Flex Serv Manuf J (2013) 25:286–309
DOI 10.1007/s10696-011-9109-0

A two-stage stochastic programming approach for new tape-out allocation decisions for demand fulfillment planning in semiconductor manufacturing

Chen-Fu Chien · Jei-Zheng Wu · Chien-Cheng Wu

Published online: 26 July 2011

Abstract Demand fulfillment and capacity utilization directly affects customer satisfaction, market growth, and the profitability of the company in the semiconductor industry. These characteristics boost the significance of allocating various customer demands to a number of wafer fabrication facilities (fabs) with different capacity configurations. Before volume production, the introduction of new semiconductor product, namely new tape-out (NTO), requires extremely sophisticated and lengthy qualification with high-cost masks and pilot runs in the qualified fabs. Thus, the NTO allocation will affect future product mix of the qualified fabs, and the flexibility to fulfill the volume demands of the allocated NTOs in the corresponding fabs. This research aims to construct a two-stage stochastic programming (2-SSP) demand fulfillment model. The first stage considers NTO allocation decisions to a number of qualified fabs before the corresponding demand volume is realized. The second stage allocates the capacity to fulfill demand requirements based on the results of four options of capacity reconfiguration: (1) qualifying a product to more than one fab (share); (2) physically transferring a set of masks for a product from one fab to another, where a requalification is required (transfer); (3) moving tools from under-loaded fabs to over-utilized fabs (backup); and (4) utilizing different technologies to capacity inside a fab to support the technology with insufficient capacities (exchange). Both the share and transfer options require long

C.-F. Chien
Department of Industrial Engineering and Engineering Management,
National Tsing Hua University, Hsinchu, Taiwan
e-mail: cfchien@mx.nthu.edu.tw

J.-Z. Wu (✉)
Department of Business Administration, Soochow University, Taipei, Taiwan
e-mail: jzwu@scu.edu.tw

C.-C. Wu
Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan

lead time for qualification, whereas the backup and exchange options can be accomplished within a planned timeframe. A numerical study based on real settings is conducted to estimate the validity of the proposed 2-SSP model via values of stochastic solution (VSS) and expected values of perfect information (EVPI). The results showed the benefits of adopting 2-SSP models, especially in an environment with high-demand fluctuation. Furthermore, the proposed 2-SSP can provide near-optimal solutions similar to those of deterministic models with perfect information.

Keywords New tape-out allocation · Capacity planning · Demand fulfillment planning · Stochastic programming · Manufacturing strategy · Decisions under uncertainty

1 Introduction

In the semiconductor industry that is very capital intensive, semiconductor companies strive to enhance their capital effectiveness via managing demand fulfillment and capacity utilization to maintain their competitive advantage (Leachman et al. 2007; Wu and Chien 2008). Following Moore's Law that the number of transistors fabricated on an integrated circuit (IC) will be doubled approximately every 1 or 2 years (Moore 1965), semiconductor manufacturing companies have been continuously developing new technology nodes and investing most advanced tools and facilities to fulfill the demands for producing advanced products. In particular, the cost of building one modern 300 mm wafer fabrication facility (fab) is 4 billion US dollars, and the price for a single machine installation may range from 3 to 10 million US dollars. On average, equipment depreciation costs account for approximately 60% of total production costs.

On the other hand, it takes a long lead time between 6 and 12 months for equipment procurement especially when the demand highly fluctuates (Wu et al. 2005; Chien et al. 2010). For semiconductor manufacturers, planning the capacity and fulfilling various customer demands are very critical, since they will affect customer satisfaction and capacity utilization, ultimately affecting corporate profitability and growth (Chien and Zheng 2011). On the other hand, demand fulfillment is very challenging, due to planning the release of numerous products, along with the high level of uncertainty associated with the demand, and the requisite of long lead-time for qualifying a specific product in a fab. Poor management causes strife between demand and capacity among fabs or technologies, and sales are lost while a capacity remains idle.

To avoid this strife and to enhance capacity utilization, semiconductor companies should improve order allocation for various products, and corresponding capacity configurations of the allocated fabs. In particular, order allocation includes the allocation of the new tape-outs (NTOs) of forthcoming products as well as the allocation adjustment of existing products in light of their demand changes. Since the IC layout design was stored in magnetic tapes decades ago, NTO means to hand off the file of new product design to make a lithography mask for mass production in fab (Mouli and Winstead 2007). NTO is a time-consuming process that includes the creation of expensive masks, qualifications and pilot runs to ascertain the production capability in the selected fab for potential volume production later. NTO allocation is a crucial leading factor for proactively managing future demand profiles in the fabs. However, not every NTO will generate volume production for the product later and the generated volumes are also varied among the products with different life cycles. Little research has been done to examine the characteristics of NTO and address the NTO allocation.

Indeed, NTO allocation can be optimized to form a desirable expected demand profile via allocating more NTOs to an anticipated under-loaded fab in the future and fewer NTOs to the other fabs. Furthermore, the masks are increasingly expensive as the device feature is continuously shrunk and not sharable between the fabs. Thus, poor allocation of NTOs will cause reallocation of some NTO that requires another long qualification lead time and tremendous engineering efforts in another fab and extra cost for remaking or revising the masks. In practice, most companies rely on time-consuming deterministic solutions and manual adjustments, resulting in loss of consistent quality, since combinatorial complexities are involved in the adjustments of NTO allocations.

This study constructs a demand fulfillment planning framework to facilitate the decision-making process regarding NTO allocations, with considering capacity backup and product reallocation to alter short-term capacity configuration, to fulfill the demands and minimize the costs. Since the planning is based on forecast demands, the problem is formulated to cope with demand uncertainty and to find an optimal solution under different versions of forecasts, i.e., scenarios. In particular, we constructed a two-stage stochastic programming (2-SSP) demand fulfillment model that allocates NTOs to a number of qualified fabs before the corresponding demand volume is realized and then matches fab capacity with various demands, while employing various capacity reconfiguration options.

The remainder of this paper is organized as follows: Sect. 2 defines the NTO allocation and demand fulfillment issues, and related literature. Section 3 presents the 2-stage stochastic programming model, and demonstrates how it addresses the issues. Section 4 introduces a study based on synthetic data aligning with actual settings in a globally recognized semiconductor fab located in Taiwan, to examine the validity of the proposed approach and to analyze the suggested decisions under different situations of demand variation. Section 5 presents a conclusion with a discussion on future research directions.

2 Problem statement and related studies

2.1 New tape-out allocation, capacity configuration, and production

New tape-out represents the stage where the IC designer sends a new circuit to the manufacturer for wafer fabrication in the semiconductor industry. The name is derived from past conventions, where the design of a new product was sent to the manufacturer via tapes. For semiconductor manufacturers, NTOs introduce the latest products with the potential for new demand, thereby changing the future demand profile of fabs. Since demand volume is highly uncertain, multiple forecasters provide projections according to the information they obtain from different industry experts. For example, sales and marketing professionals may provide two contradictory demand forecasts. For example, in low season, the sales demand forecast tends to be affected by short customer behavior and thus may be pessimistic, while the marketing demand forecast is based on long term economic trends and thus may be optimistic. In addition, demand planners collect information from sales and marketing departments to form a compromise forecast. Different forecast versions in this study represent "scenarios" with varied occurrence probabilities. If no stochastic model is applied, one scenario must be consented for NTO allocation based on group decisions.

Each semiconductor product must undergo the process of mask creation and pilot run to qualify for fab production capability prior to entering the volume production stage. The lead time is approximated to be two quarters, i.e., 6 months, on average. A set of masks is essential to implement the qualification process for each new product, so that the pilot run of products can be executed. The pilot run manufactures products in small volumes for the purpose of qualification. The capacity it occupies is so minuscule that it can almost be ignored. The qualification is intended to certify the capability of fabs, to properly manufacture products conforming to its specifications and functionality requirements. The masks are the means to enabling follow-up actions. Masks are expensive, and can range from several hundred thousand dollars to more than a million US dollars per set. In addition, masks are created according to the feature of the processes and tools of the fab.

New tape-out allocation can select the ideal fab to produce the specific NTO where multiple fabs are concerned. It chooses a fab for each NTO for production before the masks can be created, and the pilot run process can begin. The fact that the creation of masks is dependent on the feature of fabs implies that sharing masks between fabs is unfeasible, since most fabs do not possess identical tools and homogeneous manufacturing processes. Poor NTO allocation will result in a disparity between demand and capacity, raising the need for product reallocation and consequently resulting in additional costs for creating or revising masks. Therefore, NTO allocation is influential to increase future demand and lower potential additional costs, the results of which can either be disastrous, or allow a company to increase its profit margin substantially.

The share and transfer options are two modes of adjusting the allocation of existing products to increase demand. To "share a product," which entails the manufacture of a product in more than one fab, an additional set of masks will be created for the product, subsequently increasing costs. All fabs sharing a product are qualified to manufacture the same product and enjoy the production flexibility by adjusting production volume between these fabs. To "transfer a product" signifies switching from the fab a product was originally assigned to; its masks will be physically transferred to the new destination, where the new fab will manufacture the product. A number of revisions on the masks may be required once the product has been transferred, to fit the process and tools of the target fab. The cost is typically lower than the sharing option, since only certain parts of the masks need reworking. However, transferring does not have the additional benefit of increasing flexibility that the sharing option possesses.

The capacity backup and exchange options are approaches to configuring the capacity of different technologies and fabs to fulfill demand according to spare capacity. The concept of inter-fab backups is to shift capacity from under-loaded fabs to over-utilized ones, consequently relaxing the tight production schedule. Overloaded fabs lend their machines to under-loaded fabs, where the main cost lies on transportation. The semiconductor manufacturer in this study can benefit from the effect of clustering fabs because fab locations are generally close, requiring little time to transport machines between fabs for the purpose of inter-fab backup.

Inner-fab backup, also called "capacity exchange," utilizes the capacity of different technologies inside a fab to support the technology with insufficient capacity. A fab shares a number of common tools between several technologies, and other dedicated tools are devoted to technologies that are more specific. The combination of the capacity of dedicated tools and the allocated working time of common tools affect the capacity of each technology. Thus, when the working time of common tools allocated to a technology increases, the capacity will increase accordingly, if not bound by dedicated tools. The different cycle times for producing special technologies result in a disparity for capacity exchange. Therefore, the capacity exchange rate is significant for estimating capacity more accurately. In practice, capacity backup or exchange takes a number of days, negligible when compared to a planned timeframe, e.g., a quarter.

Considering both the capacity configuration of fabs and their capability to manufacture products, the planned capacity configuration indicates the feasibility for allocating products to fabs for production. With the planned capacity configuration and wafer demand forecast, the manufacture of products can be determined, giving rise to production costs for fulfilling demand, along with potential penalties for failing to meet that demand. This research considers the following expenses as targets to be minimized: mask, capacity backup, exchange, production, and penalty for unmet demands.

2.2 Related works

The NTO allocation problem is similar to the issue of product-to-plant allocation in the automotive industry, enunciated by Inman and Gonsalvez (2001). The trouble lies in introducing new automotive models to plants because the modification of a plant requires heavy financing and long lead time to qualify a new product. The problem becomes more complex with regard to high-volume products, which may involve the capacity of more than a single plant. Other related studies regarding the product-to-plant allocation problem has been developed for the automotive industry. Fleischmann et al. (2006) developed a planning model for product allocation to global production sites, and affected capacity investment. Kauder and Meyr (2009) also developed a strategic network-planning model for the automotive industry, particularly for premium cars.

Sharing products is similar to the concept of "chaining" where a chain is the direct or indirect connection formed between a group of products and plants, according to product assignment decisions (Jordan and Graves 1995). The sharing option proposed in this research also connects fabs by manufacturing the same product in different fabs, for flexibility in deciding the production allocation for a certain product. Jordan and Graves (1995) demonstrated that utilizing the chaining method can enhance limited flexibility to a near total level. This research, however, differentiates itself from the works of Jordan and Graves (1995), Inman and Gonsalvez (2001), and Kauder and Meyr (2009) by addressing additional applications of transfer, inter-fab backup options, inter-fab exchange options, and multiple demand scenarios.

Rather than assuming that the demand is deterministic, stochastic programming (SP) can treat the demand distribution as a set of discrete scenarios in the model, minimizing expected costs across all scenarios. Common SP methods include the two-stage SP model and the multi-stage SP model. The decisions in the two-stage SP can be halved into two categories: decisions in the first stage remain unaffected by uncertain conditions, signifying that the decisions were made prior to the emergence of said uncertain conditions. Therefore, NTO allocation is the first-stage decision because it must be made prior to the observation of the realized demand. The second-stage decisions are dependent on the first-stage decisions and the scenarios. They engage in a recourse action to compensate for the first-stage decisions, minimizing expected costs under scenarios once the uncertain variables have been disclosed. The second-stage problem is also called the recourse problem. The production decisions belong to the second-stage because the demand realization and allocation results must be considered, to decide how each fab will fulfill demand. In contrast, multi-stage SP is a generalized form allowing revisions of decisions based on the uncertainties revealed during each time period. The multi-stage scenarios of times can be independent or dependent, influencing the complexity of computation and the effectiveness of modeling. This study focused on two-stage SP, where NTO demand scenarios correspond with multiple forecasts from different forecasters, such as sales, marketing, and demand-planning departments. To request multi-stage scenarios across a number of times from forecasters is unfeasible in practice, whereas generating them alongside a scenario tree is computationally intractable. A more detailed discussion on two-stage and multi-stage SP models and their comparisons can be found in Huang and Ahmed (2005), who also state that the value of multi-stage SP is augmented by addressing the gap between the objectives of two-stage and multi-stage SP models.

The performance of stochastic models is commonly measured via the expected value of perfect information (EVPI), and the value of stochastic solution (Birge and Louveaux 1997). Let $\mathbf{S}$ be the set of scenarios and denote the optimal solution of a deterministic model under scenario $s \in \mathbf{S}$ by $\mathrm{DM}(s) = \arg\min\limits_{x} z(x,s)$. Assuming that complete information is obtainable prior to the formation of a decision, we have the optimal wait-and-see solution, $\mathrm{WS}(s) = \mathrm{DM}(s)$, of a deterministic model under scenario $s \in \mathbf{S}$. The corresponding objective value of $\mathrm{WS}(s)$ is $z(\mathrm{DM}(s),s) = z(\mathrm{WS}(s),s)$. The expected value of the WS solutions can be computed as $z_{\mathrm{WS}} = \sum\limits_{s \in \mathbf{S}} P_s \cdot z(\mathrm{WS}(s),s) = E_{\mathbf{S}} z(\mathrm{WS}(\mathbf{S}),\mathbf{S})$ where P_s represents the probability of occurrence of scenario s. Conversely, the optimal value of the so-called here-and-now (HN) solution, corresponding to the recourse problem, is defined as $z_{\mathrm{HN}} = \min_x E_{\mathbf{S}} z(x, \mathbf{S})$ where the solution is denoted by $\mathrm{HN}(\mathbf{S}) = \arg z_{\mathrm{HN}}$. Subsequently, the EVPI is defined as $z_{\mathrm{HN}} - z_{\mathrm{WS}}$, the difference between the obtained value of the HN solution considering uncertain circumstances versus the expected value of WS solutions obtained when perfect information is available. A decision maker should reject a proposal if the price of the information is greater than the EVPI.

Alternatively, the benefit of “value of the stochastic solution” lies in its capability to compare the expected value (EV) solution against the HN solution. The EV solution is the optimal solution of the mean value problem, i.e., $\mathrm{EV}(\bar{s}) = \arg \min_x z(x, \bar{s})$ where the set of expected values of all scenarios is considered the input scenario, denoted by $\bar{s} = E(\mathbf{S})$. Considering realization, the expected objective value of applying the EV can be further defined as $z_{\mathrm{EV}} = \sum_{s \in \mathbf{S}} P_s \cdot z(\mathrm{EV}(\bar{s}), s) = E_{\mathbf{S}}[z(\mathrm{EV}(\bar{s}), \mathbf{S})]$. Subsequently, the conventional value of the stochastic solution is defined as $z_{\mathrm{EV}} - z_{\mathrm{HN}}$.

Instead of measuring the conventional value of the stochastic solution with respect to the mean value problem, this study is concerned with the value of the stochastic solution, hereafter abbreviated as the VSS, with respect to the deterministic model solution, $\mathrm{DM}(s)$, under a given scenario $s \in \mathbf{S}$. In practice, each forecaster has a distinct view, and the reality is likely similar to the proposed scenarios. Conversely, when employing the deterministic model, one scenario must be consented for NTO allocation based on group decisions. Simply utilizing expected demand can hardly persuade decision makers to take further action, due to the lack of causal explanations. Defining $\mathrm{VSS}(s) = \sum_{s' \in \mathbf{S}} P_{s'} \cdot z(\mathrm{DM}(s), s') - z_{\mathrm{HN}} = E_{\mathbf{S}}[z(\mathrm{DM}(s), \mathbf{S})] - z_{\mathrm{HN}}$ yields $\mathrm{VSS} = E_{\mathbf{S}}[\mathrm{VSS}(\mathbf{S})]$.

A number of SP applications are in practice concerning capacity planning in the semiconductor industry and other industries. Barahona et al. (2005) presented a stochastic model of capacity planning under demand uncertainty for semiconductor manufacturing. They tried minimizing the expected unmet demand with a two-stage stochastic model, to decide which tools to procure under capacity and budget constraints. Chen et al. (2002) developed an SP model for technology and capacity planning in an environment involving multiple products, stochastic demands, and technology alternatives. They also proposed a solution procedure based on the Lagrangian method restricting simplicial decomposition. Christie and Wu (2002) presented a multi-stage SP model for strategic-capacity planning for a semiconductor manufacturer; the model determined the quantity of distinct technology configured in each facility and for each time period. Geng et al. (2009) considered demand and capacity uncertainty when constructing an SP model. It showed that the results of decisions were optimal according to the variation of capacity. Geng and Jiang (2009) summarized characteristics of 2-SSP applications to semiconductor industries, including objectives of unmet demand, production costs, allocation costs, inventory costs, target utilization, supply preferences, constraints of tool purchase budget and capacity, as well as decisions regarding tool purchase, wafer starts, work assignment, and the allocation of wafer operations to different tools. Other studies relevant to the semiconductor industry focus on the allocation of products to tools instead of fabs (Bilgin and Azizoglu 2009; Chung et al. 2006; Wang et al. 2007). Studies have proposed solution techniques such as Lagrangian relaxation and decomposition-based branch-and-bound strategies to manage the large-scale mixed-integer program. In particular, the L-shaped method formulates 2-SSP as a dual block-angular structure with which one may perform a Dantzig-Wolfe decomposition of the dual or a Benders' decomposition of the primal to accelerate computation (Birge and Louveaux 1997).

Nevertheless, relevant studies have failed to present a model incorporating NTO allocation, product share and transfer options, capacity backup and exchange options, and production decisions. Without an integrated model, the decision-maker cannot evaluate a critical NTO allocation plan capable of demonstrating flexibility.

3 The two-stage stochastic programming model for NTO allocation decisions

3.1 Terminologies and notations

Sets	
$\mathbf{E}$	Set of existing products
$\mathbf{F}$	Set of all fabs considered
$\mathbf{F}_i$	Set of fabs capable of producing technology required by product i
$\mathbf{F}_k$	Set of fabs capable of producing technology k
$\mathbf{K}$	Set of technologies considered
$\mathbf{K}_j$	Set of technologies fab j is capable of producing
$\mathbf{N}$	Set of NTOs (new products)
$\mathbf{S}$	Set of demand scenarios
$\mathbf{T}$	Set of times within planning horizon, $\mathbf{T} = \{1, 2, \ldots, T\}$
Decision variables	
x_{ij}	Whether to allocate NTO i to fab j, $i \in \mathbf{N}, j \in \mathbf{F}_i, x_{ij} \in \{0, 1\}$. The first-stage decision variable. This decision is formed once in the model, when the action time is aligned with the first production of the NTO
y^s_{ij}	Whether to share existing product i to fab j under scenario s, $i \in \mathbf{E}, j \in \mathbf{F}_i, s \in \mathbf{S}, y^s_{ij} \in \{0, 1\}$
z^s_{ijlt}	Whether to transfer product i from fab j to fab l at time t under scenario s, $i \in \mathbf{E} \cup \mathbf{N}, j, l \in \mathbf{F}_i, j \neq l, t \in \mathbf{T}, s \in \mathbf{S}, z^s_{ijlt} \in \{0, 1\}$
pf^s_{ijt}	Capability to manufacture product i by fab j at time t under scenario s, $i \in \mathbf{E} \cup \mathbf{N}, j \in \mathbf{F}_i, t \in \mathbf{T}, s \in \mathbf{S}, \mathrm{pf}^s_{ijt} \in \{0, 1\}$
q^s_{ijt}	The quantity of demand of product i fulfilled by fab j at time t under scenario s, $i \in \mathbf{E} \cup \mathbf{N}, j \in \mathbf{F}_i, t \in \mathbf{T}, s \in \mathbf{S}, q^s_{ijt} \geq 0$
u^s_{it}	The quantity of demand of product i not satisfied at time t under scenario s, $i \in \mathbf{E} \cup \mathbf{N}, t \in \mathbf{T}, s \in \mathbf{S}, u^s_{it} \geq 0$
c^s_{jkt}	The amount of capacity of technology k in fab j at time t under scenario s, $j \in \mathbf{F}, k \in \mathbf{K}_j, t \in \mathbf{T}, s \in \mathbf{S}, c^s_{jkt} \geq 0$
ce^s_{jhkt}	The amount of capacity exchanged from technology h to technology k in fab j at time t under scenario s, $j \in \mathbf{F}, h, k \in \mathbf{K}_j, h \neq k, t \in \mathbf{T}, s \in \mathbf{S}, \mathrm{ce}^s_{jhkt} \geq 0$

cb^s_{jlkt} The amount of capacity of technology k in fab j physically relocating (backups) some of its equipment (thereby capacity) to fab l at time t under scenario s, $k \in \mathbf{K}, j, l \in \mathbf{F}_k, j \neq l, t \in \mathbf{T}, s \in \mathbf{S}, \mathrm{cb}^s_{kjlt} \geq 0$

Parameters

P_s Probability of occurrence of scenario s, $P_s \geq 0$, $\sum_{s \in \mathbf{S}} P_s = 1$

PT_{ik} Whether product i is manufactured with technology k, $i \in \mathbf{E} \cup \mathbf{N}, k \in \mathbf{K}, \mathrm{PT}_{ik} \in \{0, 1\}$

FT_{jk} Whether fab j is capable of producing technology k, $j \in \mathbf{F}, k \in \mathbf{K}, \mathrm{FT}_{jk} \in \{0, 1\}$

PFA_{ij} Capability of fab j to manufacture product i starting from time 1, $i \in \mathbf{E}, j \in \mathbf{F}_i, \mathrm{PFA}_{ij} \in \{0, 1\}$. Particularly, PFA_{ij} represents the value of pf^s_{ijt} when t is equal to zero.

CA_{jkt} Capacity of technology k in fab j at time t, $j \in \mathbf{F}, k \in \mathbf{K}_j, t \in \mathbf{T}, \mathrm{CA}_{jkt} \geq 0$

D^s_{it} Forecast demand of product i at time t under scenario s, $i \in \mathbf{E} \cup \mathbf{N}, t \in \mathbf{T}, s \in \mathbf{S}, D^s_{it} \geq 0$

LS_i Lead time to share product i with another fab, $i \in \mathbf{E} \cup \mathbf{N}, \mathrm{LS}_i \geq 0$

LT_i Lead time to transfer product i to another fab, $i \in \mathbf{E} \cup \mathbf{N}, \mathrm{LT}_i \geq 0$

ER_{hk} Capacity exchange rate from technology h to technology k, $h, k \in \mathbf{K}, h \neq k, \mathrm{ER}_{hk} \geq 0$

CEL_k Upper limit of capacity for technology k to be exchanged to other technologies, $k \in \mathbf{K}, \mathrm{CEL}_k \geq 0$

CBL_j Upper limit of capacity for fab j to back up other fabs, $j \in \mathbf{F}, \mathrm{CBL}_j \geq 0$

CE_{jh} Cost for fab j to exchange capacity from technology h to other technologies, $j \in \mathbf{F}, h \in \mathbf{K}_j, \mathrm{CE}_{jh} \geq 0$

CB_{jk} Cost for fab j to back up technology k of other fabs, $k \in \mathbf{K}, j \in \mathbf{F}_k, \mathrm{CB}_{jk} \geq 0$

CM_i Cost of making a new set of masks for product i, $i \in \mathbf{E} \cup \mathbf{N}, \mathrm{CM}_i \geq 0$

CT_i Cost of transferring product i, $i \in \mathbf{E} \cup \mathbf{N}, \mathrm{CT}_i \geq 0$

CP_{ij} Unit cost of manufacturing product i by fab j, $i \in \mathbf{E} \cup \mathbf{N}, j \in \mathbf{F}_i, \mathrm{CP}_{ij} \geq 0$

CU_i Unit penalty for unsatisfied demand of product i, $i \in \mathbf{E} \cup \mathbf{N}, \mathrm{CU}_i \geq 0$

3.2 Assumptions

1. Capacity expansion decision is formed in advance, and is thus not considered in this model. Ongoing capacity expansions are reflected by the time-indexed capacity parameter, CA_{jkt}.
2. No sharing or transferring decisions in progress.
3. Inventory and backlog are not considered, since this study focused on wafer foundry that is make-to-order without inventory, while backlog will become deferred demand.
4. All parameters are constant and do not vary with the volume of products to manufacture.
5. The time lag for backup and exchange are negligible, since each of them only cost a number of days, compared to monthly or quarterly based time slots. Fabs clustered within an area with a diameter of 10-km is normal in a mega-fab environment.

This study focuses on make-to-order production systems, such as fabrication foundry services, where inventory and backlog are commonly negligible in the capacity and demand-planning context. To further take nonlinear and varying parameters, such as costs and lead times, into taken, one may consider incorporations of nonlinear relations among capacity, cycle time, work-in-process, utilization, and throughput based on open-network queue models (Bitran and Tirupati 1989; Bard et al. 1999; Kim and Uzsoy 2008) as a future research direction. A literature review can be found in Pahl et al. (2007).

3.3 Model construction

The demand fulfillment problem is constructed as a two-stage stochastic model under discrete uncertain scenarios. The decisions in the model include the first-stage NTO allocation, and the second-stage product sharing and transferring options, capacity backup and exchange options, and the production plan for each product.

$$\text{Min} \sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{F}_i} \mathrm{CM}_i x_{ij} + \sum_{s \in \mathbf{S}} P_s (\mathrm{CO}_s + \rho_s) \tag{1}$$

$$\begin{aligned} \mathrm{CO}_s = {} & \sum_{i \in \mathbf{E}} \sum_{j \in \mathbf{F}_i} \mathrm{CM}_i y^s_{ij} + \sum_{i \in \mathbf{E} \cup \mathbf{N}} \sum_{j \in \mathbf{F}_i} \sum_{l \in \mathbf{F}_i, l \neq j} \sum_t \mathrm{CT}_i z^s_{ijlt} + \sum_{j \in \mathbf{F}} \sum_{h \in \mathbf{K}_j} \sum_{k \in \mathbf{K}_j, k \neq h} \sum_t \mathrm{CE}_{jhk} \mathrm{ce}^s_{jhkt} \\ & + \sum_{k \in \mathbf{K}} \sum_{j \in \mathbf{F}_k} \sum_{l \in \mathbf{F}_k, l \neq j} \sum_t \mathrm{CB}_{kjl} \mathrm{cb}^s_{jlkt} + \sum_{i \in \mathbf{E} \cup \mathbf{N}} \sum_{j \in \mathbf{F}_i} \sum_t \mathrm{CP}_{ij} q^s_{ijt}, \quad \forall s \in \mathbf{S} \end{aligned} \tag{2}$$

$$\rho_s = \sum_{i \in \mathbf{E} \cup \mathbf{N}} \sum_t \mathrm{CU}_i u^s_{it}, \quad \forall s \in \mathbf{S} \tag{3}$$

$$\sum_{j \in \mathbf{F}_i} x_{ij} \geq 1, \quad \forall i \in \mathbf{N} \tag{4}$$

$$\sum_{j \in \mathbf{F}_i} q^s_{ijt} + u^s_{it} = D^s_{it}, \quad \forall i \in \mathbf{E} \cup \mathbf{N},\ t \in \mathbf{T},\ s \in \mathbf{S} \tag{5}$$

$$\begin{aligned} c^s_{jkt} = {} & \mathrm{CA}_{jkt} + \sum_{h \in \mathbf{K}_j, h \neq k} \mathrm{ER}_{hk} \mathrm{ce}^s_{jhkt} - \sum_{h \in \mathbf{K}_j, h \neq k} \mathrm{ce}^s_{jkht} + \sum_{l \in \mathbf{F}_k, l \neq j} \mathrm{cb}^s_{ljkt} - \sum_{l \in \mathbf{F}_k, l \neq j} \mathrm{cb}^s_{jlkt}, \\ & \forall j \in \mathbf{F},\ k \in \mathbf{K}_j,\ t \in \mathbf{T},\ s \in \mathbf{S} \end{aligned} \tag{6}$$

$$\sum_{k \in \mathbf{K}_j} \sum_{l \in \mathbf{F}_k, l \neq j} \mathrm{cb}^s_{jlkt} \leq \mathrm{CBL}_j \sum_{k \in \mathbf{K}} \mathrm{CA}_{jkt}, \quad \forall j \in \mathbf{F},\ t \in \mathbf{T},\ s \in \mathbf{S} \tag{7}$$

$$\sum_{h \in \mathbf{K}_j, h \neq k} \mathrm{ce}^s_{jkht} \leq \mathrm{CEL}_k \mathrm{CA}_{jkt}, \quad \forall j \in \mathbf{F},\ k \in \mathbf{K}_j,\ t \in \mathbf{T},\ s \in \mathbf{S} \tag{8}$$

$$\sum_{i \in \mathbf{E} \cup \mathbf{N}} \mathrm{PT}_{ik} q^s_{ijt} \leq c^s_{jkt}, \quad \forall j \in \mathbf{F},\ k \in \mathbf{K}_j,\ t \in \mathbf{T},\ s \in \mathbf{S} \tag{9}$$

$$\mathrm{PFA}_{ij} = \mathrm{pf}^s_{ijt} + \sum_{l \in \mathbf{F}_i, l \neq j} z^s_{ijlt}, \quad \forall i \in \mathbf{E},\ j \in \mathbf{F}_i,\ s \in \mathbf{S},\ t = 1 \tag{10}$$

$$\mathrm{pf}^s_{ijt} = 0, \quad \forall i \in \mathbf{N},\ j \in \mathbf{F}_i,\ s \in \mathbf{S},\ t \leq \mathrm{LS}_i \tag{11}$$

$$x_{ij} = pf^{s}_{ijt} + \sum_{l\in\mathbf{F}_i, l\neq j} z^{s}_{ijlt}, \quad \forall i\in\mathbf{N},\ j\in\mathbf{F}_i,\ s\in\mathbf{S},\ t=\mathrm{LS}_i+1 \tag{12}$$

$$\mathrm{pf}^{s}_{ij,t-1} = \mathrm{pf}^{s}_{ijt} + \sum_{l\in\mathbf{F}_i, l\neq j} z^{s}_{ijlt}, \quad \forall i\in\mathbf{E},\ j\in\mathbf{F}_i,\ s\in\mathbf{S},\ 1<t\leq\mathrm{LT}_i \tag{13}$$

$$\begin{gathered}\mathrm{pf}^{s}_{ij,t-1} + \sum_{l\in\mathbf{F}_i, l\neq j} z^{s}_{ilj,t-\mathrm{LT}_i} = \mathrm{pf}^{s}_{ijt} + \sum_{l\in\mathbf{F}_i, l\neq j} z^{s}_{ijlt}, \\ \forall i\in\mathbf{E},\ j\in\mathbf{F}_i,\ s\in\mathbf{S},\ t>\mathrm{LT}_i,\ t\neq\mathrm{LS}_i+1\end{gathered} \tag{14}$$

$$\begin{gathered}\mathrm{pf}^{s}_{ij,t-1} + y^{s}_{ij} + \sum_{l\in\mathbf{F}_i, l\neq j} z^{s}_{ilj,t-\mathrm{LT}_i} = \mathrm{pf}^{s}_{ijt} + \sum_{l\in\mathbf{F}_i, l\neq j} z^{s}_{ijlt}, \\ \forall i\in\mathbf{E},\ j\in\mathbf{F}_i,\ s\in\mathbf{S},\ t>\mathrm{LS}_i+1\end{gathered} \tag{15}$$

$$\mathrm{PFA}_{ij} + y^{s}_{ij} \leq 1, \quad \forall i\in\mathbf{E},\ j\in\mathbf{F}_i,\ s\in\mathbf{S} \tag{16}$$

$$q^{s}_{ijt} \leq \mathrm{pf}^{s}_{ijt} D^{s}_{it}, \quad \forall i\in\mathbf{E}\cup\mathbf{N},\ j\in\mathbf{F}_i,\ s\in\mathbf{S},\ t\in\mathbf{T} \tag{17}$$

The objective is to minimize the expected costs and penalties under different scenarios, as shown in Eq. 1. Equation 2 displays the expected costs, consisting of mask creation costs, mask revision costs, production costs, capacity backup costs, and exchange costs. The costs arising from demand allocation decisions mainly entail mask creation and revision costs. Mask creation costs occur because of the NTO allocation decisions and product sharing decisions. Mask revision costs correlate with the transfer of products. The cost of capacity decisions involves backup and exchange to adjust the capacity configuration of fabs.

The penalty function penalizing unmet demand is as Eq. 3. The penalty for unmet demand varies between products, to represent that the importance of each product is not identical. The penalty occurs in both new products and existing products, whenever a demand is not fulfilled.

Equation 4 displays the restriction that every NTO should be allocated to at least a fab. In addition, each NTO would possibly be assigned to more than one fab, which in actuality constitutes the sharing decisions of NTOs. Equation 5 assures the demand consistency, which is intended to separate the total demand of each product into satisfied demand, and unmet demand. The constraint assures the satisfied demand, while the unmet demand will equal total demand for every product, including both NTOs and existing demands in every period. This feature restricts demand from reaching total satisfaction because the overall capacity might be insufficient to fulfill demand. It avoids the unfeasibility of the model, and provides the base for the unmet demand penalty function.

Equation 6 defines available capacity after utilizing the capacity backup and capacity exchange option. The available capacity of a technology changes when capacity is exchanged from other technologies for sustenance, experiencing a decrease when it supports other technologies. Furthermore, the capacity backup and exchange options can also cause the available capacity rate to rise or descend. When different technologies exchange the capacity, the amount of capacity received might not equate to the amount donated from other technologies because it is multiplied by a capacity exchange rate. The received capacity could either be higher or lower than the amount donated. This is because the original capacity plan is calculated based on a certain product mix assumption. If the product mix realization does not follow the assumption, capacity will be different. However, setting up machines occupies little time, since the greatest alteration lies in the assumption.

Equations 7 and 8 provide the maximum proportion for the backup and exchange option to utilize capacity. Equation 7 requires that the amount of capacity backup be limited to a certain ratio. Since the backup uses the capacity of one fab to support another, to be without limitation would be unfeasible for preventing depletion of the capacity of a fab. Capacity backup usually transfers tools between the fabs to change capacities, and the limitation signifies that only part of these tools can be transferred to support other fabs. Equation 8 also limits capacity exchange decisions, to prevent exhausting the capacity of one technology to sustain another.

Equation 9 restricts the total production volume to less than the available capacity after capacity backup and exchange. It ensures that the production volume remains in the range achievable by fabs, while product allocation and capacity backup approaches.

The feasibility constraints (10–13) limit the ability of fabs to manufacture products. Equation 10 states that the ability to manufacture existing products in the first period is dependent on the original allocation, and can be altered by product transfer decision. When a transfer occurs, the product is removed from fab j and relocated to fab l (i.e., the masks are moved from fab j to fab l), without involving other fabs. Equations 11 and 12 are the constraints for NTOs. NTO allocation affects the ability to manufacture products, and the ability of fabs to produce them is only after a lead time to qualify, which is equal to the time period required to share products. Equation 13 states that production feasibility remains as constant as the previous period when t is smaller than LT_i. The only possibility to alter the feasibility is to transfer the product to another fab.

Equation 14 represents the production feasibility that will be affected by receiving a product transferred from another fab. Equation 15 considers product-sharing decisions for existing products, meaning that if a product i is shared to a fab, the influence on the feasibility of the fab to manufacture it will be realized after a lead time LS_i. Finally, Eq. 16 restrains a product from being assigned to a fab which is not able to produce it. Equation 17 limits the production volume to less than the demand if the fab is capable of producing it. Otherwise, the production quantity will be equal to zero. The number of total binary variables equals to the sum of number of NTO allocation, transfer, share, and capability decision variables, or equivalently $\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{F}_i} 1+\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{E}}\sum_{j\in\mathbf{F}_i} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{E}\cup\mathbf{N}}\sum_{j,l\in\mathbf{F}_i,j\neq l} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{E}\cup\mathbf{N}}\sum_{j\in\mathbf{F}_i} 1$. The number of total continuous variables equals to the sum of number of production quantity, unmet demand, amount of capacity exchanged, and amount of backup capacity decision variables, $\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{E}\cup\mathbf{N}}\sum_{j\in\mathbf{F}_i} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{N}} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{j\in\mathbf{F}}\sum_{k\in\mathbf{K}_j} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{j\in\mathbf{F}}\sum_{h,k\in\mathbf{K}_j,h\neq k} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{k\in\mathbf{K}}\sum_{j,l\in\mathbf{F}_k,j\neq l} 1$. On the other hand, the number of effective constraints is the number of constraints summing over Eqs. 4–17, i.e., $\sum_{i\in\mathbf{N}} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{E}\cup\mathbf{N}} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{j\in\mathbf{F}} 1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{j\in\mathbf{F}}\sum_{k\in\mathbf{K}_j} 3+\sum_{s\in\mathbf{S}}$

$\sum_{i\in\mathbf{E}}\sum_{j\in\mathbf{F}_i}2+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{F}_i}1+\sum_{t\in\mathbf{T}}\sum_{s\in\mathbf{S}}\sum_{i\in\mathbf{E}\cup\mathbf{N}}\sum_{j\in\mathbf{F}_i}1$. As the estimation of problem scale shown, the proposed model can only deal with limited number of products, technologies, fabs, scenarios, and times. When the problem size grows much larger, we need to incorporate with specialized methods, such as the L-shaped method (Birge and Louveaux 1997), to find solutions. On the other hand, most of stochastic integer programming problems for capacity decisions are *NP-hard* (Ahmed and Sahinidis 2003; Barahona et al. 2005). Therefore, efficient approximate solutions or heuristic strategy are particularly required for large-scale problems.

4 Numerical study

4.1 Experimental design

Based on real settings, the numerical study conducts experiments to evaluate the performance of the stochastic model against the deterministic model. The proposed models are solved by LINGO on a PC, with Intel Core 2 Duo 3.00 GHz CPU and 3.24 GB RAM. The synthetic data include three technologies, three fabs, and thirty products over a 6-quarter planning horizon, i.e., $t = 1, 2, \ldots, 6$, generated according to the opinions of the domain expert, to simulate actual scenarios. Among the three fabs, only Fab 2 is capable of processing all three technologies, whereas the other two fabs are incapable of processing one technology, e.g., as shown in Table 1, Fab 1 and Fab 3 cannot process Technology 3 and Technology 1, respectively. In addition, five of the thirty products are NTOs to be allocated, and the other 25 are existing products. More specifically, in Tables 2 and 3, products D01 to D05 are NTOs, while D06 to D30 are existing products. The tables provide corresponding technologies, allocations of existing products, and demand forecasts. The setting does not include the case of qualifying a product for more than one fab to reflect a leading company who has utilized its flexible clustering fabs.

Table 1 Capacity settings (CA_{jkt}) (unit: 1,000 wafers)

Fab	Technology	Quarter					
		1	2	3	4	5	6
1	1	35	34	34	36	37	37
	2	31	33	35	35	36	36
	3	–	–	–	–	–	–
2	1	31	29	31	32	33	32
	2	28	28	30	29	31	30
	3	28	27	30	30	32	33
3	1	–	–	–	–	–	–
	2	33	33	36	36	37	36
	3	29	29	32	31	33	32

Table 2 Demand setting for NTOs $\left(D_{it}^{s}, i \in \mathbf{N}\right)$ (unit: 1,000 wafers)

Product	Technology	Share lead time	Transfer lead time	Quarter					
				1	2	3	4	5	6
D01	1	2	1	0.0	0.0	5.0	6.7	8.6	9.8
D02	2	2	2	0.0	0.0	6.0	7.4	8.8	10.1
D03	2	1	1	0.0	0.0	6.2	8.1	9.9	10.4
D04	2	2	1	0.0	0.0	0.0	5.5	7.9	9.0
D05	3	2	1	0.0	0.0	0.0	6.0	7.5	8.5

Table 3 Demand setting for existing products $\left(D_{it}^{s}, i \in \mathbf{E}\right)$ (unit: 1,000 wafers)

Product	Technology	Fab	Share lead time	Transfer lead time	Quarter					
					1	2	3	4	5	6
D06	1	1	1	1	5.4	3.5	0.0	0.0	0.0	0.0
D07	2	1	1	1	7.2	5.2	3.5	0.0	0.0	0.0
D08	3	3	1	1	4.9	4.1	3.6	0.0	0.0	0.0
D09	2	2	2	1	9.9	7.5	6.8	5.0	0.0	0.0
D10	1	1	1	1	7.2	5.0	3.9	3.6	3.0	0.0
D11	3	2	2	1	12.0	9.8	6.6	6.2	4.5	0.0
D12	1	2	3	2	12.3	11.2	8.9	8.8	7.5	7.5
D13	1	2	1	1	10.4	9.1	8.4	7.5	7.6	6.8
D14	2	1	2	1	9.8	8.9	8.0	8.2	7.0	6.7
D15	1	2	1	1	10.9	9.6	11.1	10.9	10.7	8.1
D16	2	3	1	1	14.1	15.3	13.3	12.0	12.4	11.1
D17	2	3	2	1	11.8	10.5	10.1	9.4	9.7	8.9
D18	3	3	1	1	10.9	10.2	10.0	8.7	9.4	9.4
D19	2	3	1	1	12.5	12.4	14.4	12.2	11.6	10.9
D20	3	3	1	1	9.2	9.4	11.4	9.9	8.3	9.3
D21	1	1	1	1	6.3	7.5	7.1	6.7	7.4	7.3
D22	2	1	2	2	12.0	13.7	14.8	15.0	14.7	14.1
D23	2	2	1	1	9.5	10.4	10.8	10.0	10.6	11.5
D24	3	2	1	1	11.6	12.1	13.0	12.7	14.2	15.3
D25	1	1	1	1	6.5	7.4	8.2	9.6	7.5	8.6
D26	3	3	1	1	6.1	7.5	9.9	9.3	11.2	10.3
D27	1	1	3	2	5.0	5.9	6.7	8.4	9.2	10.7
D28	1	2	2	2	0.0	5.0	5.8	7.6	9.5	10.8
D29	3	2	2	1	0.0	5.5	7.6	9.5	10.7	12.5
D30	2	2	1	1	0.0	7.3	7.5	8.6	9.4	10.8

The demand of existing products is generated with the assumption that several products will opt out of the time period while a number of them with stable demand survive. For instance, in Table 3, D06 to D11 are products that will opt out of the time period in the planning horizon, D12 to D19 are in a decreasing trend, and D20 to D30 are experiencing an increase in demand. The aggregated demand of each technology is set at near capacity, since the capacity is assumed to expand according to the demand forecast, shrinking the gap between demand and capacity.

Expenses required for planning include the mask cost, production cost, capacity backup cost, capacity exchange cost, and the penalty for unmet demand. The costs are set under the suggestion of domain experts to make their relative scale reasonable. The costs for creating a new set of masks and revising existing masks are different because only parts of the masks require reworking. The cost of new masks is incurred in conjunction with NTO allocation or the sharing of a product, and the revision cost occurs with the transference of a product. The costs also vary with the technologies required to manufacture the product. Costs of creating masks increase when more advanced technology is necessary. Table 4 shows the costs of creating new masks and revising them. The production costs change depending on the technology and fabs, since the capability of each fab is not identical. The costs for advanced technologies are also higher than mature technologies. Table 5 lists the production costs. In addition, the penalty of unmet demand is set at double the rate of the average production cost, i.e., $1,770 for Technology 1, $2,300 for Technology 2, and $2,950 for Technology 3. The costs for capacity backup and capacity exchange are assumed to be proportionate to the production costs. Capacity backup entails transportation costs between the fabs, and the backup cost is assumed to equal 10% of the production cost. Capacity exchange does not calculate actual costs because since the tools do not require transportation, the costs only involve a number of management issues, and is thus set as 3% of production costs.

Tables 6 and 7 show the capacity backup and exchange costs. Additional information required includes the capacity exchange rate, the limit of capacity

Table 4 Mask costs setting ($USD per mask set)

Option	Technology 1	Technology 2	Technology 3
New masks (CM_i)	300,000	600,000	800,000
Revisions (CT_i)	120,000	280,000	350,000

Table 5 Production costs setting (CP_{ij}) ($USD)

Fab	Technology 1	Technology 2	Technology 3
1	850	1,100	–
2	920	1,200	1,500
3	–	1,150	1,450

Table 6 Backup costs setting (CB_{jk}) ($USD per backup)

Fab	Technology 1	Technology 2	Technology 3
1	85	110	–
2	92	120	150
3	–	115	145

Table 7 Exchange costs setting (CE_{jh}) ($USD per exchange)

Fab	Technology 1	Technology 2	Technology 3
1	25.5	33	–
2	27.6	36	45
3	–	34.5	43.5

Table 8 Capacity exchange rate setting (ER_{hk})

Technology	1	2	3
1	1	0.8	0.6
2	1.2	1	0.7
3	1.3	1.2	1

Table 9 Lead time setting for share and transfer

Technology	Share (LS_i)			Transfer (LT_i)	
	1Q	2Qs	3Qs	1Q	2Qs
1	0.8	0.2	0.0	0.9	0.1
2	0.6	0.3	0.1	0.8	0.2
3	0.1	0.7	0.2	0.5	0.5

backup and exchange, and the lead time to share or transfer products. The limit of capacity backup is set as 15% for each fab, and the limit of exchange is set as 30% for Technology 1, 25% for Technology 2, and 20% for Technology 3. Table 8 shows the capacity exchange rate. Table 9 presents a randomly generated lead time set for each product to be shared and transferred, according to the probability suggested by the domain expert, to be applied in each simulation instance. An additional constraint for generating lead time is that sharing occupies more time than transferring. The costs and lead time are also set under the suggestion of the domain expert, to make their relative scale more relevant. NTO allocation involves critical planning in the strategy level. Most of the input data are sensitive and confidential. To oblige the non-disclosure agreement, all presented data have been transformed to relative values, which will not affect the generality for further explanation.

A scenario presents the future demand of each product during the planning horizon. To evaluate the performance of the stochastic model, demand scenarios are created to enable the creation of new profiles, by multiplying the original demand of each product against the volume and variation factors. The factors are randomly generated, and set at three levels: high, medium, and low. For the volume factor, *Fvol*, the high level has a range between 105 and 130% of the original demand volume. The range of medium volume is between 80 and 105%, and the range of low volume is between 55 and 80%. For the variation factor, *Fvar*, the high-level range is between 75 and 125% of the original demand. The ranges for medium level and low level are between 85 and 115%, and between 95 and 105%, respectively.

The combination of volume and variation enable the creation of new scenarios. According to the levels set for these factors, random numbers are generated for each product in a demand scenario, i.e., the demand in the forecasting horizon of each product is multiplied by a demand volume factor and a demand variation factor. For instance, the generated demand for Product D19 in the high volume level and medium variation level is (12.5, 12.4, 14.4, 12.2, 11.6, 10.9) × *Fvol* × *Fvar* where *Fvol* and *Fvar* are random numbers generated from [1.05, 1.30] and [0.85, 1.15], respectively. These factors change the relative demand volume among products, but the increasing or decreasing trend remain unaffected.

Based on these settings, the values of applying the two-stage SP model in NTO decisions will be evaluated from two different angles. Firstly, among different demand variation levels, the differences between EVPIs and the VSSs will be examined, while considering different demand volume levels as scenarios of which probabilities are set equally, e.g., 1/3. For the high demand variation level, for example, we will measure the corresponding EVPI and the VSS, with respect to three scenarios, including high, medium, and low volume levels, denoted by $\mathbf{S}\mu = \{\mathrm{H}\mu, \mathrm{M}\mu, \mathrm{L}\mu\}$ with corresponding probabilities $P_{\mathrm{H}\mu} = P_{\mathrm{M}\mu} = P_{\mathrm{L}\mu} = 1/3$. Secondly, EVPIs and the VSSs will be measured under different demand variation scenarios, say $\mathbf{S}\sigma = \{\mathrm{H}\sigma, \mathrm{M}\sigma, \mathrm{L}\sigma\}$ with $P_{\mathrm{H}\sigma} = P_{\mathrm{M}\sigma} = P_{\mathrm{L}\sigma} = 1/3$ for each demand volume level. Every problem for each demand volume or variation level replicates 10 instances. After monitoring the experiment results, regarding the above setting, the coefficient of variation of the average total cost of ten instances for each solution is less than 5%, one of the common simulation convergence criteria (Mackie and Cooper 2009).

4.2 Results and discussion

This section discusses the decisions under different scenario realizations, to estimate the validity of the proposed approach. Tables 10 and 11 summarize the VSS in the numerical study, which shows that SP is more beneficial when demand volume and variation are set at a high level. Table 12 shows the EVPI under different levels of variation and volume, and indicates that perfect information is more valuable when the variation is high, compared to the change in volume. Based on the summarized results, reasons why SP models are useful in high demand variation environments should be studied. This can be accomplished by analyzing the cost structure comparisons under different scenarios (Table 13), along with the portion of different decisions, including transfer, share, backup, exchange, and production

Table 10 VSS under different levels of variation ($USD for six quarters)

Variation	VSS under different volumes			VSS
	High	Medium	Low	
High	148,627	37,665	140,223	108,838
Medium	69,945	41,397	81,740	64,360
Low	107,737	64,067	49,235	73,679

Table 11 VSS under different levels of volume ($USD for six quarters)

Volume	VSS under different variations			VSS
	High	Medium	Low	
High	35,398	62,006	31,673	43,026
Medium	34,110	18,037	36,391	29,513
Low	5,151	16,308	30,598	17,352

Table 12 Average EVPI of SP model ($USD for six quarters)

Level demand	High	Medium	Low
Volume	68,889	37,119	14,292
Variation	216,329	103,129	168,619

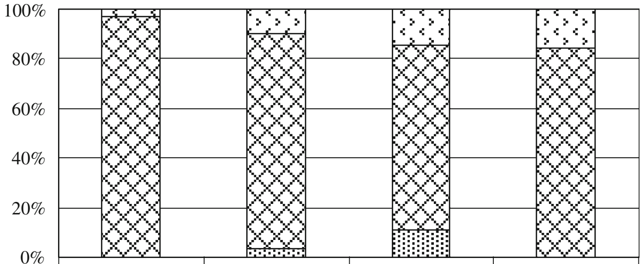
(Figs. 1, 2, 3). The coefficients of variation for total costs are around or lower than 50%.

The cost values in Table 13 represent the differences between the HN solutions of 2-SSP models, DM solutions of deterministic models lacking perfect information, and WS solutions of deterministic models. In particular, a scenario depicted by a “scenario” means that the HN solution provides the objective values. The smaller the value in Table 13, the lesser the costs compared to WS solutions. Table 13 shows that when scenario realization is in the high demand volume level, SP suggests the HN solutions with higher unmet penalties, higher backup costs, lower exchange costs, lower production costs, while the costs of sharing and transferring remain unchanged. In Fig. 1, the WS solutions adopted in the high demand scenario mainly involve capacity backup and exchange, with no actions on share and transfer options. A similar decision pattern can be discerned from the HN solutions. The reason that share and transfer decisions are seldom made when the demand volume is relatively high is that the act of sharing and transferring products accomplishes little under this situation, since the problem falls under the shortage of overall capacity. High demand volume renders capacity almost fully utilized, and only little capacity is left idle, therefore, by using the backup and exchange options, the capacity can be utilized to sustain the unmet demand. Since the available capacity is limited, the cost of backup and exchange is equal to the cost of sharing or transferring products, and thus, its priority is higher.

When the demand is at a medium volume, almost every approach is employed, while the capacity decisions still assume a greater proportion (Fig. 2). In this situation, overall capacity mostly suffices to fulfill all demand. However, due to the imbalance of demand distributed among fabs and technologies, the capacity configuration must be adjusted to fit the demand profile. Capacity backup and exchange are more popular, since their execution is simple and their costs are lower due to the short lead time before the decisions take effect. In addition, some sharing and transferring decisions are also made to alter the future demand profile, to alleviate the imbalance of demand. Although the HN solutions suggest additional backup actions, thereby yielding extra backup costs, these further costs cancel out because all the other costs have been lessened.

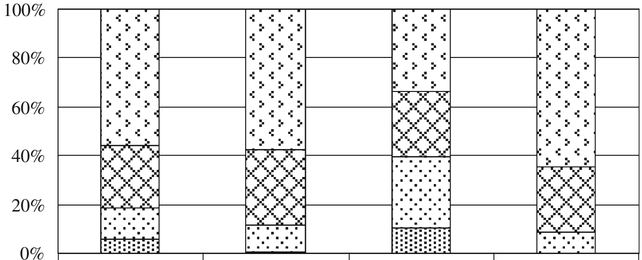
Table 13 Cost structure of optimal decisions compared with perfect information under different scenarios in the case of high demand variation ($USD for six quarters)

Scenario realization	High volume			Medium volume			Low volume		
Scenario considered	Medium	Low	Scenario	High	Low	Scenario	High	Medium	Scenario
Share costs	0	0	0	120,000	680,000	−30,000	90,000	90,000	90,000
Transfer costs	84,000	306,000	0	196,000	351,000	−12,000	666,000	514,000	374,000
Backup costs	167,414	322,716	281,019	223,672	−674,889	585,093	−118,191	−161,416	−161,416
Exchange costs	−484,819	−533,178	−721,072	−56,122	−83,517	−6,997	−55,984	15,789	−29,676
Production costs	−821,363	−1,176,018	−1,354,837	−34,532	375,391	−462,664	182,541	115,429	180,242
Unmet penalty	1,242,951	1,638,490	1,935,659	−118,515	−136,338	−18,329	0	0	0
Total costs	188,182	558,010	140,770	330,503	511,647	55,102	764,366	573,802	453,149
Coefficient of variation for total costs	0.48	0.36	0.32	0.41	0.21	0.17	0.52	0.33	0.29



	High Volume	Medium Volume	Low Volume	Scenario
⊡ Backup	3%	10%	14%	16%
⊠ Exchange	97%	87%	75%	84%
⊡ Share	0%	0%	0%	0%
▦ Transfer	0%	3%	11%	0%

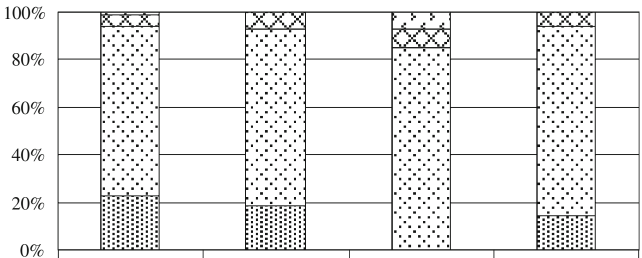
Fig. 1 Decisions under the high demand volume level



	High Volume	Medium Volume	Low Volume	Scenario
⊡ Backup	56%	57%	34%	65%
⊠ Exchange	26%	31%	26%	27%
⊡ Share	13%	11%	29%	9%
▦ Transfer	6%	0%	10%	0%

Fig. 2 Decisions under the medium demand volume level

Figure 3 depicts counter-intuition results. When demand is low, few capacity backup and exchange decisions are necessary, and more share and transfer decisions are made. In this case, demand can be fulfilled with the existing capacity configuration. Backup and exchange can be beneficial for fulfilling more demand, but the effect is less significant than sharing and transferring. Sharing and transferring products help reduce costs in a low demand scenario because the cost saved by manufacturing products in a fab with lower production costs is higher than the cost for sharing or transferring. This situation occurs when production unit costs between fabs experience significant difference, thereby enabling the saved production cost to cover the mask costs when the demand volume of a particular product is high enough. In the synthetic scenario, this condition is valid for a number of products, hence, the decision of sharing and transferring are used extensively in the low demand scenario. The importance of NTO allocation can also be gauged by the decisions made under the low demand scenario because decision



	High Volume	Medium Volume	Low Volume	Scenario
□ Backup	1%	0%	7%	0%
☒ Exchange	5%	7%	8%	6%
⊡ Share	71%	74%	85%	79%
▦ Transfer	23%	19%	0%	14%

Fig. 3 Decisions under the low demand volume level

makers are likely to reject sharing and transferring proposals after considering similar situations. In this circumstance, the HN solutions rearrange costs to outperform DM solutions without perfect information, and their contributions are marginal.

The aforementioned findings suggest that the success of 2-SSP models rely on its ability to provide similar solution patterns with WS solutions. The run time for a deterministic model ranges from 2 to 8 s. For stochastic models with 6,720 binary variables, 4,570 continuous variables, and 13,580 constraints, the run time is between 10 and 20 min when both demand volume and demand variation factors are set at high or low levels. When one of the factors is set at medium level for a stochastic model, up to 28 h are required to obtain results. The medium level settings represent the situation when capacity and demand are close. Therefore, more practicability and trade-offs should be inputted during optimization, yielding longer computation time.

5 Conclusion

To enhance capacity utilization and improve demand fulfillment, this research has developed a 2-SSP approach to support NTO allocation decisions for demand fulfillment planning, to achieve a robust solution under uncertain conditions. The first stage considers the NTO allocation decisions, before the corresponding demand volume is realized. The second stage allocates the capacity to meet the demand requirement based on the results of four options of capacity reconfiguration, including share, transfer, backup, and exchange. The results employing realistic data have validated practical viability and decision quality of this approach under different scenarios, which are more capital effective than present deterministic solutions. Adopting 2-SSP models in a high demand variation environment is beneficial because the 2-SSP can provide solution patterns similar to optimal solutions of deterministic models with perfect information.

To implement the proposed model, data used in the model should be carefully maintained to ensure that the decision is made based on credible information. In practice, different departments can participate in the decision-making process regarding the allocation of products and the planning of capacity. Inter-departmental communication is therefore a significant trait for intelligence sharing and making informed decisions, though a lot of time and manpower must be allocated for the attainment of such a structure. In addition, because different departments would make dissimilar decisions, reaching optimized decisions for the benefit of the company would remain a challenge. The proposed model integrates the decision-making process among departments, to reach solutions deemed most profitable for a company, with the ability to replace the original manual labor. In addition, the stochastic model possesses the ability to function under uncertain conditions. The implementation of the proposed approach can save the case company allotted manpower while shortening decision-making time, and still able to make the most reliable choice, which is apt to achieve the optimization required.

This study focuses on a make-to-order production system. However, in the make-to-stock environment, e.g., in memory IC fabs, inventory may be used to fulfill the demand of different companies. Some products might also be manufactured in advance when the forecasted demand is high. Thus, inventory and backlog policy can also be modeled, with the risk of overproduction considered. Many NTOs are abandoned after qualification, i.e., only several new products that pass for qualification will yield new demand, since a good number of products are abandoned after the qualification process. The ratio of successful NTOs entering the market, called hit rate in practice, can be utilized to gain a more accurate estimation of demand. Customer opinions can also be additional inputs when allocating products. Furthermore, the model can incorporate customer opinions and priorities to improve customer satisfaction. Sensitivity analysis can be performed on the upper limits of capacity backup and exchange, to examine the benefits of changing the number of dedicated tools. To prevent long run time when solving the stochastic model, special solution techniques, such as the Lagrangian relaxation or decomposition-based branch-and-bound strategies, such as the L-shaped method, can be employed to solve problems within a reasonable timeframe.

Acknowledgments This study was partially supported by National Science Council, Taiwan (NSC97-2221-E-007-111-MY3; NSC100-2410-H-031-011-MY2; NSC99RA13).

References

Ahmed S, Sahinidis N (2003) An approximation scheme for stochastic integer programs arising in capacity expansion. Oper Res 51(3):461–471

Barahona F, Bermon S, Günlük O, Hood S (2005) Robust capacity planning in semiconductor manufacturing. Nav Res Logist 52(5):459–468

Bard JF, Srinivasan K, Tirupati D (1999) An optimization approach to capacity expansion in semiconductor manufacturing facilities. Int J Prod Res 37(15):3359–3382

Bilgin S, Azizoglu M (2009) Operation assignment and capacity allocation problem in automated manufacturing systems. Comput Ind Eng 56:662–676

Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer, New York

Bitran GR, Tirupati D (1989) Capacity planning in manufacturing networks with discrete options. Ann Oper Res 17:119–136

Chen ZL, Li S, Tirupati D (2002) A scenario-based SP approach for technology and capacity planning. Comput Oper Res 29(7):781–806

Chien CF, Zheng J (2011) Mini-max regret strategy for robust capacity expansion decisions in semiconductor manufacturing. J Intell Manuf. doi:10.1007/s10845-011-0561-1

Chien CF, Chen Y, Peng J (2010) Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle. Int J Prod Econ 128(2):496–509

Christie RME, Wu SD (2002) Semiconductor capacity planning: stochastic modeling and computational studies. IIE Trans 34(2):131–143

Chung SH, Huang CY, Lee AHI (2006) Capacity allocation model for photolithography workstation with the constraints of process window and machine dedication. Prod Plan Control 17:678–688

Fleischmann B, Ferber S, Henrich P (2006) Strategic planning of BMW's global production network. Interfaces 36(3):194–208

Geng N, Jiang Z (2009) A review on strategic capacity planning for the semiconductor manufacturing industry. Int J Prod Res 47(13):3639–3655

Geng N, Jiang Z, Chen F (2009) Stochastic programming based capacity planning for semiconductor wafer fab with uncertain demand and capacity. Eur J Oper Res 198(3):899–908

Huang K, and Ahmed S (2005) The value of multi-stage stochastic programming in capacity planning under uncertainty. Online http://www2.isye.gatech.edu/sahmed/SCPP9.pdf. Accessed 10 March 2011

Inman RR, Gonsalvez DJA (2001) A mass production product-to-plant allocation problem. Comput Ind Eng 39(3):255–271

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. Manage Sci 41(4):577–594

Kauder S, Meyr H (2009) Strategic network planning for an international automotive manufacturer. OR Spectr 31(3):507–532

Kim S, Uzsoy R (2008) Exact and approximate algorithms for capacity expansion problems with congestion. IIE Trans 40:1185–1197

Leachman RC, Ding SW, Chien CF (2007) Economic efficiency analysis of wafer fabrication. IEEE Trans Autom Sci Eng 4(4):501–512

Mackie KR, Cooper CD (2009) Landfill gas emission prediction using Voronoi diagrams and importance sampling. Environ Model Softw 24:1223–1232

Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8):114–117

Mouli C, Winstead CH (2007) Tapeout Execution System (TES), a key enabler of DFM/co-optimization. International Symposium on Semiconductor Manufacturing (ISSM) DM-O-165

Pahl J, Voß S, Woodruff DL (2007) Production planning with load dependent lead times an update of research. Ann Oper Res 153(1):297–345

Wang KJ, Wang SM, Yang SJ (2007) A resource portfolio model for equipment investment and allocation of semiconductor testing industry. Eur J Oper Res 179(2):390–403

Wu JZ, Chien CF (2008) Modeling strategic semiconductor assembly outsourcing decision based on empirical settings. OR Spectr 30(3):401–430

Wu SD, Erkoc M, Karabuk S (2005) Managing capacity in the high-tech industry: a review of literature. Eng Econ 50(2):125–158

Author Biographies

Chen-Fu Chien is a Tsinghua Distinguished Professor in National Tsing Hua University (NTHU). He received BS (with Phi Tao Phi Honor) with double majors in Industrial Engineering and Electrical Engineering from NTHU. He received MS of Industrial Engineering and PhD of Operations Research and Decision Sciences from UW-Madison. He was a Fulbright Scholar in UC Berkeley and also received the PCMPCL Training in Harvard Business School. From 2005 to 2008, he had been on-leave as the Deputy Director of Industrial Engineering Division in Taiwan Semiconductor Manufacturing Company (TSMC), which is the world largest semiconductor foundry. His research efforts center on decision analysis, modeling and analysis for semiconductor manufacturing, manufacturing strategy, and data mining. Dr. Chien has received six invention patents on semiconductor manufacturing and published two books, more than 90 journal papers and a number of case studies in Harvard Business School. He has been invited to give keynote and plenary talks in various conferences including IEEM, APIEMS, C&IE, IML and leading universities worldwide. His keynote speech in IEEM 2006 was then invited by IIE to publish as a feature article in *Industrial Engineer* (2007 Feb, pp. 47–49). Dr. Chien received the Distinguished Research Award and Tier 1 Principal Investigator (Top 3%) from NSC, Distinguished University-Industry Collaborative Research Award from the Ministry of Education, University Industrial Contribution Awards from the Ministry of Economic Affairs, Distinguished University-Industry Collaborative Research Award and Distinguished Young Faculty Research Award from NTHU, Distinguished Young Industrial Engineer Award, Best IE Paper Award, and the IE Award from Chinese Institute of Industrial Engineering (CIIE), Best Engineering Paper Award and Distinguished Engineering Professor by Chinese Institute of Engineers in Taiwan. He is now an Associate Editor of *IEEE Transactions on Automation Science and Engineering*, Area Editor of *Flexible Services and Manufacturing Journal*, and an Advisory Board Member of *OR Spectrum.*

Jei-Zheng Wu is an Assistant Professor at Department of Business Administration, Soochow University (SCU), Taiwan. He received his PhD and MS in Industrial Engineering and Engineering Management from NTHU. He received BS with double majors in Business Administration and Mathematics from National Taiwan University. His professional experience includes Adjunct Professor at NTHU, Yuan Ze University, Postdoctoral researcher at NTHU, and co-op at IBM Thomas J. Watson Research Center. He received the Best Paper Award at the CIIE 2010 Annual Meeting and the Young Scientist Prize at the Intelligent Manufacturing & Logistics Systems International Conference in 2008. His main research interests include manufacturing strategy, supply chain management, decision analysis, meta-heuristics, and decision support systems. He has served as Guest Editor for a number of journals including *OR Spectrum, Flexible Services and Manufacturing Journal, and Journal of Quality.*

Chien-Cheng Wu is an engineer at TSMC, Taiwan. He received his MS in Industrial Engineering and Engineering Management from NTHU. His main research interests include modeling and analysis for semiconductor manufacturing, production management, and decision analysis.