

A Postponement Model for Demand Management

Ananth V. Iyer • Vinayak Deshpande • Zhengping Wu

Krannert School of Management, Purdue University, West Lafayette, Indiana 47907

Krannert School of Management, Purdue University, West Lafayette, Indiana 47907

Singapore Management University, 469 Bukit Timah Road, Singapore 259756

aiyer@mgmt.purdue.edu • deshbandev@mgmt.purdue.edu • zpwu@smu.edu.sg

In this paper, we analyze demand postponement as a strategy to handle potential demand surges. Under demand postponement, a fraction of the demands from the “regular” period are postponed and satisfied during a “postponement” period. This permits capacity to be procured to satisfy the postponed demands. A reimbursement per unit is paid to customers whose demands are postponed. The basic idea is that by preempting stockouts through demand postponement, we can reduce overall stockout costs. We formulate and solve a two-stage capacity planning problem under demand postponement. We propose a power range class of distributions to capture the nature of demand surges. We establish the scalability and conjugate properties of the power range distributions under demand postponement, which leads to a tractable analysis of the problem. We analytically solve the problem of determining the optimal regular and postponement period capacities, and the demand splitting rule to minimize the supplier’s expected cost. We show that (a) the value of postponement may be significant depending on cost and demand parameters, (b) a postponement strategy may lead to reduced investment in initial capacity, and (c) it may be optimal to do no demand postponement over a range of demands even after observing a higher demand signal. We then relax several model assumptions and provide results for these extensions. We conclude with managerial insights.

(Demand Postponement; Capacity Planning; Value of Information)

1. Introduction

In this paper, we focus on demand postponement as a tool to manage demand surges, i.e., situations where demand is likely to exceed short-term capacity. Under demand postponement, a fraction of the demands from the “regular” period are postponed and satisfied during a “postponement” period. This permits capacity to be procured to satisfy the postponed demands. A reimbursement per unit is paid to customers whose demands are postponed. The basic idea is that by preempting stockouts or backorders through demand postponement, we may reduce overall expected costs.

The analytical framework of demand postponement can be applied to a variety of practical situations

spanning multiple disciplines. We first discuss an example from the utility industry to provide one motivating context. In §1.1, we expand our model to accommodate examples from a number of other industry contexts. During the power crisis in California, innovative management tools such as interruptible power contracts were used by the utility industry (Emshwiller 2001, Thurm 2001, <http://www.flexyourpower.ca.gov>). Consider the problem faced by a manager of an electrical grid (supplier) that supplies industrial customers. The supplier has to plan for short-term capacity (e.g., supply for an eight-hour block of a day on a repetitive basis). The supplier uses his forecast of demand to buy capacity at a prespecified unit price in advance of this time block. If the

demand for this time block falls short of capacity, the supplier will have unused capacity for which he is already committed. However, if the demand for this time block exceeds the committed capacity, the supplier has to buy extra capacity from the spot market at a much higher price to satisfy demand. At the beginning of this time block, the supplier updates his forecast of demand for this time block. Based on the revised forecast, he can contact his industrial customers and get them to “postpone demand” to a later time in return for a reimbursement per unit postponed.

The supplier thus announces a fraction, $(1 - \beta)$, of demand that will be postponed to a nonregular production period (henceforth called postponement period) prior to the demand realization. Implementation of this demand postponement announcement may require the industrial customers to reschedule some of their operations to the postponement period and thus reduce their demand for the regular production period. The supplier reimburses the customers based on a predetermined unit postponement cost (which may or may not cover all the customers’ true costs). The demand realization then takes place for the customers. The supplier supplies the customers the earlier announced fraction (β) of the demand in the regular production period and the remaining $(1 - \beta)$ fraction of the demand in the later postponement period. Thus, postponing 50% of the demand would correspond to supplying power for four hours during the regular production period and supplying power for four hours during a nonregular production period (e.g., a night shift). Once a β is announced, we do not focus on the associated scheduling problem across multiple customers. For example, if there are two customers and $\beta = 0.5$, we do not model such details as when customer 1 is supplied power and when customer 2 is supplied power for four hours.

The postponement mechanism can be implemented as follows. Assume that the supplier decided to postpone a certain fraction $(1 - \beta)$ of the total demand. The specific manner in which demand postponement occurs can follow the following two possible schemes:

(a) *Postpone a fraction of demand for each customer:* Every unit of demand is split with β delivered in

the regular period and $1 - \beta$ delivered in the postponement period. In this scheme, every customer is affected and has a fraction of his demand postponed. The utility example described above provides such an example.

(b) *Postpone all demand for a fraction of customers:* A fraction $1 - \beta$ of the demand is postponed and thus delivered entirely in the postponement period, the remaining fraction β of the demand is delivered in the regular period. Note that in this case, any given customer may see his demand delivered entirely in the regular period or entirely in the postponement period depending on whether his demand was postponed or not. The fashion example, the hotel service example, and the telecommunications example, described in §1.1, provide potential contexts for this mechanism.

In §3, we provide a model structure (called “base model”) that permits tractable analysis and results. Figure 1 provides the time line of events for the base model and is described below.

(1) We model one period of demand and two periods to satisfy it—the regular period and the postponement period. A planning period exists before the realization of demand.

(2) The supplier chooses an initial capacity k_1 at the start of the planning period at a unit capacity cost of c_1 based on the demand distribution $f(x | \theta_0)$.

(3) Supplier receives an update x_0 on the demand distribution, where x_0 is a draw from the prior demand distribution.

(4) The demand distribution $f(x | x_0, \theta_0)$ is updated.

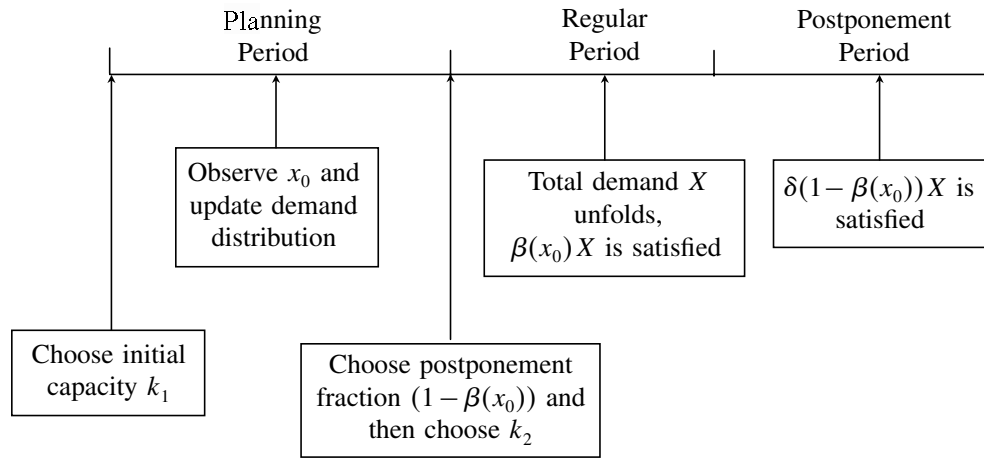
(5) A fraction $(1 - \beta(x_0))$ of the demand is chosen, and informed, that their demand will be satisfied during the postponement period. All postponed demand units are reimbursed s per unit for demand postponement. The associated customers reschedule some of their activities based on the fraction β announced.

(6) A capacity choice $k_2(x_0)$, which we refer to as k_2 for simplicity is made (because x_0 and $\beta(x_0)$ are known at the time of this decision), at a unit cost c_3 , to satisfy demand during the postponement period.

(7) Demand realization X occurs.

(8) Demand for regular period ($X_1 = \beta(x_0)X$ where X is total demand) is satisfied using k_1 and subcontracting at a cost of c_2 .

Figure 1 Time Line of Decisions



(9) Some known fraction $1 - \delta$ of the postponed demand might not be retained in the postponement period. For example, some customers may not be willing to accept the postponement arrangement, and hence, their demands are lost. Demand for postponement period ($X_2 = \delta(1 - \beta(x_0))X$) is satisfied using k_2 and subcontracting at a cost of c_4 .

For the base model, we assume that (a) there are no fixed costs associated with these commitments, (b) the only demand in this postponement period occurs from the postponed demand, (c) there is no information asymmetry of demand and there is no gaming in shortage costs, (d) all customers are paid the same unit postponement cost s , and (e) leftover capacity from the regular period cannot be used to satisfy postponement period demand.

We next provide the following extensions of our base model in §4.

(1) Of the postponed demand, an uncertain fraction of demands δ is retained in the postponement period.

(2) The postponement costs vary across customers, and hence, the cost of postponement is increasing in the demand postponed.

(3) Postponement period capacity is chosen after demand is realized.

(4) Leftover capacity from the regular period can be used to satisfy postponement period demand.

1.1. Linking the Time Line to Other Problem Contexts

We now describe how the proposed model can be used in a variety of other contexts.

(1) Consider an example from the fashion goods industry where an upstream supplier produces to satisfy demands of retailers. The supplier normally produces during the regular period such that goods can be delivered to the retailer about two to four weeks before the start of the selling season (Dalby and Flaherty 1990) using normal transport modes. This period of two to four weeks allows the retailers to position the goods in the retail stores before the selling season begins. However, suppose before the regular period production commences (i.e., during the planning period), the supplier receives signals (x_0) regarding the potential demand. He also knows that the demand during the planning period follows a percent done curve (Eppen and Iyer 1997) so that the timing of the signal can be mapped into a fraction of the total season demand. The supplier then chooses a fraction β and uses it to decide the set of customers whose orders (that corresponds to a β fraction of demand) will be satisfied using the regular period capacity.

Demand from the remaining customers, i.e., the fraction $1 - \beta$, is satisfied using the postponement period capacity. Because the postponement period

capacity results in retail delivery closer to the start of the retail season (i.e., with less than the usual time for retail deliveries), there may be additional premium transport costs to ship the products to the retailer, as well as costs associated with rush handling at the retailer for retail store delivery before the start of the season. There may, however, be some savings to the retailer due to lower holding costs. The supplier pays s per unit to cover the retailer's net cost associated with this later delivery from the postponement period production. The postponement period capacity may be at a different location than regular capacity (with perhaps different associated cost per unit $c_3 \geq c_1$ as well as higher subcontracting costs $c_4 \geq c_2$). Thus, there may or may not be an opportunity for the supplier to use leftover capacity from the regular period.

The main difference between the utility example and the fashion supplier example is that demand in this case is from retailers to the upstream supplier (and not from the consumer of electricity). The deliveries during the regular period as well as the postponement period are sold during the same selling season, starting at the end of the postponement period, thus causing no retail stockout differences. What is different about the two periods is the cost to get the goods produced (i.e., c_3, c_4 are different from c_1, c_2) as well as costs to distribute to the retailer at the start of the selling season. For example, a Harvard Business School case, Liz Claiborne and Ruentex (Dalby and Flaherty 1990), describes the role of expedited shipments closer to the start of the selling season as well as overtime payments to suppliers to get the goods to market on time. These costs would be reflected in the unit payment s to the retailers.

(2) For a service example, consider customers who consume capacity from a hotel or an airline. As the service system detects potential capacity shortages, it may start contacting potential guests to suggest a different hotel location or hotel in the case of hotels, or a flight with a different airline or a different flight in the airline case. This advance notice to guests may generate a lower cost for postponement than waiting until the customers show up at the service facility. In addition, guests who sign up to accept this offer may be permitted to pay a lower price for their flexibility, similar to the interruptible power contract in

the utility industry. Bialogorsky et al. (1999) provide examples of overselling with opportunistic cancellations in the airline and hotel industries, which is similar in spirit to our suggestion. The main difference is that in their paper, the postponement decision is made as demand is realized, whereas we suggest that the postponement decision may have to be made in the absence of complete demand information.

(3) Consider a telecommunications provider who offers customers the chance of access when they need it (regular period) potentially being required to access at a postponement period. Customers who are willing to be postponed may get to pay a lower fee for use as well as may be provided some disruption payment. This might permit the provider to manage the system more efficiently through intervening only to smooth out bottlenecks. Odlyzko (1999) suggests an approach to separate access to the same capacity type based on pricing differences to prevent Internet congestion. The main difference is that we propose this decision to suggest a delayed access would have to be made based on a signal of usage.

In our model, we focus on the following questions: (a) given a demand signal x_0 , how should the postponement decision $\beta(x_0)$ be chosen, and (b) given the capability to postpone demand, how should the initial capacity k_1 be chosen? Note that this problem has two features: (a) given an initial capacity, the demand postponement problem looks similar to a yield management problem, and (b) the initial capacity decision is similar to a newsvendor model. Our goal in this paper is to provide a complete analysis of this problem, generate comparative statics for this model, and suggest managerial implications from this model.

We formulate and solve a two-stage capacity planning problem under demand postponement. We propose a power range class of distributions to capture the nature of demand surges. We establish the scalability and conjugate properties of the power range distributions under demand postponement, which leads to tractable analysis of the problem. We analytically solve the problem of determining the optimal regular and postponement period capacities to minimize expected cost for the supplier. We also analyze how the supplier adjusts the demand splitting rule according to some prior information update.

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/236135211002010123>