Abstract Number: 015-0074

Finished-goods Inventory Study under Capacitated Postponement in Semiconductor Manufacturing

Dr. Dong Tang

Customer Fulfillment, Planning and Logistics Group, Intel Corporation

5000 West Chandler Boulevard, Chandler, Arizona 85226

Email: dong.tang@intel.com

Tel: 001-480-552-6562

POMS 21st Annual Conference

Vancouver, Canada

May 7 to May 10, 2010

Finished-goods Inventory Study under

Capacitated Postponement in Semiconductor Manufacturing

Dong Tang

Customer Fulfillment, Planning and Logistics Group, Intel Corporation

5000 West Chandler Boulevard, Chandler, Arizona 85226

Email: dong.tang@intel.com

Abstract: Semiconductor manufacturers practicing delayed differentiation or postponement usually suffer from constrained finishing capacities and therefore still keep finished-goods inventories. This paper studies base-stock inventory models with and without demand forecasting and provides a computationally efficient method to set optimal inventory targets for finished products under capacitated postponement. Computations show inventory-saving benefit quickly vanishes after capacity reaches a certain level. The value of forecasted advance-demand information to postponement is justified, but can easily be overstated. When capacity limitation becomes severe, intuitions often guide producers to build to forecast more than finishing lead time ahead. Results in this work indicate that these intuitions may be invalid under capacitated postponement, reveal that forecasted advance-demand information is valuable only when the variance of demand forecast errors is less than that of demands, and show that the optimal forecast lead time can be obtained in the same way as if capacity is unlimited.

Keywords: Semiconductor Manufacturing, Capacity, Delayed Differentiation, Forecast, Advance demand Information

1. Introduction and Background

To cope with conflicts between product diversity and inventory cost saving, many manufactures or service providers adopt delayed differentiation or postponement, which keeps the product commonality as late as possible until the product differentiation is really necessary. In such a way, a large portion of the inventory can be kept as semi-finished goods instead of finished goods and so reduction of inventory cost can be achieved. Though delayed differentiation can significantly reduce inventory cost, it doesn't completely remove the necessity of keeping finished-goods inventory largely because of the following two reasons: (a) finishing lead time, e.g. the time needed to convert semi-finished goods into finished goods. Manufacturers need finished-goods inventory to fulfill the customer orders during the finishing lead time; (b) finishing capacity. Capacitated finishing lines prevent manufacturers from fully supporting customer orders that are larger than finishing capacities, so finished-goods inventory is needed. Much of the inventory control research in the literature has been carried out around (a) to provide optimal inventory policies. Readers are referred to Zipkin [1] for a comprehensive literature review. This research focuses on (b) and studies finished-goods inventory models under finishing-line capacitated postponement. This study is of great interest to semiconductor manufactures, given the fact that a diversity of finished products are made from a common set of assembled packages, and finishing lines are often capacitated.

Manufacturers are often found using the base-stock inventory policy, which can be simply described as follows: a fixed inventory target s for a finished product is set as the base-stock level, when a customer demand D arrives we immediately fulfill it out of

the on-hand inventory and simultaneously order the same amount replenishment from the external/internal supplier. The base-stock level s is commonly determined based upon a pre-set service level or customer order fill rate. Exceptions can happen to manufactures with non-negligible manufacturing setup cost or re-order cost, in which case a (s, S) type of inventory policy prevails. Semiconductor manufactures usually adopt high volume manufacturing (HVM) methods, which make the per-unit setup cost so low that the base-stock policy suffices the inventory management purpose. Under an ideal situation where the external/internal supplier can always replenish D into the warehouse, it is seen that the inventory position will always be kept at the base-stock level of s. Difficulties emerge when the supply/finishing line is capacitated and can replenish/produce less than D in some time periods.

The base-stock inventory policy under capacitated scenarios works very similar to the conventional one: in each of the inventory review period determine the inventory replenishment order size according to the conventional base-stock policy, and produce to the replenishment order size unless it exceeds the capacity, in which case we produce to capacity. Federgruen and Zipkin [2] proved the optimality of this policy with average-cost criteria in a periodic-review single-item inventory system. One can easily see that, in order to reach the same level of service, the new required base-stock level with finishing capacities will be higher than s, which is calculated based upon an infinite capacity assumption. So will be the expected net inventory and associated inventory cost. However, questions remain as to how much higher the new base-stock level should be and what method is appropriate and efficient to determine it. Another question of great

managerial interest would be: will the expense of adding capacity be paid off by inventory cost saving and how to find the optimal capacity?

Finishing capacities often drive manufactures to build ahead according to demand forecast. When the limitation of finishing capacities becomes severe, intuitions often guide producers to build to forecast even more than finishing lead times ahead. For instance, even though finishing lead times hardly exceed 1~2 weeks, the studied semiconductor manufacturer often uses 2~3 week's worth of demand forecast to pull semi-finished goods through finishing lines when finishing capacities are moderately constrained. When finishing capacities become more constrained, a pull-in practice of using even more week's worth of demand forecast is adopted, though we are not clear whether or not this practice is rational and makes theoretical sense.

Certainly any manufacturer adopts delayed differentiation makes a variety of products. Often those different products are produced in the same manufacturing line. This fact further adds complexity to the problem of determining base-stock level for each product because a capacity allocation decision needs to be made conjunctionally. When the capacity is constrained, some manufacturers will decide a product priority list and make products on top of the list first, some manufacturers will use the so-called "share-the-pain" strategy and make the products proportionally according to their replenishment order qualities, and other manufacturers may take a hybrid of these two methods. It is not the interest of this research to provide a capacity allocation strategy. In fact, the production planning and inventory management literature has a rich body of work on the topic of optimizing capacity allocations using mathematical programming models. Readers can refer to Glasserman [3] for detailed treatment. In this research, we adopt

"share the pain", which is the most commonly used method within the studied semiconductor maker. This setting has a plausible effect that allows the decoupling of multiple products and consequently makes analysis much easier.

To address the above questions and concerns occur during the delayed differentiation practice in the studied semiconductor manufacturer, this paper investigates two types of base-stock models under capacitated postponement and is organized as follows. Section 2 studies basic models with simple demand settings, and develops a computationally efficient method to calculate optimal base-stock levels for finished products. Section 3 is devoted to an advanced and more complicated model based upon demand forecasting and forecast lead time. Several results of managerial interests are obtained analytically based on this advanced model. Section 4 carries out computational studies on both randomly generated and real-world data, and provides a managerial insight into the subtle interrelations among demand, inventory, capacity, and lead times that is lack from the analytical results. Section 5 concludes this work and provides future research opportunities.

2. Notations and Basic Models

Let's start with a basic setting where customer demands in different time periods are independent and identically distributed and the distribution function can be derived from demands observed in the past.

It has been found hard to provide a good characterization for real-world demand distributions in the semiconductor industry, even when the demand process is stationary. Most work in the production planning and inventory management literature assumes that

the demand nicely follows a continuous distribution, such as the widely-used normal distribution. However, the study in a typical semiconductor maker shows that demands at the stock-keeping unit (SKU) level hardly follow any known continuous distribution function. Even aggregated demands at the product family level are sometime hard to be mathematically described by known continuous distribution functions. A plausible way to overcome this difficulty is to use discrete demand distributions, e.g. with probability $p(x)$ that the demand will be *x* or within a range close to *x*, where *x* is a non-negative integer. Probabilities can be obtained by studying the histogram of observed demands.

In this setting, a studied manufacturer makes *n* finished products $prod_1$, $prod_2$, ..., $prod_n$, the demands for those products at any given time *t* are independent and denoted by $D_1(t)$, $D_2(t)$, ..., $D_n(t)$ respectively. The finishing capacity, denoted by C, is constant over time t. Assuming enough materials are available in the semi-finished goods inventory, we are interested in finding out the optimal base-stock level s_i^*, for any $prod_i$, that satisfies the pre-determined service level α.

By using the share-the-pain strategy, those *n* products can be decoupled and the capacity is then assigned to each product proportionally to its expected demand. Thus it suffices the purpose to analyze any one of the products. In the following discussion, the product index can be dropped by taking advantage of this convenience. We denote

D_t: Demand at t

IN_t: Net inventory in the warehouse at t

IP_t: Inventory position at t

IO_t: Total of outstanding replenishment orders at t

O_t : Replenishment order placed at t

$p(x)$: Probability of demand at t being equal to x, e.g. $Prob\{D_t = x\}$

s : Base-stock level

C : Finishing capacity

IS_t : Inventory shortfall, which measures the gap between the base-stock level and the inventory position, at t. By definition

$$IS_t = s - IP_t \tag{1}$$

The inventory position is defined by

$$IP_t = IN_t + IO_t \tag{2}$$

2.1. Basic Model with Zero Supply Lead Time

When the supply lead time is zero, there is no outstanding replenishment order, e.g. $IO_t = 0$. Therefore

$$IP_t = IN_t$$

and equation (1) becomes

$$IS_t = s - IN_t \tag{3}$$

It should be clear that the inventory shortfall is caused by the limited finishing capacity. If the finishing line has infinite capacity, IP_t will always be kept at s and so the inventory shortfall will be zero consistently. With inventory shortfall IS_t and demand D_t at t, we need to replenish $IS_t + D_t$ into the warehouse to bring up the inventory position to s. However, limited by the finishing capacity C, we can only replenish $Min\{C, IS_t + D_t\}$. Therefore, at time $t+1$ we have net inventory

$$IN_{t+1} = IN_t + Min\{C, IS_t + D_t\} - D_t$$

By basic algebra, the inventory shortfall at time $t+1$ can be written as

$$IS_{t+1} = s - IN_t - Min\{C, IS_t + D_t\} + D_t = Max\{0, IS_t + D_t - C\} \quad (4)$$

Let $X_t = D_t - C$, we thus rewrite (4) into

$$IS_{t+1} = Max\{0, IS_t + X_t\} = (IS_t + X_t)^+ \quad (5)$$

It is clear that the distribution of random variable X_t is solely determined by that of D_t since C is a constant. Without loss of generality, we assume that at the beginning of the horizon ($t = 0$), the inventory shortfall is zero, e.g. $IS_0 = 0$. Conclusion can be drawn from equation (5) that the inventory shortfall is independent of the base-stock level s and solely determined by X_t, e.g. the capacity and demand. It is important to have $E[X] < 0$, or $E[D] < C$, for the long-run inventory shortfall process to be stable. All discussions in this paper are based on this assumption.

We are interested in the long term steady-state performance of the system. Let all formerly defined variables without the *t* index denote the corresponding steady-state variables, for instance *IN* and *IS* are steady-state net inventory and inventory shortfall respectively. Given a required service level α, the optimal base-stock level can be derived from the equation below

$$\alpha = Prob\{IN \geq 0\} = Prob\{IS \leq s\} \quad (6)$$

And the expected net inventory is

$$E[IN] = s - E[IS] \quad (7)$$

Customer demands in the real world are never infinite large and we can thus estimate both the upper and lower bounds, for instance by studying historical demand data. Let

UB_D be the upper bound and LB_D be the lower bound of customer demands. By assumption, $LB_D < C$ and $UB_D > C$. Demands are discrete and naturally take values from set $\widetilde{D} = \{LB_D, LB_D + 1, LB_D + 2, ..., UB_D\}$ at any time t. Notice we can decrease the dimension of set $\widetilde{D}$ by assuming demands are discrete with positive integral step λ, so D_t takes a value from set $\widetilde{D} = \{LB_D, LB_D + \lambda, LB_D + 2\lambda, ..., UB_D\}$. For instance, the studied semiconductor maker generally measures demand with unit of one hundred, e.g. $\lambda = 100$. When the demand volume is very large, choosing an appropriate value of λ allows us to balance the efficiency and accuracy of optimal base-stock computations. We are also able to estimate the probability of D_t being equal to a given value in set $\widetilde{D}$ from demand histogram

$$p(x) = Prob\{D_t = x, x \in \widetilde{D}\}$$

To calculate the optimal base-stock level from equation (6), we need first find an appropriate and efficient method to compute the inventory shortfall distribution

$$Prob\{IS = n\},\ n \geq 0$$

and further

$$Prob\{IS \leq n\} = \sum_{i=0}^{n} Prob\{IS = i\}$$

2.2. Computation of Inventory Shortfall Distribution

In the literature, equation (5) is known as Lindley recursion (Lindley [4]), which is originated from the studies of the G/G/1 queue and used to characterize the customer waiting-time distribution. The inventory shortfall process $\{IS_t\}_{t=1, 2, 3...}$ can thus be viewed

as a Lindley process. In queuing theory, the process $\{IS_t\}_{t=1, 2, 3...}$ is closely related to the random walk $\{S_t\}_{t=1, 2, 3...}$, where

$$S_1 = 0$$

and

$$S_t = S_{t-1} + X_{t-1} = \sum_{i=1}^{t-1} X_i \ , \ \forall t > 1$$

By reflected random walk, it can be established

$$\begin{aligned} IS_t &= \left(IS_{t-1} + X_{t-1}\right)^+ \\ &= Max\left\{0, S_t - S_{t-1}, S_t - S_{t-2}, ..., S_t - S_1\right\} \\ &= Max\left\{0, X_{t-1}, X_{t-1} + X_{t-2}, ..., X_{t-1} + X_{t-2} + ... + X_1\right\} \end{aligned}$$

X_t's of different time periods are i.i.d. random variables, so we conclude that

X_{t-1} has the same distribution as $S_2 = X_1$

$X_{t-1} + X_{t-2}$ has the same distribution as $S_3 = X_1 + X_2$

…

$X_{t-1} + X_{t-2} + ... + X_2$ has the same distribution as $S_{t-1} = X_1 + X_2 + ... + X_{t-2}$

$X_{t-1} + X_{t-2} + ... + X_1$ has the same distribution as $S_t = X_1 + X_2 + ... + X_{t-1}$ (in fact they are identical)

Clearly the distribution of IS_t is identical to that of $M_t = Max\left\{S_1, S_2, ..., S_t\right\}$ (note that $S_1 = 0$). Notice that M_t is the maximum of the random walk $\{S_t\}_{t=1, 2, 3...}$ during time periods [0, t], it also represents an ascending ladder height (record value) of the random walk in periods [0, t], which is achieved at an ascending ladder point within [0, t] (not

necessarily at *t*). The random walk between any two adjacent ascending ladder points is a probabilistic replica of the random walk from time zero to the first ascending ladder point. The height difference between any two adjacent ascending ladder points is identical to the first ascending ladder height in distribution. Readers can refer to Asseuman [5] and Feller [6] for a rigid and systematic discussion.

Denote by $g^{+}(x)$ the (defective) probability that the first strong ascending ladder height of the random walk is x, $0 < x \le UB_D - C$. Denote by $g^{-}(x)$ the probability that the first weak descending ladder height of the random walk is x, $LB_D - C \le x \le 0$. Let $K = UB_D - C$, if a simple notion is necessary in the following discussion. The steady-state distribution of the inventory shortfall can be obtained as $t \to \infty$

$$\begin{aligned} Prob\{IS = n\} &= Prob\{M_\infty = n\} \\ &= \sum_{i=1}^{K} Prob\{M_\infty = n - i\} \cdot g^{+}(i) \\ &= \sum_{i=1}^{K} Prob\{IS = n - i\} \cdot g^{+}(i) \end{aligned} \tag{8}$$

Let $q_n = Prob\{IS = n\}$ and $a_i = g^{+}(i)$, we turn equation (8) into

$$q_n = \sum_{i=1}^{K} q_{n-i} \cdot a_i, \quad \forall n > 0 \tag{9}$$

The inventory shortfall will never be negative, thus

$$q_n = 0, \quad \forall n < 0 \tag{10}$$

The sum of inventory shortfall probabilities from negative infinity to positive infinity equals one, so that

$$\sum_{n=-\infty}^{\infty} q_n = 1 \tag{11}$$

Equations (9), (10) and (11) eventually lead to

$$q_0 = 1 - \sum_{i=1}^{K} a_i \tag{12}$$

To verify equation (12), we start with (11), which is equivalent to

$$q_0 + \sum_{n=1}^{\infty} q_n = 1$$

It can be seen

$$\sum_{n=1}^{\infty} q_n = \sum_{n=1}^{\infty} \sum_{i=1}^{K} q_{n-i} \cdot a_i = \sum_{i=1}^{K} a_i \cdot \sum_{n=1}^{\infty} q_{n-i} \tag{13}$$

Notice

$$\sum_{n=1}^{\infty} q_{n-i} = \sum_{n=1}^{i-1} q_{n-i} + \sum_{n=i}^{\infty} q_{n-i} = 0 + \sum_{n=i}^{\infty} q_{n-i} = \sum_{j=0}^{\infty} q_j = 1 \tag{14}$$

Inserting (14) into (13), we obtain

$$\sum_{n=1}^{\infty} q_n = \sum_{i=1}^{K} a_i$$

which implicates

$$q_0 = 1 - \sum_{n=1}^{\infty} q_n = 1 - \sum_{i=1}^{K} a_i$$

In order to compute the inventory shortfall distribution from equations (9) and (12), we first need to obtain the value of a_i, e.g. $g^+(i)$. It has been shown in queuing theory (Asseuman [5]) that

$$p(x) = g^+(x) + g^-(x) - g^+ * g^-(x) \tag{15}$$

By definition, the convolution in equation (15) is written as

$$g^+ * g^-(x) = \sum_{y=LB_D-C}^{0} g^+(x-y)\cdot g^-(y) \tag{16}$$

By inserting (16) into (15) and reformatting, it is obtained

$$g^-(x) = p(x) - g^+(x) + \sum_{y=LB_D-C}^{0} g^+(x-y)\cdot g^-(y) \tag{17}$$

and

$$g^+(x) = p(x) - g^-(x) + \sum_{y=LB_D-C}^{0} g^+(x-y)\cdot g^-(y)$$

$$= p(x) + \sum_{y=LB_D-C}^{-1} g^+(x-y)\cdot g^-(y) + g^+(x)\cdot g^-(0)$$

Therefore, $g^+(x)$ can be written as

$$g^+(x) = \frac{p(x) + \sum_{y=LB_D-C}^{-1} g^+(x-y)\cdot g^-(y)}{1-g^-(0)} \tag{18}$$

Based upon equations (17) and (18), a simple iterative algorithm **- Subroutine-RW**, which is close to the ones Grassmann and Jain [7] proposed to calculate the waiting-time distribution in GI/G/1 queues, can be devised to approach the true values of $g^+(x)$. Iteration index i is used in the algorithmic description, e.g. $g_i^+(x)$ and $g_i^-(x)$ are the values of $g^+(x)$ and $g^-(x)$ in iteration i.

Subroutine-RW

Begin

Step 1: Initialize $g_0^+(x) = 0$ and $g_0^-(x) = 0$. Set computational gap tolerance ε to a small number.

Step 2: For i =1, 2, 3, …, do:

2.1: Compute $g_i^-(x)$, $LB_D - C \le x \le 0$ (notice $g_i^+(x) = 0$ when $LB_D - C \le x \le 0$)

$$g_i^-(x) = p(x) - g_{i-1}^+(x) + \sum_{y=LB_D-C}^{0} g_{i-1}^+(x-y) \cdot g_{i-1}^-(y)$$

$$= p(x) + \sum_{y=LB_D-C}^{0} g_{i-1}^+(x-y) \cdot g_{i-1}^-(y)$$

2. 2: Calculate $g_i^+(x)$, $0 < x \le UB_D - C$ (notice $g_{i-1}^-(x) = 0$ when $0 < x \le UB_D - C$)

$$g_i^+(x) = \frac{p(x) + \sum_{y=LB_D-C}^{-1} g_{i-1}^+(x-y) \cdot g_i^-(y)}{1 - g_i^-(0)}$$

2.3: If $\sum_{x=1}^{UB_D-C} \left| g_i^+(x) - g_{i-1}^+(x) \right| < \varepsilon$, exit and report $g_i^+(x)$ as $g^+(x)$.

End

Then by equations (9) and (12), the inventory shortfall distribution can also be calculated by an iterative algorithm - **Procedure-IS**, which is described as

Procedure-IS

Begin

Step 1: Initialize $q_i = 0$, $1 - K \le i < 0$. Set computational gap tolerance φ to a small number.

Step 2: Call **Subroutine-RW** to obtain $g^+(i)$, $0 < i \le K$

Step 3: Compute

$$q_0 = 1 - \sum_{i=1}^{K} g^+(i)$$

Step 4: For n =1, 2, 3, …, do:

4.1: Compute

$$q_n = \sum_{i=1}^{K} q_{n-i} \cdot g^+(i)$$

4.2: If $1 - \sum_{i=0}^{n} q_i < \varphi$, exit and report q_i, $0 \le i \le n$, as the inventory shortfall distribution.

End

2.3. Basic Model with Positive Supply Lead Time

With some modification, the basic model can be extended to solve more realistic problems where the supply lead time or finishing lead time is a positive integer L_R. At time t, we review the begin-on-hand net inventory IN_t and inventory position IP_t, calculate the inventory shortfall IS_t, receive the replenishment order placed L_R periods ago, accept and fulfill customer demand D_t, and then place replenishment order O_t.

When we review the net inventory and inventory position at time *t*, the total of outstanding replenishment orders is expressed as

$$IO_t = \sum_{\tau=t-L_R}^{t-1} O_\tau$$

since replenishment orders placed before $t - L_R$ have been received at *t* and order O_t has not been placed yet.

By equations (1) and (2), we write the inventory shortfall as

$$IS_t = s - IN_t - \sum_{\tau=t-L_R}^{t-1} O_\tau$$

Note O_{t-L_R} has been received at $t+1$ when we review the net inventory, therefore

$$IN_{t+1} = IN_t + O_{t-L_R} - D_t$$

Limited by the finishing capacity C, replenishment order at t is $O_t = Min\{C, IS_t + D_t\}$. By basic algebra, one may verify that the inventory shortfall at $t+1$ can be written as

$$\begin{aligned} IS_{t+1} &= IS_t + D_t - Min\{C, IS_t + D_t\} \\ &= Max\{0, IS_t + D_t - C\} \end{aligned}$$

which is still in the form of equation (4). So the distribution of *IS* can still be computed by the **Procedure-IS**.

Given a service level α, the optimal base-stock level can be computed from the equation below

$$\begin{aligned} \alpha &= Prob\{IN_t + O_{t-L_R} - D_t \geq 0\} \\ &= Prob\left\{IS_t + \sum_{\tau=t-L_R+1}^{t-1} Min\{C, IS_\tau + D_\tau\} + D_t \leq s\right\} \end{aligned} \tag{19}$$

Notice $Min\{C, IS_t + D_t\} = IS_t + D_t - IS_{t+1}$, therefore

$$\begin{aligned} IS_t + \sum_{\tau=t-L_R+1}^{t-1} Min\{C, IS_\tau + D_\tau\} + D_t &= IS_t + D_t + \sum_{\tau=t-L_R+1}^{t-1}\left(IS_\tau + D_\tau - IS_{\tau+1}\right) \\ &= IS_{t-L_R+1} + \sum_{\tau=t-L_R+1}^{t} D_\tau \end{aligned} \tag{20}$$

Inserting (20) into (19), we obtain

$$\alpha = Prob\left\{IS_{t-L_R+1} + \sum_{\tau=t-L_R+1}^{t} D_\tau \leq s\right\} \tag{21}$$

It can easily be verified that random variable IS_{t-L_R+1} is independent of D_τ, $\forall \tau \in [t-L_R+1, t]$. Customer demands at different time are generally considered as i.i.d.

random variables, the distribution of lead time demand $\sum_{\tau=t-L_R+1}^{t} D_\tau$ may be obtained either computationally from D_τ or empirically from historical observations. After the inventory shortfall distribution being calculated by the **Procedure-IS**, the optimal base-stock level can then be obtained from equation (21) computationally.

The expected net inventory at *t* is given by

$$E[IN_t] = s - E[IS_t] - E\left[\sum_{\tau=t-L_R}^{t-1} O_\tau\right]$$

It is clear that when $E[D] < C$ the inventory shortfall process will be stable in the long run, and so will be the replenishment order process. In a stable inventory system, the expected demand equals the expected replenishment order, e.g. $E[D] = E[O]$. Therefore

$$E[IN] = s - E[IS] - L_R \cdot E[D] \qquad (22)$$

3. Advanced Model with Demand Forecasting and Demand Lead Time

When supply/manufacturing lead times are long, it is common for manufactures to forecast customer demands and then build to forecast. This is particularly true for semiconductor makers since wafer fabrications generally take several months and assembly and test processes commonly require one or two weeks. Build to forecast essentially helps manufactures gain demand lead time, which is a counterpart of and thus offsets the supply lead time according to Hariharan and Zipkin [8], if demand is perfect and capacity is infinite. Nevertheless, forecasted demand is hardly perfect in the real-world. Let the demand forecast at time *t* be F_t, which is expected to be materialized into

real customer demand D_{t+L_D} at time period $t+L_D$, where L_D is the demand lead time or forecast lead time. Demand D_t can be written as

$$D_t = F_{t-L_D} + e_t$$

where e_t is the demand forecast error. When the used forecasting method is appropriate, forecast errors are independent and identically distributed random variables with zero mean, also the forecast error e_t is independent of the forecast F_{t-L_D}. The variability of forecast errors, measured by $\sigma(e)$, generally increases in the forecast lead time L_D according to Magee, Copacino and Rosenfield [9].

At time t, we review the begin-on-hand net inventory IN_t and inventory position IP_t, calculate the inventory shortfall IS_t, receive the replenishment order placed L_R periods ago, accept and fulfill customer demand D_t, make demand forecast F_t for the demand that will be revealed L_D periods later, and place replenishment order O_t.

Denote by IU_t the total of forecasted but unrevealed customer demands when the net inventory and inventory position are reviewed at *t*, therefore

$$IU_t = \sum_{\tau=t-L_D}^{t-1} F_\tau$$

To fully utilize the benefit brought by advance-demand information, we redefine the inventory position at *t* as (see Tang [10] for details)

$$IP_t = IN_t + IO_t - IU_t = IN_t + \sum_{\tau=t-L_R}^{t-1} O_\tau - \sum_{\tau=t-L_D}^{t-1} F_\tau$$

It has been shown by Tang [10] that placing replenishment order $F_t + e_t$ can always keep the inventory position stationary at the base-stock level of *s* when the capacity is infinite. However, when the capacity is limited by *C*, the replenishment order size at *t* is

$$O_t = Min\{C, IS_t + F_t + e_t\}$$

It can be easily verified that

$$IN_{t+1} = IN_t + O_{t-L_R} - D_t$$

We still using equation (1) to define the inventory shortfall, therefore

$$IS_{t+1} = s - IN_{t+1} - \sum_{\tau=t-L_R+1}^{t} O_\tau + \sum_{\tau=t-L_D+1}^{t} F_\tau$$

By basic algebra, we can show

$$\begin{aligned} IS_{t+1} &= IS_t + F_t + e_t - Min\{C, IS_t + F_t + e_t\} \\ &= Max\{0, IS_t + F_t + e_t - C\} \end{aligned} \quad (23)$$

Generally speaking, forecast errors e_t and e_{t+L_D} are identical in distribution, also F_t and e_t are independent, so $F_t + e_t$ is distributionally identical to D_{t+L_D} and then D_t. Therefore, according to equation (23), the long-run inventory shortfall is stable given $E[D] < C$ and its distribution can also be calculated by the **Procedure-IS**.

Given a required service level α, the optimal base-stock level can be computed from the equation below

$$\begin{aligned} \alpha &= Prob\{IN_t + O_{t-L_R} - D_t \geq 0\} \\ &= Prob\left\{IS_t + \sum_{\tau=t-L_R+1}^{t-1} Min\{C, IS_\tau + F_\tau + e_\tau\} - \sum_{\tau=t-L_D}^{t-1} F_\tau + D_t \leq s\right\} \end{aligned} \quad (24)$$

Notice $Min\{C, IS_t + F_t + e_t\} = IS_t + F_t + e_t - IS_{t+1}$, hence

$$IS_t + \sum_{\tau=t-L_R+1}^{t-1} Min\{C, IS_\tau + F_\tau + e_\tau\} - \sum_{\tau=t-L_D}^{t-1} F_\tau + D_t$$

$$= IS_t + \sum_{\tau=t-L_R+1}^{t-1} \left(IS_\tau + F_\tau + e_\tau - IS_{\tau+1}\right) - \sum_{\tau=t-L_D}^{t-1} F_\tau + D_t$$

$$= \begin{cases} IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_\tau, & \text{if } L_R > L_D \\ IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t + \sum_{\tau=t-L_R+1}^{t+L_D-L_R} e_\tau, & \text{if } L_R \le L_D \end{cases}$$

Therefore, equation (24) becomes

$$\alpha = \begin{cases} Prob\left\{ IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_\tau \le s \right\}, & \text{if } L_R > L_D \\ Prob\left\{ IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t + \sum_{\tau=t-L_R+1}^{t+L_D-L_R} e_\tau \le s \right\}, & \text{if } L_R \le L_D \end{cases} \quad (25)$$

The net inventory at t is of the form

$$IN_t = s - IS_t - \sum_{\tau=t-L_R}^{t-1} O_\tau + \sum_{\tau=t-L_D}^{t-1} F_\tau$$

$$= s - IS_t - \sum_{\tau=t-L_R}^{t-1} O_\tau + \sum_{\tau=t}^{t+L_D-1} D_\tau - \sum_{\tau=t}^{t+L_D-1} e_\tau$$

Using the same argument in subsection 2.3, we can establish $E[D] = E[O]$ when the inventory shortfall process is stable in the long run. The expected net inventory is then written as

$$E[IN] = s - E[IS] - L_R \cdot E[O] + L_D \cdot E[D] = s - E[IS] - (L_R - L_D) \cdot E[D] \quad (26)$$

Under the infinite capacity assumption, it has been argued by Hariharan and Zipkin [8] that it is unnecessary to have the demand lead time being larger than the supply lead time. However, when finishing lines are severely capacitated, intuitions often guides

manufactures to build to forecast even more than finishing lead times ahead, e.g. $L_D > L_R$. It is not immediately clear whether or not these intuitions are valid and make mathematical sense. To simplify the analysis, we consider the most favorable case in which the advance-demand information is perfect or the demand forecast is accurate with $e_t = 0, \forall t$. Then equation (25) turns into

$$\alpha = \begin{cases} Prob\left\{ IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau \le s \right\}, & \text{if } L_R > L_D \\ Prob\left\{ IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t \le s \right\}, & \text{if } L_R \le L_D \end{cases} \tag{27}$$

When the inventory shortfall reaches its long-run steady state as $t \to \infty$, given a required service level α, equation (27) indicates that the optimal base-stock level is determined by the mean and variance of random variables $IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau$ when $L_R > L_D$ and $IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t$ when $L_R \le L_D$. Denote by $\sigma(\cdot)$ the standard deviation of a random variable and $f(\sigma(\cdot))$ the portion of optimal base-stock level determined by the random variable's variance. Note the function $f(\cdot)$ increases in σ to reflect the fact that the more fluctuation the demand is the larger base-stock level is required to meet the same service level.

When $L_R > L_D$, we may write the optimal base-stock level s^* as

$$s^* = \lim_{t\to\infty} \left\{ E\left[IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau \right] + f\left(\sigma\left(IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau \right) \right) \right\}$$

$$= E[IS] + (L_R - L_D) \cdot E[D] + \lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau \right) \right)$$

Therefore, by equation (26), the expected net inventory under the optimal base-stock policy is

$$E[IN] = \lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau\right)\right) \tag{28}$$

When $L_R \le L_D$, s^* then becomes

$$\begin{aligned} s^* &= \lim_{t\to\infty}\left\{E\left[IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t\right] + f\left(\sigma\left(IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t\right)\right)\right\} \\ &= E[IS] - (L_D - L_R)\cdot E[D] + \lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t\right)\right) \end{aligned}$$

So obtained the expected net inventory under the optimal base-stock policy

$$E[IN] = \lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1} - \sum_{\tau=t}^{t+L_D-L_R} D_\tau + D_t\right)\right) \tag{29}$$

Random variable IS_{t-L_R+1} is independent of D_τ ($\forall \tau \in [t + L_D - L_R + 1, t]$ when $L_R > L_D$ and $\forall \tau \in [t, t + L_D - L_R]$ when $L_R \le L_D$). We can thus conclude from equations (28) and (29) that the demand lead time L_D effectively offsets the supply lead time L_R in terms of reducing net inventories, just as Hariharan and Zipkin [8] showed with an infinite capacity assumption. It can also be easily verified from equations (28) and (29) that $E[IN]$ reaches its minimum, e.g. $\lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1}\right)\right)$, when $L_D = L_R$. This indicates that even if the finishing line is capacitated and the forecasted ADI is perfect, it is better not to build to forecast more than L_R time periods ahead. So we can immediately conclude that when forecasted ADI is imperfect, we shall never build to forecast more than L_R time periods ahead as well. Given the fact that the variance of forecast errors increases in the forecast lead time L_D, we would practically want $L_D < L_R$.

Considering forecast errors, with the same treatment as above we may write the optimal base-stock level as

$$s^* = \lim_{t\to\infty}\left\{E\left[IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_\tau\right] + f\left(\sigma\left(IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_\tau\right)\right)\right\}$$

$$= E[IS] + (L_R - L_D)\cdot E[D] + \lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_\tau\right)\right)$$

Then the expected net inventory under the optimal base-stock policy is

$$E[IN] = \lim_{t\to\infty} f\left(\sigma\left(IS_{t-L_R+1} + \sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_\tau\right)\right) \tag{30}$$

The inventory shortfall distribution calculated by the **Procedure-IS** is irrelevant to L_D. Random variable IS_{t-L_R+1} is independent of D_τ ($\forall \tau \in [t + L_D - L_R + 1, t]$) and e_τ ($\forall \tau \in [t - L_R + 1, t + L_D - L_R]$). Function $f(\cdot)$ increases in σ . So it is seen from equation (30) that the minimum $E[IN]$ is achieved at the point where $\sigma\left(\sum_{\tau=t-L_R+L_D+1}^{t} D_\tau + \sum_{\tau=t-L_R+1}^{t-L_R+L_D} e_t\right)$ is minimized. This indicates that it makes sense to use forecasted ADI if and only if $\sigma(e) < \sigma(D)$. This also suggests that the optimal forecast lead time that minimizes $E[IN]$ is irrelevant to the inventory shortfall. Empirically, $\sigma(e)$ of different forecast lead times can often be expressed as a function of L_D, for instance Magee, Copacino and Rosenfield [9] use

$$\sigma(e) = \eta \cdot L_D^{\beta}$$

to describe the interrelation between the forecast error variability and the forecast lead time, where η and β are constant parameters. So the optimal forecast lead time can be

analytically obtained in the same way as if the capacity is unlimited. Readers are referred to Tang [10] for details on computing the optimal forecast lead time under an infinite capacity assumption.

4. Numerical Study

Due to the lack of closed-form representations for the inventory shortfall distribution, we are limited from analytical results that can elegantly describe the interrelations among demand, inventory, capacity, and lead times. In this work, numerical approaches are used to study these relationships and provide some managerial insights. We are going to introduce two sets of experiments. The first set is composed of several cases where demands are randomly generated according to different normal distributions. The second set consists of two real-world cases from the studied semiconductor manufacture. The demands in the second set hardly follow the widely used normal distribution assumption. In fact, the historical demand pattern of one case reveals some time series properties and the demand distribution of the other case has a long right tail. For each case in the two sets of experiments, we study four capacitated scenarios where the capacity is of the form

$$C = \mu_D + \xi \cdot \sigma_D$$

where μ_D and σ_D are the mean and standard deviation of demands respectively. ξ is a parameter chosen to be corresponding to the four capacitated scenarios, e.g. $\xi = 0.5$, $\xi = 1$, $\xi = 2$, and $\xi = 3$. We further benchmark those four capacitated scenarios against an incapacitated one where $C = \infty$. The service level is set to 0.95 for all test cases. In all inventory shortfall computations, the gap tolerances ε and φ are all set to 10^{-9}. It has been observed that the **Procedure-IS** converges fast on a moderately configured

computer (Pentium M 1.8GHz processor with 1.0G RAM) and computations finished within 8 seconds for all test cases in the two experiment sets.

In first set of experiments, we shall consider both zero and non-zero supply lead time. We particularly choose the non-zero supply lead time to be 7 time periods (days) since the assembly and test throughput time in the studied semiconductor maker is generally close to this number. In this experiment set, we set up three demand groups which represent small, medium and large demand volumes. Further each demand group is composed of three scenarios where demand variances are small, medium and large relatively to demand means.

Table 1: Results from randomly generated experiment cases

			$L_R = 0$					$L_R = 7$				
			$C = \infty$	$\xi = 0.5$	$\xi = 1$	$\xi = 2$	$\xi = 3$	$C = \infty$	$\xi = 0.5$	$\xi = 1$	$\xi = 2$	$\xi = 3$
$\mu_D = 150$	$\sigma_D = 30$	s	0	73	26	0	0	1181	1206	1186	1181	1181
		$E[IS]$	0	16.52	3.69	0.17	0	0	16.52	3.69	0.17	0
		$E[IP]$	0	57	23	0	0	1181	1189	1183	1181	1181
		$E[IN]$	**0**	**57**	**23**	**0**	**0**	**131**	**140**	**133**	**131**	**131**
	$\sigma_D = 60$	s	0	149	56	0	0	1311	1364	1323	1313	1312
		$E[IS]$	0	33.18	8.07	0.58	0.03	0	33.18	8.07	0.58	0.03
		$E[IP]$	0	116	48	0	0	1311	1331	1315	1312	1312
		$E[IN]$	**0**	**116**	**48**	**0**	**0**	**261**	**281**	**265**	**262**	**262**
	$\sigma_D = 90$	s	0	233	83	0	0	1442	1534	1459	1444	1442
		$E[IS]$	0	57.39	12.05	0.80	0.01	0	57.39	12.05	0.80	0.01
		$E[IP]$	0	175	71	0	0	1442	1477	1447	1443	1442
		$E[IN]$	**0**	**175**	**71**	**0**	**0**	**392**	**427**	**397**	**393**	**392**
$\mu_D = 550$	$\sigma_D = 110$	s	0	259	96	0	0	4329	4418	4349	4331	4329
		$E[IS]$	0	56.79	13.58	0.89	0.05	0	56.79	13.58	0.89	0.05
		$E[IP]$	0	202	82	0	0	4329	4361	4335	4330	4329
		$E[IN]$	**0**	**202**	**82**	**0**	**0**	**479**	**511**	**485**	**480**	**479**
	$\sigma_D = 220$	s	0	526	193	0	0	4807	4999	4852	4814	4808
		$E[IS]$	0	119.42	27.79	1.99	0.07	0	119.42	27.79	1.99	0.07
		$E[IP]$	0	406	165	0	0	4807	4880	4824	4812	4808
		$E[IN]$	**0**	**406**	**165**	**0**	**0**	**957**	**1030**	**974**	**962**	**958**
	$\sigma_D = 330$	s	0	823	293	0	0	5287	5632	5351	5292	5288
		$E[IS]$	0	213.89	45.03	2.92	0.11	0	213.89	45.03	2.92	0.11
		$E[IP]$	0	609	248	0	0	5287	5418	5306	5289	5288
		$E[IN]$	**0**	**609**	**248**	**0**	**0**	**1437**	**1568**	**1456**	**1439**	**1438**
$\mu_D = 1000$	$\sigma_D = 200$	s	0	488	181	0	0	7870	8039	7905	7873	7870
		$E[IS]$	0	108.56	25.44	1.68	0.04	0	108.56	25.44	1.68	0.04
		$E[IP]$	0	379	156	0	0	7870	7930	7880	7871	7870
		$E[IN]$	**0**	**379**	**156**	**0**	**0**	**870**	**930**	**880**	**871**	**870**
	$\sigma_D = 400$	s	0	967	357	0	0	8741	9074	8812	8746	8741
		$E[IS]$	0	213.50	50.07	3.43	0.09	0	213.50	50.07	3.43	0.09
		$E[IP]$	0	754	307	0	0	8741	8860	8762	8743	8741
		$E[IN]$	**0**	**754**	**307**	**0**	**0**	**1741**	**1860**	**1762**	**1743**	**1741**
	$\sigma_D = 600$	s	0	1505	550	0	0	9612	10191	9710	9618	9612
		$E[IS]$	0	363.30	78.73	5.43	0.28	0	363.30	78.73	5.43	0.28
		$E[IP]$	0	1142	471	0	0	9612	9828	9631	9613	9612
		$E[IN]$	**0**	**1142**	**471**	**0**	**0**	**2612**	**2828**	**2631**	**2613**	**2612**

Results from the first experiment set are list in table 1. We use expected net inventories under the optimal base-stock policy to gauge the results since expected net inventories are directly related to expected inventory cost. Several observations can be made from table 1.

i. Under a same demand and capacity setting, expected net inventories with non-zero supply lead time are higher than those with zero supply lead time.

ii. With a same demand mean, expected net inventories very much proportionally increase in the demand standard deviation σ_D regardless of the supply lead time and capacity.

iii. Given the same demand setting, expected net inventories with capacity limitations are higher than those with infinite capacities. However, after capacities reach a certain level, e.g. $\mu_D + 2 \cdot \sigma_D$, the inventory-saving benefit quickly vanishes.

The second set of experiments consists of two real-world cases. Demand data reported here is uniformly scaled in order to protect confidential information of the studied semiconductor maker. Nonetheless, the demand properties we are interested in are still preserved. The historical demand pattern of case 1 is plotted in figure 1. Clearly it has non-negligible time series characteristics such as seasonality and trend.

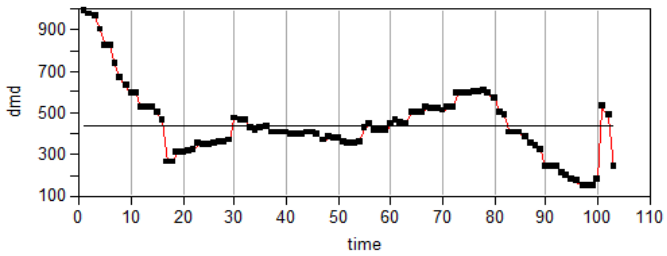

Figure 1: Historical demand pattern of case 1

The distribution and normal probability plot shown in figures 2 and 3 indicate the real-world demands in this case can hardly be described by a normal distribution.

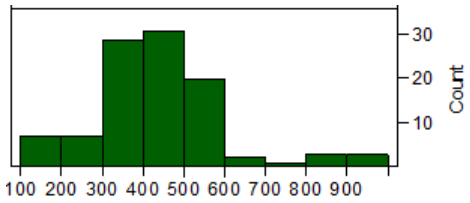

Figure 2: Demand distribution

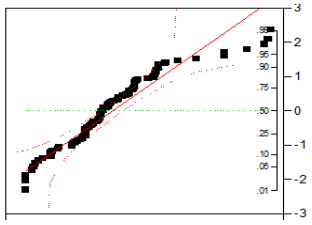

Figure 3: Demand normal probability plot

The demand mean μ_D and standard deviation σ_D of this case are 440 and 168 respectively.

The historical demand pattern of the real-world case 2 is plotted in figure 4, and the demand distribution and normal probability plot are displayed in figures 5 and 6. The demand mean μ_D and standard deviation σ_D are 147 and 209 respectively. This case represents a very typical demand situation in the studied semiconductor maker: the demand distribution is severely skewed and has a long right tail. Lognormal distributions have been found more suitable than commonly used normal distributions.

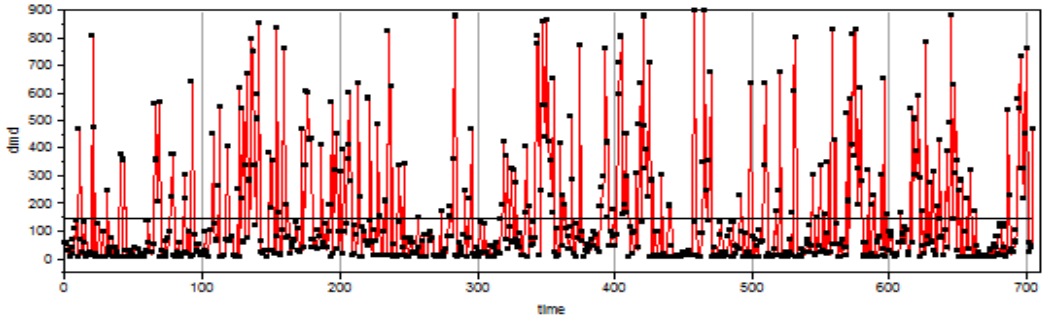

Figure 4: Historical demand pattern of case 2

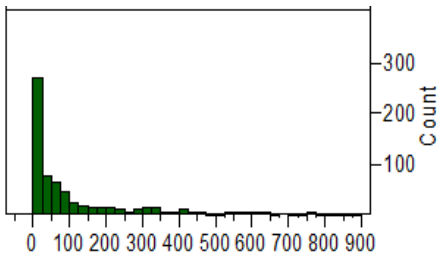

Figure 5: Demand distribution

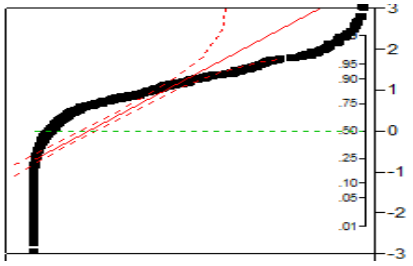

Figure 6: Demand normal probability plot

In this set of experiments, the supply lead time is still set to 7 time periods. By the gradient boosting tree (GBT) method (Friedman [11] and Schapire [12]) with historical demands and market movements as predictors, forecasted ADI is obtained and its value is analyzed by comparing net inventories computed from models with and without demand forecasting under capacitated postponement. In fact, demand forecasting in the studied semiconductor manufacture is quite challenging and the variances of forecast errors are fairly large. Nevertheless, the experiment results in table 2 still show significant benefits from forecasted ADI: by building to forecast which is made 2 time periods ahead of realization, e.g. $L_D = 2$, net inventories can be reduced by 7%~10%. Another observation can be made is that the inventory-saving benefit quickly vanishes after $\xi = 2$, just as we concluded from table 1 with randomly generated experiment cases.

Table 2: Results from real-world experiment cases

		Use No Forecasted ADI $L_D = 0$, $L_R = 7$					Use Forecasted ADI $L_D = 2$, $L_R = 7$ Case1: $\sigma(e) = 91$; Case 2: $\sigma(e) = 139$				
		$C = \infty$	$\xi = 0.5$	$\xi = 1$	$\xi = 2$	$\xi = 3$	$C = \infty$	$\xi = 0.5$	$\xi = 1$	$\xi = 2$	$\xi = 3$
Case 1 $\mu_D = 440$ $\sigma_D = 168$	*s*	3835	4021	3889	3847	3836	2891	3086	2946	2901	2891
	E[*IS*]	0	104.52	34.31	8.47	0.82	0	104.52	34.31	8.47	0.82
	E[*IP*]	3835	3916	3855	3839	3835	2891	2981	2912	2893	2890
	E[*IN*]	**761**	**842**	**781**	**764**	**761**	**690**	**781**	**711**	**692**	**690**
Case 2 $\mu_D = 147$ $\sigma_D = 209$	*s*	2044	2322	2128	2059	2046	1658	1947	1753	1676	1661
	E[*IS*]	0	156.87	56.53	12.84	1.70	0	156.87	56.53	12.84	1.70
	E[*IP*]	2044	2165	2071	2046	2044	1658	1790	1696	1663	1659
	E[*IN*]	**1013**	**1134**	**1041**	**1015**	**1013**	**926**	**1058**	**965**	**931**	**927**

5. Conclusions

To understand the interrelations among demand, inventory, capacity, and lead times under capacitated delayed differentiation, this work studied two types of bas-stock inventory models. The basic ones do not consider demand forecasting. The advanced model, which takes positive supply lead time and forecasted ADI into account, are more

complicated and realistic. A practical and computationally efficient method to calculate optimal base-stock levels was developed from the basic models and can however be applied to the advanced model. Finishing capacities often force manufactures practicing postponement to build ahead according to demand forecast. This work justified the value of forecasted ADI under capacitated postponement. However, forecasted ADI can easily be abused. When the limitation of finishing capacities becomes severe, intuitions often guide manufacturers to build to forecast even more than finishing lead times ahead. Results in this research indicate that, if demands are stationary, these intuitions are invalid and even under capacitated postponement it may not be beneficial to build to forecast more than supply lead times ahead. Closer studies on the advanced model revealed that the forecasted ADI is only useful when the variance of forecast errors is less than that of demands, and showed that the optimal forecast lead time can actually be obtained in the same way as if finishing lines are incapacitated.

General as it is, this work doesn't rule out the validity of the practice to use forecasted ADI with the forecast lead time being longer than the supply lead time, if the basic assumptions that demands are stationary and $E[D]$ is less than C do not hold. This practice applies to products in their early or late lifecycle periods. During these periods, the demands are fast ramping up or ramping down, and often manufacturing capacities are not fully committed due to the risk of demand shift. Another example where this practice may prevail is short-lifecycle products such as fashion goods. Customer demands usually concentrate in a short season and the total demand volume can far exceed the production capacity and thus producers have to make demand forecast and then build to forecast long ahead of demand materialization. In fact, lifecycle-based inventory systems

under capacitated postponement are interesting research topics with tremendous practical value. Studies on these topics will provide valuable extensions to this work.

References

[1] Zipkin P. Foundations of inventory management. New York: Mc-Graw Hill Higher Education; 2000.

[2] Federgruen A, Zipkin P. An inventory model with limited production capacity and uncertain demands, I: the average-cost criterion. Mathematics of Operations Research 1986; 11: 193-207.

[3] Glasserman P. Allocating production capacity among multiple products. Operations Research 1996; 44: 724-34.

[4] Lindley DV. The theory of a queue with a single server. Mathematical Proceedings of the Cambridge Philosophical Society 1952; 48: 277-89.

[5] Asmussen S. Applied probability and queues, 2nd ed. New York: Springer; 2003.

[6] Feller W. An introduction to probability theory and its applications (vol. 2), 3rd ed. New York: Wiley; 1968

[7] Grassmann WK, Jain JL. Numerical solutions of the waiting time distribution and idle time distribution of the arithmetic GI/G/1 queue. Operations Research 1989; 37: 141-50.

[8] Hariharan R, Zipkin P. Customer-order information, lead times, and inventories. Management Science 1995; 41: 1599-607.

[9] Magee JF, Copacino WC, Rosenfield DB. Modern logistics management - integrating marketing, manufacturing, and physical distribution. New York: Wiley; 1985.

[10] Tang D. Inventory planning with forecasted advance demand information. Working Paper; 2009.

[11] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001; 29: 1189-232.

[12] Schapire RE. The boosting approach to machine learning: an overview. In: Denison DD, Hansen MH, Holmes C, Mallick B, Yu B, editors. Nonlinear estimation and classification. New York: Springer; 2003, p. 149-72.