce customized IC's for a variety of electronic devices. The IC manufacturer is organized by a number of business units, each serving a particular market sector. We consider product managers from two different business units competing for the capacity available from wafer fabs. The PMs' decision problems can be characterized by newsvendor models, and that their decision problems differ only in the demand means. This implies that the unit profit from one wafer and the unit cost for one unsold wafer are the same for both PMs, and that the demand distribution has the same variance at the same mean value. We observe that this assumption holds well in the custom semiconductor industry; the wafer manufacturing costs are very similar because they are manufactured at the same fab and the only difference is the setup made for different circuits that are printed on the wafer. Most companies try to maintain a stable profit rate for all fab technologies; therefore, on the average the business units have similar profit margins. At the end-product level the profit rate may be different, but at the wafer level the average is quite similar. We also observed that individual orders could be accurately represented by a Normal distribution. Appealing to the central limit theorem, we further assume that the total demand is also normally distributed.

A significant portion of the demand comes from a few major customers in the form of custom-design chips unique to their products. Once a production run is completed for a customer, that particular batch of wafers cannot be used to satisfy demands for other customers. We model this case by a normally distributed demand with constant coefficient of variation γ. Therefore, the demand distribution takes the form $F_i(\xi_i \mid \theta_i, \gamma\theta_i)$ where θ_i is the mean and γ is the coefficient of variation of the demand distribution. As the expected volume of orders increases, so does the variance of the distribution. We assume that demand mean is privately known to each PM but the coefficient of variation is publicly known in the company.

The demand distribution defined this way also satisfies the assumptions in Section 2.2. Under these assumptions the optimal allocation can be represented by a simple proportional allocation rule (Cachon and Lariviere, 1999b) defined by

$$y_i^*(\theta_i) = \min\left\{ \overline{y}_i(\theta_i),\ b\frac{\theta_i}{\theta_1 + \theta_2} \right\} \qquad i = 1,2 \qquad (13)$$

for any coefficient of variation γ. When the PMs get capacity allocations in proportion to their expected demand, the company-wide total expected profits are achieved.

Recall that the prior beliefs held by the participants about the type of PM *i* is characterized by distribution function $\Phi_i(\theta_i)$ with support $\theta_i \in [\bar{\theta}_i, \underline{\theta}_i]$. We represent the belief function by the Beta distribution, which is well suited to describe a random process in the absence of relevant data (Law and Kelton, 1991). This is appropriate since the demand means at each tactical planning point is private information owned by the PMs, which is not readily extractable from historical data. However, historical or market research data *can* be used to determine the support of the belief function distribution. An important characteristic of semiconductor demand is that during a specific period of time, it could be going through a ramp-up, mature, or ramp-down phase depending on the stage of the product lifecycle and the market conditions. We use the skewness parameter in the Beta distribution to represent the expected demand patterns. A Beta density that is skewed right assigns higher probability to higher nominal demand values within its range, thus represents anticipation for demand ramp up. Conversely, a left skewed Beta density represents anticipation for demand ramp down. A Beta density with no skewness (Uniform distribution) can describe the beliefs for mature demand pattern with no noticeable ramps. This provides a straightforward setting to describe the beliefs of the participants in the capacity-allocation game. The data values used in this case study is listed in Table 2 below.

Table 2. Data Used in the Example.

Data	**Value**	**Data**	**Value**
r (unit profit)	50	θ_1 *(PM$_1$ demand mean)*	[1000,2000]
v (unit loss)	25	θ_2 *(PM$_2$ demand mean)*	[1500,2500]
b (capacity)	2550	$\Phi_1(\theta_1), \Phi_2(\theta_2)$ *(Belief functions of PM i's type)*	Beta(1,1): uniform – mature demand expected Beta(1,3): skewed right – ramp up expected Beta(3,1): skewed left – ramp down expected
γ *(cof of var)*	0.05		

The data generated in Table 2 is captures the key relationships we have observed at a semiconductor manufacturer. We set the capacity level such that at their lowest types, both PMs get the optimal allocation that solves their local problem (i.e. $y_i^* = \overline{y}_i, \forall i$).

In this numerical example, we compute the C_1, C_2 and B values under a variety of initial capacity-allocations and belief functions. The results are reported in Tables 3-5. In the company, every PM gets at least 25% of the capacity as its initial share; therefore we covered initial allocations from 25% to 75% for both PMs. We used Maple version 6.0 to carry out the computations. In order to reduce computational requirements to match to our hardware capabilities, we discretized the beta distribution for the belief functions. We divided the support of the probability distribution into 10 equal intervals and took the middle point of the interval as the value of the random variable and the probability density of the interval as the probability. We conducted pilot runs to find the number of intervals so that the accuracy loss resulting from discretization is negligible. The C_i values are found by total enumeration over the discretized values of θ_i.

First, we look at the uniformly distributed belief function case. As seen from Table 3 below, there are no initial allocation settings where q is nonnegative except the (0.75, 0.25) case. However, initial shares from (0.6, 0.4) through (0.75, 0.25) essentially falls into case (a.2) and (a.1) for PM 1 and 2, respectively, and are therefore unacceptable. In particular, at the initial share configuration of (0.75, 0.25), PM 1 is allocated a capacity that is more than her optimal newsvendor solution ($\overline{y_i}(\theta_i)$) for most of her possible types. Such initial configurations actually defeat the purpose of assigning an initial allocation.

Another point Table 3 highlights is that q is very small compared to B. Therefore, a budget shortage actually leads to a negligible amount of extra bonuses to be paid to the PMs. Alternatively, the violation of individual rationality can be acceptably low. This situation can be explained by looking at the effect of increasing θ_i on the optimal total expected profit function $z^*()$. As θ_i increases, so does $z^*()$. However, the rate of increase drops at a high rate at high values of θ_i as described by the derivative below (Cachon and Lariviere, (1999b) and by the envelope theorem):

$$\frac{\partial z^*(\theta_i,\theta_{-i})}{\partial \theta_i} = (r+v)\frac{1}{\theta_i^2}F(\frac{b}{\theta_i+\theta_{-i}},1)\,. \qquad (14)$$

This implies that $z^*()$ shows little sensitivity to θ_i, and therefore the value of information in this example is relatively low.

Next, we look at the effects of different belief functions as shown in Table 4 below. It is clear that coordination becomes easier as a higher level of demand mean is anticipated. A higher demand means an increase in the total system profits, which leads to an increase in the surrogate profit function for the PMs. Thus, they become more willing to participate to get the benefits from increasing total profits.

Table 3.Computational results for uniformly distributed belief fn. (*B*=127457).
($^{+,\ -}$ indicates the upper and lower support for the parameter respectively).

x_1,x_2	$\theta_1^{\min},\theta_2^{\min}$	C_1, C_2	C_1+C_2	q
0.25,0.75	*1050*$^{-}$, *2450*$^{+}$	95308, 31875	127183	-274
0.30,0.70	*1050*$^{-}$, *2450*$^{+}$	88933, 38250	127183	-274
0.35,0.65	*1050*$^{-}$, 2150	82559, 44624	127184	-273
0.40,0.60	1150, 1750	76430, 50993	127424	-33
0.45,0.55	1350, *1550*$^{-}$	70123, 57122	127245	-212
0.50,0.50	1650, *1550*$^{-}$	63749, 63433	127183	-274
0.55,0.45	*1950*$^{+}$, *1550*$^{-}$	57375, 69808	127183	-274
0.60,0.40	*1950*$^{+}$, *1550*$^{-}$	51000, 76183	127183	-274
0.65,0.35	*1950*$^{+}$, *1550*$^{-}$	44627, 82558	127186	-271
0.70,0.30	*1950*$^{+}$, *1550*$^{-}$	38386, 88933	127319	-138
0.75,0.25	*1950*$^{+}$, *1550*$^{-}$	33599, 95308	128907	1450

Especially with the ramp-up, case the q value, although negative, is negligibly small. Another observation is that the initial allocation value of (0.4, 0.6) induced the highest q under all of the belief functions. This happens to be very close to the optimal allocation that is based on the expected value of belief functions: $y_i^*(E[\theta_i], E[\theta_{-i}])$. The expected-value-based optimal allocations are (0.41, 0.59), (0.42, 0.58) and (0.43, 0.57) for ramp-down, mature, and ramp-up, respectively. It seems that a good rule of thumb is to set initial shares based on the expected value of beliefs.

We also look at the initial shares that ensure individual rationality and budget balance as characterized by theorem 3. As shown in Table 5, the total allocation varies between 80- 90% of the total capacity. Similar to what we have observed before, as high demand levels are anticipated, the total that needs to be allocated decreases as the q value increases, hence the budget surplus.

Table 4. Comparison for different belief functions.

x_1,x_2	q (Ramp-down)	q (Mature)	q (Ramp-up)
0.25,0.75	-437	-274	-2
0.30,0.70	-437	-274	-2
0.35,0.65	-436	-273	-1
0.40,0.60	-17	-33	-1
0.45,0.55	-377	-212	-1
0.50,0.50	-437	-274	-2
0.55,0.45	-437	-274	-2
0.60,0.40	-437	-274	-2
0.65,0.35	-434	-271	1
0.70,0.30	-301	-138	134
0.75,0.25	1287	1450	1722

Table 5. The initial capacity-allocation that satisfies conditions of Theorem 3.

Belief	x_1,x_2	Total allocated	q
Ramp down	0.35, 0.55	0.90	11803
Uniform	0.35, 0.55	0.90	12224
Ramp up	0.30, 0.50	0.80	25486

6. Conclusion

We propose a capacity-allocation game motivated by the business environments of semiconductor manufacturing. We showed that, under the profit-only bonus structure popular in this industry, the incentives of the capacity-competing PMs are not properly aligned with the corporate interest. This is due to the private information that each PM holds regarding the demands for her business unit, and the initial capacity share that is assigned to PMs before demands realize. The latter results from a common semiconductor practice known as supply-demand planning or strategic capacity planning; we show that it leads to many unexpected consequences.

We develop a capacity-allocation mechanism with two-part tariff, where the side payment (executive bonus) is a function of the PMs' reports about their private demand information. Under the proposed allocation mechanism, the PMs are induced to reveal their true demand information knowing that they must forego their initial allocations and that the system-optimum allocations will be implemented. The bonuses are paid after the profits in all business units are realized and observed. We investigate the situations when we can attain voluntary participation of the PMs and budget-balanced bonus payments by the HQ. Our results characterize the conditions under which these properties can be attained. Our main conclusion is that reducing the total initial assignment and keeping a fraction of the capacity for competition at the time of tactical planning is sufficient to balance the inefficiency due to the PMs' asymmetric information. The example in our case study indicates that the initial assignment that supports a desirable coordination could be as high as 90% of the total capacity.

In our analysis, we assume that the belief functions are common knowledge among the PMs and the HQ. Consequently, we also require that incentive compatibility and individual rationality hold in expectation. However, our assumptions may not hold true in some cases. For example, PMs who are operating in different markets may not hold a belief function regarding the demand of the others. In order to relax the common knowledge assumption we have to impose tighter restrictions on the incentive compatibility and individual rationality. Specifically, these conditions have to hold without regard to the type of the participating PMs. In this case, the analysis will differ in that the participation charges to the PMs will have to be less in order to accommodate the relaxed assumptions. Resolving this issue is reserved for future research.

ACKNOWLEDGEMENTS

We appreciate the comments provided by two anonymous referees and the guest editors, which improve the exposition of the paper. This research is supported in part by NSF grants DMI-0121395 and DMI-0432422. The authors acknowledge the generous supports from Agere Systems, Lucent Technologies, and Semiconductor Research Corporation (SRC/ISMT grant 2004-OJ-1223).

REFERENCES

BALASUBRAMANIAN S, BHARDWAJ P, (2004), “When not all conflict is bad: Manufacturing-marketing conflict and strategic incentive design”, *Management Science*, 50 (4): 489-502.

BARAHONA, F., BERMON, S., GUNLUK, O., AND HOOD, S., (2001) "Robust Capacity Planning in Semiconductor Manufacturing", IBM report RC22196.

CACHON, G.P., AND LARIVIERE, M.A., (1999a), “An equilibrium analysis of linear, proportional and uniform allocation of scarce capacity”, *IIE Transactions*, 31, pp. 835-849.

CACHON, G.P., AND LARIVIERE, M.A., (1999b), “Capacity choice and allocation: strategic behavior and supply chain performance”, *Management Science*, Vol. 45, No. 8, pp. 1091, 1108.

CAKANYILDIRIM, M., AND ROUNDY, R. O., (2002) "Evaluation of Capacity Planning Practices for the Semiconductor Industry", *IEEE Transactions on Semiconductor Manufacturing*, Vol. 15, No. 3, 331-340.

CELIKBAS, M., SHANTHIKUMAR., J.G. AND SWAMINATHAN, J.M., (1999), “Coordinating production quantities and demand forecasts through penalty schemes”, *IIE Transactions*, 31, pp. 851-864.

CHRISTIE, R. M. E., AND WU, S. D., (2002) "Semiconductor Capacity Planning: Stochastic Modeling and Computational Studies", *IIE Transactions*, Vol. 34, No. 2, 2002.

GIBBARD, A. (1973) “Manipulation for Voting Schemes”, *Econometrica,* Vol. 41, pp. 587-601.

GREEN, J. and J.J. LAFFONT. (1977). “Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods”, *Econometrica*, Vol. 45, pp. 427-438.

CRAMTON, P., AND GIBBONS, R., AND KLEMPERER, P., (1987), “Dissolving a partnership efficiently”, *Econometrica*, Vol. 55, No. 3. pp. 615-632.

GROVES, T., AND LOEB, M., (1979), “Incentives in a divisionalized firm”, *Management Science*, Vol. 25, No. 3.

HARRIS, M., KRIEBEL, C.H., AND RAVIV, A., (1982), “Asymmetric information, incentives and intrafirm resource allocation”, *Management Science*, Vol. 25, No. 6.

KARABUK, S., AND WU, S.D., (2003), “Strategic Capacity Planning in the Semiconductor Industry: A Stochastic Programming Approach”, *Operations Research*, Vol. 51, No. 6, November-December 2003, pp. 839-849.

KARABUK, S., AND WU, S. D., (2002), “Decentralizing Semiconductor Capacity Planning Via Internal Market Coordination”, *IIE Transactions*, 34, 9, 743-759.

KOUVELIS, P., AND LARIVIERE, M.A., (2000), "Decentralizing cross functional decisions: coordination through internal markets", *Management Science*, Vol. 46, No. 8, pp. 1049-1058.

LARIVIERE, M.A., (1998), "Supply Chain Contracting and Coordination with Stochastic Demand", *Quantitative Models for Supply Chain Management*, S. Tayur, M. Magazine and R. Gaeshan, (editors) Kluwer Academic Publishers, Norwell, MA.

LAW A., AND KELTON, W.D., (1992), *Simulation Modeling and Analysis*, New York, McGraw-Hill.

MAKOWSKI, L., AND MEZZETTI, C., (1993), "The possibility of efficient mechanisms for trading an invisible object", *Journal of Economic Theory*, 59, pp. 451-465.

MAKOWSKI, L., AND MEZZETTI, C., (1994), "Bayesian and weakly robust first best mechanisms: characterizations", *Journal of Economic Theory*, 64, pp. 500-519.

MALLIK, S. AND HARKER, P. T. (1998), "Coordinating Supply Chains with Competition: Capacity-allocation in Semiconductor Manufacturing", *Working Paper,* Fishman-Davidson Center, The Wharton School, University of Pennsylvania.

MCAFEE, R.P., (1991), "Efficient Allocation with Continuous Quantities", *Journal of Economic Theory,* 53, pp.51-74.

MYERSON, R. B. (1979). "Incentive Compatibility and The Bargaining Problem", *Econometrica*, 7(1):61–73.

MYERSON, R. B. (1981). "Optimal Auction Design", *Mathematics of Operations Research*, 6(1):58–73.

MYERSON, R. B. (1982). "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems" *J. Math. Econom.*, 10(1):67–81.

MYERSON, R. AND SATTERTHWAITE, M., (1983), "Efficient mechanisms for bilateral trading", *J. Econ. Theory*, 28, pp. 265-281.

PORTEUS, E.L., AND WHANG S., (1991), "On manufacturing/marketing incentives", *Management Science*, Vol. 37, No. 9.

SWAMINATHAN, J. M., (2000), "Tool capacity planning for semiconductor fabrication facilities under demand uncertainty", *European Journal of Operational Research*, 120 (3): 545-558.

SWAMINATHAN, J. M., (2002), "Tool procurement planning for wafer fabrication facilities: a scenario-based approach", *IIE Transactions*, 34, (2): 145-155.

TRENCH, W. F., (1978), Advanced Calculus, Harper & Row Publishers: New York, Hagerstown, San Francisco, London.

WILLIAMS, S.R., (1999), "A Characterization of Efficient, Bayesian Incentive compatible mechanisms", *Economic Theory*, 14, pp. 155-180.

Wu, S.D., "Supply Chain Intermediary: A Bargaining Theoretic Framework," in *Handbook of Quantitative Supply Chain Analysis*, International Series of Operations Research and Management Science, (D. Simchi-Levi, S. D. Wu, and M. Shen, eds.), Kluwer Academic Publishers, Norwell, MA.

APPENDIX

Proof: Theorem 1.

(T1.a) By definition of $y^*(\theta)$ in section 2.2.

(T1.b) Consider the payments to the surrogate profit function under the mechanism <*y**, *t*>.

(*i*) Each PM receives the total realized profits.

(*ii*) We define the *total participation charge* in expectation with regards to demand for PM *i* as follows.

$$h_i(\theta_i^{'},\theta_{-i}^{'}) = C_i - \frac{1}{2}[B + B_i(\theta_i^{'}) - B_{-i}(\theta_{-i}^{'}) - z^*(\theta_i^{'},\theta_{-i}^{'})] - \min\{0,q\}$$

The expected participation charge with respect to the type of the other PMs is:

$$E_{\theta_{-i}}[h_i(\theta_i^{'},\theta_{-i}^{'})] = C_i - \frac{1}{2}[B + B_i(\theta_i^{'}) - B - B_i(\theta_i^{'})] - \min\{0,q\} = H_i \qquad \forall \theta_i^{'}$$

It is clear that the participation charge is *lump sum* in expectation with regards to the type of the other PMs.

From (*i*) and (*ii*), <*y**, *t*> is a Groves mechanism in expectation and by the equivalence theorem, <*y**, *t*> is *Bayesian incentive compatible* (Makowski and Mezzetti 1994).

(T1.c) The mechanism pays the realized profits to the respective business units plus the extra payments defined by *t*. We show that the total extra payments sum up to be less than or equal to zero.

$$\sum_{i=1}^{n} t_i(\theta^{'}) = B - \sum_{i=1}^{n} C_i + \min\{0,\ \sum_{i=1}^{n} C_i - B\}$$

Consider the right hand side of the equation above. If total expected pay, *B*, is larger than the total expected charge, (C_1+C_2+…C_n), then the third term on the right hand side deducts the difference from the total payments to the surrogate function and makes the equation evaluate to zero. On the other hand, if there is a surplus in the expected payments, then total payments evaluate to a negative value without any adjustment and the surplus is captured by the mechanism-designer in terms of less bonus payment.

(T1.d) By participating in the capacity-allocation game, PM *i* is paid $B_i(\theta_i)$ in expectation with regards to the type of the other PMs. On the other hand, she gives up what her initial share would pay her, which is $\Pi_i(x_i,\theta_i)$. The PMs are expected utility maximizers, therefore the payments are considered as expected values with regards to demand. Consequently, without knowing PM *i*'s type, the HQ can at most charge a participation fee of C_i. The PM *i* expects to make at least C_i or more compared to not participating, at any realization of her type θ_i. Consider the lump sum participation charge H_i defined in proof of (1.b). H_i equals C_i if and only if $q \geq 0$, otherwise $H_i > C_i$ and the interim individual rationality of PM *i* is violated. □

Proof: Lemma 1.

(L1.a) Consider $x_i > 0$.

(*i*) By the envelope theorem we have for every *i*,

$$\frac{\partial C_i(\theta_i)}{\partial \theta_i} = \int_{\underline{\theta}_{-i}}^{\overline{\theta}_{-i}} \left(-(r_i + v_i) \int_{0}^{y_i^*(\theta_i,\theta_{-i})} \frac{\partial F_i(\xi_i,\theta_i)}{\partial \theta_i} d\xi_i + (r_i + v_i) \int_{0}^{x_i} \frac{\partial F_i(\xi_i,\theta_i)}{\partial \theta_i} d\xi_i \right) d\Phi(\theta_{-i}) = \Delta_i .$$

By the assumptions we made regarding the demand distribution, it is clear that Δ_i increases as θ_i increases.

(*ii*) By the first mean value theorem for integrals, there exists $\theta''_{-i} \in [\underline{\theta}_{-i}, \overline{\theta}_{-i}]$ such that $B_i(\theta_i) = z^*(\theta_i, \theta''_{-i})(\overline{\theta}_{-i} - \underline{\theta}_{-i})$ (e.g. see Trench 1978). Therefore, we can rewrite Δ_i as,

$$\Delta_i = \left(-(r_i + v_i) \int_0^{y_i^*(\theta_i, \theta''_{-i})} \frac{\partial F_i(\xi_i, \theta_i)}{\partial \theta_i} d\xi_i + (r_i + v_i) \int_0^{x_i} \frac{\partial F_i(\xi_i, \theta_i)}{\partial \theta_i} d\xi_i \right) (\overline{\theta}_{-i} - \underline{\theta}_{-i})$$

According to the mean value theorem θ''_{-i} may change value with respect to θ_i. However, we know from (*i*) that Δ_i is an increasing function of θ_i. Therefore, Δ_i increases regardless of the value of θ''_{-i}. The term Δ_i is negative as long as $y_i^*(\theta_i, \theta''_{-i}) < x_i$, hence C_i is decreasing up to the θ_i value until $y_i^*(\theta_i, \theta''_{-i}) = x_i$ where it evaluates to zero. Beyond that point, as θ_i is increased $y_i^*(\theta_i, \theta''_{-i}) > x_i$ is satisfied. Consequently, Δ_i always takes a positive value and C_i increases.

Derivation of C_i values for cases (a.1) and (a.2) is straightforward. For case (a.3): by the mean value theorem for integrals we have,

$$C_i = \Pi_i(y_i^*(\theta_i^{\min}, \theta''_{-i}), \theta_i^{\min}) + \sum_{j \in -i} \Pi_j(y_j^*(\theta_i^{\min}, \theta''_{-i}), \theta''_j) - \Pi_i(x_i, \theta_i^{\min}).$$

Since $y_i^*(\theta_i^{\min}, \theta''_{-i}) = x_i$, it follows that

$$C_i = \sum_{j \in -i} \Pi_j(y_j^*(\theta_i^{\min}, \theta''_{-i}), \theta''_j) = \int_{\underline{\theta}_{-i}}^{\overline{\theta}_{-i}} \left(\sum_{j \in -i} \Pi_j(y_j^*(\theta_i^{\min}, \theta_{-i}), \theta_j) \right) d\Phi(\theta_{-i}).$$

(L1.b) Consider $x_i = 0$.

This is a special case of the previous case where the optimal allocation is always greater than initial share, which is zero. Therefore $\theta_i^{\min} = \underline{\theta}_i$ and the second term in C_i evaluates to zero. □

Proof: Theorem 2.

Let $\overline{C}(\theta) = \sum_{i=1}^{n} z^*(\underline{\theta}_i, \theta_{-i}))$, $\overline{B}(\theta) = z^*(\theta)$. From Lemma 1, Theorem 1 and independence of types we have,

$$\sum_{i=1}^{n} C_i = \int_{\underline{\theta}}^{\overline{\theta}} \overline{C}(\theta) d\Phi(\theta) \text{ and } B = \int_{\underline{\theta}}^{\overline{\theta}} \overline{B}(\theta) d\Phi(\theta)$$

If we can show that $\overline{C}(\theta) \geq \overline{B}(\theta)\ \forall \theta$, then this implies that $\sum_{i=1}^{n} C_i \geq (n-1)B$.

(*i*) At the lowest values of types $\underline{\theta}$ we have,

$\overline{C}(\underline{\theta}) = nz^*(\underline{\theta}) \geq \overline{B}(\theta) = (n-1)z^*(\underline{\theta})$.

(*ii*) At any $(\theta_1, \theta_2, ... \theta_n)$

$$\frac{\partial \overline{C}(\theta)}{\partial \theta_i} = \sum_{j \in -i} \left(-(r_i + v_i) \int_0^{y_i^*(\theta_1, \theta_2, ... \theta_{j-1}, \underline{\theta}_j \theta_{j+1}, ... \theta_n)} \frac{\partial F_i(\xi_i, \theta_i)}{\partial \theta_i} d\xi_i \right)$$, and

$$\frac{\partial \overline{B}(\theta)}{\partial \theta_i} = -(n-1)(r_i + v_i) \int_0^{y_i^*(\theta_1, \theta_2, ... \theta_n)} \frac{\partial F_i(\xi_i, \theta_i)}{\partial \theta_i} d\xi_i \ .$$

By the assumptions we made regarding the demand distribution function, $y_i^*(\theta_1, \theta_2, ... \theta_{j-1}, \underline{\theta}_j, \theta_{j+1}, ..., \theta_n) > y_i^*(\theta_1, \theta_2, ... \theta_n) \quad \forall i, j$. Therefore, we can conclude that at any $(\theta_1, \theta_2, ... \theta_n)$, $\overline{C}(\theta)$ increases at a higher rate than $\overline{B}(\theta)$. With (*i*), this completes the proof. □

Proof: Lemma 2

(a) *if part*: assume $\theta_i^{\min} = \overline{\theta}_i,\ \forall i$;

By Lemma 1(a)(3) $y_i^*(\overline{\theta}_i, \theta_{-i}^{''}) = x_i,\ \forall i$. Also by assumption A2, we have: $\sum_{i=1}^{n} y_i^*(\overline{\theta}_i, \overline{\theta}_{-i}) = b$.

Since $\Pi_i() > 0,\ \forall i,\ \theta_j^{''} < \overline{\theta}_j,\ j \in \text{-}i$. With assumption A3, this implies that

$y_i^*(\overline{\theta}_i, \theta_{-i}^{''}) > y_i^*(\overline{\theta}_i, \overline{\theta}_{-i}),\ \forall i$. Therefore, $\sum_{i=1}^{n} y_i^*(\overline{\theta}_i, \theta_{-i}^{''}) > b$.

only if part: if $\sum_{i=1}^{n} x_i > b$, then $\sum_{i=1}^{n} y_i^*(\overline{\theta}_i, \theta_{-i}^{''}) > b$. Hence $\theta_i^{\min} = \overline{\theta}_i,\ \forall i$.

(b) The argument is same as (a) for the stated types. □

Proof: Theorem 3.

By Theorem 1, Lemma 1 and independence of types we have,

$$\sum_{i=1}^{n} C_i = \int_{\underline{\theta}}^{\overline{\theta}} \left(\sum_{i=1}^{n} \left(\sum_{j\in -i} \Pi_j (y_j^*(\underline{\theta}_i, \theta_{-i}), \theta_j) \right) \right) d\Phi(\theta)$$

$$(n-1)B = \int_{\underline{\theta}}^{\overline{\theta}} \left(\sum_{j=1}^{n} (n-1)\Pi_j (y_j^*(\theta), \theta_j) \right) d\Phi(\theta) .$$

By the assumptions we made regarding the demand distribution function, $y_j^*(\theta_1,\theta_2,...\theta_{i-1},\underline{\theta}_i,\theta_{i+1},...,\theta_n) > y_j^*(\theta_1,\theta_2,...\theta_n)$ $\forall i, j$, therefore, at any $(\theta_1,\theta_2,...\theta_n)$ the expression inside the integral of $\sum_{i=1}^{n} C_i$ is greater than the expression inside the integral of (n-1)*B*.

□