Omega 39 (2011) 481–492

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Managing finished-goods inventory under capacitated delayed differentiation

Dong Tang *,1

Customer Fulfillment, Planning and Logistics Group, Intel Corporation, 5000 West Chandler Boulevard, Chandler, AZ 85226, USA

A R T I C L E I N F O

Article history:
Received 12 December 2009
Accepted 20 October 2010
Available online 30 October 2010

Keywords:
Capacity
Delayed differentiation
Forecasted advance-demand information

A B S T R A C T

Delayed differentiation or postponement is widely advocated to mitigate conflicts between product diversity and inventory cost savings. Manufacturers practicing postponement often suffer from severely constrained finishing capacities and noticeable finishing lead times. Therefore, inventories are still needed for finished products. Using the concept of inventory shortfall, 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 the capacity reaches a certain level. The value of forecasted advance-demand information (ADI) to postponement is justified, but can easily be overstated. Finishing capacities usually force manufacturers to build ahead according to demand forecast. When capacity limitation becomes severe, intuitions often guide producers to build to forecast even more than finishing lead times ahead. Results of this research indicate that these intuitions may be invalid and build to forecast more than finishing lead times ahead may not be a good practice. Further studies reveal that under capacitated postponement the forecasted advance-demand information is useful 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 the capacity is unlimited.

1. Introduction and background

Delayed differentiation or postponement maintains the product commonality as late as possible until product differentiation is necessary. It has been conceived as an effective way in both academia and industry to deal with the conflicts between product diversity and inventory cost savings. A classical delayed differentiation case study was reported by Bruce [1] on Benetton. By reversing the sequence in which yarn is knitted and dyed, Benetton successfully postponed the sweater color selection until the seasonal fashion is known. Thereafter, stories of successful postponement were told in various industry segments [2–6]. Zinn and Bowersox [7] realized that there are different types of postponement and classified them into five categories: shipping, labeling, packaging, assembly, and manufacturing. Further, for each category they developed a cost model that helps managers to justify the postponement. Abundant quantitative studies in the literature [8–12] showcase models that focus on product/process redesign to implement postponement by assessing the tradeoff between the benefits of inventory pooling and redesign costs. Lee and Tang [13] went a step further to develop a model for evaluating the benefits and costs associated with postponement achieved through a broader range of approaches such as standardization, modular design and process restructuring. In a similar effort, Ernst and Kamrad [14] explored the value of different supply chain structures in the context of modularization and postponement, and showed that it is of operational advantage to make combined modularization and postponement decisions. Recently, Hallgren and Olhager [15] did an empirical research on the relationship between volume and product mix flexibility, a combination that postponement concerns. They found that different levels of volume and mix flexibility combinations have significant impact on the operational performance.

Though understanding the benefits of postponement, many industrial practitioners, especially semiconductor manufacturers, often find it is too costly, sometimes unrealistic, to enable postponement by sweeping changes that include redesigning and modularizing their products/processes. Small and incremental changes to the existing processes are more welcomed. In the studied semiconductor manufacturer, for instance, a "fuse-to-order" type of postponement, which suggests making a small change to the test/finishing flow by holding components at the semi-finished goods inventory after the test stage and fusing them into finished products according to orders, gained lots of traction because no additional expensive equipment or product redesign is needed. Delayed differentiation can significantly reduce inventory costs, but it does not completely remove the necessity of keeping finished-goods inventory largely because of finishing lead time and finishing capacity. When implementing postponement, this

* Tel.: +1 480 552 6562 (work); +1 217 721 5311 (cell).
E-mail addresses: dong.tang@intel.com, td.rabi@gmail.com
1 Present address: 2130 East Saltsage Drive, Phoenix, AZ 85048, USA.

0305-0483/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.omega.2010.10.006

semiconductor manufacturer faced a hard problem to set finished-goods inventory targets given the fact that its finishing line capacities are usually limited.

Finishing capacities often drive manufacturers to build ahead according to demand forecasts. When the limitation of finishing capacities becomes severe, intuition often guides producers to build to the forecast even over a longer period than the finishing lead time. 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 additional weeks in the demand forecast is adopted, though we are not clear whether or not this practice is rational and makes theoretical sense.

To address the above problems and concerns during the delayed differentiation practice in the studied semiconductor manufacturer, this paper investigates finished-goods inventory control problems with and without forecasted advance-demand information (ADI) under capacitated postponement. It supplements the postponement literature with simple and computationally efficient algorithms to set finished-goods inventory targets. This work is closely related to the abundant inventory studies that deal with both capacity and demand uncertainty [16–20]. Most of the literature uses stochastic mathematical programming models to study the inventory problems and emphasizes the strategies of capacity allocation and structures of the optimal inventory policies. Differing from them, this research captures finishing capacity in an inventory shortfall process that is universally applicable to the inventory problems with or without lead times and forecasted ADI, and derives simpler models that provide managerial insight into the interrelations among demand, inventory, capacity, and lead times under a delayed differentiation environment.

Other related topics in the literature include advance-demand information and information sharing in supply chain and inventory management [21–27]. The vast majority of these studies assume that the ADI shared between different supply chain stages is perfect and aim at justifying its value in reducing the total supply chain and inventory costs. Chen and Chen [28] developed a model for the multiple-item budget-constraint newsboy inventory problem considering a reservation policy in which a discount rate is provided to those customers who are willing to make a reservation, e.g. provide ADI. Their research lead to an algorithm that determines both optimal order quantity and discount rate to achieve the maximal total expected profit under a limited budget. Acknowledging that advance-demand information is usually imperfect, Tan et al. [29] took imperfect ADI into a dynamic cost model through a "demand realization probability" in a periodic-review inventory system with no fixed order setup cost. They concluded that an optimal inventory control policy is of a state-dependant order-up-to type. Özer and Wei [30] used additive forecast updates to model ADI in a capacitated production system. They investigated the value of imperfect ADI and proved the optimality of a modified base-stock policy, if there are zero fixed costs. Chen and Lee [31] advanced further to model the imperfect ADI obtained from a series of evolving demand forecasts by the Martingale Model of Forecast Evolution (MMFE) process in a two-stage supply chain. Starting with a generalized order-up-to inventory policy, they derived a total supply chain cost model to explore the value of information sharing. With the optimality of the base-stock policy for capacitated production-inventory systems with zero fixed costs being proved by various papers [21,27,29,30], this work aims at including imperfect ADI of different forecast lead times into base-stock inventory models. By modeling the demand forecast error as an increasing function of the forecast lead time, we are able to carry out quantitative discussion on the optimal forecast lead time and the pull-in practice presented earlier. This modeling method is relevant and appropriate for the studied semiconductor manufacturer due to the long throughput time from wafer starts to finished products, demands are forecasted in both daily and weekly buckets over the horizon of several months. The projected demands of different forecast lead times and their historical accuracy data are immediately available. Also this semiconductor manufacturer adopts high volume manufacturing (HVM) methods, which make the per-unit setup cost so low that the base-stock policy generally suffices for the inventory management purpose.

Certainly any manufacturer who 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 a 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 literature has a rich body of work on the topic of optimizing capacity allocations using mathematical programming models. Readers can refer to Glasserman [19] 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 keeps the analysis simple.

The rest of 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 lacking from the analytical results. Section 5 concludes this work and provides future research opportunities.

2. Notations and basic models

Let us 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 is difficult 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, our 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 difficult 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.