Incorporating Delay Mechanism in Ordering Policies in Multi-Echelon Distribution Systems

Kamran Moinzadeh and Yong-Pin Zhou

University of Washington Business School

{kamran, yongpin}@u.washington.edu

June, 2005; revised July, 2006; December 2006

Abstract

In this paper, we devise a framework for obtaining the optimal ordering policy in a single location, continuous review inventory system with arbitrary inter-demand times. We show that it is optimal to order at demand arrival epochs only if the inter-demand time has constant or decreasing failure rate (CFR or DFR). When the inter-demand time has increasing failure rate (IFR), we show that the optimal policy is to delay the order. We then extend this policy to multi-echelon distribution systems consisting of one supplier and many retailers. Both decentralized and centralized systems are considered. We derive expressions and procedures for the evaluation of total cost and the computation of optimal delay in all settings considered. More importantly, we study the impact of our delay policy in all the settings. The numerical results indicate that for the single location model, the optimal delay can reduce the total cost significantly. Results from the single location model can be applied to the decentralized multi-echelon system, where the upstream supplier acts as a single location system. The supplier order delay can also have a significant impact (either positive or negative) on the retailers' total cost as well as the system's total costs. Finally, the impact of supplier order delay is minimal in the centralized multi-echelon setting. We offer an intuitive explanation for this observation.

1. Introduction

The literature on single location continuous review inventory models is rich and many studies have developed frameworks for analyzing and computing optimal or near optimal parameters of the well known (*s,S*) and (*Q,R*) policies under different settings (see Hadley and Whitin 1963, Tijms 1971, Sivlazian 1974, Nahmias 1976, Sahin 1979 and 1982, Zipkin 1986a &b, Zheng 1992, and Federgruen and Zheng 1992, *etc.*). It is well known that in single location systems under continuous review and Poisson demand, the (*Q*,*R*) policy, where orders are placed only at demand epochs, is optimal. However, when demand follows an arbitrary renewal distribution, the (*Q,R*) policy is no longer optimal as order epochs may not coincide with demand epochs.

In this paper, we consider continuous review inventory systems where demand forms a renewal distribution with arbitrary inter-demand distribution, and introduce an order delay policy referred to as the (*Q*,*R*,*T*) policy. Under this policy, a replenishment order will be placed *T* time units after inventory position reaches *R*, or when the next demand occurs (and the inventory position decreases to *R*-1), whichever occurs first. It can be easily verified that the classical (*Q*,*R*) policy is a special case of the (*Q*,*R*,*T*) policy for *T*=0. The question we ask in this paper is not whether the class of (*Q*,*R*,*T*) policies will outperform the class of (*Q*,*R*) policies, but rather by how much. More importantly, we analyze this policy in a multi-echelon setting as well. Finally, we investigate the effect of system parameters (cost, leadtime, demand, *etc.*) and system configuration (single location *vs.* multi-echelon, centralized *vs.* decentralized, *etc.*) on the magnitude of cost savings.

We will show that in the single echelon setting, when the inter-demand time exhibits decreasing or constant (*i.e.,* Poisson demand) failure rate (DFR or CFR), it is indeed optimal never to delay the order. However, when inter-demand time has increasing failure rate (IFR), then by delaying and observing the elapsed time since last demand, the system can get more "*up-to-date*" information on the time until next demand and improve its performance by delaying the order placement. We show that the magnitude of such improvements can be quite significant. An

important application of this model is to the upstream site (supplier or warehouse) in a distribution system, where due to downstream (retailers) batch-ordering, (see Duermeyer and Schwarz 1981, Nahmias 1981, Lee and Moinzadeh 1987, Moinzadeh and Lee 1986, Svoronos and Zipkin 1988 and Axsater 1990 for examples), inter-order times from each of the downstream sites is IFR. For instance, when external demand at the downstream locations is Poisson, the inter-order times from each site is Erlang, which is IFR. Clearly, in such settings, it may be optimal for the up-stream installation to delay its order placement.

In this paper, we also study order delay in a multi-echelon setting, both decentralized and centralized. In a decentralized distribution system where retailers order in batches larger than one, the supplier operates like a single location model facing IFR demand and uses an order delay policy. Clearly, the single location results suggest that the supplier will be better off. However, the impact of the supplier order delay on the retailers is more complex. For instance, if the supplier keeps the same R as that in the classical (Q,R) policy and uses a delay T, then the retailers will be worse off as their leadtimes will be stochastically larger. But, if the possibility of an order delay causes the supplier to increase its R to $R+1$ and then use a delay T, then the delay policy used by the supplier can actually *benefit* the retailers. In a numerical experiment, we observe that the impact on the retailer costs can be significant.

In a centralized serial inventory system, it is well known that echelon-based inventory policy is optimal (see Clark and Scarf 1960, Axsater and Rosling 1993 and Chen 1998). For centralized distribution systems with multiple retailers and one supplier (*i.e.,* warehouse), the general form of the optimal policy is not known. In all previous works, (Q,R)-type policies have been employed as the order policy of choice in all locations and the optimal parameter values are found for both the retailers and the supplier. For examples, see Duermeyer and Schwarz (1984), Lee and Moinzadeh (1987), Moinzadeh and Lee (1986), Svoronos and Zipkin (1988) and Axsater (1990). In this paper, we will consider a similar system to those mentioned above, employ the (Q,R,T) order policy at the supplier and study its impact on such systems. Numerical results suggest that the order delay has

minimal impact on the system cost. We provide an intuitive explanation based on the results in the decentralized setting.

To summarize, the contribution of this paper is two fold: First, in a single location setting, we use a delay policy and identify the conditions when it is optimal. Second and more important, we apply the order delay policy to both the centralized and decentralized distribution systems, and compare them with systems that do not use order delay. Furthermore, we identify the settings where employment of delay policies at the supplier is most critical in such distribution systems.

The only papers known to the authors that deal with general, non-Poisson, demand process are Schultz (1989), Moinzadeh (2001) and Katircioglu (1996). Schultz (1989) studies the concept of order delay in an (S-1,S) inventory setting where S, the maximum stocking level, is set at one. Moinzadeh (2001) considers a single location inventory system in the (S-1,S) setting where demand follows an arbitrary renewal distribution. He proposes an improved order policy, which is based on delaying the order placement in such systems; that is, an order will be placed T time units after a demand occurs. The optimal delay, T, however, is fixed and not determined in real time. This policy is suboptimal as it does not take into account the demand activities during the delay period, T. Even so, it is shown that this order policy with delay can result in significant cost savings compared with the traditional (S-1,S) policy. The method of study in this paper is based on Moinzadeh (2001), and provides a significant improvement.

Katircioglu (1996) also considers the single location version of our problem. Although the results are similar, our method of analysis is different. Furthermore, unlike Moinzadeh (2001) and Katircioglu (1996), we consider in this paper the order delay policy in distribution systems consisting of one supplier and multiple retailers, both centralized and decentralized, and study the effect of such a delay on system costs. We identify the settings when employment of delay policies at the supplier is most critical in such distribution systems.

The rest of the paper is organized as follows: In Section 2, we study the delay policy in a single-echelon setting and identify the conditions when it is optimal. Then in Section 3, we expand

our analysis to a multi-echelon distribution system. In Section 4, we numerically study the impact of the delay policies on costs for single location, both centralized and decentralized multi-echelon distribution systems. We close in Section 5 by summarizing the contribution of this study and suggesting avenues for future research. All the proofs can be found in the appendix.

2. A Single Location Model

We study a continuous review inventory system where demand is of unit size and follows a renewal process with a mean rate λ. Lead time is a constant, L, and any excess demand is backordered. We propose the following dynamic delayed ordering policy, which we call the *delayed-order policy*:

When the current system inventory position is r and it has been t time units since last demand occurrence, an order of Q will be placed $\tau_r(t)$ *time units from now.*

Note that the *delayed-order-policy* is the optimal class of policies in such systems as at any inventory level, the ordering decision is examined at any point in time. We will show that the optimal policy parameters under this policy will be such that the resulting policy operates like a (Q,R,T) policy. For fixed Q, the optimal value of $\tau_r(t)$ can be found by minimizing the average total cost rate when inventory position is r and an order will be placed. We now derive an expression for this average cost rate. Consider a unit in an order, say, the k^{th} unit ($k=1,\ldots,Q$). This unit will be assigned to the $(r+k)^{\text{th}}$ future demand. The expected cost associated with this unit, consisting of holding and backorder costs, will be:

$$E_{Y^t_{r+k}}\left\{h\left[Y^t_{r+k}-\left(L+\tau_r(t)\right)\right]^+ + \pi\left[\left(L+\tau_r(t)\right)-Y^t_{r+k}\right]^+\right\},$$

where Y^t_{r+k} denotes the time until the $(r+k)^{\text{th}}$ future demand when the last demand occurred t time units ago (all necessary notations are defined in Table 1).

Table 1: Notation for the single location model

Q	fixed order quantity
R	maximum inventory position level that triggers an order
$\tau_r(t)$	order delay when inventory position is r and time t have elapsed since last demand (to simplify notation, the subscript will sometimes be suppressed)
L	constant order leadtime
h	unit holding cost per unit time
π	unit backorder cost per time unit
X	random variable representing the *i.i.d.* inter-demand times
$f(\cdot)/F(\cdot)$	PDF/CDF of X
$Z_k = \sum_{i=1}^{k} X_i$	k-fold convolution of X for $k>0$, $Z_0=0$, $Z_k=-Z_{-k}$ for $k<0$
Y_1^t	random "residual life" of X when t time units have elapsed; *i.e.,* time until the next demand when the last demand occurred t time units ago
$g_1^t(\cdot)/G_1^t(\cdot)$	PDF/CDF of Y_1^t
$Y_k^t = Y_1^t + Z_{k-1}, \forall k \geq 1$	time until the k^{th} future arrival, when t time units have elapsed since last demand epoch for $k>0$, $Y_0^t=-t$, $Y_k^t=-t+Z_k$ for $k\leq -1$
$G_k^t(x)$	CDF of Y_k^t
$H_{i,j}^t(x) = \sum_{k=i}^{j} G_k^t(x)$	The sum of the G_k functions for $k = i$ to j

We will fix Q throughout the paper; then for each r and t, we find the optimal delay $\tau_r(t)$ which minimizes the average total cost rate when inventory position is r and an order will be placed as follows:

$$\min_{\tau_r(t)} \frac{\lambda}{Q} \sum_{k=r+1}^{r+Q} E_{Y_k^t} \left\{ h\left[Y_k^t - \left(L+\tau_r(t)\right)\right]^+ + \pi\left[\left(L+\tau_r(t)\right) - Y_k^t\right]^+ \right\}. \tag{1}$$

Finally, once the optimal delay $\tau_r(t)$ is fully characterized, we find the optimal value of r for the delayed order policy. Note that since $\tau_r(t)$ is updated continuously for all r and t, an order will be placed only when $\tau_r(t)=0$ (when $\tau_r(t)>0$, its exact value is not important; an order will not be placed). The objective function in (1) is convex in $\tau_r(t)$, so it suffices to analyze the first order derivative of the objective function (where $H_{i,j}^t(x) = \sum_{k=i}^{j} G_k^t(x)$):

$$FO = \frac{\lambda}{Q}\sum_{k=r+1}^{r+Q}\left[-h\cdot \Pr\left(Y_k^t > L+\tau_r(t)\right)+\pi\cdot \Pr\left(Y_k^t \le L+\tau_r(t)\right)\right] = \frac{\lambda}{Q}\left[-hQ+(h+\pi)H_{r+1,r+Q}^t\left(L+\tau_r(t)\right)\right]. \tag{2}$$

All the proofs of the Lemmas and Theorems to follow can be found in the Appendix.

Lemma 1

(1) When $r+Q\le 0$, it is optimal not to delay orders. That is, $\tau_r(t)\equiv 0$.

(2) When $r+Q>0$, the optimal delay satisfies $\tau_r(t) = \left[\left(H_{r+1,r+Q}^t\right)^{-1}\left(\frac{hQ}{\pi+h}\right)-L\right]^+$.

Lemma 2

(1) When X has DFR, $\tau_r(t)$ is non-decreasing in t.

(2) When X has CFR, $\tau_r(t)$ is constant in t.

(3) When X has IFR, $\tau_r(t)$ is non-increasing in t.

The first two statements in Lemma 2 imply that if it is optimal not to order at a demand epoch, it is optimal not to order as long as the inventory position remains the same (*i.e.,* $\tau_r(0)>0 \Rightarrow \tau_r(t)>0, \forall t$). Therefore, we have:

Theorem 1 If the inter-demand time distributions have DFR or CFR, then it is optimal to place orders only at demand epochs.

Corollary 1 When the demand process follows a Poisson process, it is optimal to order only at demand epochs.

Now we focus on the more interesting case of IFR. Examples of IFR demand abound in practice. For example, uniform and Erlang distributions have IFR, and for some parameter combinations, the Weibull distribution also has IFR.

Lemma 3 Suppose the inter-demand time has IFR. For any r, $\tau_r(t)$ is decreasing in t, and $\lim_{t\to\infty} \tau_{r+1}(t) \geq \tau_r(0)$.

In Theorem 2, we characterize the optimal policy parameters under the proposed delayed policy.

Theorem 2 Let $\underline{r} = \min\{r:\ \tau_r(0) > 0\}$.

(a) If $\tau_{\underline{r}}(t) = 0$ for some finite t, then let $R=\underline{r}$ and $T = \inf\{t:\ \tau_R(t) = 0\}$.

(b) If $\tau_{\underline{r}}(t) > 0$ for all finite t, then let $R=\underline{r}-1$ and $T = 0$.

The following policy is the optimal delayed-order policy:

(1) When $r>R$, it is optimal not to order for all t.

(2) When $r=R$, it is optimal not to order for $t<T$; and it is optimal to order at T.

(3) When $r<R$, it is optimal to order for $t=0$ (*i.e.,* no delay).

The results in Lemma 3 and Theorem 2 are illustrated in Figures 1 and 2 where the behavior of $\tau_r(t)$ is depicted as a function of t for different values of r. In Figures 1 and 2, the two (a) panels correspond to (2) in Theorem 2, and the two (b) panels correspond to (3) in Theorem 2.

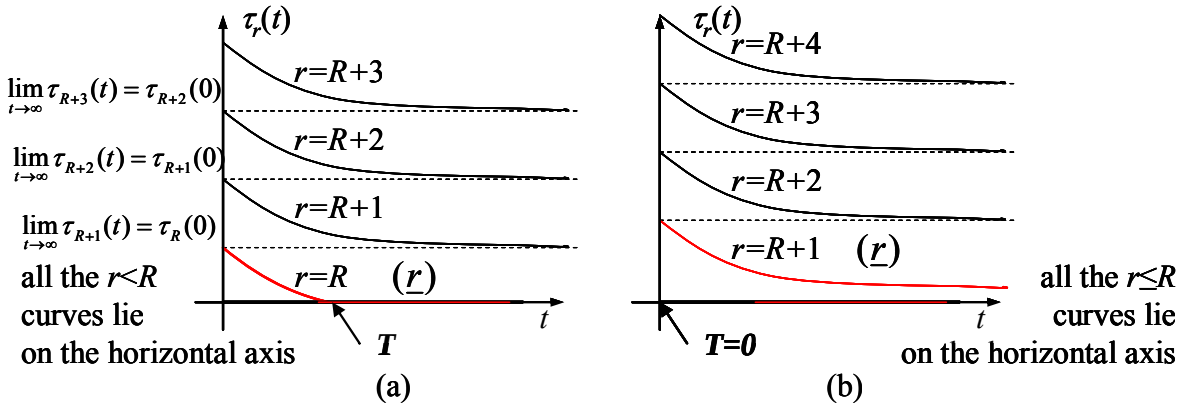


Figure 1: The $\tau_r(t)$ Curves for $\lim_{t\to\infty} \tau_{r+1}(t) = \tau_r(0)$

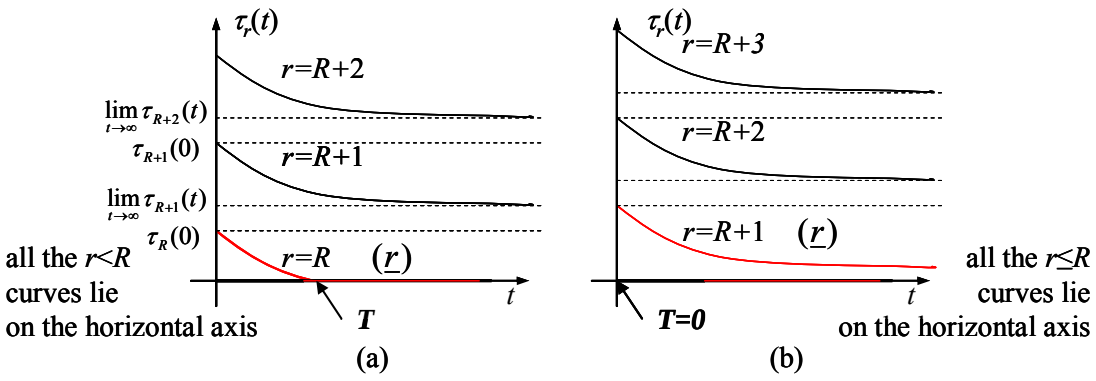


Figure 2: The $\tau_r(t)$ *Curves for* $\lim_{t\to\infty}\tau_{r+1}(t) > \tau_r(0)$

Numerically, one needs to calculate $\tau_r(0)$ for the various values of r, and use Theorem 2 to find R. Moreover, for this R, if the optimal T is positive, the first order condition of optimality can be used to find the T such that $\tau_R(T)=0$; that is, $\dfrac{hQ}{\pi+h} = \sum_{k=R+1}^{R+Q} G_k^T(L) = H_{R+1,R+Q}^T(L)$. The right hand side of the equation is decreasing in T, so a numerical method could be used to solve this equation. An important implication of Theorem 2 is that if the system starts out with a very high inventory position r, then the curve would never intersect with the horizontal axis (case 1 in Theorem 2). Then no orders will be placed until the next demand epoch, and so on, until the inventory position finally drops down to R (case 2). So effectively, we are reducing the inventory position from r to R. Similarly, when the system starts out with very low inventory position r (case 3), successive orders of size Q will be placed right away to bring the inventory position to above R.

When the inventory position reaches R, an order will be placed after T time units or when the next demand arrives, whichever occurs first. Therefore, when the system is in steady state, an order will be placed only in two situations:

(1) inventory position is R-1; or

(2) inventory position is R, and it is been T time units since last demand.

This clearly demonstrates that the optimal delayed-order policy is a (Q,R,T) policy. Now, the expected total cost rate of a (Q,R,T) policy for any fixed Q, which consists of the average cost rate if demand occurs before T is expired and the average cost rate if T expires before an occurrence of a demand, can be expressed as:

$$\int_T^\infty \left\{ \sum_{\substack{R+1\le k\le R+Q \\ k>0}} E_{Z_{k-1}}\left[h\left(Z_{k-1}+x-T-L\right)^+ + \pi\left(L+T-x-Z_{k-1}\right)^+ \right] + \sum_{\substack{R+1\le k\le R+Q \\ k\le 0}} \pi E_{Z_{-k}}\left[Z_{-k}+T+L\right] \right\} f(x)dx$$
$$+\int_0^T \left\{ \sum_{\substack{R+1\le k\le R+Q \\ k>0}} E_{Z_{k-1}}\left[h\left(Z_{k-1}-L\right)^+ + \pi\left(L-Z_{k-1}\right)^+ \right] + \sum_{\substack{R+1\le k\le R+Q \\ k\le 0}} E_{Z_{-k}}\left[\pi\left(L+x+Z_{-k}\right)\right] \right\} f(x)dx \tag{3}$$

The optimal delay T can be found by applying Lemma 1 and Theorem 2.

The first order derivative is as follows:

$$FO_2 = \int_T^\infty \left\{ \sum_{\substack{R+1\le k\le R+Q \\ k>0}} \left[-h\Pr\left(Z_{k-1}+x-T-L>0\right) + \pi\Pr\left(L+T-x-Z_{k-1}>0\right)\right] + \sum_{\substack{R+1\le k\le R+Q \\ k\le 0}} \pi \right\} f(x)dx$$
$$-\left\{ \sum_{\substack{R+1\le k\le R+Q \\ k>0}} E_{Z_{k-1}}\left[h\left(Z_{k-1}-L\right)^+ + \pi\left(L-Z_{k-1}\right)^+ \right] + \sum_{\substack{R+1\le k\le R+Q \\ k\le 0}} \pi E_{Z_{-k}}\left[Z_{-k}+T+L\right] \right\} f(T)$$
$$+\left\{ \sum_{\substack{R+1\le k\le R+Q \\ k>0}} E_{Z_{k-1}}\left[h\left(Z_{k-1}-L\right)^+ + \pi\left(L-Z_{k-1}\right)^+ \right] + \sum_{\substack{R+1\le k\le R+Q \\ k\le 0}} E_{Z_{-k}}\left[\pi\left(L+T+Z_{-k}\right)\right] \right\} f(T)$$
$$= \bar{F}(T)\left[\pi Q - (h+\pi) \sum_{\substack{R+1\le k\le R+Q \\ k>0}} \int_0^\infty \Pr\left(Z_{k-1}+y-L>0\right) g_1^T(y)dy \right]$$
$$= \bar{F}(T)\left[\pi Q - (h+\pi) \sum_{\substack{R+1\le k\le R+Q \\ k>0}} \bar{G}_k^T(L) \right]$$
$$= \bar{F}(T)\left[\pi Q - (h+\pi) \sum_{R+1\le k\le R+Q} \bar{G}_k^T(L) \right]$$
$$= \bar{F}(T)\left[-hQ + (h+\pi) H_{R+1,R+Q}^T(L) \right]. \tag{4}$$

Comparing equation (4) with equation (2), we find that for any optimal (*Q,R,T*) policy where FO_2=0 in (4), setting $\tau_R(T) = 0$ will also satisfy *FO*=0 in (2). Therefore the optimal *R*, *T* values are the same for both the (*Q*,*R*,*T*) policy and the delayed-order policy.

Remark 1: The model in Moinzadeh (2001) is a special case of our delayed-order policy model for the (*S*-1,*S*) setting. Moreover, Moinzadeh (2001) assumes that the order delay, τ, is static, fixed, and not allowed to change in real time. In the event that many demand orders occur after inventory position reaches *S*-1, and before τ has expired, which would have triggered an order under our policy before τ, the policy in Moinzadeh (2001) would still place an order after the fixed τ. Clearly this is less effective.

Remark 2: Katircioglu (1996) studies the (*Q*,*R*,*T*) policy using a different approach. He first analyzes a single unit in the order, then uses induction to extend it to multiple units, and, later, infinite time horizon. Like Moinzadeh (2001), Katircioglu (1996) studies only the single-echelon problem. Our analysis of the multi-echelon distribution system is presented in the next section.

3. A Multi-Echelon Distribution System

Now we will apply our (*Q*,*R*,*T*) policy to a distribution system. We will follow the classic setting for distribution systems (see Svoronos and Zipkin 1986, Axsater 1990 and Moinzadeh 2002). In such a system, there are *N* identical retailers and one supplier. Demand at each retailer follows a Poisson process with rate λ; as a result, each retailer follows a (*Q*,*R*) policy when placing orders. Excess demand is backordered at the retailers. Upon placement of an order, if the supplier has stock, it will fill the order immediately; otherwise, the retailer's order will be delayed until the supplier has sufficient stock to fill the order. It is assumed that the transit time from the supplier to the retailer is fixed and constant *L*. The supplier orders its stock from an outside source, which is assumed to have

ample stock. We assume that the supplier's order quantity, $\hat{Q}$, is an integer multiple of retailer's order quantity, Q; that is, $\hat{Q}=mQ$ (m=1,2,..). Finally, the order leadtime from the outside source to the supplier is assumed to be a constant $\hat{L}$. (Throughout this section, we will apply the ˆ notation to all the parameters, variables, and functions associated with the supplier.)

In this section, we first analyze the demand process at the supplier, and show that it follows an IFR process. Then, we will examine the supplier's use of the delayed-order policy both in the centralized and decentralized settings.

Since demand at the retailers is Poisson demand and a (Q,R) policy is employed by the retailers, each retailer's inter-order time has an Erlang distribution with mean Q/λ and shape parameter Q (denoted as Erlang($\lambda/Q,Q$)). We start by examining a random epoch when the supplier receives a retailer order. Without loss of generality, assume that Retailer 1 just placed the order. Thus, the inventory position at Retailer 1 will be $R+Q$ just after order placement. It is well known in the literature that the inventory positions at the other retailers are uniformly distributed on {R+1, R+2, …, R+Q}. The following three Lemmas will show that the random time until the next order at the supplier by any retailer has IFR.

Lemma 5 Any Erlang distribution has IFR.

Lemma 6 At any retailer order epoch, all the other retailers have IFR time until next order.

Lemma 7 The minimum of independent IFRs is also IFR.

Together, the three Lemmas imply that the inter-demand time seen by the supplier is IFR.

Proposition 1 If the retailer order size, Q, is greater than 1, then the inter-order time at the supplier has IFR.

Since retailers see Poisson demand, they will order only at their own demand epochs. It is the supplier who should optimally introduce delay into its own ordering.

Centralized Setting In the centralized setting, the retailers and the supplier belong to the same company. Therefore the objective is to find the optimal inventory policy parameters for the retailers and the supplier jointly, in order to minimize total supply chain cost, consisting of holding and penalty cost at the retailers and holding cost at the supplier. We assume that the retailers are identical and they all use a (*Q*,*R*) policy. We assume that the supplier follows a $(\hat{Q}, \hat{R}, \hat{T})$ policy. When retailers are identical, under the standard assumptions made here which is in line with previous work in the literature (see Svoronos and Zipkin 1986, Chen 1998 and Axsater 1993), it is reasonable to set the supplier's reorder point to be a multiple of *Q* (see Zipkin 2000, pp. 320 for a discussion). Thus, we assume that $\hat{R} = \hat{r}\,Q$. Recall that a delayed-order policy for the supplier means:

When the inventory position reaches $\hat{R}$*, the supplier places an order of quantity* $\hat{Q}$*,* $\hat{T}$ *time units in the future or when the next retailer order is received, whichever occurs first.*

Let us now follow a supplier order. This shipment will be used to satisfy the $(\hat{r}+1)^{th}$, $(\hat{r}+2)^{th}$, …, $(\hat{r}+m)^{th}$ retailers' future orders. There are two costs associated with this order:

- C_1, the supplier's expected holding cost (when a shipment arrives at the supplier before its corresponding retailer order) per unit in an order.
- C_2, the expected retailer holding (when units arrive at the retailer before their corresponding demands arrive at the retailer) and penalty cost (when units arrive at the retailer after their corresponding demands arrive at the retailer) per unit in an order.

We will analyze the two costs one by one:

Let $\hat{V}_j = \begin{cases} \left[\hat{Z}_{j-1} + (\hat{Z}_1 - \hat{T})^+ - \hat{L}\right]^+ & \text{if } j > 0 \\ 0 & \text{if } j \le 0 \end{cases}$, then $\hat{V}_j$ is the time that the supplier order arrives at the supplier before the $(j-\hat{r})^{th}$ future retailer order shows up, where $j = \hat{r}+1, \hat{r}+2, \ldots, \hat{r}+m$. Consequently,

$$C_1 = \sum_{j=\hat{r}+1}^{\hat{r}+m} \hat{h}E[\hat{V}_j]/m = \sum_{j=\hat{r}+1}^{\hat{r}+m} \hat{h}E\left[\hat{Z}_{j-1} + (\hat{Z}_1 - \hat{T})^+ - \hat{L}\right]^+ / m .$$

Moreover, let

$$\hat{W}_j = \begin{cases} \left[\hat{L} - (\hat{Z}_1 - \hat{T})^+ - \hat{Z}_{j-1}\right]^+ & \text{if } j > 0 \\ \hat{Z}_{-j} + \min\left\{\hat{T}, \hat{Z}_1\right\} + \hat{L} & \text{if } j \le 0 \end{cases}. \tag{5}$$

Then $\hat{W}_j$ is the time that the supplier order arrives at the supplier after the $(j-\hat{r})^{th}$ future retailer order shows up, where $j = \hat{r}+1, \hat{r}+2, \ldots, \hat{r}+m$. ($W$ is also called the *retard*). When a retailer orders from the supplier, if it becomes the j^{th} order in the supplier order, then the total leadtime for this retailer's order is $L + \hat{W}_j$. Define the time it takes for the $(i-R)^{th}$ future customer order to arrive by Z_i, $i = R+1, R+2, \ldots, R+Q$. Then Z_i is an Erlang random variable with parameters $\lambda/\hat{\imath}$ and i. Clearly, the retailer order arrives ahead of the corresponding customer order if and only if $Z_i > L + \hat{W}_j$. Therefore, the holding cost is $h\left[Z_i - L - \hat{W}_j\right]^+$ and the penalty cost is $\pi\left[L + \hat{W}_j - Z_i\right]^+$. Thus,

$$C_2 = \sum_{\substack{i=R+1,R+2,\ldots,R+Q \\ j=\hat{r}+1,\hat{r}+2,\ldots,\hat{r}+m}} \left\{hE\left[Z_i - L - \hat{W}_j\right]^+ + \pi E\left[L + \hat{W}_j - Z_i\right]^+\right\} / (mQ) .$$

Theorem 3 In the centralized setting, the expected total system cost rate can be expressed as:

$$TC = N\lambda\left(C_1 + C_2\right) =$$

$$N\lambda\left\{\sum_{j=\hat{r}+1}^{\hat{r}+m} \hat{h}E\left[\hat{Z}_{j-1} + (\hat{Z}_1 - \hat{T})^+ - \hat{L}\right]^+ / m + \sum_{\substack{i=R+1,R+2,\ldots,R+Q \\ j=\hat{r}+1,\hat{r}+2,\ldots,\hat{r}+m}} \left\{hE\left[Z_i - L - \hat{W}_j\right]^+ + \pi E\left[L + \hat{W}_j - Z_i\right]^+\right\} / (mQ)\right\}.$$

We can then use Theorem 3 to search for the optimal R and T for the supplier's (Q,R,T) policy.

Remark 3: When the retailers are non-identical, the analysis becomes very complicated. One may need to keep track of each sub-batch as Forsberg (1996) and Axsater (2000) did. We will not carry out that analysis in this paper, and consider this a challenging but important future research topic.

Decentralized Setting In the decentralized setting, the retailers and the supplier belong to different companies. Therefore they each find the optimal inventory policy parameters to minimize total inventory cost for themselves. Specifically the supplier chooses the optimal (Q,R) policy parameters to minimize its own holding and penalty costs, while the retailers chooses their own optimal (Q,R,T) policy parameters to minimize their own holding and penalty costs.

Since demand at the retailers is stationary Poisson, it is optimal for the retailers to use the traditional (Q,R) ordering policy. However, to derive the optimal parameter, the retailer needs to use $L+W$ as the random order leadtime, where W is given in (5).

No matter what value of R the retailers use, their orders to the supplier have Erlang inter-order time distribution. According to Proposition 1, the supplier sees IFR demand process, so it is optimal for the supplier to use the (Q,R,T) ordering policy. The derivation of the operating characteristics for the supplier is similar to that of the centralized case, so the detailed are omitted.

Remark 4: When the retailers are non-identical but independent, Lemmas 5-6 continue to hold. Thus it is still optimal for the supplier to delay its ordering.

4. Numerical Analysis

In this section we test the effectiveness of (Q,R,T) policy by comparing it with the classical (Q,R) policy via a numerical experiment. We present results for the single location setting first in Section 4.1; then we focus on the multi-echelon distribution setting in Section 4.2. In doing so, we

study the decentralized and the centralized multi-echelon distribution systems in Sections 4.2.1 and 4.2.2, respectively.

4.1 Single Location Model

We now study the effectiveness of the order delay policy compared with the classical (Q,R) policies in a single location inventory system discussed in Section 2. To compare the effectiveness of the (Q,R,T) and the classical (Q,R) policy, we compute the percent deviation as:

$$\%\ deviation \ = \ 100 * \frac{Average\ Total\ Cost(Q,R,T) - Average\ Total\ Cost(Q,R)}{Average\ Total\ Cost(Q,R)}$$

In our numerical experiment, we only focus on the non-trivial IFR case, and assume that the inter-demand times follow an Erlang(λ/k, k). The choice of Erlang distribution serves as the building block for the multi-echelon problem, which is analyzed in more details in the sections that follow. Note that the squared coefficient of variance (CV^2) of Erlang is $1/k$. In order to see the impact of the variance, not the mean, of the inter-order time, we normalize the mean of the Erlang distributions considered to unity (*i.e.,* $\lambda = k$), and vary k. We vary the leadtime L, the holding and penalty cost parameters (h and π), and the resultant critical ratio $\frac{\pi}{\pi + h}$, which approximates the service level. Specifically, we let $\lambda = k \in \{2, 4, 8\}$, $L \in \{0.1, 0.5, 1, 2, 5\}$, h=1, $\frac{\pi}{\pi + h} \in \{0.6, 0.8, 0.9, 0.95\}$, and $Q \in \{1,2,4,8\}$. Results of the numerical tests are shown in Figure 3.

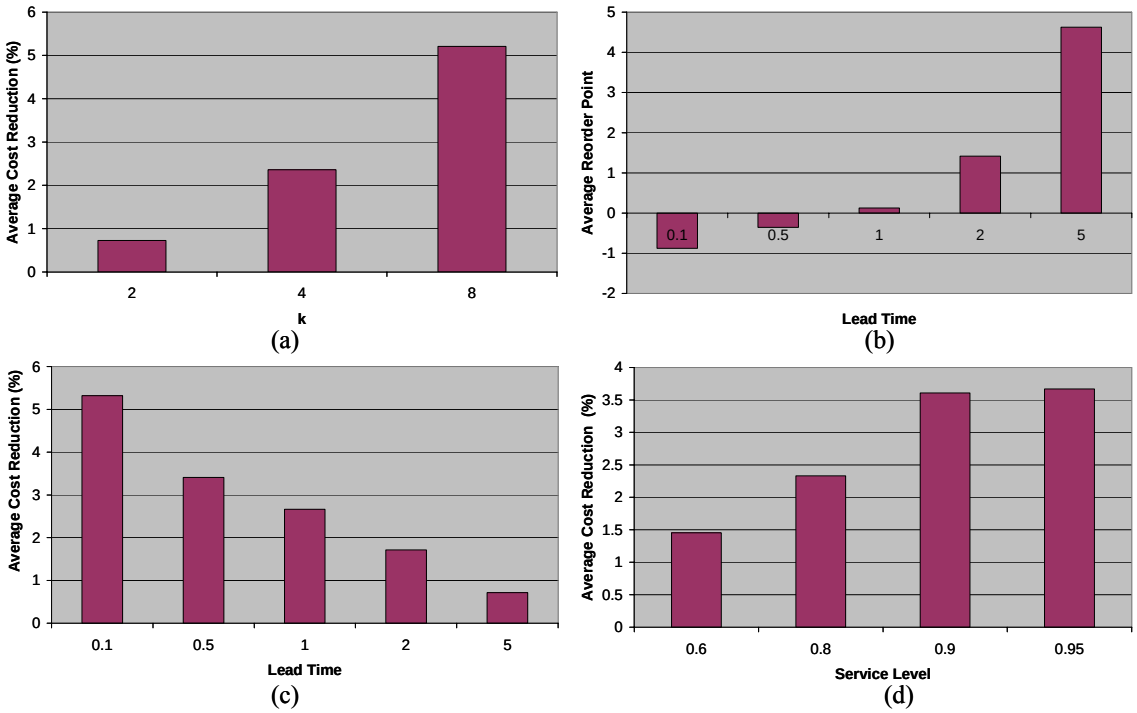


Figure 3: Comparison of delayed and no-delay policies: single location setting

Note that the inter-demand time has an Erlang distribution of mean 1. The squared coefficient of variation (CV^2), however, is $1/k$. It is quite clear that as k increases, the CV of the inter-demand time decreases, and the demand arrivals become more "predictable". In such cases, the order delay policy will result in larger savings over the (Q,R) policy (see Figure 3(a)). In one extreme, when k=1, the inter-demand time is Poisson, and there will be no cost savings whatsoever by delaying the order. In the other extreme, when the inter-demand time is constant (CV=0 or k=∞), the optimal delay will perfectly match the supply with demand resulting in no holding and penalty costs. This observation was made by both Katircioglu (1996) and Moinzadeh (2001). This results in the most cost savings.

From Figure 3(b), we observe that as the order leadtime increases, the optimal reorder point also increases. Thus, for each replenishment order cycle, the time until corresponding demand arrival is composed of more inter-demand times. The convolution of more inter-demand times results in less

variable "time until corresponding demand" which results in lower % cost difference between the order delay policy and the (*Q,R*) policy. This observation can be verified in Figure 3(c). From Figure 3(c), it is worth noting that when leadtimes are very small (one tenth of the average inter-demand time), the average cost savings by the ordering delay is substantial.

In Figure 3(d), service level (represented by $\pi/(\pi+h)$) is higher and the cost reduction is increasing in service level. Theoretically, however, cost reduction may decrease in service level when service level is low. Since they are unusual in practice, we do not perform any analysis for low service levels.

Remark 5: The cost savings in Figure 3 are all average numbers for a fixed parameter value. For individual cases, the cost saving can be large. Since the cost difference is larger when leadtime is smaller and *k* is larger, the highest cost saving, 48.0%, occurs, not surprisingly, at k=8, L=0.1, and π=9. The magnitude of cost savings is consistent with those presented in Katircioglu (1996), indicating that the results are indeed very general, not specific to a particular inter-demand time distribution.

4.2 A Multi-Echelon Distribution System

4.2.1 Decentralized Setting

In this section, we study a decentralized distribution system discussed in Section 3. In such a system, without having to be concerned with the retailer costs, the supplier can choose its delay order to significantly reduce its average total cost (as shown in Section 4.1). However, we aim to investigate the effect of the supplier's decision on the retailers' costs as well as the total system cost. In our numerical experiment, we vary the following parameters: for the retailers, $N \in \{2, 4, 8, 16\}$, h=1, $\pi/(\pi+h)$=0.9, L=1, $Q \in \{2,4,6\}$; for the supplier, we considered, $\hat{L} \in \{0.1, 0.5\}$, $\hat{h} \in \{0.2,$

0.5}, $\hat{\pi}/(\hat{\pi}+\hat{h}) \in \{0.8, 0.9\}$, and $m \in \{1, 2, 4, 8\}$. We also normalize the mean of the retailers' inter-demand time to unity (*i.e.,* λ=1).

In the decentralized setting, when the supplier uses order delays to reduce its cost, it is natural to think that this will increase the retailers' costs. However, this is not necessarily true. Suppose that without using the order delay, it is optimal for the supplier to reorder at R; then with the delay the supplier could do one of two things: use the same R and find T, or increase R by 1 and then find T. Therefore, the impact of the order delay on the retailers can be either positive (when R is increased by one) or negative (when R is kept the same). Thus, even in the decentralized distribution settings, the delayed order policy can be used to lower costs for both the supplier and the retailers, and hence the system.

In all, we considered a total of 384 cases. As can be seen in Figure 4 and Table 2, when averaging over all the cases, both the retailers and the system see their total cost go up. The small values of the overall averages do not tell the whole story, however. In most cases there is no difference between the delayed and non-delayed policies, but when there is a difference, the % cost differences can be quite large in either direction. A further look at the breakdown of the cost difference reveals more information. When parameters are such that the order delay by the supplier causes the retailers' (system's) total cost to go up, the average increase is quite significant, 9.799% (7.313%). On the other hand, when the delay causes the retailers' (system's) costs to decrease, on average the decrease is relatively small, 0.345% (0.375%). Furthermore, the maximum and minimum % cost differences are significant for the retailers, the supplier, as well as the system as a whole. Given that the supplier order delay policy can have a significant impact for all the parties involved, we next study the impact of system parameters in these systems.

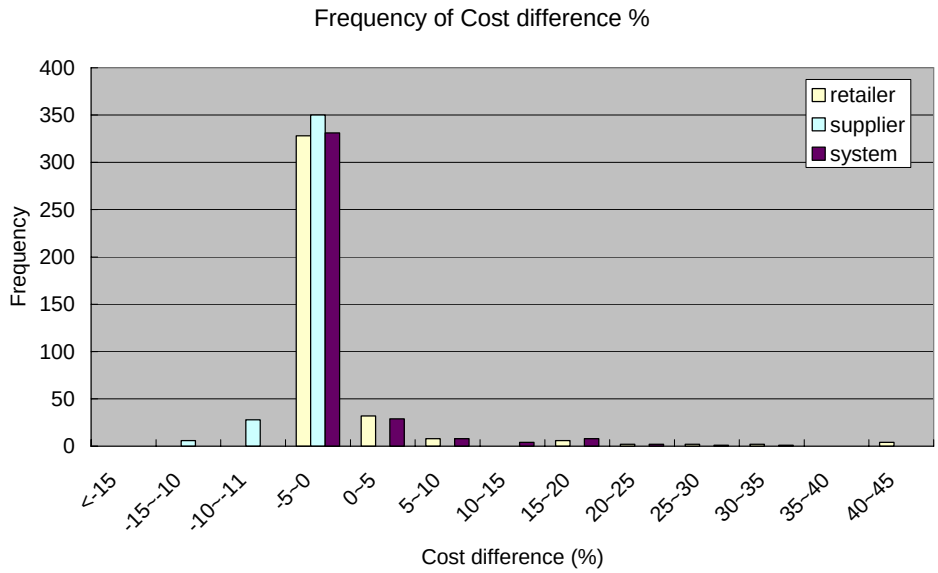


Figure 4: Frequency of cost improvement (delayed vs no delay) for the multi-echelon decentralized setting

Table 2: Percent cost difference from delayed to no-delay policy: multi-echelon decentralized setting

	Retailer	Supplier	System
Average of all cases	1.353%	-1.012%	0.919%
Average of positive	9.799%	N/A	7.313%
Average of negative	-0.345%	-2.680%	-0.375%
Max %	43.568%	0%	31.595%
Min %	-3.504	-13.562%	-4.565%

Figure 5 shows how the cost differences for the retailers, the supplier, and the system change with respect to Q, $\hat{L}$, and $\hat{\pi}/(\hat{\pi}+\hat{h})$, respectively. Not surprisingly, similar to the single location setting, the supplier's savings increases as the retailer's order quantity gets larger making supplier's inter-demand times more predictable. The behavior of retailer's cost, as well as the system's, seems to go in the opposite direction. And the higher the cost savings for the supplier, the higher the cost increases for the retailers and the system.

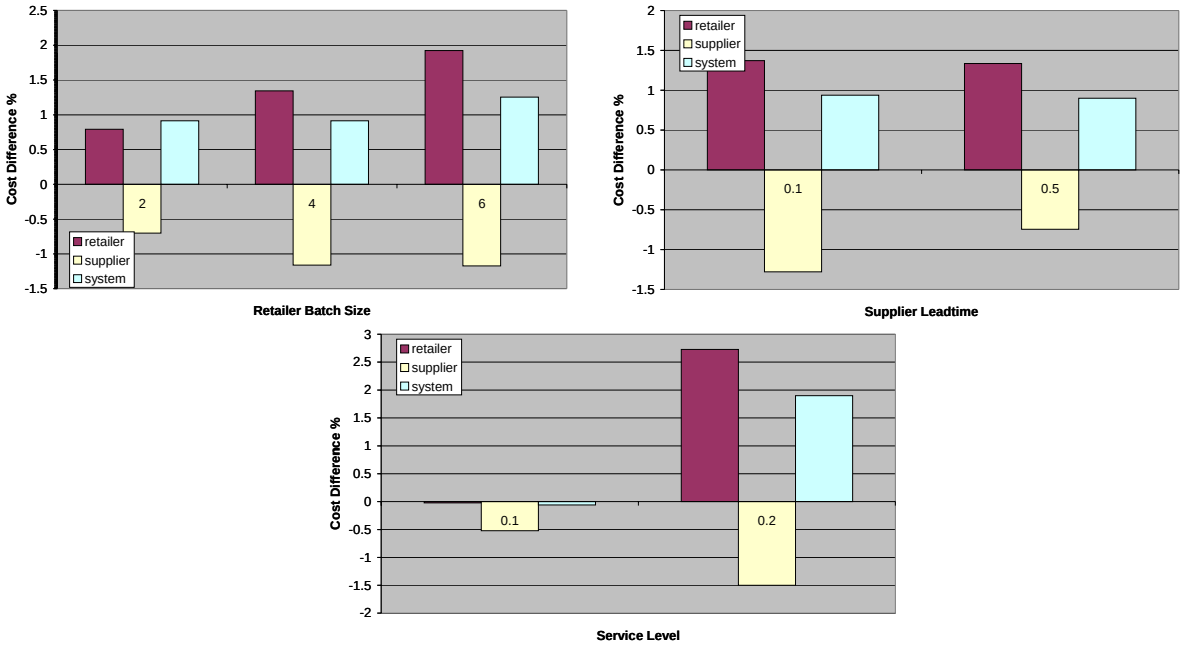
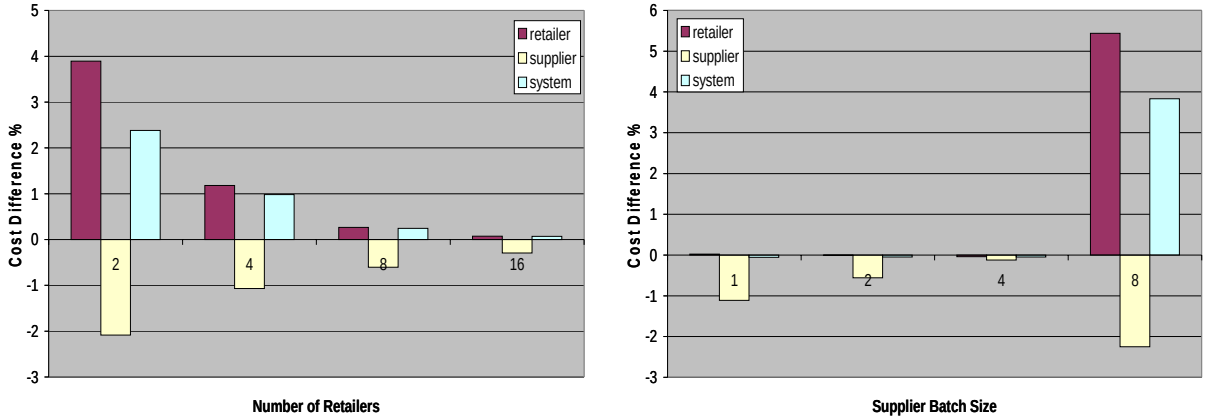


Figure 5: Average cost differences between delay policy and no-delay policy: multi-echelon decentralized setting

Next we analyze the impact of two other parameters: N, the number of retailers, and m, the supplier's batch size expressed in terms of the number of retailer's batch size. As we only consider identical retailers, when N gets large, the total order stream to the supplier becomes more "regular". In the limit when $N \to \infty$, the supplier's demand stream becomes Poisson, and the optimal actions are not to delay any orders (*i.e.,* $T \equiv 0$). Therefore, it is reasonable to expect that the cost difference is larger when N is smaller. This observation is verified in Figure 6(a). It is interesting to note again that it seems the higher the cost savings for the supplier, the higher the cost increases for the retailers and the system.



(a) (b)

Figure 6: Average cost differences between delay policy and no-delay policy: multi-echelon decentralized setting

The impact of supplier order delay as m is varied is less clear. As can be seen in Figure 6(b), as m increases, the supplier's savings will first decrease and then increase. This can be explained as follows: As the number of sub-batches gets large (*e.g.,* m=8), supplier's batch size will get large. Thus, the supplier's cost will be dominated by the holding cost due to this large order quantity. Therefore, better timing of when an order is placed becomes more critical and hence, the supplier's savings goes up. Similarly, when the supplier's batch size is small (*e.g.,* m=1), a more accurate timing of order placement is critical in balancing the holding and the shortage costs.

4.2.2 Centralized Setting

We now consider the impact of the delayed order policy in centralized distribution systems. In the multi-echelon distribution system, all the retailers face Poisson demand, so it is optimal for them not to delay orders. However, the order delay is optimal for the supplier, who faces an IFR inter-demand time (see Proposition 1 and Theorem 4). In this section, we analyze how the supplier delay as well as various system parameters affects the total system costs.

Similar to the decentralized setting, we study a comprehensive set of parameters by fixing h=1, L=1, and varying the values of λ, Q, and π for the retailers, and the values of N, m, $\hat{L}$, and $\hat{h}$ for the supplier. We choose not to present the detailed results here, however, as the percentage difference of total costs is minimal, and the overall average is small. We offer two explanations for this: First, in the centralized distribution system, retailers' total holding cost is a dominant component of the total cost, especially for large N and Q, and optimal delay order at the supplier has minimal impact on this holding cost. Second, as was shown in the previous section, whenever the supplier introduces a delay in the order, it may have a negative impact on the retailers (stochastically increasing the retailer leadtime) if not accompanied by an increase in the supplier reorder point $\widehat{R}$. In many of the cases we considered, the delay is not accompanied by the increase in $\widehat{R}$. Therefore, the cost decrease at the supplier is offset by the cost increase at the retailers. As a whole, system cost

improves only very slightly. Note that this is also consistent with our observation in decentralized systems where the total system cost difference between delayed order and (Q,R) policies was minimal.

5. Conclusion

Traditional inventory models assume either Poisson demands or that orders can only be placed at demand arrival epochs. In this paper we study a general demand process for a single location inventory system, and show that it is optimal to order at demand arrival epochs only if the inter-demand time has constant or decreasing failure rate (CFR or DFR). We also show that when the inter-demand time has increasing failure rate (IFR), the optimal policy is to delay the order. This improves the classical (Q,R) policy to the (Q,R,T) policy where T represents the delay in ordering.

We then extend this policy to the decentralized multi-echelon distribution setting. When the retailers face *i.i.d.* Poisson demands and use the (Q,R) ordering policy, their orders to the supplier follow an Erlang process. We show that the aggregation of all these order streams also has IFR, therefore it is optimal for the supplier to use the (Q,R,T) policy for ordering. In the case of a centralized multi-echelon distribution system, we show that when the two echelons have identical unit holding cost, it is again optimal for the supplier to delay its order in the (Q,R) framework.

In all three settings we give explicit expressions and procedures for the evaluation of total cost and the computation of optimal delay T. Using these we carry out comprehensive numerical tests and find that for the single location model the optimal delay can reduce the total cost significantly. The cost savings are the highest when leadtime is short and inter-demand time is more variable (high CV). The cost savings, as a function of the desired service level, follow a cyclic pattern. We provide an intuitive explanation for such behavior.

In the decentralized multi-echelon distribution system, we also find that the order delay can have significant impact on the retailers and the supplier separately. We find parameter combinations that have the most impact and provide intuitive explanations. Given the decentralized setting, it is

obvious that the supplier always benefits from the order delay. The impact on the retailer, however, is not immediately clear. We show that the supplier order delay can either benefit or hurt the retailers, but mostly the negative impact is larger in magnitude than the positive ones.

In the centralized multi-echelon distribution system, however, we find that the order delay adopted by the supplier has minimal impact on the total system cost. This has to do with the fact that holding cost at all the retailers tend to dominate the total cost and, for given Q, they are not affected by the order delay. Moreover, sometimes the order delay that benefits the supplier may have negative impact on the retailers (this is made clearer in the decentralized cases), offsetting the impact on the total system cost. This (negative) result suggests that the (Q,R) policies commonly used in distribution inventory systems, while theoretically suboptimal, are very good numerically.

To summarize, based on our analysis, the proposed delay policy has more impact on the average cost savings in single location settings with more predictable inter-demand times, short leadtimes and high service levels. In centralized distribution systems, our findings indicate that the policy has little impact on total system's cost. However, when the system is decentralized, then the supplier can benefit by adopting the delay policy; this is usually achieved at retailer's expense. In addition to conditions stated for the single location situation, this benefit is more significant in distribution systems with few retailers.

It is possible to analyze how the decentralized multi-echelon distribution system can be coordinated to perform like the centralized one. One possible mechanism for achieving the coordination is to impose an appropriate backorder penalty cost on the supplier. Given that the supplier's order delay results in a continuous spectrum of the service level on [0,1), it is always possible to find the appropriate penalty cost. We do not study this issue in this paper as the coordination issue generally applies to the distribution system, with or without order delays, and its analysis exceeds the scope of this paper. It is certainly a very interesting and promising topic for future research. Another interesting extension to this work will be to consider the case when retailers are not identical. As a result, in such situations, retailers' order quantities will be different. This

makes the analysis of the operating characteristics of the system complicated. In fact, there are only a few studies that deal with such distribution systems even under the traditional *(Q,R)* policies (e.g., Forsberg 1997, Axsater 2000).

References

1. Axsater, S., "Simple solution procedures for a class of two-echelon inventory problems," *Operations Research*, 38 (1990), 64-69.
2. Axsater, S., "Exact analysis of continuous review (R,Q) policies in two-echelon inventory systems with compound Poisson demand," *Operations Research*, 48 (2000), 686-696.
3. Axsater, S. and K. Rosling, "Installation and echelon stock policies for multilevel inventory control," *Management Science*, 39 (1993), 1274-1280.
4. Barlow, R. and F. Proschan, Statistical theory of reliability and life testing. Holt, Rinehart and Winston, Inc. (1976).
5. Chen, F., "Echelon Reorder Points, Installation Reorder Points, and the Value of Centralized Demand Information," *Management Science* 44 (1998), S221-S234.
6. Clark, A. and H. Scarf, "Optimal Policies for a Multi-Echelon Inventory Problem," *Management Science*, 6, 475-490 (1960).
7. Deurmeyer, B. and L. Schwarz, "A model for the analysis of system service level in warehouse/retailer distribution systems: the identical retailer case," L. Schwarz, ed. *Multilevel Production/Inventory Control*. Chapter 13 (TIMS Studies in Management Science 16). Elsevier, New York (1981).
8. Federgruen, A. and Y.S. Zheng, "An Efficient Algorithm for Computing an Optimal (r,Q) Policy for Continuous Review Stochastic Inventory Systems," *Operations Research*, 40 (1992), 808-812.
9. Forsberg, R. "Exact evaluation of (R,Q)-policies for two-level inventory systems with Poisson demand," *European Journal of Operational Research*, 96 (1996), 130-138.

10. Hadley, G. and T.M. Whitin, Analysis of inventory systems. Prentice Hall (1963).
11. Lee, H. and K. Moinzadeh, "Two-parameter approximations for multiechelon repairable inventory models with batch ordering policy," *IIE Transactions*, 19 (1987), 140-149.
12. Katircioglu, K., Essays in inventory control. Ph.D. dissertation. The University of British Columbia (1996).
13. Moinzadeh, K and H. Lee, "Batch size and stocking levels in multiechelon repairable systems," *Management Science*, 32 (1986), 1567-1581.
14. Moinzadeh, K., "An improved ordering policy for continuous review inventory systems with arbitrary inter-demand time distributions," *IIE Transactions*, 33 (2001), 111-118.
15. Moinzadeh, K., "A multi-echelon inventory system with information exchange," *Management Science*, 48 (2002), 414-426.
16. Nahmias, S., "On the Equivalence of Three Approximate Continuous Review Inventory Model," *Naval Research Logistics Quarterly*, 23 (1976), 31-38.
17. Nahmias, S., "Managing reparable Item Inventory Systems: A Review," in L.B. Schwarz (Ed.), Multi-Level Production/Inventory Control Systems: Theory and Practice, TIMS Studies in Management Sciences, 16, North Holland, Amsterdam (1981).
18. Sahin, I., "On the Stationary Analysis of Continuous Review (s,S) Inventory Systems with Constant Leadtimes," *Operations Research*, 27 (1979), 717-729.
19. Sahin, I., " On the Objective Function Behavior in (s,S) Inventory Models," *Operations Research*, 30 (1982), 709-724.
20. Schultz, C.R., "Replenishment Delays For Expensive Slow-Moving Items," *Management Science*, 35 (1989), 1454-1462.
21. Sivlazian, B.D., "A Continuous Review (s,S) Inventory System with Arbitrary Interarrival Distribution between Unit Demand," *Operations Research*, 22 (1974), 65-71.
22. Svoronos, A. and P. Zipkin, "Estimating the performance of multi-echelon inventory systems," *Operations Research*, 37 (1988), 57-72.

23. Tijms, H.C., Analysis of (s,S) Inventory Models, Mathematical Centre Trachts, 40, Mathematical Centrum, Amsterdam (1972).

24. Zheng, Y.S., "On Properties of Stochastic Inventory Systems," *Management Science*, 38 (1992), 87-103.

25. Zipkin, P., "Stochastic Leadtimes in Continuous Review Inventory Systems," *Naval Research Logistics Quarterly*, 35 (1986 a), 763-774.

26. Zipkin, P., "Inventory Service Level Measures: Convexity and Approximations," *Management Science*, 32 (1986 b), 975-981.

27. Zipkin. P., Foundations of Inventory Management. The McGraw-Hill Companies (2000).