International Journal of Production Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tprs20

Evaluating capacity pooling strategy in semiconductor manufacturing: a productivity perspective study

Wen-Chih Chen [a] & Chen-Fu Chien [b]

[a] Department of Industrial Engineering and Management , National Chiao Tung University , Taiwan

[b] Department of Industrial Engineering and Engineering Management , National Tsing Hua University , Taiwan

Published online: 28 Jun 2010.

To cite this article: Wen-Chih Chen & Chen-Fu Chien (2011) Evaluating capacity pooling strategy in semiconductor manufacturing: a productivity perspective study, International Journal of Production Research, 49:12, 3635-3652, DOI: 10.1080/00207543.2010.492799

To link to this article: http://dx.doi.org/10.1080/00207543.2010.492799

International Journal of Production Research
Vol. 49, No. 12, 15 June 2011, 3635–3652

Taylor & Francis
Taylor & Francis Group

Evaluating capacity pooling strategy in semiconductor manufacturing: a productivity perspective study

Wen-Chih Chen[a]* and Chen-Fu Chien[b]

[a]*Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan;* [b]*Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan*

(*Received 30 September 2009; final version received 12 April 2010*)

Recently, increasing attention has been focusing on the concept of the borderless fab, which expands capacity through a manufacturing strategy rather than capital investments. In a borderless fab, the capacity of several wafer fabs is pooled, and partially completed wafers are allowed to move from one fab to another. This paper proposes a model to evaluate the potential benefits of adopting capacity pooling from the macro-viewpoint. We demonstrate our model using actual full-scale fab-level operational data, and the result reveals that capacity pooling can improve monthly capacity by 3% on average.

Keywords: manufacturing strategy; capacity pooling; borderless fab; performance evaluation; semiconductor manufacturing

1. Introduction

The global semiconductor industry is highly competitive. Semiconductor companies compete on the basis of reductions in the cost per transistor through a combination of advances in technology nodes as well as efficiency and scale in manufacturing. Continual technological innovation results in shorter product life-cycles and requires firms to respond rapidly to fluctuating demand. Thus, decision-making to meet market requirements is often challenging. Capacity is an essential factor to strengthen a firm's competitive edge and ensure long-term success. Larger capacity provides more services/products within a shorter timeframe to satisfy faster delivery for products with higher variability and complexity, and may also imply lower marginal costs due to economies of scale. However, capacity expansion that depends on investing in new fabrication facilities (fabs) and/or tools is astonishingly expensive, and the long lead times tend to produce great uncertainty. Fine-tuning individual fabs to increase their productivity is one alternative, but may prove difficult for mature processes and products.

Recently, increasing attention has been focusing on the concept of a borderless fab, in which the capacity of several wafer fabs is pooled, and partially completed wafers are allowed to move between fabs (e.g., Gan *et al*. 2007). The borderless fab is achievable when a semiconductor manufacturer has fabs located within reasonable proximity. This requirement is practically feasible in the semiconductor industry in general, which is capital-intensive, with high barriers to entry and a few key manufacturers with many

*Corresponding author. Email: wenchih@faculty.nctu.edu.tw

ISSN 0020–7543 print/ISSN 1366–588X online

DOI: 10.1080/00207543.2010.492799
http://www.informaworld.com

production facilities. The industry's clustering effect, seen, for example, in Taiwan's Hsinchu Science and Industrial Park, provides an attractive environment for implementing the borderless fab concept.

Capacity pooling, in particular inventory pooling, has been studied in the context of operations management. For example, Corbett and Rajaram (2006) document the pooling benefits under different product dependence. Benjaafar *et al*. (2005) present a general analytical model, in which production and inventory systems are modelled independently, to investigate issues of pooling in production-inventory systems. The results yield several useful managerial insights about the benefits of pooling. In semiconductor manufacturing, intra-fab tool grouping to share capacities, particularly as a backup for unexpected machine breakdown, has been well addressed (e.g., Chien and Hsu 2006, Chien *et al*. 2007). Recently, there has been another line of research addressing the inter-fab capacity pooling and sharing (e.g, Gan *et al*. 2007, Wu *et al*. 2009).

Previous research has noted the quantitative benefits of capacity pooling in the semiconductor manufacturing industry. For example, Gan *et al*. (2007) build a discrete event simulation model to evaluate borderless fab performance, and report that the aggregated cycle time of pooling fabs can be reduced in different operational settings. Other authors have attempted to resolve scheduling and/or routing problems in semiconductor manufacturing under the capacity pooling scenario (e.g., Wu *et al*. 2009), i.e., they study the operational optimisation given the adoption of capacity pooling. Small-scale cases with detailed operational assumptions are used to demonstrate the effectiveness of proposed algorithms, and the results show that the capacity pooling scenario leads to shortened time cycles and higher throughput.

This paper proposes a model to analyse the potential gains from adopting the borderless fab concept especially with respect to productivity. Clearly, whether or not to employ a manufacturing strategy such as the borderless fab is a key decision. We believe that the model described in this paper provides outstanding quantitative evaluation to assist strategic decision-making. Unlike the simulation studies (e.g., Gan *et al*. 2007) that specify highly detailed production environments and demand intensive engineering, technical backgrounds and advanced computational power, the proposed model is based on non-parametric productivity analysis approaches, also known as data envelopment analysis (DEA) (Charnes *et al*. 1978, Banker *et al*. 1984). This type of approach takes a macro-viewpoint of large and complex wafer fabrication processes, and is more consistent with the nature of strategic decision-making. We can simulate the possible 'behaviour' of a fab operating under a specified strategy and summarise its performance according to a pre-specified objective. The prediction is based on ideal, not average, performance. Since ideal performance resembles capacity-related decision-making, it can reveal the optimum potential benefits of a firm's candidate strategies.

In addition, we use actual full-scale fab-level data representing the overall operations of a firm to demonstrate our approach. The results of our analysis show that about 50% of cases have more than 4% improvement on average when capacity pooling is adopted. The results are far more meaningful than studies that rely on small-scale data. Emanating from the characteristics of the semiconductor manufacturing processes we study, we propose an analytical model to represent aggregate product design specifications, which is a significant contribution to the productivity analysis and DEA methodology literature.

It should be pointed out that, although motivated by a real need in semiconductor manufacturing with a specific purpose, we approach the problem and present our model in a general symbolic form; the model is not case-dependent. There is enough flexibility to

extend the approach to other applications and industries. For example, TFT-LCD (thin film transistor and liquid crystal display) and printed circuit board (PCB) manufacturing share similar characteristics, including re-entrant process and multi-site production (see, e.g., Rau *et al.* 2005), and have the same problems. It can also apply as an aid to outsourcing decisions, such as 'should a firm outsource a portion of its final products or a portion of intermediate products (processes)?' The former option corresponds to no capacity pooling while the latter is an extension of capacity pooling.

The remainder of this paper is organised as follows. Section 2 introduces notions of conventional non-parametric productivity analysis as the technical background for the study. Section 3 proposes a model to analyse general design-related activities. Section 4 addresses the fabrication processes, and models the fabrication capability using the models addressed in Sections 2 and 3. Section 5 evaluates the potential benefits of the capacity pooling strategy by comparing the results of two output-maximisation models. Section 6 shows the empirical results based on a real case in Taiwan. Section 7 concludes.

2. Methodology background – production technology

Our model relies on non-parametric productivity analysis methods to 'simulate' the feasible behaviour based on observed data from the macro-viewpoint. We will first introduce the technical background of productivity analysis in this section, and propose a new model to analyse design specifications. The topics presented are not tied directly to semiconductor manufacturing; the models proposed can be easily adapted to other industries.

Conventional productivity analyses utilise technology to describe a process transforming a set of inputs, I, into a set of outputs, J. Technology is represented by its production possibility set:

$$T \equiv \left\{(x, y) \in \Re_+^{|I|} \times \Re_+^{|J|} : x \text{ can produce } y\right\},$$

where $x \in \Re_+^{|I|}$ is the input vector and $y \in \Re_+^{|J|}$ is the output vector. The output level set $P(x)$ is the collection of feasible outputs produced by given input vector x, namely $P(x) = \{y : (x, y) \in T\}$. Similarly, the input level set given output y is $L(y) = \{x : (x, y) \in T\}$. Some basic axioms are imposed on the technology as its underlying characteristics:

(A1) strong disposability of inputs and outputs: $(x, y) \in T$ if $(x', y') \in T$, $x \geq x'$ and/or $y \leq y'$.

(A2) returns to scale:

(1) constant returns to scale (crs): $(x, y) \in T$ implies $\alpha(x, y) \in T$ for $\forall \alpha \in [0, \infty)$.
(2) non-increasing returns to scale (nirs): $(x, y) \in T$ implies $\alpha(x, y) \in T$ for $\forall \alpha \in [0, 1]$.
(3) variable returns to scale (vrs): $(x, y) \in T$ implies $\alpha(x, y) \in T$ for $\forall \alpha \in \{1\}$.

(A3) convexity: $(x, y) \in T$ and $(x', y') \in T$, then $\alpha(x, y) + (1 - \alpha)(x', y') \in T$, $\forall \alpha \in [0, 1]$.

(A1) states that using more resources to produce fewer outputs than an achieved observation is always doable. (A2) is related to the feasibility due to scaling up or down of (x, y). (A3) utilises the straightforward engineering interpolation as an approximation.

In reality, the underlying technology T is unknown but can be estimated based on a set of observations R and the belief of axioms. Adopting (A1) and (A2), the estimated production possibility set with respect to a single observation (x_r, y_r), $r \in R$, is:

$$T_r(rs) = \left\{(x, y) : x \geq \alpha x_r, 0 \leq y \leq \alpha y_r, \alpha \in \Gamma(rs)\right\}$$

where:

$$\Gamma(crs) = [0, \infty);$$
$$\Gamma(nirs) = [0, 1];$$
$$\Gamma(vrs) = \{1\}.$$

$\Gamma(rs)$ characterises the returns to scale, and inequalities represents the strong disposability. Considering all data in R and following (A1) and (A2), the estimated production technology is as follows, according to the presence of adopting convexity (A3):

$$T(rs, \psi) = \left\{(x, y) : x \geq \sum_{r \in R} \alpha_r x_r \lambda_r;\ y \leq \sum_{r \in R} \alpha_r y_r \lambda_r;\ \alpha_r \in \Gamma(rs);\ \lambda_r \in \Psi(\psi), r \in R\right\}$$

where:

$$\Psi(nc) = \left\{\lambda_r \in \Re_+^{|R|} :\ \sum_{r \in R} \lambda_r = 1;\ \lambda_r \in \{0, 1\}, r \in R\right\}$$
$$\Psi(c) = \left\{\lambda_r \in \Re_+^{|R|} :\ \sum_{r \in R} \lambda_r = 1;\ \lambda_r \geq 0, r \in R\right\}.$$

$\Psi(\psi)$ indicates whether (A3) convexity is imposed; $\psi = c$ is a case adopting convexity, and $\psi = nc$ does not assume convexity.

Consequently, imposing assumptions of (A1) to (A3) and their combinations characterises different production technologies. For example, it is clear that allowing $\tau_r = \alpha_r \lambda_r$, $T(crs, c)$ can be rewritten as:

$$T(crs, c) = \left\{(x, y) : x \geq \sum_{r \in R} x_r \tau_r;\ y \leq \sum_{r \in R} y_r \tau_r;\ \tau_r \geq 0, r \in R\right\}.$$

This is the standard expression of *crs* technology estimated by convexity. Following the same procedure:

$$T(vrs, c) = \left\{(x, y) : x \geq \sum_{r \in R} x_r \tau_r;\ y \leq \sum_{r \in R} y_r \tau_r;\ \sum_{r \in R} \tau_r = 1;\ \tau_r \geq 0, r \in R\right\},$$

is the expression of standard *vrs* technology. Following only (A1), *vrs* and non-convexity leads to standard free disposal hull (FDH) technology:

$$T(vrs, nc) = \left\{(x, y) : x \geq \sum_{r \in R} x_r \tau_r;\ y \leq \sum_{r \in R} y_r \tau_r;\ \sum_{r \in R} \tau_r = 1;\ \tau_r \in \{0, 1\}, r \in R\right\}.$$

Furthermore, we note that the selection of production technologies addressed above is case-dependent and also based on making certain assumptions. The careful use of assumptions will yield better approximation of underlying technology[1].

3. Design technology

Motivated by real need, we propose a model to analyse input-output transformation processes related to product design specifications. Consider another process also transforming inputs $x \in \Re_+^{|I|}$ to products $y \in \Re_+^{|J|}$ and denote it as D, i.e.:

$$D \equiv \left\{(x, y) \in \Re_+^{|I|} \times \Re_+^{|J|} : x \text{ can produce } y\right\}.$$

Unlike production technology T addressed above, any input-output bundle $(x, y) \in D$ is designed, not manufactured. For example, (x, y) may represent overall input requirements based on bill of material (BOM) for a family of products sharing common parts, or a recipe for a wafer type. Hereafter, D is termed as the design technology.

We refer to the fundamental component of the design as the atom, which is the basic element and cannot be decomposed. For example, an atom could be the BOM list for a single product, or a recipe for a specific wafer product. Leontief (1953) pioneered this concept regarding a special case with only a single output; paraphrasing him, let us suppose that an input rate vector $\pi \in \Re_+^{|I|}$ to produce a single unit output $y \in \Re_+$ is given. The maximum output produced by the quantity vector of input $q \in \Re_+^{|I|}$ can be determined by:

$$\max_{\alpha} \left\{\alpha \in \Re_+ : \alpha\pi \leq q\right\}.$$

The information $(\pi, y) \in \Re_+^{|I|} \times \Re_+$ is an atom.

In this paper, D, collecting feasible product specifications, represents not only a collection of atoms, but can be interpreted as the results of both atoms and the scale (quantity) of the products, e.g., the resource-product requirements for any customer order. Suppose atoms A and B represent two product specifications: the component requirements with respect to an order of 1000 units of A and 500 units of B is a feasible product specification and is collected in D. The fundamental axiom for D is the additivity property:

(A4) additivity: $(x, y) \in D$ and $(x', y') \in D$ implies $(x, y) + (x', y') \in D$.

The additivity property generalises D from a set of different atoms to product requirements, and aggregates product requirements from the design aspect. Consider $(x, y) \in D$; we have $(x, y) + (x, y) = 2(x, y) \in D$ according to (A4). Thus, any quantity as the multiplier of an atom is also in set D. Extending this idea, design technology D can be estimated providing a single observed record $(x_r, y_r) \in D$ as:

$$D_r(Z_+) = \left\{(x, y) : x = \alpha x_r; 0 \leq y = \alpha y_r; \alpha \in Z_+\right\},$$

where α are non-negative integers, and $\alpha = 0$ leads to $(\mathbf{0}, \mathbf{0})$ and implies that nothing can provide nothing. In addition, $\alpha = 1$ reveals the transformation (the design) consisting of one unit of (x_r, y_r), and $\alpha > 1$ suggests that there are α times of (x_r, y_r). Since $D_r(Z_+)$ is derived from (A4), the only possibility of scaling down ($\alpha < 1$) is $(\mathbf{0}, \mathbf{0})$.

Now consider a set of records, each with identical input-output structure, that are observed and thus are feasible. These records are based on experienced units (batches)

according to a specific process design BOM or recipes, but not necessarily the basic presence of an atom. For example, (x_r, y_r), $r \in R$ can be the results of a 1000 units of a particular product. Extending $D_r(Z_+)$, the estimated design technology based on a data set R is:

$$D(Z_+) = \cup_{r\in R} D_r(Z_+)$$
$$= \left\{(x,y) : x = \sum_{r\in R} \alpha_r x_r \lambda_r;\ y = \sum_{r\in R} \alpha_r y_r \lambda_r;\ \alpha_r \in Z_+, \lambda_r \in \{0,1\}, r \in R\right\}, \quad (1)$$

where λ_r indicates whether or not record r is added as part of the estimation. In fact, (1) assumes that no decomposition of observed records is allowed, but is simply a combination of $D_r(Z_+)$. Let $\tau_r = \alpha_r \lambda_r$, (1) can be rewritten as:

$$D(Z_+) = \left\{(x,y) : x = \sum_{r\in R} x_r \tau_r;\ y = \sum_{r\in R} y_r \tau_r;\ \tau_r \in Z_+, r \in R\right\}. \quad (2)$$

$D_r(Z_+)$ and $D(Z_+)$ are conservative estimations not allowing the decomposition of (x_r, y_r), $r \in R$. This assumption typically applies to cases when $(x_r,\ y_r)$ is an atom, or represents a basic batch of various products. For cases in which all observed records are aggregate data and significantly larger than an atom, assuming fractional decomposition is reasonable for practical simplicity. Therefore, we can further assume the continuity of input-output data; the estimated technology based on a single record $r \in R$ can be expressed as:

$$D_r(crs) = \left\{(x,y) : x = \alpha x_r;\ 0 \le y = \alpha y_r;\ \alpha \in \Gamma(crs)\right\}.$$

In fact, constant returns to scale is the variety of (A4) assuming continuity. Similar to the derivation of (2), the estimation of D allowing continuity based on entire R is:

$$D(crs) = \cup_{r\in R} D_r(crs)$$
$$= \left\{(x,y) : x = \sum_{r\in R} \alpha_r x_r \lambda_r;\ y = \sum_{r\in R} \alpha_r y_r \lambda_r;\ \alpha_r \in \Gamma(crs), \lambda_r \in \{0,1\}, r \in R\right\}$$
$$= \left\{(x,y) : x = \sum_{r\in R} x_r \tau_r;\ y = \sum_{r\in R} y_r \tau_r;\ \tau_r \in \Re_+, r \in R\right\}.$$

It should be noted that if a single product specification record, say r, is given, $D_r(Z_+)$ or $D_r(crs)$ is used as the approximation. Set R can be interpreted as historical data, or, most important in practice, as a set of product specifications; for example, each record $r \in R$ may represent a single product (output) and its necessary components or processes (inputs). Therefore, $D(Z_+)$ and $D(crs)$ are used as the overall possible aggregated input-output bundles. D represents possible input-output bundles, which are the results of design, not production; there is no tolerance or waste. Thus $D(Z_+)$ and $D(crs)$ do not apply (A1); there are only equalities in $D(Z_+)$ and $D(crs)$. Unlike T, design technologies only exist crs if returns to scale is assumed; other types of returns to scale are impossible.

4. Wafer fabrication process

This section presents the wafer fabrication process as a two-stage process and models it using the technologies addressed in Sections 2 and 3. The proposed model describes and predicts the capability of a fabrication process according to its past experience.

4.1 *Two-stage wafer fabrication process*

Wafer fabrication transforms resources (inputs), e.g., labour, equipment, etc., into outputs. Various types of wafers are the physical final outputs produced and delivered by a fab. Wafer fabrication from resources to final delivered products can be described as a two-stage transformation process as shown in Figure 1 (Chen and Chien 2009).

The first stage (Stage 1) is a process providing various types of masking layer (outputs) by consuming resources. For clarity, we can interpret the output of Stage 1, layers, as a service provided rather than a physical product. This is a standard production process and is related to productivity.

The second stage (Stage 2) shows the process of transforming layers, the outputs of Stage 1, to wafers. Different types of wafer products require different masking layer processes. Design for manufacturability (DFM) is an important new engineering concept, including a set of methodologies, of designing products in a manner that simplifies manufacture. The number of layers is the number of processes required in wafer manufacturing. A larger number of wafer outputs with fewer layers are always preferred since it means fewer efforts in manufacturing. Manufacturability results from engineering supports, particularly in R&D activities of design, or manufacturing, such as the manufacturing recipe. It is also affected by business strategies, i.e., allocating products to each fab. Obviously, products with fewer layers per wafer are easier to produce; therefore, the fab assigned to such products will likely achieve superior performance. In general, manufacturability is closely related to the product specifications.

Stage 1 is a typical input-output transformation in the productivity literature, and thus we can use production technology $T(rs, \psi)$ based on observed records as an estimation to characterise the process. As described above Stage 2 concerning manufacturability means that the related input-output bundles are designed, not produced, and thus inefficiency or

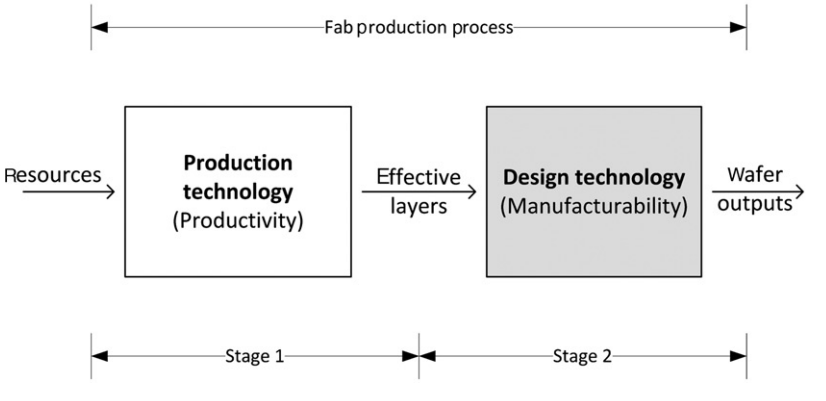

Figure 1. The fabrication process.

mix changes are impossible. Design technology D can be used to characterise this process. The outputs of Stage 1 (layers) will be transformed to the final products – wafers – via design technology.

4.2 *Technology for wafer fabrication*

To formalise the concept, we consider a two-stage process with intermediate products (e.g., layers), which are the outputs of Stage 1 and the inputs of Stage 2. Let I be a set of inputs (resources), J a set of intermediate outputs (e.g., layers in the fabrication) and O the final outputs (e.g., wafers). The two-stage process thus utilises inputs $x \in \Re_+^{|I|}$ to produce intermediate outputs $y \in \Re_+^{|J|}$ in Stage 1 and transforms y to the final outputs $z \in \Re_+^{|O|}$ in Stage 2. As addressed in Section 4.1, Stage 1 technology of the wafer fabrication process is production-related and can be described using T. Stage 2 technology is manufacturability (product specification), which is related to design technology, and can be described using D. In a two-stage process, intermediate products y link Stages 1 and 2, and the overall process can be expressed by integrating Stage 1 and Stage 2 as:

$$F = \left\{(x, y, z) \in \Re_+^{|I|} \times \Re_+^{|J|} \times \Re_+^{|O|} : (x, y) \in T, (y, z) \in D\right\}.$$

It should be noted that F indeed assumes that every y produced will be completely transformed to z without loss. However, F can be easily modified to relax the assumption by adding dummy intermediate products, $y' \in \Re_+^{|J|}$, and then we have $(x, y) \in T$, $(y', z) \in D$ and $y \geq y'$.

Consequently, F defines the underlying capability of a wafer fabrication process; namely, it determines the feasible results of consuming resources to provide final products. Suppose R is the observed record set, and each record $r \in R$ consists of $(x_r, y_r, z_r) \in \Re_+^{|I|} \times \Re_+^{|J|} \times \Re_+^{|O|}$. Typically, each record r, a historical record for a specific time period r, tracks how resources (x_r) are used to produce intermediate products (y_r) and subsequently converted into final products (z_r). It should be noted that (x_r, y_r, z_r) is the aggregated information in period r, and not for a single product. F can be estimated in practice through the estimations of T and D corresponding to R.

5. Strategy evaluation

We evaluate the potential benefits of the capacity pooling strategy by comparing the results of two output-maximisation models presented in this section. Both models reveal maximum total wafer outputs delivered from the viewpoint of management: one adopts the borderless fab strategy while the other uses the current strategy. Both models utilise the estimated F proposed in Section 4.2 to describe the production and design technology of individual fabs and to represent the input-output transformation capability, i.e., the feasible region of the transformation. Together with a given objective function, we simulate the best possible results for both settings. That is, we determine and compare the ideal capacity for two strategies, and the difference in the results will reveal the potential benefits of adopting the new strategy.

5.1 *Models*

Consider that management wishes to maximise wafer outputs for a particular wafer recipe or a summary of certain recipes (y_0, z_0). (y_0, z_0) represents the given layer-wafer bundle corporate-wise, where z_0 is the quantity vector of different wafer outputs, and y_0 are the necessary number of layer processes associated with z_0. For example, (y_0, z_0) can be the monthly summary of various product recipes, or a recipe for one wafer type. Further, assume that S collects different fabs and the available resources at fab $s \in S$ are x_0^s. Hereafter the superscripts represent the index of the production site. The problem corresponding to the proposed output-maximising objective can be restated as: management tries to maximise wafer outputs with the required recipe configuration (y_0, z_0) using resources x_0^s at fab s. The corporate decision model corresponding to the capacity non-pooling strategy is:

$$\max_{y^s, z^s, \lambda_r^s, \theta^s} \sum_{s \in S} \theta^s \tag{B-O}$$

subject to for each fab $s \in S$:
production technology:

$$x_0^s \geq \sum_{r \in R^s} x_r^s \lambda_r^s \tag{3}$$

$$y^s \leq \sum_{r \in R^s} y_r^s \lambda_r^s \tag{4}$$

$$\sum_{r \in R^s} \lambda_r^s = 1, \quad \lambda_r^s \geq 0, \quad r \in R^s; \tag{5}$$

design technology:

$$y^s = \theta^s y_0 \tag{6}$$

$$z^s = \theta^s z_0 \tag{7}$$

$$y^s \geq 0.$$

R^s is the collection of wafer production data $(x_r^s, y_r^s, z_r^s) \in \Re_+^{|I|} \times \Re_+^{|J|} \times \Re_+^{|O|}$, $r \in R^s$, at fab s. Each (x_r^s, y_r^s, z_r^s) has identical interpretations to those addressed in Section 4.2, but now represents fab s. Using the observed data, constraints (3) to (5) specify the production technology of fab s, determining feasible y^s using available resources x_0^s. Constraints (6) and (7) characterise design technology, i.e., the product specifications, for fab s. Volume of wafer subject to the product recipe (y_0, z_0) is to be maximised as $\theta^s(y_0, z_0)$ for each fab s. To achieve this goal, fab s should provide $y^s = \theta^s y_0$ as (6). The optimal solution θ^{s*} implies that θ^{s*} times of (y_0, z_0) can be produced in fab s; the total maximum wafer outputs are thus $\sum_{s \in S} \theta^{s*}$ times of z_0 for corporate.

There are $|S|$ groups for all constraints, and each group represents one fab. Variable returns to scale and convexity which are commonly seen in the literature are assumed and thus $T(vrs, c)$'s are used for individual fabs. We assume heterogeneous production technology for fabs, and technology T is estimated by observations from only the corresponding fab R^s. Fabs with homogeneous production technology can also be assumed if necessary; in those cases, $R = \bigcup_{s \in S} R^s$ is used to estimate T. As mentioned

earlier, other assumptions about (A2) and (A3) are possible, but are case-dependent and rely on subjective judgements. Constraints (6) and (7) indeed adopt $D_0(crs)$, which estimates D based on a single record and allows decomposition of given records. $D_0(Z_+)$ can be used when fraction of products is not allowed, and this assumption is particularly useful when (y_0, z_0) is the atom or simple sum of atoms. Further, constraints (6) and (7) are unnecessary and can be removed from the model as:

$$\max_{\lambda_r^s,\theta^s}\left\{\sum_{s\in S}\theta^s : x_0^s \geq \sum_{r\in R^s} x_r^s\lambda_r^s, \theta^s y_0 \leq \sum_{r\in R^s} y_r^s\lambda_r^s, \sum_{r\in R^s}\lambda_r^s = 1, \lambda_r^s \geq 0, r\in R^s, s\in S\right\}. \quad (8)$$

The optimal quantity of wafer outputs can be obtained by the optimal solution $\sum_{s\in S}\theta^{s*}$ of Model (8) as $(\sum_{s\in S}\theta^{s*})z_0$.

The following model (BL-O) determines the maximum wafer outputs with the same product recipes (y_0, z_0) and resource x_0^s when adopting the borderless fab strategy:

$$\max_{z,y^s,\lambda_r^s,\phi} \phi \quad \text{(BL-O)}$$

Subject to for each fab $s \in S$:
production technology:

$$x_0^s \geq \sum_{r\in R^s} x_r^s\lambda_r^s \quad (9)$$

$$y^s \leq \sum_{r\in R^s} y_r^s\lambda_r^s \quad (10)$$

$$\sum_{r\in R^s}\lambda_r^s = 1, \quad \lambda_r^s \geq 0,\ r\in R^s; \quad (11)$$

corporate-wise wafer outputs
design technology:

$$\sum_{s\in S} y^s = \phi y_0 \quad (12)$$

$$z = \phi z_0 \quad (13)$$

$$y^s \geq 0,\ s\in S.$$

Identical to Model (B-O), production technology for individual fabs is specified by $T(vrs, c)$ as constraints (9) to (11). Borderless fab pools the capacity of the layer processes to deliver the final products as (12) and (13); $\sum_{s\in S} y^s$ shows that production capacity is pooled. Constraints (12) and (13) also specify the design technology in response to the requirement (y_0, z_0) as $D_0(crs)$. The optimal value ϕ^* indicates $(100 \times \phi^*)\%$ of z_0 will be delivered overall. Further, (13) is unnecessary and can be removed from the model to provide the same optimal solution as:

$$\max_{y^s,\lambda_r^s,\phi}\left\{\phi : x_0^s \geq \sum_{r\in R^s} x_r^s\lambda_r^s,\ y^s \leq \sum_{r\in R^s} y_r^s\lambda_r^s,\ \sum_{r\in R^s}\lambda_r^s = 1,\ y^s \geq 0,\ \lambda_r^s \geq 0,\ r\in R^s,\ s\in S; \sum_{s\in S} y^s = \phi y_0\right\}.$$

It should be noted that $\sum_{s\in S} y^s$ indicates that the layer processes are pooled, but (BL-O) does not reveal where the final products are delivered, i.e., which fab processes the last layer required. However, the overall allocation of different layer types can be observed from the optimal solution y^{s*}.

$\phi^* z_0 - (\sum_{s\in S}\theta^{s*})z_0 = (\phi^* - \sum_{s\in S}\theta^{s*})z_0$ shows the difference in capacity after adopting the borderless fab strategy given the product specification of (y_0, z_0). $\phi^* - \sum_{s\in S}\theta^{s*}$ is the proportional increment in the benefits while maintaining the same wafer product mix. Therefore, the percentage improvement due to borderless fab strategy:

$$\frac{\left(\phi^* - \sum_{s\in S}\theta^{s*}\right)}{\sum_{s\in S}\theta^{s*}} \tag{14}$$

is the measure. Higher values of (14) suggest more significant benefits due to the borderless fab. $(\phi^* - \sum_{s\in S}\theta^{s*})z_0\mathbf{1}$ is the improvement in terms of total wafer outputs in a particular period, where $\mathbf{1}$ is the $|O| \times 1$ vector of 1's.

Model (B-O) assumes all fabs produce the same product mix as required by management. In reality this assumption may be invalid since the overall product demand can be allocated to fabs to improve their outputs. Product allocation is the decision to enhance the use of individual independent capacities, and can be seen as fine-tuning the process under the non-pooling circumstance. In addition, price information should be considered to incorporate management and individual fabs' performance, while recognising that prices may vary dramatically over time. Whether or not to pool capacity is a long-term decision; and product allocation is short-term and so we can disregard it. However, we note that other objectives, such as profit maximisation, are also possible, but may require additional information, e.g., prices and costs. Moreover, unlike simulation studies requiring strong engineering technical background and costly software, our models (B-O) and (BL-O) are based on standard linear programming (LP) problems, and can be implemented by standard packages such as Lingo and Cplex. Microsoft Excel also provides an add-in package, called solver, to solve LP problems.

5.2 *Numerical example*

We present a simple hypothetical example to illustrate the proposed models and visualise the benefits of capacity pooling. Suppose there are two fabs ($S = \{U, V\}$) using the same units, say 100, of a resource to provide two types of layer process ($J = \{Y_1, Y_2\}$). Each fab has three historical records ($R^U = \{A, B, C\}$ and $R^V = \{L, M, N\}$), which are listed in Table 1. Further assume that there is only one type of wafer product requiring both Y_1 and

Table 1. Values for the numerical example.

	Collected records						Bordered		Borderless	
	Fab U			Fab V			Fab U	Fab V	Fab U	Fab V
	A	B	C	L	M	N	a	b	c	N
Resource	100	100	100	100	100	100	100	100	100	100
Y_1	80	120	40	100	50	150	127.3	109.4	101.6	150
Y_2	150	50	90	40	120	80	89.1	76.6	96.1	80

Y_2 processes. A recipe needs processing Y_1 and Y_2 50 and 35 times, respectively, to produce one unit of wafer output, i.e., $y_0 = (50, 35)$ and $z_0 = 1$. The corresponding Model (B-O) is as follows:

$$\max \theta^U + \theta^V$$

subject to:

$$\begin{aligned}
&100 \geq 100\lambda_A + 100\lambda_B + 100\lambda c \\
&y_1^U \leq 80\lambda_A + 120\lambda_B + 40\lambda c \\
&y_2^U \leq 150\lambda_A + 50\lambda_B + 90\lambda c \\
&\lambda_A + \lambda_B + \lambda c = 1 \\
&y_1^U = 50\theta^U, \quad y_2^U = 35\theta^U \\
&z^U = 1\theta^U \\
&100 \geq 100\lambda_L + 100\lambda_M + 100\lambda_N \\
&y_1^V \leq 100\lambda_L + 50\lambda_M + 150\lambda_N \\
&y_2^V \leq 40\lambda_l + 120\lambda_M + 80\lambda_N \\
&\lambda_L + \lambda_M + \lambda_N = 1 \\
&y_1^V = 50\theta^V, \quad y_2^V = 35\theta^V \\
&z^V = 1\theta^V \\
&\lambda_A, \lambda_B, \lambda c, y_1^U, y_2^U \geq 0 \\
&\lambda_L, \lambda_M, \lambda_N, y_1^V, y_2^V \geq 0.
\end{aligned}$$

The corresponding (BL-O) is:

$$\max \phi$$

subject to:

$$\begin{aligned}
&100 \geq 100\lambda_A + 100\lambda_B + 100\lambda c \\
&y_1^U \leq 80\lambda_A + 120\lambda_B + 40\lambda c \\
&y_2^U \leq 150\lambda_A + 50\lambda_B + 90\lambda c \\
&\lambda_A + \lambda_B + \lambda c = 1 \\
&100 \geq 100\lambda_L + 100\lambda_M + 100\lambda_N \\
&y_1^V \leq 100\lambda_L + 50\lambda_M + 150\lambda_N \\
&y_2^V \leq 40\lambda_l + 120\lambda_M + 80\lambda_N \\
&\lambda_L + \lambda_M + \lambda_N = 1 \\
&y_1^U + y_1^V = 50\phi \\
&y_2^U + y_2^V = 35\phi \\
&z = 1\phi \\
&\lambda_A, \lambda_B, \lambda c, y_1^U, y_2^U \geq 0 \\
&\lambda_L, \lambda_M, \lambda_N, y_1^V, y_2^V \geq 0.
\end{aligned}$$

Optimal solutions of the two LP problems above are $\theta^{U*} = 1.27$, $\theta^{V*} = 1.09$, and $\phi^* = 2.56$, which suggests that improvement level is 6.3% according to (14). Detailed optimal solutions for the maximum numbers of layer processed are listed in the last four columns of Table 1.

Figure 2 visualises the benefits of capacity pooling in this example, in which the x-axis and y-axis are the values of Y_1 and Y_2, respectively. According to our approach, the feasible production technology of Fab U, using 100 units of resource, is the area surrounded by x-axis, y-axis and the solid lines. Records A and B are on the boundaries of the feasible region, representing the ideal performance, while C, with poor performance, is within the region. This shows that our models simulate the feasible behaviour by ideal, not average, performance. Similarly, dashed lines represent the boundaries of feasible production for Fab V. Under a bordered fab strategy, identical output mix is required for both fabs, represented by ray OP; the optimal layers processed are points a and b for Fab U and Fab V, respectively. We observe that the slopes of line segments AB and MN are different. The slopes represent the marginal rate of transformation, i.e., the rate at which one output must be sacrificed to produce a single extra unit of another output by using the same level of resources. Fab U will benefit by reallocating product (layer type) mix to process more Y_2, while Fab V should process Y_1 as much as possible. The optimal solutions under a borderless fab strategy are points c and N for Fabs U and V, respectively (Figure 2, Table 1), which indicate both fabs also reallocate their product mix (from a to c and from b to N). Consequently, the corporate-wise layers processed improve from (236.6, 165.7) to (251.6, 176.1), and wafer output increases by 6.3%. One fab may be more productive in processing a particular layer type; a careful allocation of the mix of layer types optimally can generate more benefits. Capacity pooling allows the flexibility to reallocate the different layers processed to achieve higher productivity.

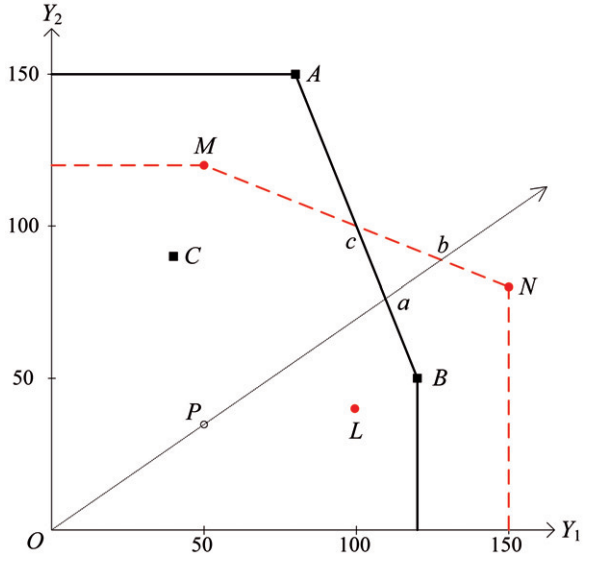

Figure 2. A two-output illustration.

6. Case study

This section reports a case study based on real data from a leading semiconductor manufacturer to demonstrate the effectiveness of the proposed approach. Our data set represents the overall full-scale fab-level operations, and provides a complete picture of the operations. The case study evaluates the potential benefits of adopting new capacity pooling strategies using the proposed models, particularly by comparing Models (B-O) and (BL-O).

6.1 *Data*

We study two fabs ($S = \{A, B\}$) located within a few miles of one another that are owned by the same semiconductor manufacturer. They are 8-inch fabs, which are more stable compared to 12-inch or advanced fab, making a borderless fab strategy ideal. We consider four resources: headcount, equipment, space and time, i.e., $I = \{$headcount, equipment, space, time$\}$. The first three resources are commonly found in other studies. Time, which is not as straightforward as the first three, is the total time used in production. Given the same level of labour, equipment and space, more layers require longer times. Similarly, for example, less labour (equipment) generally results in longer times to generate the required layers, and this reveals the substitutability among resources. For practical implementation, the detailed definitions of resources can be found in Chen and Chien (2009).

We consider three different layer processing types – polysilicon layers, metal layers and the others – as intermediate products, i.e., $J = \{$poly, metal, other$\}$. The three layer types are fundamental processes for all wafer products. Layer data is collected as number of layers processed per month. The four categories of wafer outputs are 0.13 μm, 0.18 μm, 0.25 μm and 0.35 μm. Hence, $O = \{0.13, 0.18, 0.25$ and $0.35\}$. Wafer outputs are measured as pieces of equivalent 8-inch wafers. A single record (x_r^s, y_r^s, z_r^s) consists of monthly information about resources (x_r^s), effective layers processed (y_r^s) and effective wafer products delivered (z_r^s) for a particular month r at fab s. We collect records for 39 consecutive months at each fab, for a total of 78 records (39 for each fab, i.e., $|R^s| = 39$)[2].

6.2 *Results and discussion*

To evaluate the potential gains, we compare (B-O) and (BL-O) based on (14). The same parameters (R^s, x_0^s and (y_0, z_0)) are applied to both models. R^s determines the possible resource-output transformation while x_0^s and (y_0, z_0) define the environmental settings, which are dynamic due to internal resource supply and external demand. We must compare both models considering the dynamic environment, i.e., testing different scenarios in terms of resources and product configurations (x_0^s and (y_0, z_0)).

To represent different scenarios, 39 monthly resource configurations for fabs (x_r^s's) are considered. We use each (y_r^s, z_r^s) ($r \in R^s$ and $s \in S$) to represent possible corporate-wise overall product parameters for $(y_0, z_0) = (y_r^s, z_r^s)$. There are 78 scenarios regarding product specifications, 39 from each fab. Therefore, 3042 (39×78) different resource and product design combinations can be used as the testing scenarios. In addition, since (y_r^s, z_r^s) is the aggregate data and can be assumed decomposable, *crs* is assumed to estimate D.

We first present the potential output differences by adopting the borderless fab strategy to a particular case. Tables 2 and 3 list the resource configurations for two fabs ($x_0^s, s \in S = \{A, B\}$) and the overall product specifications ((y_0, z_0)). There are slight

Table 2. Resource configuration for two fabs.

Resources	Headcount	Equipment	Space	Time
Fab A	2222	76150	19548	2192033
Fab B	1973	74950	19000	2195735

Table 3. Product configuration.

Layer type			Wafer type			
Poly	Metal	Other	0.13	0.18	0.25	0.35
127913	379577	1957906	0	43429	27070	22719

Table 4. Production outputs under different strategies.

		Layers processed (K layers)			Wafers produced (K pieces)			
	Fab	Poly	Metal	Other	0.13	0.18	0.25	0.35
Bordered	A	77.2 (25.5)[a]	229.2 (0)	1182.14 (73.73)	0	26.2	16.3	13.7
	B	78.8 (0)	233.7 (90.7)	1205.45 (167.67)	0	26.7	16. 7	14.0
	Sum[b]	156.0	462.9	2387.6	0	53.0	33.0	27.7
Borderless	A	102.7 (0)	229.2 (0)	1255.87 (0)	n/a	n/a	n/a	n/a
	B	69.2 (0)	280.8 (24.6)	1374.8 (0)	n/a	n/a	n/a	n/a
	Sum	171.9	510.0	2630.7	0	58.4	36.4	30.5

Notes: [a]excess capacity;
[b]total used capacity.

resource differences between Fabs A and B, but their individual underlying production capabilities cannot be revealed explicitly. Table 4 compares the optimal capacity for the requested products (Table 3) based on (B-O) and (BL-O). The results show that capacity pooling can increase wafer outputs by 10.28%. Using a bordered strategy, the metal layer process (229.2 K layers) and poly layer process (78.8 K layers) are the bottlenecks for Fabs A and B, respectively, resulting in significant excess capacities for the other layer masking processes in both fabs. For example, Fab A has excess capacity of 25.5 K layers for the poly layer process, which is about one third of the total capacity. Capacity pooling allows trans-shipment between fabs, and excess capacity is better utilised for both fabs (Table 4).

We observe that Fab A has identical total capacities (used plus excess) for each layer type under both strategies while Fab B shows inconsistent results (Table 4). This observation reveals that the improvements of Fab A derive from better use of the excess capacity. On the other hand, Fab B benefits by reallocating the output (layer type) mix,

e.g., poly layers drop from 78.8 K to 69.2 K while other layer types increase from 1373.1 K (1205.45 K + 167.67 K) to 1374.8 K. At a macro level, capacity pooling provides more productive loading mix and better use of excess capacities, and thus more wafer outputs. Consequently, the capacity pooling benefits would be more significant if there are more differences in fabs' excess capacity status and optimal layer mix. In addition, the results suggest that simply adding up available capacity from individual fabs as the overall capacity is improper although this approach is intuitive and easy to use. As observed, output mix reallocation should be considered.

Different scenarios may give different results. Table 5 summarises the percentage improvement due to the borderless strategy for the 3042 instances. The improvement measure is based on (14), using the best of current strategy as the comparing basis, not the initial status. The average improvement for the 3042 scenarios is 2.15% and the maximum improvement is 16.65%. The improvement is equivalent to an additional 3097 pieces of 8-inch wafer monthly on average, or 18,136 pieces as the maximum once the borderless fab strategy is adopted (Table 5). The first quartile of improvement is 0% and the median is 0.96%. The numbers imply a large proportion of the scenario instances without significant improvement or even no improvement. Fifty percent of the instances have improvement less than 1%. However, the findings do not imply the failure of borderless fab, but rather reveal that there is no need for transporting between two fabs in these circumstances. In other words the current strategy can handle (y_0, z_0) and x_0^s, $s \in S$, as well as borderless. A borderless fab strategy allows incomplete wafers to be transported from one fab to another, but such transportation is not necessary. The capacity pooling utilises inter-fab logistics to expand the overall corporate capacity. A central decision unit, not a fab, plans and controls detailed operations.

We further investigate the instances in which transshipments do exist. Table 6 presents the 1495 (out of 3042) instances with improvement no less than 1%, which are about 49% of the scenario instances studied. The average improvement becomes 4.19% or 5987 pieces of 8-inch wafer. We conclude that if current capacity strategy cannot sufficiently support the need, borderless fab can improve the overall corporate capacity about 4% (or around 6000 8-inch wafer outputs per month) on average.

Table 5. Summary of improvements.

	No. of obs.	Mean	Min	Q1	Median	Q3	Max
Improvement (%)	3042	2.15	0	0	0.96	3.44	16.65
8-inch wafer outputs (pieces)[a]	3042	3097	0	0	1505	5065	18,136

Note: [a]additional outputs due to capacity pooling.

Table 6. Summary of improvements (>1%).

	No. of obs.	Mean	Min	Q1	Median	Q3
Improvement (%)	1495	4.19	1.0	2.07	3.47	5.47
8-inch wafer outputs (pieces)[a]	1495	5987	873	3000	5200	8166

Note: [a]additional outputs due to capacity pooling.

We observe that the improvement in capacity from employing the borderless fab strategy comes with some costs. Additional inter-fab transportation expenses, including carrier investment and operation expenses, are required. Therefore, typically we decide to adopt the borderless fab strategy only when the potential gains in wafer outputs are significant. The 'cutting edge' of significance is case-dependent and subjective. Rather than making a call on whether to apply capacity pooling, we reiterate that this paper provides reliable quantitative information to assist the decision-making process. Although we investigate the effect of pooling only two fabs, the model is more general for any number of fabs. Further, in the implementation of (B-O) and (BL-O), a sufficient number of records produce better estimation of technologies.

7. Conclusion

This paper presented a model to evaluate the capacity pooling, 'borderless fab', strategy in semiconductor manufacturing. The model provides quantitative information as an aid to decision-making. The model is based on non-parametric activity analysis and is a macro-level evaluation, which closely resembles strategic decision-making. Our study used a full-scale data set representing real aggregate fab operations, while previous research has relied only upon small-scale cases to demonstrate effectiveness. We found that capacity pooling can improve monthly capacity by 3% on average.

In summary, the major contributions of this paper are three-fold. First, we presented a tool to answer an important question in the semiconductor industry: 'How much benefit can be gained from adopting the borderless fab strategy?' Rather than promoting adoption of capacity pooling, we provide quantitative information for key decision-making. Second, the generic model proposed, termed design technology, described and analysed high-level product design specifications. Third, our case study was based on a fab-level data set collected from the past operations of a leading semiconductor manufacturer. The empirical results thus are more meaningful and provide more salient insights than earlier research.

The implementation of capacity pooling is not free, but, for example, is accompanied by logistics costs and delay in batching to transport. Other hurdles to adopting capacity pooling include the need to invest heavily in information technology, such as integrating manufacturing execution system (MES). Further, trade-offs regarding quality issues are required. For example, the displacements (i.e., overlay errors) between layers often should be reduced to enhance the yield; this engineering concern results in a constraint for capacity pooling and we suggest it requires further studies. We note that the potential benefits should be studied prior to undertaking a cost-benefit analysis, and that our model fulfils this function from the perspective of productivity rather than alternative selection.

A recent trend in the industry is twin-fab construction. Instead of building a large fab at one time, two fabs can be built separately in the same location, considering the firm's fiscal status and market demand. If the trend holds, productivity gains due to capacity pooling can be realised and possible negative concerns can be minimised. For example, the transportation costs for capacity pooling can be reduced significantly in a twin-fab setting. The two fabs have the same MES, and the material handling systems can integrate both facilities for optimum control. The findings of our empirical study provide quantitative information to support this concept for new facility investments.

Our approach, which is deterministic, static, and based on historical data, does provide a quick, macro-level evaluation for capacity pooling. We recognise that its limitations – that the detailed dynamic behaviour of the systems are not revealed and future demand uncertainty and technology changes are not fully incorporated – should be addressed in future research. However, as mentioned, the proposed model can be easily applied to other cases and industries. The proper use of the assumptions we have discussed will yield better approximations of the underlying production technology, which typically requires domain expertise and subjective judgements. We also suggest that a fruitful area of research will be the development of a systematic procedure.

Acknowledgements

This research was supported in part by the National Science Council, Taiwan (NSC 97-2221-E-009-113) and Taiwan Semiconductor Manufacturing Company (96A0280J8).

Notes

1. Please refer to Fare *et al*. (1994) for comprehensive information.
2. However, we are not able to reveal any detailed information due to manufacturer confidentiality.

References

Banker, R.D., Charnes, A., and Cooper, W.W., 1984. Some models for estimating technical and scale inefficiency in data envelopment analysis. *Management Science*, 30 (9), 1078–1092.

Benjaafar, S., Cooper, W.L., and Kim, J.-S., 2005. On the benefits of pooling in production-inventory systems. *Management Science*, 51 (4), 548–565.

Charnes, A., Cooper, W.W., and Rhodes, E., 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2 (6), 429–444.

Chen, W.-C. and Chien, C.-F., 2009. Measuring relative performance of wafer fabrication operations: a case study. *Journal of Intelligent Manufacturing*, doi: 10.1007/s10845-009-0302-x. [In press].

Chien, C.-F. and Hsu, C., 2006. A novel method for determining machine subgroups and backups with an empirical study for semiconductor manufacturing. *Journal of Intelligent Manufacturing*, 17 (4), 429–440.

Chien, C.-F., *et al*., 2007. Construct the OGE for promoting tool group productivity in semiconductor manufacturing. *International Journal of Production Research*, 45 (3), 509–524.

Corbett, C.J. and Rajaram, K., 2006. A generalization of the inventory pooling effect to nonnormal dependent demand. *Manufacturing and Service Operations Management*, 8 (4), 351–358.

Fare, R., Grosskopf, S., and Lovell, C.A.K., 1994. *Production frontiers*. Cambridge, UK: Cambridge University Press.

Gan, B.P., *et al*., 2007. Analysis of a borderless fab using interoperating AutoSched AP models. *International Journal of Production Research*, 45 (3), 675–697.

Leontief, W.W., 1953. *The structure of the American economy*. New York: Oxford University Press.

Rau, H., Chu, Y.-H., and Cho, K.-H., 2005. Layer modelling for the inspection allocation problem in re-entrant production systems. *International Journal of Production Research*, 43 (17), 3633–3655.

Wu, M.C., Chen, C.F., and Shih, C.F., 2009. Route planning for two wafer fabs with capacity-sharing mechanisms. *International Journal of Production Research*, 47 (20), 5843–5856.