An Order Fulfillment Model With Periodic Review Mechanism in Semiconductor Foundry Plants

Chi Chiang and Hui-Lan Hsu

***Abstract*—In today's globalization competition, manufacturing firms are using an order fulfillment system to give available-to-promise (ATP) capacity efficiently. An ordinary order fulfillment system will plan capacity based on forecasts and assign ATP quotas to incoming orders. Its basic idea is to enhance capacity utilization and avoid poor customer service. However, in the semiconductor industry, demand is highly volatile, and a make-to-order (MTO) manufacturer often runs the risk of cancelled committed demands. In this research, we propose an integrated order fulfillment model for a MTO semiconductor foundry fab to maximize corporate profit. Specifically, we suggest a periodic allocation review mechanism to reallocate unused ATP quotas. We examine the model performance based on different data sets. Results showed that capacity utilization and profitability are improved substantially with the periodic review mechanism, especially when demand forecast is not reliable.**

***Index Terms*—Available-to-promise, linear programming, management information systems, order fulfillment, production management.**

I. Introduction

In the semiconductor foundry industry, fabrication is activated in response to an actual order, i.e., companies run make-to-order (MTO) operations, and inventory cannot be used to smooth the demand as in make-to-stock (MTS) manufacturing. In this industry, over billions of capital expenditure is invested per year and capacity planning and utilization are vital to the success of a company. One of the biggest challenges is to avoid that a high-margin customer cannot obtain required capacity because a low-margin customer has booked the capacity earlier. As the average wafer fabrication flow time is around three months, it is not practical for customers to wait this long and then be informed about the exact delivery date. The ability to respond quickly to customer orders is important in gaining the competitiveness in this industry [1]. Capacity planning managers often use software such as JDA i2 or SAP APO to allocate capacity to customers based on their demand forecasts. After demand is confirmed with a customer, the corresponding available-to-promise (ATP) capacity is committed accordingly. The ATP system follows to promise orders with these committed ATP quotas. An effective order fulfillment system will help a semiconductor manufacturer plan for ATP capacity and assign them to incoming orders in a timely manner; it also enables a manufacturer to accurately estimate the order completion date [2]. Customers can thus obtain the ordered products on time and start the following production or promotion plans.

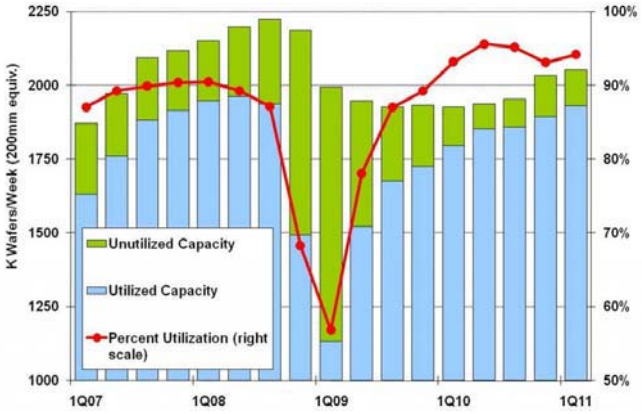

Fig. 1. Capacity utilization of the semiconductor industry (source: silicon semiconductor [3]).

When long-term capacity falls short of demand, a foundry company can choose to increase capacity by building new plants and/or purchasing new equipment. However, in the short term when there is large demand and expanding capacity is impossible, an order fulfillment system should be in place for a foundry company to achieve high capacity utilization. Fig. 1 shows the utilized and unutilized capacity of the semiconductor industry from 2007 to 2010 according to Silicon Semiconductor [3], where capacity was measured by converting the various output into 200mm equivalent wafers. It is seen that the utilization rate is usually around 90%, except in the financial crisis time period of 2008/Q4 to 2009/Q2. If the utilization rate exceeds 90%, foundry capacity may become tough to plan for or manage [4], [5]. It seems from Fig. 1 that foundry companies constantly face the challenge of capacity planning.

Driven by Moore's Law, the semiconductor industry has continued technology migrations and wafer size enlargement to maintain technology innovation and cost reduction per transistor to penetrate into other segments for component substitution and thus achieve unparalleled growth [6]. Because of decreased line width, advanced or newer process technologies are more difficult and complex to be applied in new IC design. IC design failure can easily spoil an entire project and related capacity plans in IC design houses. Characterized

Manuscript received January 30, 2014; revised July 15, 2014; accepted July 18, 2014. Date of publication July 25, 2014; date of current version November 10, 2014.

The authors are with the Department of Management Science, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: adeline.ms98g@nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSM.2014.2342493

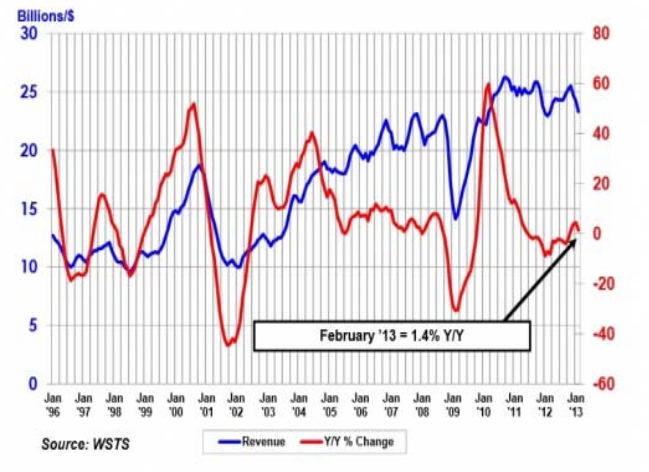

Fig. 2. World semiconductor trade statistics (source: semiconductor industry association [8]).

by short life cycle and advanced technologies, the current digital consumer era fluctuates rapidly as shown in Fig. 2. With less historical information available, forecasting customer demand for advanced technologies is more difficult than for old ones. Furthermore, demand forecasting is usually based on expert opinion and requires sales managers to visit customers constantly. To obtain sufficient capacity, customers may overstate their demand requirement. Chien *et al*. [7] conducted an empirical study and validated the practical viability of their proposed forecasting model for a leading semiconductor foundry company. They found that the highest and lowest mean absolute percentage errors (MAPE) of a major consumer product are 97.55% and 6.35%, respectively. It seems that foundry companies also face an increasing challenge in demand forecasting.

In this study, we propose a three-stage order fulfillment model, which includes allocation planning, order promising and periodic allocation review, to release unutilized capacity in the MTO environment. In the semiconductor foundry industry, capacity is usually categorized as committed and uncommitted after allocation planning [2]. As mentioned above, if committed capacity is not booked timely by a customer order, the capacity is not utilized and thus wasted. Also, the actual order quantity and its required delivery date may be different from those in the original demand forecast. These situations cannot be handled by the use of inventory as in MTS manufacturing but will considerably affect ATP capacity consumption. Also, in the real world, customer demand varies quickly but is updated often passively. Consequently, an active periodic review mechanism should be included in an order fulfillment model to reallocate the unutilized capacity to other coming orders. Such a review mechanism shall increase capacity utilization, especially when customer forecast is not reliable. The leftover "uncommitted capacity" should also be used to enhance the order fulfillment rate when promising customer orders. The proposed order fulfillment model can be employed either on the real-time order processing mode or on the batch processing mode. The contribution of this study is to present an integrated model to bridge the order fulfillment system with a legacy logistics collaboration platform (i.e., B2B platform) as well as initiate a reviewing mechanism which will release the uncommitted capacity to satisfy those orders without allocation. Potential benefit is the improved profit and capacity utilization as compared to an ordinary order fulfillment system without periodic allocation review.

The rest of this paper is organized as follows. In next section, we will review the literature and models relevant to our study. In Section III, we propose an order fulfillment model with an allocation review mechanism. In Section IV, we present some computational results. Finally, Section V concludes this paper.

II. Literature Review

Conventional ATP models refer to the mechanism of committing the inventory to customers [9]. They are commonly implemented in ERP systems and applied in the MTS environment. Nowadays, advanced ATP models are generally built based on availability of supply chain resources, including raw materials, work-in-process, finished goods, manufacturing and distribution capacity. Advanced models are practical in the MTO environment as well. In this section, we will review related ATP models in the MTO environment as well as different order processing modes in order fulfillment systems.

In the MTO environment, the uncertainty in the semiconductor industry is typically the timing and size of customer orders. Chen *et al*. [10] provided a quantity and due date quoting ATP model; Chen *et al*. [11] provided a batch ATP model with consideration of manufacturing flexibility and customer preference. They showed that the optimal batching interval size can bring the system closer to global optimality, but they did not consider the multi-factory operation. Jung *et al*. [12] devised an optimized ATP system for a MTO company. Also, Robinson and Carlson [13] presented a real time order promising model in a mix MTO and MTS manufacturing environment; Ebadian *et al*. [14] proposed a five-step decision structure for the order entry stage that improves production planning and profit in MTO environments. In addition, Zhang and Tseng [15] extended the Chen *et al*.'s ATP model [11] by considering customer flexibility in the order commitment process for high mix low-volume production. The above studies, however, concentrated only on ATP methods and algorithms and neglected the integration of allocation planning and ATP systems.

Kilger and Schneeweiss [2] classified order fulfillment situations and presented the simple rules that can be applied in both the allocation planning and ATP capacity consumption. Noh *et al*. [16] presented an approach for reserving capacity for urgent orders in a MTO system and showed its impact on system profit through a simulation experiment; Pibernik and Yadav [17] developed a service-level based MTO system to determine capacity reservation that could meet the due date requirement of important customers; Ponsignon and Mönch [18] proposed heuristic approaches for solving master planning problems in semiconductor manufacturing; Tsai and Wang [19] developed a multi-site three-stage ATP model with different cost structure that is appropriate for

MTO manufacturing. Also, Meyr [20] proposed a deterministic linear programming model for ATP allocation and consumption in the lighting industry and showed that customer segmentation can indeed improve profits substantially. Recently, Chiang and Wu [4] investigated an order fulfillment model with the joint effect of order size and margin in a MTO environment; Lin *et al*. [21] developed a scenario-based two-stage stochastic programming model to seek a capacity allocation and expansion policy which was robust to demand uncertainties. Kallrath *et al*. [22] introduced how raw silicon is transformed to microchips and how SAP APO is implemented in the semiconductor industry. However, reviewing the unutilized capacity and integrating SAP APO with a B2B collaboration platform has not been included in the design of SAP APO. Most of these studies assume that the demand forecast input is reliable, and focus on the optimization of ATP systems. Little attention has been paid to capacity utilization improvement if reliable forecasts are not available and some committed capacity is not consumed eventually. In this study, we propose a solution to the problem of releasing unutilized capacity in due time if reliable forecasts are not available by using an order fulfillment system with a periodic review mechanism.

There are also some research works on different processing modes for ATP systems. Customers usually expect an immediate response for their order query and are not willing to wait long for an order promise. However, companies can obtain the benefit of global optimization through batch processing if all the information can be collected over a period of time. Fischer [23] proposed an ATP model with allocation planning and showed the relative advantages of the batch order processing compared to the single order processing for a practical case in the lighting industry. Chen *et al*. [11] showed that the optimal batching interval size can bring the system closer to global optimality. Pibernik [24] compared different ATP consumption rules for managing the stockout situations of a pharmaceutical company and suggested changing from a single order to a batch order processing mode if shortage is foreseeable. Framinan and Leisten [25] reviewed different ATP systems and categorized the timing of order processing (batch or real time) as one of the ATP related decisions. Also, Meyr [20] implemented his models by simulating three scenarios, i.e., global optimization (GO), single order processing (SOP), batch order processing (BOP), with consideration of penalty costs for loss of goodwill, and showed that real-time order fulfillment performs better than batch fulfillment if demand forecast is reliable. In this study, we will examine the comparative advantage of the two order processing modes, which is similar to the comparison of SOP and BOP in Meyr.

III. Model

In this section, we first describe the modeling background and then present an integrated three-stage order fulfillment model with an allocation review mechanism.

A. Modeling Background

The engineering chain of the semiconductor industry is as follows: an IC design house produces IC design, a foundry service provider performs IC manufacturing, an assembly house focuses on IC assembly, and a test house conducts IC testing [26]. To achieve fast time-to-market, an IC design house relies on effective allocation planning for sufficient capacity from foundry companies. To help customers, i.e., IC design houses and integrated device manufacturers (IDM), focus on their core competency, many foundry companies provide premium on-line service. For instance, TSMC provides access to data updated three times a day on a wafer lot's status in fabrication, assembly and testing, final testing, ordering and shipping [27]; it promotes the use of a logistics collaboration platform by customers to shorten time to delivery. To enhance the communication between a foundry company and its customers, we will apply this B2B collaboration idea in our allocation review mechanism.

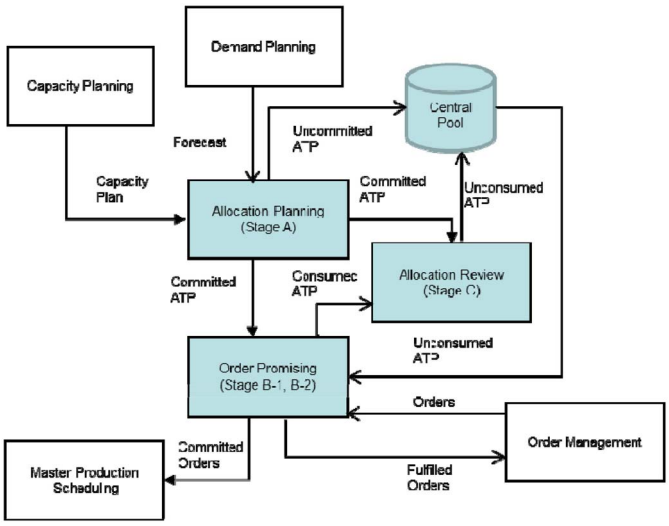

Fig. 3. Order fulfillment system in semiconductor foundry plants.

Semiconductor foundry manufacturing consists of four main processes: wafer manufacturing (or fabrication), circuit probing, assembly, and final test. Among the four, the wafer fabrication process is the most critical one with longest flow time. It is a complex process, including wafer cleaning, oxidation, deposition, lithography, etching, diffusion, ion implantation, metallization, inspection, and measurement. It takes around three months to complete the whole fabrication. In this study, we examine the wafer fabrication process only and propose a three-stage order fulfillment model (see Fig. 3). These stages are discussed below and will be formulated by linear programming in 3.2.

To begin with, on the top left of Fig. 3, capacity is planned periodically in terms of technology codes, i.e., specifications of technology's factory routing, average layer cycle time, throughput, and other manufacturing information to generate projected production output targets; to streamline production as well as commit customers with exact delivery date and quantity, the production output targets are planned into daily slots (details of the capacity planning are outside of the scope of this research), which become one of the two inputs of the allocation planning stage (stage A). The other input of stage A is from demand planning. The demand forecast of a foundry company is usually based on regional sales input

and requires sales managers to visit customers frequently. Also, when planning several months ahead, customers have difficulties in specifying exact product requirements to their foundry companies (which may include processing details for hundreds of wafer fabrication steps). Thus, customers usually provide only critical technology code information along with their demand forecast. The forecast data is combined and aggregated as monthly demand, which will be reviewed and revised internally by top managers according to marketing insights and strategy decisions; it will then be disaggregated to weekly demand based on historical trends and domain knowledge, and divided proportionally into daily demand. Next, capacity can be allocated for maximized profit in stage A. According to Kilger and Schneeweiss [2], allocation planning is to reserve "committed ATP" capacity for customers in the medium term (i.e., from 6^{th} to 18^{th} month); thus it is usually performed monthly. It is generally run with daily granularity to achieve a quick response for customer order confirmation [22]. Customers will be informed of their committed ATP quotas after this planning. They can communicate with regional sales managers if the committed ATP quotas cannot satisfy their demand before the allocation plan is finalized. Once the plan is released to customers, committed ATP quotas will be secured and only new customer demand and net change will be considered in the next planning run. In this study, we allocate the wafer fabrication capacity to demand with requested technology code for a specific customer in a specific plant. These are committed ATP quotas and the rest unallocated is uncommitted ATP capacity, which will be released to the central pool for open consumption in the order promising stage i.e., stage B. The committed ATP capacity is an input of the stage B. The primary purpose of stage B is to determine the committed quantity and delivery date for each order placed. When a customer places an order with required quantity and delivery date, the order promising stage will assign available capacity to it. It may be necessary to use the uncommitted ATP capacity at the central pool if the required quantity exceeds the committed ATP capacity. We will develop a linear programming model for order promising. The master production scheduling (MPS) follows to optimize daily production and customers are thus informed of their delivery quantity and date. We use B-1 to denote the initial run of stage B for promising orders before the allocation review.

Next, we discuss in detail how customer orders consume ATP capacity as well as the proposed periodic review mechanism, i.e., stage C. A pull-based ATP system is usually implemented in the semiconductor foundry industry. To help capacity and financial planners with more intuitional judgment, ATP capacity is typically planned and consumed by projected production output targets according to prespecified yield and fabrication time parameters [28]. If an order is received, a check is triggered to search for available commited ATP capacity according to order required date. If there is available capacity, the commited ATP capacity is consumed and the order finish date is obtained by subtracting the finished goods processing time from the order required date. The wafer start date is then determined by subtracting the fabrication time from the order finish date. There are also at least three days required for material preparation and equipment setup (order preparation time). Fig. 4 shows the foundry fabrication and ATP capacity consumption time frame, which includes the frozen horizon (whose duration equals the sum of order preparation time and fabrication time) and the planning horizon. Customers were allocated ATP capacity when they gave demand forecast to a semiconductor company. Later when they place orders, the order required date they requested will be in the planning horizon and the committed ATP capacity is thus consumed. As time passes, the committed capacity will be frozen for booking if their corresponding wafer finish dates are in the frozen horizon. Committed but unconsumed ATP capacity in the frozen horizon perishes. Hence, how to utilize the committed but unconsumed capacity before it perishes (the red shaded area in Fig. 4) is a critical issue for a MTO semiconductor foundry company.

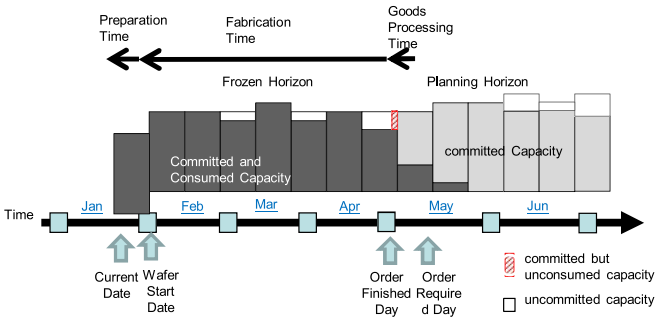

Fig. 4. Foundry fabrication and ATP consumption time frame.

Fig. 5 shows the allocation review mechanism, which is performed every day (real-time processing) or every five days (batch processing) to search for committed but unconsumed ATP capacity. If committed but unconsumed ATP records are found, the detailed customers' names, unconsumed quantities with specified technology codes, plant and time period of those records are put into a cut-off list. It is assumed that the cut-off list can be obtained by executing a scheduled program automatically. The cut-off list is then notified to the relevant customers via the B2B platform mentioned above. On the other hand, foundry customers need to monitor the product life cycle and control the inventory well so that they can immediately respond to the requests issued by the platform. Once customers decide to give up their committed ATP quotas, they can express this intension via the B2B platform. These unconsumed quotas on the cut-off list will be released to the central pool and open for consumption at stage B on the cut-off day (i.e., stage B will be run again to consume the newly released capacity and we denote this re-run as stage B-2). Whether or not IC design house buyers manage the outsourcing capacity well determines the production flexibility and the time-to-market capability [29]. Hence, IC design house buyers usually keep the updated inventory and foundry capacity reservation information. We assume that the review mechanism in Fig. 5 is so efficient that it will be completed within one day, though it should be noted that its actual duration can be adjusted to reflect the real industrial practice.

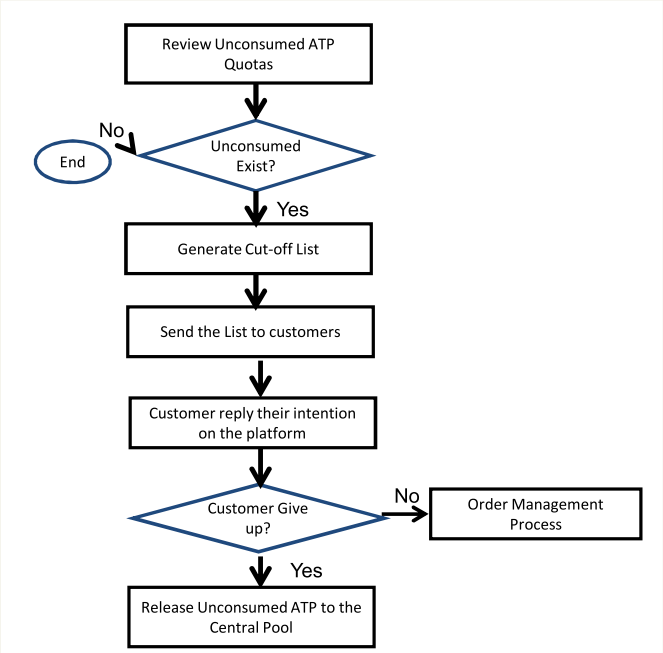

Fig. 5. Allocation review mechanism.

B. Order Fulfillment Model

We describe the proposed three-stage order fulfillment model in details. Specifically, we formulate them mathematically with linear programming. We assume that materials arrive according to their planned schedules. Since on-time delivery is usually one of the manufacturing objectives for foundry companies [30], we do not consider the inventory holding cost at stages A and order promising before allocation review (B-1) as in some of the previous research [19], [31]. However, we include it at stage B-2 to best utilize the stringent capacity. We ignore the backorder cost in the whole model because it is generally difficult to estimate in the semiconductor industry. In addition, it's important to have customer's agreement before any orders can be backlogged. If customers agree on the backlogging of their orders, they may agree to change the required date to consume the remaining unused capacity. Such orders, usually not seen in the demand forecast, have no capacity allocated to them and thus penalty for late delivery need not be charged.

An order may consist of multiple order items, and each order item i associated with customer c involves only one product, which is to be satisfied with predefined technology code g at factory f in required time period t; we denote this order item quantity as QTY_{cfgti}, and use the set I_{cfgt} to be the collection of order items from customer c with technology code g at factory f in required time period t. Thus, capacity is allocated and consumed by each order item, rather than by each order, according to its required date. The planning horizon is assumed to be long enough for planning ATP capacity to meet customer requirements.

1) Stage A: We consider a finite planning horizon that spans over T periods (e.g., days). Suppose each demand includes only one customer c. We use D_{cfgt} to denote the customer c's demand forecast that can be satisfied with technology code g at factory f and in time period t. We assume that D_{cfgt} should be greater than or equal to the sum of QTY_{cfgti}, i.e., $Dcfgt \geq \sum_{i \in I_{cfgt}} QTY_{cfgti}$. Demand is to be served with capacity CAP_{fgt} that is supplied at factory f with technology code g in period t. The goal is to find the committed ATP profile "ALP_{cfgt}" with forecasted margin M_{cfgt} that can best utilize capacity so that the total profit is maximized. The per wafer forecasted margin M_{cfgt} can be obtained by subtracting the per wafer manufacturing cost from the per wafer forecasted selling price for demand D_{cfgt}. Note that the same technology code g may have a different forecasted margin for a different customer c, at a different factory f, or in a different period t in a wafer foundry company. In a foundry plant, the expensive implanter equipment often represents the bottleneck operation and thus its capacity needs to be taken into account in particular at stage A. The bottleneck machine consumption (in hours) per wafer and total bottleneck machine consumption limit are denoted by CMW_{cfgt} and CM_{fgt}, respectively. Also, we use CAP^u_{fgt} to stand for the surplus capacity not committed to any customers, which is obtained after running stage A's model. We assume that unsatisfied order items are lost and unmet forecasts are ignored for the rest of the horizon. Table I shows the notation of stage A's model.

TABLE I
NOTATION OF THE ALLOCATION PLANNING MODEL

Indices	
c	index for customers; c = 1,2,……..C;
f	index for factories; f = 1,2,……..F;
g	index for technology codes; g = 1,2,…..G;
t	index for time periods; t = 1,2,……T;
i	index for order item;
Parameters	
D_{cfgt}	demand forecast (in wafers) from customer c required to be manufactured at factory f with technology code g in time period t
CAP_{fgt}	available capacity (in wafers) supplied at factory f with technology code g in time period t
CAP^u_{fgt}	capacity (in wafers) not committed at factory f with technology code g in time period t
M_{cfgt}	per wafer forecasted margin associated with demand D_{cfgt}
QTY_{cfgti}	required quantity (in wafers) from order item i for customer c at factory f with technology code g at time period t
CMW_{fgt}	bottleneck capacity consumption (in hours) per wafer associated with technology code g at factory f in time period t
CM_{fgt}	maximum available capacity (in hours) of the bottleneck machine at factory f with technology code g in time period t
R_{cfg}	an indicator which takes on the value of 1 if customer c's products with required manufacturing technology code g are supported at factory f
Decision Variables	
ALP_{cfgt}	the amount of capacity (in wafers) committed to demand D_{cfgt}

Formally, the basic allocation planning model is expressed as follows:

Maximize

$$Z = \sum_{c=1}^{C} \sum_{\mathrm{f}=1}^{\mathrm{F}} \sum_{\mathrm{g}=1}^{G} \sum_{t=1}^{T} M_{cfgt} \cdot ALP_{cfgt} \cdot R_{cfg} \tag{1}$$

TABLE II
ADDITIONAL NOTATION USED IN THE ATP CAPACITY CONSUMPTION MODEL

Indices	
$t_d(i)$	due date of order item i
Parameters	
C_{fgt}	the unconsumed ATP capacity (in wafers) at stage C at factory f with technology code g in time period t
PM_{cfgti}	per wafer margin associated with QTY_{cfgti}
HC_{fi}	per wafer per time period holding cost of finished products at factory f for order item i
Decision Variables	
ATP_{cfgti}	the amount of committed ATP capacity ALP_{cfgt} assigned to QTY_{cfgti}
$ATPC_{cfgti}$	the amount of unconsumed ATP capacity C_{fgt} assigned to QTY_{cfgti}

Subject to

$$\sum_{c=1}^{C} ALP_{cfgt} \bullet R_{cfg} \le CAP_{fgt} \quad \forall f, g, t \tag{2}$$

$$ALP_{cfgt} \bullet R_{cfg} \le D_{cfgt} \quad \forall c, f, g, t \tag{3}$$

$$ALP_{cfgt} \bullet R_{cfg} \ge \sum_{i \in I_{cfgt}} QTY_{cfgti} \quad \forall c, f, g, t \tag{4}$$

$$\sum_{c=1}^{C} ALP_{cfgt} \bullet CMW_{fgt} \bullet R_{cfg} \le CM_{fgt} \quad \forall f, g, t \tag{5}$$

$$ALP_{cfgt} \geqq 0 \quad \forall c, f, g, t \tag{6}$$

$$R_{cfg} \in \{0, 1\} \quad \forall c, f, g \tag{7}$$

where the binary variable R_{cfg} is used to denote whether or not customer c's products with technology code g can be manufactured at factory f. A customer's product was qualified in a specific plant with requested technology code at the product engineering phase. It's quite cost and time consuming to change the manufacturing plant or technology code. Thus, a customer places an order after the product was qualified already. Constraint (2) ensures that the ATP capacity committed to customers is less than or equal to the available capacity. Constraint (3) specifies that the committed ATP quantity cannot exceed the customer demand forecast, while constraint (4) states that it should at least satisfy the order item quantity placed. Since allocation planning is to reserve capacity for demand forecast in the medium term when only a small number of orders arrive at this stage, $CAP_{fgt} \ge \sum_{c=1}^{C} \sum_{i \in I_{cfgt}} QTY_{cfgti}$ is usually true. Constraint (5) requires that the total consumption of bottleneck machine hours should be less than or equal to the maximum bottleneck capacity. The uncommitted ATP capacity CAP^{u}_{fgt} is equal to $CAP_{fgt} - \sum_{c=1}^{C} ALP_{cfgt}$ and becomes an input of the following stages.

2) Stage B: Committed ATP capacity is consumed at this stage over a time horizon of T periods (e.g., days). We assume that this consumption is processed at the order item level, as mentioned above at the beginning of Section III-B. Also at this stage, we need to have ALP_{cfgt} and CAP^{u}_{fgt}, which are decided at stage A, and per wafer margin PM_{cfgti}, which is obtained by subtracting the per wafer manufacturing cost from the per wafer selling price for each order item QTY_{cfgti}. Table II shows the additional notation used at stage B.

The ATP capacity consumption model is expressed by B-1 or B-2.

B-1:

Maximize

$$Z = \sum_{c=1}^{C} \sum_{f=1}^{F} \sum_{g=1}^{G} \sum_{t=1}^{T} \sum_{i \in I_{cfgt}} \left(ATP_{cfgti} \cdot PM_{cfgti}\right) + \sum_{c=1}^{C} \sum_{f=1}^{F} \sum_{g=1}^{G} \sum_{t=1}^{T} \sum_{i \in I_{cfgt}} \left(ATPC_{cfgti} \cdot PM_{cfgti}\right) \tag{8}$$

Subject to

$$\sum_{i \in I_{cfgt}} ATP_{cfgti} \le ALP_{cfgt}, \quad \forall c, f, g, t, \tag{9}$$

$$\sum_{c=1}^{C} \sum_{i \in I_{cfgt}} ATPC_{cfgti} \le CAP^{u}_{fgt}, \quad \forall f, g, t \tag{10}$$

$$ATP_{cfgti} + ATPC_{cfgti} \le QTY_{cfgti} \quad \forall i \in I_{cfgt}, c, f, g, t \tag{11}$$

$$\left(\sum_{c=1}^{C} \sum_{i \in I_{cfgt}} (ATP_{cfgti} + ATPC_{cfgti})\right) \bullet CMW_{fgt} \le CM_{fgt}, \forall f, g, t \tag{12}$$

$$ATP_{cfgti} \ge 0, \ \forall i \in I_{cfgt}, c, f, g, t \tag{13}$$

$$ATPC_{cfgti} \ge 0, \forall i \in I_{cfgt}, c, f, g, t \tag{14}$$

The objective function (8) is to find optimal ATP_{cfgti} and $ATPC_{cfgti}$ that maximize the total profit. In constraint (11), the requested quantity QTY_{cfgti} is either met (i.e., consumed) by committed or uncommitted capacity. Constraint (9) states that the sum of consumed ATP quantities ATP_{cfgti} should be less than or equal to committed ATP capacity ALP_{cfgt}. Notice that ALP_{cfgt} is hard pegged to specified customers and thus cannot be consumed by other customers. Also, constraint (10) ensures that the sum of $ATPC_{cfgti}$ should be less than or equal to the uncommitted ATP capacity CAP^{u}_{fgt} in the central pool, which is not reserved by any customers and thus is open for consumption. Constraint (12) requires that the total consumption of bottleneck machine hours should be less than or equal to the maximum bottleneck capacity. This stage B model is applied to the batch processing mode. When run on the real-time mode, the above model assigns ATP_{cfgti} and/or $ATPC_{cfgti}$to an incoming order immediately (i.e., in a first-come-first-serve, or FCFS, manner) and computes its profit.

B-2:

Maximize

$$Z = \sum_{c=1}^{C} \sum_{f=1}^{F} \sum_{g=1}^{G} \sum_{t=1}^{T} \sum_{i \in I^{u}_{cfgt}} \left(ATPC_{cfgti} \cdot PM_{cfgti}\right) - \sum_{c=1}^{C} \sum_{f=1}^{F} \sum_{g=1}^{G} \sum_{t=1}^{td(i)} \sum_{i \in I^{u}_{cfgt}} \left(ATPC_{cfgti} \cdot HC_{fi}[t_d(i) - t]\right) \tag{15}$$

Subject to

$$\sum_{t=1}^{t_d(i)} ATPC_{cfgti} \le QTY_{cfgti}, \quad \forall i \in I^{u}_{cfgt}, c, f, g \tag{16}$$

TABLE III
DIFFERENCES BETWEEN STAGES B-1 AND B-2

Stage	B-1	B-2
Planning Objective	maximize profit	maximize profit minus earliness cost
Demand	order items I_{cfgt} from customers	remaining unsatisfied order items I^u_{cfgt} from B-1
Supply	committed ATP or surplus capacity from stage A	surplus or newly released capacity from stage C
Execution Timing	after stage A	after stage C
Order Confirmation	immediately after execution	immediately after execution

$$\sum_{c=1}^{C} \sum_{i \in I^u_{cfgt}} ATPC_{cfgti} \le C_{fgt}, \quad \forall f, g, t \tag{17}$$

$$\sum_{c=1}^{C} \sum_{i \in I^u_{cfgt}} ATPC_{cfgti} \bullet CMW_{fgt} \le CM_{fgt}, \quad \forall f, g, t \tag{18}$$

$$ATPC_{cfgti} \ge 0, \forall i \in I^u_{cfgt}, c, f, g, t, \tag{19}$$

Please note that when unconsumed ATP quotas on the cut-off list are released to the central pool, they will be reallocated to fulfill unsatisfied order items. The above stage B-2, instead of B-1, will be run again for the reallocation. The objective function (15) maximizes the total profit which equals the profit obtained from the consumed committed ATP capacity deducted by the cost attributed from earliness, which is computed by multiplying $ATPC_{cfgti}$ by its associated HC_{fi} and elapsed time between the order item completion date t and due date $t_d(i)$. In constraint (16), QTY_{cfgti} represents the remaining, if any, unsatisfied order item quantity updated from constraint (11), and we use the set I^u_{cfgt} to denote the collection of such order items from customer c with predefined technology code g at factory f in required time period t. Thus, constraint (16) means that the total $ATPC_{cfgti}$ quantity before the due date of an order item should be less than or equal to the order item quantity QTY_{cfgti}, and C_{fgt} in (17)is to be calculated in expression (22) below at stage C. Table III describes the differences between stages B-1 and B-2 in more detail.

3) Stage C: Forecasting demand in the semiconductor industry is difficult (as described in Section I) and incurs considerable uncertainty. We plan to reduce this uncertainty through the allocation review mechanism at stage C. We plan to review the status of ATP capacity consumption periodically and release unconsumed capacity for further utilization. It's not feasible to review the consumption status in a truly real time basis; we thus plan to review it every day (real-time processing) or every five days (batch processing). First, we need to decide the cut-off target date or period, denoted as *td*, which is obtained by adding order preparation time and fabrication time to the current date. If the real-time processing mode is used, *td* is a single day; while if the batch mode is used, *td* includes four additional following days (i.e., a total of five days). For example, suppose that three days are required for order preparation and ninety days are needed for wafer fabrication. If the current date is Jan. 28, 2013 and the review is triggered in the morning, then the cut-off target date is May 1, 2013 in the case of daily review. In other words, any ATP capacity before May 1 is frozen already, and an order placed on Jan. 28 can consume capacity only on and after May 1 (i.e., its earliest wafer out date is May 1). The unconsumed capacity of May 1, 2013 will expire on Jan. 29 due to insufficient time for wafer fabrication determined by the manufacturing flow time. As time moves on over the planning horizon, only emergency orders with a higher selling price are allowed to consume the unused capacity in the frozen horizon (details of the emergency order processing are outside of the scope of this research). If committed but unconsumed ATP quantities (after communicating with customers) for those days in *td* are found, the detailed customers' information (including unconsumed capacity) forms a cut-off list, as described in Section III-A. The unconsumed quotas on the cut-off list will be released to the central pool and open for re-consumption at stage B-2.

We use $BQTY_{cfgt}$ to represent the sum of consumed ATP capacity ATP_{cfgti} which is committed to customer c at factory f with technology code g in time period t, and $BCQTY_{cfgt}$ to be the sum of consumed uncommitted ATP capacity $ATPC_{cfgti}$ assigned to customer c at factory f with technology code g in time period t. Thus, by definition,

$$BQTY_{cfgt} = \sum_{i \in I_{cfgt}} ATP_{cfgti} \quad \forall c, f, g, t; t \in td, \tag{20}$$

$$BCQTY_{cfgt} = \sum_{i \in I_{cfgt}} ATPC_{cfgti} \quad \forall c, f, g, t; t \in td, \tag{21}$$

We, in fact, aggregate all the consumed ATP capacity originally hard pegged to a customer by the cut-off target date as $BQTY_{cfgt}$ in expression (20), and aggregate all the consumed uncommitted ATP capacity originally from the central pool by the cut-off target date as $BCQTY_{cfgt}$ in (21). Hence, C_{fgt}, the main parameter at stage C, can be computed by

$$C_{fgt} = \left(\sum_{c=1}^{C} ALP_{cfgt} - \sum_{c=1}^{C} BQTY_{cfgt} \right) + \left(CAP^u_{fgt} - \sum_{c=1}^{C} BCQTY_{cfgt} \right) \quad \forall f, g, t; t \in td \tag{22}$$

TABLE IV
EXPERIMENT DATA OF CAPACITY

Month	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	Total
Technology 1	40950	40950	40950	40950	40950	40950	42950	42950	42950	43550	43550	43550	505200
Technology 2	46800	46800	46800	46800	46800	46800	46800	46800	46800	46800	46800	46800	561600
Technology 3	46200	46200	46200	46200	46200	46200	46200	46200	46200	46200	46200	46200	554400
Total	**133950**	**133950**	**133950**	**133950**	**133950**	**133950**	**135950**	**135950**	**135950**	**136550**	**136550**	**136550**	**1621200**

TABLE V
EXPERIMENT DATA OF DEMAND FORECAST

Month	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	Total
Customer 1	**50400**	**49300**	**50400**	**50650**	**50400**	**50000**	**50400**	**50650**	**50800**	**51400**	**50500**	**50800**	**605700**
Technology 1	14400	13500	14400	14400	14400	14200	14400	14650	14400	14800	14400	14400	172350
Technology 2	20100	20100	20100	20100	20100	20100	20100	20100	20100	20100	20100	20100	241200
Technology 3	15900	15700	15900	16150	15900	15700	15900	15900	16300	16500	16000	16300	192150
Customer 2	**52200**	**52200**	**52100**	**52200**	**52200**	**51900**	**53100**	**52400**	**53050**	**53100**	**53100**	**53100**	**630650**
Technology 1	15150	15150	15250	15150	15150	14850	15450	15250	16050	15650	15450	15150	183700
Technology 2	22800	22800	22600	22800	22800	22800	22800	22900	22800	23100	23400	23700	275300
Technology 3	14250	14250	14250	14250	14250	14250	14850	14250	14200	14350	14250	14250	171650
Customer 3	**38400**	**38500**	**38700**	**38750**	**38600**	**37900**	**38600**	**38500**	**38800**	**38400**	**39200**	**38750**	**463100**
Technology 1	10800	10900	10950	11000	11000	10300	11000	10900	11100	10800	11600	10800	131150
Technology 2	12300	12300	12300	12450	12300	12300	12300	12300	12300	12300	12300	12300	147750
Technology 3	15300	15300	15450	15300	15300	15300	15300	15300	15400	15300	15300	15650	184200
總計	**141000**	**140000**	**141200**	**141600**	**141200**	**139800**	**142100**	**141550**	**142650**	**142900**	**142800**	**142650**	**1699450**

Note in the above expression that committed capacity ALP_{cfgt}, consumed ATP quantities $BQTY_{cfgt}$, and $BCQTY_{cfgt}$ are all summed up over customers.

IV. EXPERIMENTS

We will provide a case study of a semiconductor company to examine the proposed model. The data is collected from a foundry plant with some modifications. The experiment design and corresponding data are described in Section IV-A. The effects of the allocation review mechanism are discussed in Section IV-B. The performance of the proposed model when run on the real-time and batch order processing modes is also examined in Section IV-C.

The proposed model can be solved to optimality by any LP software or optimization solver. For this research, we used the modeling and optimizing language Lingo 13.0 and the simple database Microsoft Access 2010. Also, all the computation was executed on a personal computer with an Intel Core i7-2600 2.80 GHz processor and 8 GB RAM, operated by the Microsoft Windows 7 professional system. While stages B and C in the real-time mode were solved within a second, all stages in the batch mode attained their optimal solutions in a second as well.

A. Experiment Design

The environment where the company operates is a classical MTO. For simplicity in the following experiments, we reduced the size and scope of the actual data. We consider three customers, 273 orders items, three technology codes at a factory, a planning horizon of twelve months at stage A and ten time periods (i.e., days) at stages B and C. Tables IV and V list the experiment data of capacity and demand forecast from 6^{th} to 17^{th} months. To investigate the performance of the proposed model, we use the same capacity data set, demand forecast data set, but different order sets with various MAPE values (to be defined below). Tables VI–VIII show the three order sets with MAPE values of 15, 30 and 70, which were collected to consume the committed ATP capacity in the first ten time periods of the 6^{th} month. We also compare the performance of batch and real time order processing modes under the same MAPE value in the second experiment.

The total demand forecast within the planning horizon exceeds the total supply by 4.8%. The demand forecast in the first ten time periods of the 6th month also exceeds the actual total order quantity by 6.44% (as semiconductor foundry customers often report higher demand to sales managers to ensure sufficient capacity). Although the total capacity and total order quantity are very close, shortage exists (because of unanticipated demand fluctuation) in this case study and the order fulfillment rate, defined by the percentage of customer order quantities fulfilled from committed ATP capacity, is highly dependent on demand forecasting accuracy. MAPE is one of the most common methods to compute forecasting errors and given by

$$\text{MAPE} = \frac{100\%}{n} \cdot \sum_{t=1}^{T} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \tag{23}$$

where n = number of periods

y_t = forecasted quantity in time period t

y_t = actual ordering quantity in time period t

[32]. A smaller MAPE value means a more accurate forecast for a specified period of time. Lewis suggested the following interpretation of MAPE values:

- less than 10 percent is highly accurate forecasting
- between 10 and 20 percent is good forecasting
- between 20 and 50 percent is reasonable forecasting
- greater than 50 percent is inaccurate forecasting

In the semiconductor industry, it's difficult to maintain highly accurate forecasting continuously, as described in

TABLE VI
ORDER SET WITH MAPE = 15

Time Period	01	02	03	04	05	06	07	08	09	10	Total
Customer 1	1700	1750	1350	1450	1800	1300	1350	1950	850	1800	15300
Technology 1	400	500	400	100	500	400	400	650	150	600	4100
Technology 2	700	650	450	650	700	800	650	800	450	750	6600
Technology 3	600	600	500	700	600	100	300	500	250	450	4600
Customer 2	1200	1050	1800	1700	2250	1200	1500	1850	1000	2250	15800
Technology 1	250	300	300	600	800	600	200	800	300	800	4950
Technology 2	700	450	900	800	700	500	800	150	300	800	6100
Technology 3	250	300	600	300	750	100	500	900	400	650	4750
Customer 3	1350	1300	1000	1500	1600	1100	800	1350	1100	2100	13200
Technology 1	550	350	350	700	600	500	100	300	250	100	3800
Technology 2	100	500	250	300	800	500	100	650	600	100	3900
Technology 3	700	450	400	500	200	100	600	400	250	1900	5500
Total	4250	4100	4150	4650	5650	3600	3650	5150	2950	6150	44300

TABLE VII
ORDER SET WITH MAPE = 30

Time Period	01	02	03	04	05	06	07	08	09	10	Total
Customer 1	2000	1400	1400	1300	1800	1200	2000	1950	1350	1600	16000
Technology 1	450	350	250	200	500	300	850	650	750	400	4700
Technology 2	800	650	450	650	700	800	850	800	350	750	6800
Technology 3	750	400	700	450	600	100	300	500	250	450	4500
Customer 2	1150	1550	1800	1350	2100	950	1400	1550	1200	1800	14850
Technology 1	200	300	600	400	800	600	100	800	300	750	4850
Technology 2	700	450	600	650	700	250	800	200	300	800	5450
Technology 3	250	800	600	300	600	100	500	550	600	250	4550
Customer 3	1450	1550	1250	1300	1700	1150	700	1250	1850	1250	13450
Technology 1	550	350	650	400	700	350	200	200	250	150	3800
Technology 2	100	400	200	400	800	500	100	650	600	100	3850
Technology 3	800	800	400	500	200	300	400	400	1000	1000	5800
Total	4600	4500	4450	3950	5600	3300	4100	4750	4400	4650	44300

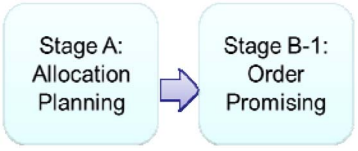

Fig. 6. Base model.

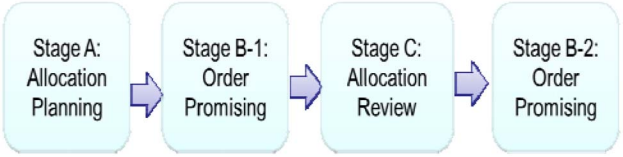

Fig. 7. Proposed model.

Section I. Thus, we exclude highly accurate forecasting and only compare good, reasonable, and inaccurate forecasting in our experiments.

B. Benefit of Allocation Review Mechanism

In order to investigate the impact of the allocation review mechanism on the overall profit, we consider the following two models in the experiment. We assume the use of the batch processing mode (i.e., orders are firstly collected and then processed together to find an optimal solution).

1) Base Model: In this model (Fig. 6), we assume no allocation review mechanism. Allocation planning is first performed; then the order promising stage (stage B-1) follows.

2) Proposed Model (Base Model With Allocation Review Mechanism): In the proposed model (Fig. 7), the allocation review mechanism (stage C) is added to release the unconsumed ATP capacity. Then, order promising (stage B-2) is executed to consume the newly released ATP quotas if there are unsatisfied orders. By integrating the three stages in the order fulfillment system, a new solution is obtained with higher profit and capacity utilization rate, defined by the percentage of installed capacity actually used for production during a period of time.

The data set described in Section IV-A is used for experiment under the base model and proposed model. The experimental results are shown in Table IX (detailed results of stage

TABLE VIII
ORDER SET WITH MAPE = 70

Time Period	01	02	03	04	05	06	07	08	09	10	Total
Customer 1	1750	1000	1250	1600	1450	1600	2050	1450	1350	1250	14750
Technology 1	800	350	250	900	550	1000	850	600	350	500	6150
Technology 2	200	250	300	250	300	550	900	350	750	300	4150
Technology 3	750	400	700	450	600	50	300	500	250	450	4450
Customer 2	1000	1550	1600	1500	2150	900	1000	1300	1600	1500	14100
Technology 1	300	450	600	650	900	250	300	350	450	600	4850
Technology 2	200	300	600	400	800	300	100	700	400	750	4550
Technology 3	500	800	400	450	450	350	600	250	750	150	4700
Customer 3	1650	1850	1600	1400	2100	1250	750	1150	1950	1750	15450
Technology 1	200	300	500	600	900	750	250	350	700	350	4900
Technology 2	550	700	650	300	900	200	200	200	450	200	4350
Technology 3	900	850	450	500	300	300	300	600	800	1200	6200
Total	4400	4400	4450	4500	5700	3750	3800	3900	4900	4500	44300

TABLE IX
EXPERIMENT RESULTS

Model		Stage	MAPE Value		
			15	30	70
Proposed Model	Base Model	Stage A: Committed ATP quantity	41350	41350	41350
		Stage B-1: Consumed committed ATP quantity	36450	36450	32500
		Stage C: Released quantity	8200	8200	12150
		Stage B-2: Consumed surplus capacity quantity	5850	5350	8900
Total Consumption			42300	41800	41400

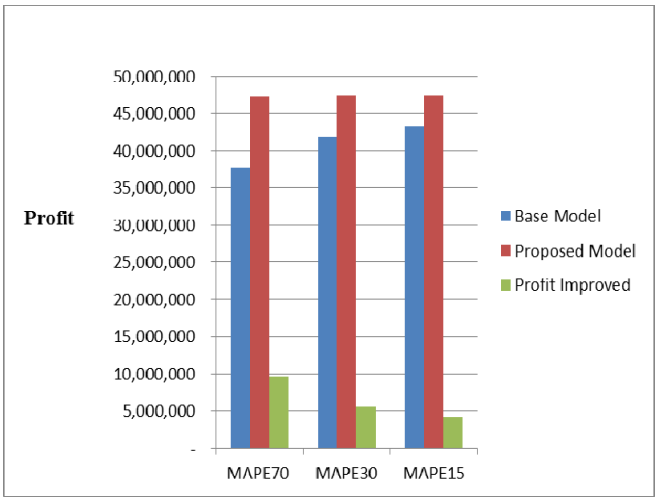

Fig. 8. Effect of the proposed model on profit.

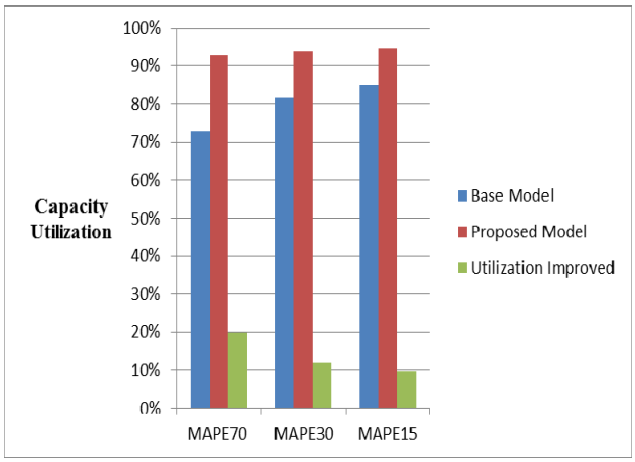

Fig. 9. Effect of the proposed model on capacity utilization rate.

A can be seen in the appendix). As expected, adding the allocation review mechanism to the order fulfillment system is advantageous as compared to the ordinary order fulfillment system. The overall profit improvement is rather significant, especially when MAPE values are larger (see Fig. 8).

In addition, Figs. 9 and 10 show the capacity utilization rate and order fulfillment rate under the two models. The proposed model performs better than the base model on all MAPE values. More importantly, the improvement on capacity utilization and order fulfillment rate is more salient when MAPE is larger. This shows the importance of the allocation review mechanism. Under the ordinary order fulfillment, order rejection is common when capacity is not available for promising. However, as demand fluctuates, allocation review is needed to release committed but unconsumed capacity for reallocation.

C. Batch Processing Versus Real-Time Processing

A real-time order promising system processes an incoming order immediately (i.e., FCFS) and usually responds to customers more in time and thus is used more often in practice than the batch processing mode. Nevertheless, some of the previous studies have shown that real-time processing of arriving orders is hardly the best way for order fulfillment in shortage situations. Collecting groups of transactions for a period of time and processing them in a batch can improve benefit [11], [22], [24]. Batch processing actually optimizes ATP consumption since it processes orders as if there were perfect knowledge of customer demand for the whole batch time interval. In this section, we also run the proposed model (as well as the ordinary model) on the real-time mode and compare the performance of the two processing modes. We will use the

TABLE X
PERFORMANCE COMPARISON BETWEEN BATCH MODE AND REAL-TIME MODE

	process mode	Profit (thousands)	capacity utilization	order fulfillment rate
Base Model	real-time	41876	81.63%	82.28%
	batch	41926	81.63%	82.28%
Proposed Model	real-time	47385	93.62%	94.36%
	batch	47450	93.62%	94.36%
	difference	0.14%	0.00%	0.00%

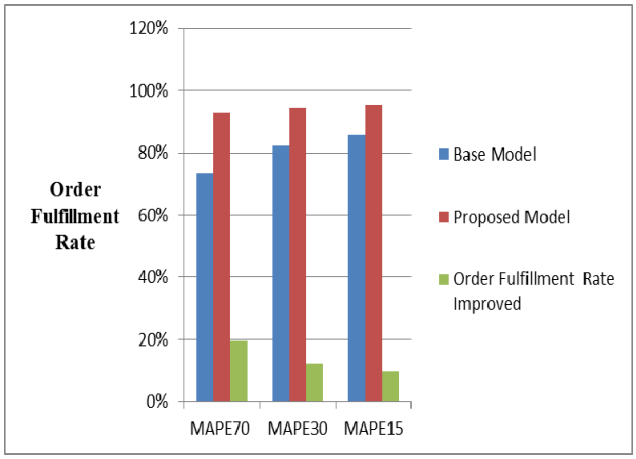

Fig. 10. Effect of the proposed model on order fulfillment rate.

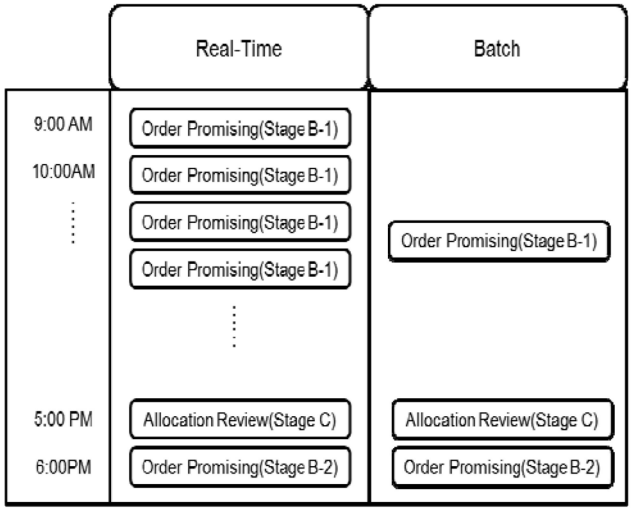

Fig. 11. Model solving time frame.

data set with MAPE value 30 in the experiment. As described in Fig. 5 and Section III-A, when the proposed model runs on the real-time mode, a scheduled program will be triggered to review committed but unconsumed ATP capacity every day and send the cut-off list to customers via the B2B platform. Once customers give up their reservation, those unconsumed ATP quotas will be released to the central pool for open consumption at the reviewing date. Fig. 11 illustrates how the model is solved in both order processing modes.

When an actual order item quantity differs from the committed ATP quantity, capacity shortage or surplus occurs. In the surplus situation, ATP capacity will be released to the central pool at stage C; however, in the shortage situation, orders are either satisfied by uncommitted ATP originally in the central pool or by the newly released ATP capacity. Since the ATP capacity in the central pool is not committed to any customers, the sequence of orders arriving each day certainly impacts the total corporate profit if more than two orders compete for the same ATP capacity in the real-time mode. On the contrary, the batch processing mode optimizes ATP consumption within each daily slot over the entire batch processing interval, and the length of the batch interval, i.e., the review frequency, will not affect the overall model performance (thus, if batch processing is performed every two days instead of every five days, the total profit will remain the same). It is expected that the profit obtained on the real-time mode will be lower than that on the batch processing mode, which serves as a benchmark of order fulfillment systems.

Table X shows the profit, capacity utilization, and order fulfillment rate under the two order processing modes. Capacity utilization and order fulfillment rate are exactly the same because these two modes have the same capacity, demand forecasts, and order quantities. With regard to the profit obtained, the proposed model performs better than the base model whether on the real-time or batch processing mode. However, the profit obtained on the batch mode is only slightly higher than that on the real-time mode in the proposed model (0.14%). This result reveals that the proposed model can allocate capacity nearly optimally even on the real-time mode.

V. CONCLUSION

In this study, we propose an order fulfillment model with a periodic review mechanism to reallocate unused capacity. As in an ordinary order fulfillment model, customers are allocated with ATP capacity that are later consumed when an actual order is placed. But unlike an ordinary order fulfillment model, the proposed model will periodically review the unused ATP capacity and reallocate them to orders for higher capacity utilization. The experiment shows that the proposed order fulfillment model performs better than an ordinary model, especially when forecast errors are large. Further

experiments also show that running the proposed model on the real-time mode may provide near-to-optimal solutions for an MTO semiconductor foundry company. This result may be useful to foundry managers as order promising systems (i.e., ATP systems) are often run on the real-time mode in practice.

With the severe competition in the semiconductor industry nowadays, foundry companies regard on-time delivery as one of the major competitive advantages as they cannot afford lost sales due to tardy delivery and/or poor customer service. This research aims to achieve on-time delivery and high order fulfillment rate for ATP consumption. Further research is possible for at least two directions. First, some capacity is still not utilized and some orders are not satisfied in this study. The proposed order fulfillment model may be modified to incorporate the emergency order processing to consume the unutilized capacity. Second, we have assumed that committed ATP capacity is hard pegged to customer orders. It is possible to include a "swap" mechanism of exchanging ATP capacity for different order items so as to utilize capacity more flexibly.

REFERENCES

[1] Z. K. Weng, "Strategies for integrating lead time and customer-order decisions," *IIE Trans.*, vol. 31, no. 2, pp. 161–171, Feb. 1999.

[2] C. Kilger and L. Schneeweiss, "Demand fulfillment and ATP," in *Supply Chain Management and Advanced Planning*, H. Stadtler and C. Kilger, Eds. Berlin, Germany: Springer, 2005, ch. 9.

[3] Silicon Semiconductor. (2010). *IC Capacity Nears 100%* [Online]. Available: http://www.siliconsemiconductor.net/article/73335-IC-capacity-nears-100.php

[4] D. M. H. Chiang and A. W. D. Wu, "Discrete-order admission ATP model with joint effect of margin and order size in a MTO environment," *Int. J. Prod. Econ.*, vol. 133, no. 2, pp. 761–775, Oct. 2011.

[5] S. V. Sridharan, "Managing capacity in tightly constrained systems," *Int. J. Prod. Econ.*, vol. 56–57, pp. 601–610, Sep. 1998.

[6] R. Leachman, S. Ding, and C. Chien, "Economic efficiency analysis of wafer fabrication," *IEEE Trans. Autom. Sci. Eng.*, vol. 4, no. 4, pp. 501–512, Oct. 2007.

[7] C. F. Chien, Y. J. Chen, and J. T. Peng, "Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle," *Int. J. Prod. Econ.*, vol. 128, no. 2, pp. 496–509, Dec. 2010.

[8] Semiconductor Industry Association. (2013, Apr.). *World Semiconductor Trade Statistics* [Online]. Available: http://www.siliconsemiconductor.net/article/77120-SIA-Memory-boosts-global-semiconductor-sales.php

[9] T. E. Vollmann, W. L. Berry, and D. C. Whybark, *Manufacturing Planning Control Systems,* 4th ed. New York, NY, USA: McGraw-Hill, 1997.

[10] C. Y. Chen, Z. Y. Zhao, and M. O. Ball, "Quantity and due date quoting available to promise," *Inf.. Syst. Front.*, vol. 3, no. 4, pp. 477–488, 2001.

[11] C. Y. Chen, Z. Y. Zhao, and M. O. Ball, "A model for batch advanced available-to-promise," *Prod. Oper. Manage.*, vol. 11, no. 4, pp. 424–440, 2002.

[12] H. Jung, I. Song, B. J. Jeong, and W. Yoo, "An optimized ATP (available-to-promise) system for make-to-order company in supply chain environment," *Int. J. Ind. Eng.*, vol. 10, no. 4, pp. 367–374, 2003.

[13] A. G. Robinson and R. C. Carlson, "Dynamic order promising: Real-time ATP," *Int. J. Integr. Supply Manag.*, vol. 3, no. 3, pp. 283–301, 2007.

[14] M. Ebadian, M. Rabbani, F. Jolai, S. A. Torabi, and R. Tavakkoli-Moghaddam, "A new decision-making structure for the order entry stage in make-to-order environments," *Int. J. Prod. Econ.*, vol. 111, no. 2, pp. 351–367, Feb. 2008.

[15] Q. Zhang and M. M. Tseng, "Modeling and integration of customer flexibility in the order commitment process for high mix low volume production," *Int. J. Prod. Res.*, vol. 47, no. 22, pp. 6397–6416, 2009.

[16] S. J. Noh, S. C. Rim, and H. S. Lee, "Meeting due dates by reserving partial capacity in MTO firms," in *Systems Modeling and Simulation: Theory and Applications* (Lecture Notes in Artificial Intelligence), vol. 3398, D. K. Baik, Ed. Berlin, Germany: Springer-Verlag, 2005, pp. 568–577.

[17] R. Pibernik and P. Yadav, "Dynamic capacity reservation and due date quoting in a make-to-order system," *Nav. Res. Logist.*, vol. 55, no. 7, pp. 593–611, 2008.

[18] T. Ponsignon and L. Mönch, "Heuristic approaches for master planning in semiconductor manufacturing," *Comput. Oper. Res.*, vol. 39, no. 3, pp. 479–491, Mar. 2012.

[19] K. M. Tsai and S. C. Wang, "Multi-site available to promise modeling for assemble-to-order manufacturing: An illustration on TFT-LCD manufacturing," *Int. J. Prod. Econ.*, vol. 117, no. 1, pp. 174–184, Jan. 2009.

[20] H. Meyr, "Customer segmentation, allocation planning and order promising in make-to-stock production," *OR Spectr.*, vol. 31, pp. 229–256, Jan. 2009.

[21] J. T. Lin, C. H. Wu, T. L. Chen, and S.H. Shih, "A stochastic programming model for strategic capacity planning in thin film transistor-liquid crystal display (TFT-LCD) industry," *Comput. Operat. Res.*, vol. 38, no. 7, pp. 992–1007, 2011.

[22] M. E. Fischer, *Available to Promise: Aufgaben und Verfahren im Rahmen des Supply Chain Management*. Regensburg, Germany: Roderer, 2001.

[23] J. Kallrath and T. I. Maindl, Eds., "Planning in semiconductor manufacturing," in *Real Optimization With SAP APO*. Berlin, Germany: Springer, 2006, ch. 5.

[24] R. Pibernik, "Managing stock-outs effectively with order fulfillment systems," *J. Manuf. Technol. Manag.*, vol. 17, no. 6, pp. 721–736, 2006.

[25] J. M. Framinan and R. Leisten, "Available-to-promise (ATP) systems: A classification and framework for analysis," *Int. J. Prod. Res.*, vol. 48, no. 11, pp. 3079–3103, 2010.

[26] F. T. Cheng, Y. L. Chen, and Y. C. Chang, "Engineering chain: A novel semiconductor engineering collaboration model," *IEEE Trans. Semicond. Manuf.*, vol. 25, no. 3, pp. 394–407, Aug. 2012.

[27] (2013, Jun. 21). Taiwan Semiconductor Manufacturing Company, *TSMC's eFoundry* [Online]. Available: http://www.tsmc.com/chinese/dedicatedFoundry/services/eFoundry.htm

[28] Y. F. Hung and R. C. Leachman, "A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations," *IEEE Trans. Semicond. Manuf.*, vol. 9, no. 2, pp. 257–269, May 1996.

[29] C. E. Lee and S. C. Hsu, "Outsourcing capacity planning for an IC design house," *Int. J. Adv. Manuf. Technol.*, vol. 24, nos. 3–4, pp. 306–320, Aug. 2004.

[30] (2013, Jun. 21). Taiwan Semiconductor Manufacturing Company *Manufacturing Excellence* [Online]. Available: http://www.tsmc.com/english/dedicatedFoundry/manufacturing/efficiency.htm

[31] G. I. Zobolas, C. D. Tarantilis, and G. Ioannou, "Extending capacity planning by positive lead times and optional overtime, earliness and tardiness for effective master production scheduling," *Int. J. Prod. Res.*, vol. 46, no. 12, pp. 3359–3386, 2008.

[32] C. D. Lewis, *Industrial and Business Forecasting Method*. London, U.K.: Butterworth, 1982.

Chi Chiang, photograph and biography not available at the time of publication.

Hui-Lan Hsu, photograph and biography not available at the time of publication.