Sharing and Lateral Transshipment of Inventory in a Supply Chain with Expensive Low-Demand Items

Jovan Grahovac • Amiya Chakravarty
A. B. Freeman School of Business, Tulane University, New Orleans, Louisiana 70118-5669
jovan.grahovac@tulane.edu • *akc@tulane.edu*

The emergence of carriers that deliver items to geographically dispersed destinations quickly and at a reasonable cost, combined with the low cost of sharing information through networked databases, has opened up new opportunities to better manage inventory. We investigate these benefits in the context of a supply chain in which a manufacturer supplies expensive, low-demand items to vertically integrated or autonomous retailers via one central depot. The manufacturer's lead time is assumed to be due to the geographical distance from the market or a combination of low volumes, high variety, and inflexible production processes. We formulate and solve an appropriate mathematical model based on one-for-one inventory policies in which a replenishment order is placed as soon as the customer withdraws an item. We find that sharing and transshipment of items often, but not always, reduces the overall costs of holding, shipping, and waiting for inventory. Unexpectedly, these cost reductions are sometimes achieved through *increasing* overall inventory levels in the supply chain. Finally, while sharing of inventory typically benefits all the participants in decentralized supply chains, this is not necessarily the case—sharing can hurt the distributor or individual retailers, regardless of their relative power in the supply chain.

(*Multi-Echelon Systems; Transshipment; Approximation in Inventory Models*)

1. Introduction

Manufacturers of construction equipment, aircraft, and other complex machinery face the particularly difficult task of maintaining good customer service while controlling their inventory holding costs. A random failure of just one component frequently makes an entire piece of expensive equipment unusable, which in turn results in costly downtime for the customer. Hence, much pressure is put on the entire supply chain, leading back to the manufacturer, to quickly resupply any failed parts. The situation is often complicated by a lack of clear-cut responsibility and coordination within the supply chain to provide adequate reserves of spare parts.

Obviously, the highest level of customer service is provided if the retailer or service center closest to the customer always has the necessary part in inventory. However, this arrangement requires that each retailer maintain vast inventories of different parts, which in turn results in excessive holding costs—particularly in light of the fact that many of these parts are rarely needed. Typically this problem is alleviated by one or more distribution centers located upstream in the supply chain, which, in addition to other possible roles (e.g., economies of scale in the transportation of high-demand items), provide a depth of stock and support individual retailers with infrequently needed items. As a result, rather than move directly from the manufacturer to the final customer, items pass through a

0025-1909/01/4704/0579$5.00
1526-5501 electronic ISSN

sequence of stages or *echelons* at which they may be temporarily stored, depending on the overall demand situation in the supply chain.

In this paper, we limit our attention to items that are infrequently needed and relatively costly to hold. While these items usually do not account for most of the overall costs of holding, shipping, and handling inventory (i.e., high-volume items tend to), at present they seem to be gaining importance in many contexts. This is a natural consequence of the frequently observed trends of increasing variety, customization, and complexity of products, improving reliability of individual components, and the increasing competitive significance of customer service.

Several recent trends in the business environment seem to be creating new opportunities to better manage low-demand items in the supply chain. The emergence of competing carriers who make fast deliveries to dispersed locations at a reasonable cost has made it possible to (1) speed up the movement of inventory by outsourcing the transportation function and, once this is done, (2) efficiently move items not only along the traditional predetermined paths but also between any two points in the system. Importantly, these changes have been accompanied by rapid decreases in the cost of sharing information through networked databases, which in turn has made it possible to instantaneously locate urgently needed items anywhere in the supply chain.

As a result of the above changes in the business environment, low-demand items can effectively be shared across the lowest echelon in the supply chain. This possibility seems particularly attractive if the manufacturer is geographically distant from the market, or if a combination of low volumes and high variety of parts makes resupply lead times exceedingly long (Narus and Anderson 1996). These benefits, however, may be difficult to attain in nonintegrated supply chains if retailers (or service centers closest to the customer) refuse to share the information about their inventory with the manufacturer and other interested parties.

In this paper, we investigate potential benefits of sharing and transshipment of expensive, low-demand items in the supply chain. We begin by reviewing the relevant literature in §2. In §3, we introduce our model and the notation that is used throughout the rest of the paper. In §4, we present our solution methodology and evaluate its accuracy using simulation. In §5, we use the developed methodology to explore the benefits of inventory sharing in fully integrated and decentralized supply chains. We summarize our results and give a few additional remarks in §6.

2. A Literature Review: Multi-Echelon Systems with Low-Demand Items

It is widely recognized in the literature that the type of inventory we are interested in here is appropriately managed using one-for-one control policies in which a replenishment order is issued as soon as an item is demanded. These policies are commonly known as $(S-1, S)$ policies, where S is the total number of items on hand and on order that is maintained constant. Several classic papers have developed the basic solutions for these policies in multi-echelon systems. These early works focused on military applications and the problem that certain equipment failures can be repaired at the respective base, while others require that parts be sent to or procured from a regional center or depot. Building on the single-echelon model proposed by Feeney and Sherbrooke (1966) and assuming compound Poisson demand and arbitrary distribution of resupply times, Sherbrooke (1968) developed the well-known METRIC approximation for multi-echelon systems. Allowing for the more general case with batching of orders at the depot, but restricting the model to deterministic resupply times and simple Poisson demand, Simon (1971) derived the stationary distributions of stock on hand and backorders at each facility. Shanker (1981) generalized Simon's results to the compound Poisson demand, and Graves (1985) developed an alternative solution and an improved approximation that can be used in efficient algorithms for large multi-echelon systems. Finally, Svoronos and Zipkin (1991) developed a methodology that allows stochastic lead times in multi-echelon systems.

The decreasing cost of computing made it possible to analyze an array of managerial problems that arise in commercial supply chains as well. Using the METRIC approximation and a real-life data set, Muckstadt and Thomas (1980) argued that systems in which each location is managed using a single-echelon model can be dramatically inferior to those in which inventory policies take advantage of the overall multi-echelon structure. Using the same data set, Hausman and Erkip (1994) found that, with a more appropriate choice of the objective function and a wider range of parameter values, the penalty for the autarchic model of control was less than 5%, which could be justified by the simplicity of managerial authority and organizational control.

The above works either did not consider expediting of orders or assumed that all replenishment orders were regular and all backorders were expedited. Building upon the single-echelon model developed by Moinzadeh and Schmidt (1991), Aggarwal and Moinzadeh (1994) showed that the linkage between backorders and expediting is frequently suboptimal in multi-echelon systems as well. Moinzadeh and Aggarwal (1997) further developed this argument by showing that multi-echelon systems can be substantially improved if the remaining lead times for outstanding orders are taken into account in deciding whether or not to expedite.

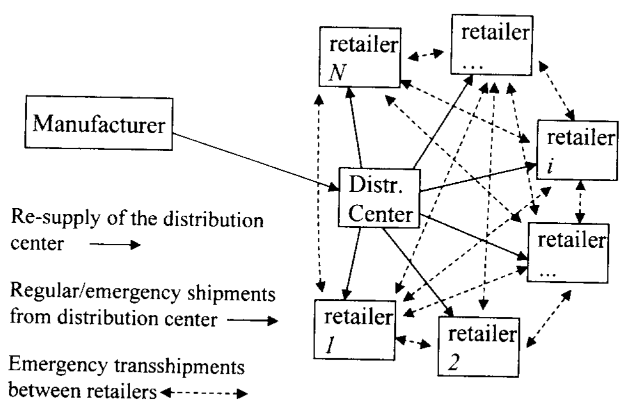
None of the above works considered the possibility of lateral transshipment between retailers (or bases). This problem was studied in the context of various batch and periodic policies by, among others, Hoadley and Heyman (1977), Karmarkar and Patel (1977), and Cohen et al. (1986). The first to tackle continuous one-for-one inventory policies in multi-echelon systems with transshipment were Dada (1984), Bowman (1986), Slay (1986), and Lee (1987). While Dada assumed pooling of all lower-echelon stock on hand *and* on order, the other three authors considered partial pooling of bases based on physical proximity, a setup similar to that used by Cohen et al. (1986). In this set up, if a base cannot satisfy a demand, the item is emergency transshipped from a nearby base, and this nearby base places a regular replenishment order with the central depot. While Lee's (1987) approximate solution builds on the approach by Graves (1985), the method developed by Bowman (1986) and later improved by Axsäter (1990) is more in the spirit of the well-known METRIC approximation.

While our work builds on, and hence is similar to Axsäter's (1990), it exhibits several important differences. One is methodological and makes our technique easily applicable to the case in which individual locations in the lowest echelon face different levels of demand. Another difference is that, in the spirit of work by Aggarwal and Moinzadeh (1994) for systems without transshipment, our model allows emergency ordering not only when there is a stockout, but at arbitrarily chosen levels of net stock. Finally, some differences stem from the fact that a typical military system and most commercial supply chains are operated in different ways and with different priorities. We clarify these operational differences in the following section.

3. Sharing and Transshipment of Inventory: Description and Notation

We consider a system in which inventory flows from the manufacturer to N service centers or retailers via one distribution center, as shown in Figure 1. All retailers and the distributor use continuous review one-for-one inventory policies. Lead times from the manufacturer to the distributor are assumed to be stochastic, independent, and to have a relatively long average duration L. Lead times from the distribu-

Figure 1 **Supply Chain with One Distribution Center, N Retailers, and Lateral Transshipment**



tor to any of the retailers consist of the deterministic transportation time and random delays that occur when the demand at the distribution center is backlogged. The deterministic transportation time is T for regular and F for emergency orders. We also assume that items can be emergency transshipped between retailers in time F. The distributor ships all orders immediately on request unless it is out of stock. Any backlogged orders are shipped on a first-come first-served basis as soon as they become available.

We assume that the demand processes at all retailers are Poisson. When the demanded item is *in stock*, the demand is filled and a replacement order is issued at the same time. If the item is out of stock, an order is placed, and the customer has to wait until the item arrives. Retailer orders can be either regular or emergency. Retailers place regular orders as long as their net stock is above a predetermined level K, which we call the "emergency trigger." Regular orders are always sourced from the distributor and arrive with a delay T. An emergency order is placed whenever a retailer's net stock is drawn to or below K. In this situation, the retailer first checks inventories throughout the system. If the distributor has items in stock, the emergency order is sourced from it, and the item arrives with a delay F. If the distributor is out of stock but the item can be located at another retailer willing to share it (i.e., that retailer has more than $K+1$ items), an emergency transshipment is requested. Emergency transshipments are sourced from a randomly selected retailer that can provide the item, and the traveling time is again F. Each transshipment causes the retailer that provided the item to place a regular replenishment order with the distributor. Finally, if all the locations are out of stock or in the emergency mode, the retailer at which the demand occurred places an emergency order that is backlogged at the distribution center. In this case, the distributor emergency ships the first item that becomes available for shipment to the respective retailer (i.e., all other pending orders by the retailer become emergency orders as long as the retailer is in the emergency mode).

The above system differs from those without lateral transshipments (e.g., Muckstadt and Thomas 1980, Hausman and Erkip 1994, Aggarwal and Moinzadeh 1994, Moinzadeh and Aggarwal 1997) in that retailers provide additional depth of stock to each other. Our system also differs from those studied by Lee (1987) and Axsäter (1990) in that it allows for emergency shipments or transshipments before or after a retailer is out of stock (i.e., we relax the constraint $K=-1$). In addition, Lee (1987) and Axsäter (1990) partition service centers into groups, and they assume that items are laterally transshipped within the group before they are sought from the depot. Their model is more descriptive of military systems in which (1) bases are relatively densely distributed so that the nearest base is usually much closer than the central depot, (2) all bases have inherent and flexible logistics capabilities, and (3) it is desirable from the system cost and objective standpoints to use these capabilities to secure fast relief from a neighboring base before the necessary item arrives from the depot. By contrast, our assumptions are more appropriate for some commercial supply chains in which, especially if the retailers are independent, items tend to be more expensive to procure from another retailer than from the distributor. Moreover, given the frequent overnight or other fixed-time guaranteed delivery times over vast geographical areas and the usual policy of locating the distribution center near the shipping company's transportation hub (Narus and Anderson 1996), there is little to be gained by transshipping within the lowest echelon first. Finally, the described arrangement also avoids unnecessary transactions and movement of items in the supply chain.

It is important to note that the system described above does not take advantage of the information about remaining lead times for outstanding orders in deciding whether to place emergency orders. Moreover, the system does not consider current inventory levels in choosing the retailer from which to source emergency transshipments. These are two weaknesses in the model, since system performance can be expected to improve if additional relevant information is used in decision making. We will return to this issue in §5, where we argue that these two concerns do not invalidate the basic results and insights that our model provides.

To facilitate the analysis, we introduce the notation that will be used throughout the rest of the paper.

We will use subscripts $i = 1, \ldots, N$ to denote retailers and $i = 0$ to denote the distribution center. We also define the following constants and variables:

- h_i = inventory holding cost at location i;
- S_i = stocking level at location i;
- K_i = emergency trigger inventory level at retailer i;
- E_i^j = emergency indicator variable for retailer i, $E_i^j = 0$ if $j > K_i$, and $E_i^j = 1$ if $j \le K_i$;
- A^j = stockout indicator variable, $A^j = 0$ if $j > 0$, and $A^j = 1$ if $j \le 0$;
- S = overall shareable stocking level of inventory available to all locations in the system $S = S_0 + \sum_{i=1}^{N}(S_i - K_i - 1)$;
- B_i = average number of backorders at location i;
- M_i = average number of items in transit in response to the demand at retailer i, and M, the average total number of items in transit to retailers (note that items in transit from the manufacturer to the distribution center are not included), $M = \sum_{i=1} M_i$;
- $1/\mu_i$ = average lead time for regular orders at retailer i;
- $1/\eta_i$ = average lead time for emergency orders at retailer i;
- λ_i = market demand rate at retailer i, and λ_0, demand rate at the distribution center, $\lambda_0 = \sum_{i=1}^{N} \lambda_i$;
- δ_i = the rate of transshipment requests from other retailers to retailer i occurring while retailer i has items in stock that it is willing to share with other retailers;
- π_i^j = steady-state probability that net stock at retailer i is j units, with $j < 0$ representing backorders;
- β_i = probability that retailer i obtains the item through regular shipment from the distributor;
- α_i = probability that retailer i obtains the item from another retailer through an emergency transshipment;
- W_i = average customer waiting time at retailer i;
- c_W = customer waiting cost per unit of time;
- c_T = cost of regular shipments in the system; and
- c_F = cost of emergency shipments and transshipments.

4. Solution Methodology

In this section, we develop a solution methodology for the supply chain with transshipments described above. In particular, we focus on characterizing probability distributions π_i^j for any given supply chain defined by the number of retailers N, stocking levels S_i, emergency triggers K_i, demand rates λ_i, and shipping times L, T, and F. Once determined, probability distributions π_i^j make it straightforward to evaluate the performance of the system in terms of the overall cost of holding, shipping, and waiting for inventory. Because our solution is close in spirit to that of Axsäter (1990), we will comment on the main differences between the two along the way.

Following the approach developed by Sherbrooke (1968), we observe that the well-known result due to Palm (1938) implies that the number of items on order at the distribution center is Poisson distributed with the parameter $\lambda_0 L$ as long as successive resupply times are independent. Hence, we write the expected number of backorders at the distribution center as

$$B_0 = \sum_{l=S_0}^{\infty} (l - S_0) \cdot e^{-\lambda_0 L} \frac{(\lambda_0 L)^l}{l!}. \tag{1}$$

The average lead time for *regular* orders at retailer i is then

$$\frac{1}{\mu_i} = T + \frac{B_0}{\lambda_0}. \tag{2}$$

To find a good approximation for the average lead time for *emergency* orders, we observe that this time is equal to F as long as items are available from the distributor or some other retailer. Items are available at the distribution center as long as the number of items on order, l, is less than S_0. When the distributor is out of stock, the emergency order can still arrive in time F if the item can be transshipped from some other retailer. We write this condition as $S_0 \le l < S - M$, where M is the expected number of items in transit to the retailers. The intuition is that the distributor "owes" $l - S_0$ items to the retailers and, in addition, an average number of M items that have left the distribution center are still in transit to the retailers. Given that the overall stocking level of shareable inventory at all retailers is $S - S_0$, items will be available for transshipment as long as $S - S_0 > l - S_0 + M$. Hence, we use the approximation that, on average, orders are backlogged at the distribution center if $l \geq S - M$. Given that the total number of backlogged orders at the distribution center is $l - S_0$, we write the following approximation of the average lead time for emergency orders at retailer i:

$$\frac{1}{\eta_i} = F + \frac{1}{\lambda_0} \sum_{l \geq S-M} (l - S_0) \cdot e^{-\lambda_0 L} \frac{(\lambda_0 L)^l}{l!}. \tag{3}$$

Items in transit include those in regular and emergency shipment from the distributor to the retailers as well as those in emergency transshipment between retailers. This pipeline stock is caused by the transportation delays T and F, and in effect implies that the system usually has fewer items than the sum of stocking levels S_i to cope with the demand. To find M_i, we observe that an arriving customer generates only one regular order if net stock at retailer i is greater than $K_i + 1$. This case happens with probability β_i. An emergency transshipment as well as one regular order are generated when the customer lowers the net stock to K_i or below while (1) the distributor is out of stock and (2) another retailer j has positive net stock that is greater than $K_j + 1$. This case happens with probability α_i. Finally, a single emergency order from the distribution center is generated in all other cases, that is, with probability $(1 - \beta_i - \alpha_i)$. Hence, following Little's law,

$$\begin{aligned} M_i &= \lambda_i[\beta_i T + \alpha_i(T + F) + F(1 - \beta_i - \alpha_i)] \\ &= \lambda_i[(\beta_i + \alpha_i)T + (1 - \beta_i)F]. \end{aligned} \tag{4}$$

Before writing the steady-state equations for net stock at retailer i, we note that the effective demand rate is $\lambda_i + \delta_i$ as long as the retailer's net stock is positive and greater than $K_i + 1$. Otherwise, the effective demand rate is λ_i. The incremental demand term δ_i is generated by other retailers that are in the emergency mode while the distribution center is out of stock. Hence, δ_i depends on the stocking and emergency trigger levels S_j and K_j throughout the system. Assuming mutual independence among all inventories, the overall expected demand for transshipments in the system when retailer i's net stock is above $K_i + 1$ can be written as

$$\sum_{k \neq i} \alpha_k \lambda_k. \tag{5}$$

To find a good approximation for probability α_k, observe that when retailer k is in the emergency mode, the number of outstanding orders at the distribution center is expected to be at least $S_k - K_k - M_k$. Given that emergency orders at retailer k will be transshipped as long as the distribution center is out of stock while the overall number of outstanding orders is sufficiently small so that an item can be found at another retailer, we write the following expression:

$$\alpha_k = (1 - \beta_k) \frac{\sum_{max\{S_0, S_k - K_k - M_k\} \leq l < S-M} e^{-\lambda_0 T} \frac{(\lambda_0 L)^l}{l!}}{\sum_{l \geq S_k - K_k - M_k} e^{-\lambda_0 T} \frac{(\lambda_0 L)^l}{l!}}. \tag{6}$$

Assuming that each retailer is a small part of the system, and thus its inventory is independent from that at the distribution center, we could approximate α_k by multiplying $(1 - \beta_k)$ with a simple sum of Poisson probabilities over $S_0 \leq l < S - M$. However, the above result becomes useful when retailers maintain higher stocking levels than the distribution center (i.e., $S_k > S_0$).

To find δ_i, observe that if retailer i has items in stock available for transshipment, the expected overall number of retailers with some stock available for sharing is

$$1 + \sum_{k \neq i} \beta_k. \tag{7}$$

Assuming that all these retailers are equally likely to provide items for transshipment, we write

$$\delta_i = \frac{\sum_{k \neq i} \alpha_k \lambda_k}{1 + \sum_{k \neq i} \beta_k}. \tag{8}$$

The above discussion characterizes the demand and replenishment processes for inventory at retailer i, which in turn can be used to find stationary probabilities π_i^j. If retailer i is not in the emergency mode, the resupply time is defined by Equations (1) and (2). If the retailer is in the emergency mode, the demand rate is λ_i. However, the demand rate for the nonemergency mode and the resupply time for the emergency mode are more complicated because we need probabilities β_i and α_i to use Equations (3), (4), (6), and (8). Unfortunately, β_i and α_i depend precisely on the stationary probabilities π_i^j that we are trying to calculate. We resolve this problem by developing a simple iterative procedure in the following subsection.

4.1. Iterative Calculation of Stationary Probabilities π_i^j

Following a methodology similar to that used by Axsäter (1990), we note that it is straightforward to write the following steady-state equations for inventory on hand and backorders at retailer i:

$$\pi_i^{S_i}[\lambda_i + (1 - A^{S_i} E_i^{S_i - l})\delta_i] = \pi_i^{S_i - l}[(1 - E_i^{S_i - l})\mu_i + E_i^{S_i - l}\eta_i]. \tag{9}$$

For $k = 1, 2, \ldots$ we have

$$\begin{aligned}\pi_i^{S_i-k}[\lambda_i + (1 - A^{S_i-k} E_i^{S_i-k-l})\delta_i &+ k(1 - E_i^{S_i-k})\mu_i + kE_i^{S_i-k}\eta_i] \\ &= \pi_i^{S_i-k+l}[\lambda_i + (1 - A^{S_i-k+l} E_i^{S_i-k})\delta_i] \\ &+ \pi_i^{S_i-k-l}(k+1)[(1 - E_i^{S_i-k-l})\mu_i \\ &+ E_i^{S_i-k-l}\eta_i].\end{aligned} \tag{10}$$

Equations (9) and (10) make it possible to express all stationary probabilities in terms of $\pi_i^{S_i}$ as

$$\pi_i^{S_i-k} = \pi_i^{S_i}\frac{\prod_{l=0}^{k-1}[\lambda_i + (1 - A^{S_i-l}E_i^{S_i-l-1})\delta_i]}{k!\prod_{l=1}^{k}[(1 - E_i^{S_i-l})\mu_i + E_i^{S_i-l}\eta_i]}, \quad k = 1, 2, \ldots. \tag{11}$$

Given that the stationary probabilities sum to one, we find

$$\pi_i^{S_i} = \frac{1}{1 + \sum_{k=1}^{\infty}\frac{\prod_{l=0}^{k-1}[\lambda_i + (1 - A_i^{S_i-l}E_i^{S_i-l-1})\delta_i]}{k!\prod_{l=1}^{k}[(1 - E_i^{S_i-l})\mu_i + E_i^{S_i-l}\eta_i]}}. \tag{12}$$

We can now formulate the following simple iterative procedure for calculating π_i^j. In the first iteration, we assume $\beta_i = 1$ and $\alpha_i = 0$ for all retailers i, and use Equations (1) through (4), (6), and (8) to obtain all the inputs necessary to calculate π_i^j using Equations (11) and (12). Subsequently, we use the obtained stationary probabilities π_i^j to calculate the new values of β_i

$$\beta_i = \sum_{j=K_i+2}^{S_i} \pi_i^j. \tag{13}$$

New values of α_i are then found using Equation (6), and we repeat the above steps until π_i^j, β_i, and α_i converge.

We tested the developed methodology on a set of numerical experiments similar to that used by Lee (1987) and Axsäter (1990). As shown in Table 1, our results for α_i and β_i closely parallel those obtained using simulation for a wide range of values. Although not directly comparable because we study systems with different priorities and patterns of transshipment, our results appear very similar in quality to those obtained by Lee (1987) and Axsäter (1990). Furthermore, all of the solutions presented in Table 1 converged in 2 to 11 iterations, with most requiring between 3 and 5 iterations. The simplicity of the calculations, combined with the fast convergence, translated into essentially instantaneous calculation of the results using general purpose spreadsheet software and a personal computer.

While our analysis is similar to Axsäter's (1990), one important difference stems from the different ways in which our two systems are operated. This is reflected in Equations (3) and (4) for the average emergency resupply times, the demand for transshipments (6) and (8), and the form of steady-state Equations (9) and (10). In addition, our iterative method makes the same set of equations readily applicable to the most general case in which retailers face different demand rates λ_i and emergency orders are not triggered by stockouts but by arbitrarily chosen net stock levels K_i. In contrast, the methodology developed by Axsäter (1990) is noniterative but requires changes in the model and hence a different set of equations whenever a retailer with different λ_i is introduced or if the link between emergency orders and stockouts is removed. In the following section, we use our computationally efficient methodology to

Table 1 **Numerical Results for $N = 10$ Identical Retailers and Shipping Times $L = 4$, $T = 2$, and $F = 0.5$**

			α_i		β_i	
λ_i	S_0	S_i	Simul.	CG	Simul.	CG
	0	1	0.23	0.24	0.77	0.76
0.04	1	1	0.15	0.16	0.85	0.84
	2	1	0.08	0.06	0.90	0.89
	0	1	0.34	0.35	0.66	0.65
		2	0.05	0.05	0.95	0.95
0.06	2	1	0.17	0.14	0.82	0.81
		2	0.02	0.02	0.98	0.98
	4	1	0.05	0.03	0.88	0.88
		2	0.00	0.00	0.99	0.99
	0	1	0.45	0.45	0.54	0.55
		2	0.08	0.08	0.92	0.92
	2	1	0.26	0.25	0.72	0.71
0.08		2	0.03	0.04	0.97	0.96
	4	1	0.11	0.07	0.83	0.82
		2	0.01	0.01	0.98	0.98
	6	1	0.02	0.02	0.86	0.85
		2	0.00	0.00	0.99	0.99
	2	1	0.37	0.35	0.62	0.61
		2	0.06	0.06	0.94	0.94
0.1	4	1	0.18	0.14	0.76	0.75
		2	0.03	0.02	0.97	0.97
	6	1	0.06	0.04	0.82	0.81
		2	0.01	0.00	0.98	0.98
		2	0.16	0.15	0.84	0.84
	5	3	0.03	0.03	0.97	0.97
0.2		4	0.00	0.01	1.00	0.99
		1	0.12	0.10	0.67	0.66
	10	2	0.03	0.02	0.94	0.93
		3	0.01	0.00	0.99	0.99

			α_i		β_i	
λ_i	S_0	S_i	Simul.	CG	Simul.	CG
		2	0.35	0.32	0.63	0.62
	12	3	0.11	0.10	0.89	0.88
		4	0.02	0.03	0.98	0.97
		2	0.09	0.07	0.81	0.80
0.40	18	3	0.02	0.02	0.95	0.95
		4	0.00	0.00	0.99	0.99
		2	0.00	0.00	0.84	0.84
	30	3	0.00	0.00	0.96	0.96
		4	0.00	0.00	0.99	0.99
		3	0.41	0.39	0.58	0.58
	14	4	0.15	0.17	0.85	0.83
		5	0.05	0.06	0.95	0.94
		3	0.17	0.15	0.81	0.80
0.60	21	4	0.05	0.05	0.94	0.94
		5	0.01	0.01	0.99	0.98
		2	0.00	0.01	0.74	0.73
	35	3	0.00	0.00	0.91	0.90
		4	0.00	0.00	0.98	0.97
		4	0.28	0.31	0.72	0.68
	20	5	0.11	0.13	0.89	0.87
		6	0.04	0.05	0.96	0.95
		7	0.01	0.02	0.99	0.98
		3	0.22	0.18	0.71	0.71
	30	4	0.09	0.07	0.89	0.89
0.80		5	0.04	0.03	0.96	0.96
		6	0.02	0.01	0.98	0.99
		3	0.02	0.02	0.83	0.83
	40	4	0.02	0.01	0.93	0.94
		5	0.01	0.00	0.98	0.98

Note. Stocking levels S_i and demand rates λ_i are varied to cover a wide range of fill rates β_i and probability of transshipment α_i.

study the benefits of sharing and transshipment of inventory in vertically integrated and decentralized supply chains.

5. Inventory Sharing in Centralized and Decentralized Supply Chains: Results

In this section, we evaluate the performance of the supply chain in terms of the overall transportation, inventory holding, and customer waiting costs. Hence we write the following function for the total cost per unit of time:

$$TC = \sum_{i=0}^{N} h_i S_i + \sum_{i=1}^{N} \lambda_i [\alpha_i c_T + (1 - \beta_i)(c_F - c_T)] + \sum_{i=1}^{N} c_W B_i. \tag{14}$$

The first summation represents all inventory holding costs. The second summation is the total additional cost stemming from the emergency shipments and transshipments in the supply chain. Note that the base cost of regular shipments from the distributor cannot be avoided; we therefore exclude it from the performance measure TC. Finally, the third summation represents all waiting costs that the customers transfer to the supply chain. Essentially, we assume that there is a fixed cost c_W for each unit of time that the customer has to wait for the desired item. This cost may be a prenegotiated dollar discount, an estimate of goodwill, or a combination of the two. Note also that the average backlog at retailer i needed in (14) is easily calculated based on π_i^j:

$$B_i = \sum_{k=S_i}^{\infty} (k - S_i)\pi_i^{S_i - k}. \qquad (15)$$

Furthermore, once we have B_i, we can easily find the expected customer waiting time $W_i = B_i/\lambda_i$, which can be used as a good measure of customer service at retailer locations.

Before we compare the performance of the supply chains with and without transshipment, we note that the case without transshipment is a special case of our model and easily solved using a method similar to that developed by Muckstadt and Thomas (1980). We need only to hold δ_i and α_i equal to zero and modify the expected lead time for emergency orders to

$$\frac{1}{\eta_i} = F + \frac{B_0}{\lambda_0}. \qquad (3')$$

Because convexity results are extremely difficult to prove for both the system with and without lateral transshipment, the results that we present below were obtained through comprehensive search over various values of S_i and K_i.

5.1. Sharing and Transshipment of Inventory in Centralized Supply Chains

In the centralized supply chain, a single decision maker minimizes Expression (14) by choosing stocking levels $S_i, i = 0, \ldots, N$ and emergency triggers $K_i, i = 1, \ldots, N$. Table 2 summarizes the results for numerical experiments conducted to evaluate the impact of sharing and transshipment of inventory on overall system costs (14). The parameters are chosen to be representative of low-demand items in the supply chains of global manufacturers of aircraft or large construction and mining equipment. The results are presented for illustrative purposes and general insight rather than to depict solutions to any specific problems.

As could be expected, whether transshipment of inventory is allowed or not, retailer stocking levels tend to increase as (1) holding costs h go down, (2) penalty for customer waiting c_W increases, and (3) demand rates λ_i increase. However, due to the small numbers involved in the optimization problem, retailer stocking levels are stable for small changes in these three parameters. As sufficiently large changes in the three parameters eventually trigger increases in retailer stocking levels, we observe rapid improvements in terms of customer waiting times W_i.

A different and somewhat counterintuitive result holds for the stocking level at the distributor. The above three monotonicity results hold for relatively *minor* changes in h, c_W, and λ_i, and for the accompanying stable stocking levels at the retailers. However, as larger changes in the three parameters trigger increases in retailer stocking levels, we can observe *decreases* in the distributor's stocking level. The intuition is that as the distributor's stocking level increases and approaches the number of retailers, it becomes optimal to push this inventory to the retailers and hence improve customer service while keeping the overall holding cost constant. This relocation of inventory, however, decreases the distributor's stocking level S_0.

Comparing the systems with and without transshipment, we observe that, in both cases, no inventory of extremely expensive, low-demand items is held at the retailer locations. These results are marked by single asterisks in Table 2. Given the absence of inventory at the lowest echelon, there are no benefits to sharing and transshipping inventory and the two systems display exactly the same performance levels in terms of TC.

Starting from this common point, however, retailer stocking levels in the system with transshipment tend to increase *at least as fast* as those in the system without transshipment when (1) holding costs h go down, (2) penalty for customer waiting c_W increases, and (3) demand rates λ_i increase. In other words, "front-line" inventories are more valuable if they can be shared and transshipped. Hence, retailer stocking levels in the supply chain with transshipment are *greater than*

Table 2 Optimal Policies for a Centralized Supply Chain with $N = 10$ Identical Retailers, Shipping Times $L = 30$, $T = 7$, and $F = 2$, and Shipping Costs $C_F = 500$ and $C_T = 50$

			With Transshipment				Without Transshipment			
C_W	λ_0		$h = 2$	$h = 10$	$h = 50$	$h = 200$	$h = 2$	$h = 10$	$h = 50$	$h = 200$
		S_0, S_i, K_i	0, 1, −1	*3, 0, −1	*2, 0, −1	*1, 0, −1	1, 1, −1	*3, 0, −1	*2, 0, −1	*1, 0, −1
	0.02	TC	21.5	60.9	142.5	303.4	23.6	60.9	142.5	303.4
		$(\alpha_i)W_i$	0.07, 0.07	0.00, 2.19	0.00,3.35	0.00, 9.44	0.13	2.20	3.35	9.44
		S_0, S_i, K_i	2, 1, −1	**0, 1, −1	*3, 0, −1	*2, 0, −1	3, 1, −1	**4, 0, −1	*3, 0, −1	*2, 0, −1
	0.05	TC	27.0	109.1	267.4	613.0	29	124.6	267.4	613.0
500		$(\alpha_i)W_i$	0.04, 0.06	0.18, 0.18	0.00, 3.80	0.00, 7.62	0.08	2.50	3.80	7.62
		S_0, S_i, K_i	4, 1, −1	0, 1, −1	*5, 0, −1	*3, 0, −1	6, 1, −1	4, 1, −1	*5, 0, −1	*3, 0, −1
	0.1	TC	37.3	136.5	462.3	1081.1	39.6	156.7	462.3	1081.1
		$(\alpha_i)W_i$	0.04, .0.10	0.35, 0.38	0.00, 3.35	0.00, 8.72	0.09	0.25	3.35	8.72
		S_0, S_i, K_i,	5, 2, −1	6, 1, −1	**2, 1, −1	*6, 0, −1	7, 2, −1	8, 1, −1	**9, 0, −1	*6, 0, −1
	0.2	TC	55.5	202.4	753.4	1971.9	58.2	220.8	820.6	1971.9
		$(\alpha_i)W_i$	0.02, 0.02	0.12, 0.22	0.48, 0.78	0.00, 6.82	0.03	0.27	2.81	6.82
		S_0, S_i, K_i	1, 1, −1	0, 1, −1	*2, 0, −1	*2, 0, −1	2, 1, −1	1, 1, −1	*2, 0, −1	*2, 0, −1
	0.02	TC	23.4	103.7	242.8	542.8	25.3	115.6	242.8	542.8
		$(\alpha_i)W_i$	0.03, 0.03	0.07, 0.07	0.00, 3.35	0.00, 3.35	0.03	0.13	3.35	3.35
		S_0, S_i, K_i	3, 1, −1	0, 1, −1	*4, 0, −1	*3, 0, −1	4, 1, −1	3, 1, −1	*4, 0, −1	*3, 0, −1
	0.05	TC	31.3	122.8	470.8	1002.1	33.3	139.1	470.8	1002.1
2,000		$(\alpha_i)W_i$	0.01, 0.04	0.18, 0.18	0.00, 2.48	0.00, 3.80	0.04	0.08	2.48	3.80
		S_0, S_i, K_i	2, 2, −1	3, 1, −1	0, 1, −1	*5, 0, −1	4, 2, −1	5, 1, −1	4, 1, −1	*5, 0, −1
	0.1	TC	47.0	162.4	592.9	1714.2	49.9	179.8	753.9	1714.2
		$(\alpha_i)W_i$	0.02, 0.01	0.08, 0.13	0.35, 0.38	0.00, 3.35	0.01	0.13	0.25	3.35
		S_0, S_i, K_i	7, 2, −1	8, 1, −1	5, 1, −1	**2, 1, −1	9, 2, −1	10, 1, −1	8, 1, −1	**9, 0, −1
	0.2	TC	59.9	256.6	893.3	2864.5	63.0	275.1	1022.6	3012.5
		$(\alpha_i)W_i$	0.01, 0.01	0.04, 0.16	0.19, 0.29	0.47, 1.04	0.01	0.16	0.27	2.81
		S_0, S_i, K_i	1, 1, −1	0, 1, −1	*2, 0, −1	*2, 0, −1	2, 1, −1	1, 1, −1	*2, 0, −1	*2, 0, −1
	0.02	TC	23.3	103.3	237.8	537.8	25.2	115.5	237.8	537.8
		$(\alpha_i)W_i$	0.03, 0.03	0.07, 0.07	0.00, 3.35	0.00, 3.35	0.03	0.13	3.35	3.35
		S_0, S_i, K_i	3, 1, −1	0, 1, −1	*4, 0, −1	*3, 0, −1	4, 1, −1	3, 1, −1	*4, 0, −1	*3, 0, −1
	0.05	TC	30.3	120.5	458.3	989.6	32.9	138.5	458.3	989.6
2,000		$(\alpha_i)W_i$	0.01, 0.04	0.18, 0.18	0.00, 2.48	0.00, 3.80	0.04	0.08	2.48	3.80
($C_F = 250$)		S_0, S_i, K_i	2, 2, −1	3, 1, −1	0, 1, −1	*5, 0, −1	4, 2, −1	5, 1, −1	3, 1, 0	*5, 0, −1
	0.1	TC	46.6	159.2	584.0	1689.2	49.8	177.9	743.9	1689.2
		$(\alpha_i)W_i$	0.02, 0.01	0.08, 0.13	0.35, 0.38	0.00, 3.35	0.01	0.13	0.37	3.35
		S_0, S_i, K_i	7, 2, −1	8, 1, −1	5, 1, −1	**2, 1, −1	9, 2, −1	9, 1, 0	7, 1, 0	**9, 0, −1
	0.2	TC	59.1	248.8	879.9	2839.5	62.4	260.9	981.1	2962.5
		$(\alpha_i)W_i$	0.01, 0.01	0.04, 0.16	0.19, 0.29	0.47, 1.04	0.01	0.08	0.23	2.81

Note. Single asterisks indicate equal performance of the systems with and without transshipment. Double asterisks indicate the cases in which the system with transshipment has higher overall inventory.

or equal to those in the supply chain without transsipment. On the other hand, inventory at the distribution center becomes relatively less valuable once transshipment of items is introduced. Hence, the distributor's stocking level in the case with transshipment tends to be *equal to or smaller than* that in the case without transshipment.

Interestingly, the two opposite trends combined with the small numbers involved in the optimization result in the possibility that the system with transshipment may hold *more inventory* than the system without transshipment. We mark these relatively rare instances with double asterisks in Table 2. In these instances, as well as all other cases in which the two

Table 3 Optimal Policies in a Centralized Supply Chain with Three Big (Subscript b) and Seven Small (Subscript s) Retailers Accounting for 60% and 40% of the Overall Demand, Respectively

		With Transshipment				Without Transshipment			
λ_0		$h = 2$	$h = 10$	$h = 50$	$h = 200$	$h = 2$	$h = 10$	$h = 50$	$h = 200$
	S_0, S_b, S_s	1, 1, 1	1, 1, 0	*2, 0, 0	*2, 0, 0	2, 1, 1	3, 1, 0	*2, 0, 0	*2, 0, 0
0.02	TC	24.0	86.9	242.8	542.8	25.8	99.5	242.8	542.8
	$K_b, W_b(\alpha_b)$	−1,0.06 (0.06)	−1,0.09 (0.07)	−1, 3.35(0.00)	−1,3.35 (0.00)	−1,0.05	−1, 0.03	−1, 3.35	−1, 3.35
	$K_s, W_s(\alpha_s)$	−1,0.02 (0.02)	−1,2.53 (0.45)	−1,3.35 (0.00)	−1,3.35 (0.00)	−1,0.02	−1,2.19	−1,3.35	−1,3.35
	$\mu/\eta, EP$	0.14, 1.7%	0.17, 4.9%	0.40, 3.7%	0.40, 1.7%	0.40, 0.8%	0.30, 3.8%	0.40, 3.7%	0.40, 1.7%
	S_0, S_b, S_s	1, 2, 1	1, 1, 1	2, 1, 0	*3, 0, 0	3, 2, 1	3, 1, 1	4, 1, 0	*3, 0, 0
0.05	TC	32.1	127.6	405.2	1002.1	34.4	142.6	464.5	1002.1
	$K_b, W_b(\alpha_b)$	−1,0.01 (0.02)	−1,0.19 (0.19)	−1,0.24 (0.08)	−1,3.80 (0.00)	−1,0.00	−1,0.16	−1,0.09	−1,3.80
	$K_s, W_s(\alpha_s)$	−1,0.06 (0.06)	−1,0.07 (0.07)	−1,3.23 (0.42)	−1,3.80 (0.00)	−1,0.05	−1,0.05	−1,2.48	−1,3.80
	$\mu/\eta, EP$	0.09, 2.8%	0.09, 2.8%	0.26, 2.9%	0.43, 2.2%	0.43, 0.8%	0.43, 0.9%	0.33, 2.1%	0.43, 2.2%
	S_0, S_b, S_s	4, 2, 1	1, 2, 1	**0, 1, 1	*5, 0, 0	6, 2, 1	4, 2, 1	**6, 1, 0	*5, 0, 0
0.1	TC	41.5	167.2	607.2	1714.2	44.2	186.8	692.4	1714.2
	$K_b, W_b(\alpha_b)$	−1,0.01 (0.00)	−1,0.07 (0.10)	−1,0.51 (0.48)	−1,3.35 (0.00)	−1,0.01	−1,0.03	−1,0.17	−1,3.35
	$K_s, W_s(\alpha_s)$	−1,0.06 (0.02)	−1,0.16 (0.16)	−1,0.32 (0.30)	−1,3.35 (0.30)	−1,0.05	−1,0.15	−1,2.51	−1,3.35
	$\mu/\eta, EP$	0.20, 3.7%	0.07, 3.7%	0.06, 3.4%	0.40, 2.6%	0.33, 2.3%	0.51, 0.8%	0.33, 3.1%	0.40, 2.6%
	S_0, S_b, S_s	5, 3, 2	6, 2, 1	2, 2, 1	7, 1, 0	7, 3, 2	9, 2, 1	8, 1, 1	8, 1, 0
0.2	TC	61.2	232.5	890.1	2799.0	64.3	252.3	1041.8	2919.6
	$K_b, W_b(\alpha_b)$	−1,0.01 (0.01)	−1,0.05 (0.04)	−1,0.20 (0.27)	−1,0.66 (0.11)	−1,0.01	−1,0.04	0,0.24	0,0.24
	$K_s, W_s(\alpha_s)$	−1,0.01 (0.00)	−1.0.13 (0.07)	−1,0.39 (0.34)	−1,3.65 (0.31)	−1,0.01	−1,0.12	−1,0.16	−1,3.57
	$\mu/\eta, EP$	0.14, 2.6%	0.17, 4.0%	0.08, 3.4%	0.37, 2.0%	0.49, 0.9%	0.36, 2.0%	0.42, 5.5%	0.42, 3.1%

Note. Customer waiting cost is $c_W = 2{,}000$, shipping times are $L = 30$, $T = 7$, $F = 2$, and shipping costs are $C_F = 500$ and $C_T = 50$. Single asterisks indicate equal performance of the two systems. Double asterisks indicate the cases in which the system with transshipment has higher overall inventory.

systems' optimal stocking levels are different, the system with transshipment consistently outperforms the other. As can be calculated from Table 2, the difference in terms of the performance measure TC is between 5% and 13% in most cases, with a maximum of 22%.

We observe values $K_i = -1$ in more than 95% of the cases in Table 2. The emergency mode thus almost completely coincides with stockouts at retailer locations for the class of items that we study. Naturally, as found by Aggarwal and Moinzadeh (1994), this relationship breaks down for sufficiently large changes in the parameters. For example, emergency orders become more attractive as their cost decreases, and we indeed observe in the lowest section of Table 2 that high-demand rates tend to result in $K_i = 0$ for the supply chain without transshipment. In these cases, retailers use the emergency mode to replenish their last unit of inventory instead of waiting for backorders to occur. In contrast, the link between the emergency mode and stockouts appears to be very stable for the supply chain with transshipment.

The above findings are confirmed if we vary the retailer size, as shown in Table 3. As in Table 2, single asterisks designate the cases in which the ability to transship makes no difference, while double asterisks mark the cases in which sharing and transshipment lead to higher overall inventory in the supply chain. Notably, if a retailer holds any items for sharing, this number is typically one or two. This result in turn implies that the actual inventory levels of candidates to supply items for transshipments must be close; for this reason, our assumption of random sourcing of in transshipments in §3 does not seem particularly strong or troubling.

We also stated in §3 that possibly costs could be further reduced if, as suggested by Moinzadeh and Aggarwal (1997), emergency orders were placed only if they could arrive faster than the outstanding regular orders. Obviously, such avoidance of emergency ordering is not an option if items are so expensive or rarely needed that they are not held at retailer locations. For all other cases, the ratios between emergency and regular resupply times (μ/η) in Table 3 indicate that the avoidance of emergency ordering is significantly more likely in the supply chain without transshipment. The reason is that, as implied by (3) and (3′), transshipments radically reduce average lead times for emergency orders because many of these orders are shipped immediately from an alternative source rather than backlogged at the distribution center. Especially when combined with high stocking levels for cheaper and more frequently demanded items (which implies many outstanding orders when there is a stockout), the high μ/η ratios for the supply chain without transshipment indicate that a large fraction of emergency orders could indeed be a avoided through the real-time use of information about outstanding orders.

It is important to note, however, that the fraction of the overall cost that is incurred on emergency shipping and transshipment (values *EP* in Table 3) is relatively small. In fact, any savings from the avoidance of emergency ordering appear unlikely to surpass the cost of holding just one additional item in the supply chain. In other words, the performance of the supply chains that we are interested in here is driven primarily by inventory holding and customer waiting costs. The avoidance of emergency orders, while important, seems secondary to decisions about stocking levels and whether or not to share inventory. Moreover, because such real-time use of information is aimed at lowering the costs rather than changing state probabilities, it is likely to supplement rather than invalidate our model and results.

Finally, Table 3 confirms the result that any reduction in inventory made possible by the sharing and transshipment of items occurs at the distribution center. In contrast, retailers tend to hold more inventory in the system with transshipment. The opposite directions in which stocking levels at the distribution center and retailers tend to move signal potential problems in decentralized supply chains in which retailers may feel threatened and refuse to share their inventories with other parties.

5.2. Sharing and Transshipment of Inventory in Decentralized Supply Chains

To analyze the decentralized system, we assume that the distribution center and the retailers each absorb their own holding cost as well as a portion of the customer waiting cost and the cost of emergency shipment and transshipment of items. We use parameter $R(0 \le R \le 1)$ to denote the relative share of the customer waiting and emergency shipping costs that is absorbed by the retailer. The distribution center incurs the remaining part $(1-R)$, and parameter R can thus be seen as a measure of the relative power of the participants in the supply chain. We write the following optimization problem for the distribution center:

$$\underset{S_0}{\text{Minimize}}\ C_0 = h_0 S_0 + (1-R)c_W \sum_{i=1}^{N} B_i + (1-R) \times \sum_{i=1}^{N} \lambda_i [\alpha_i c_T + (1-\beta_i)(c_F - c_T)] \quad (16)$$

subject to

$$\{S_i, K_i\} = \arg\max\{C_i = h_i S_i + R c_W B_i + R\lambda_i \times [\alpha_i c_T + (1-\beta_i)(c_F - c_T)]\}, \quad i = 1, \ldots, N. \quad (17)$$

The above formulation holds whether or not transshipment is allowed because we can hold $\alpha_i = 0$ and use Equation (3′) and the procedure outlined in the beginning of §5 to solve the case without transshipment. The formulation implies that the distributor acts as a Stackelberg leader that chooses stocking level S_0 knowing the retailer's response function arising from its own minimization of cost C_i over stocking levels S_i and emergency triggers K_i. We call these policies at the distribution center and the retailers *locally optimal* because all parties minimize their own narrowly defined objective functions.

Our numerical results are given in Table 4. We vary R while holding all other parameters the same as before. Because we keep $c_W = 2{,}000$ and $c_F = 500$, it is instructive to compare Table 4 with the middle section of Table 2. The distributor's stocking levels in Table 4 are consistently lower than or equal to those for the centralized supply chain (middle section of Table 2).

Table 4 Locally Optimal Policies in a Decentralized Supply Chain with 10 Identical Retailers With and Without Transshipment

			With Transshipment				Without Transshipment			
R	λ_0		$h=2$	$h=10$	$h=50$	$h=200$	$h=2$	$h=10$	$h=50$	$h=200$
		S_0, S_i, K_i	**0, 1, −1	0, 1, −1	2, 0, −1	1, 0, −1	**2, 1, −1	1, 1, −1	2, 0, −1	1, 0, −1
	0.02	C_0, C_i	2.6, 2.1	2.6, 10.1	200.0, 4.3	470.6, 11.6	4.9, 2.0	13.9, 10.2	200.0, 4.3	470.6, 11.6
		$(\alpha)E(W)$	(0.07) 0.07	(0.07) 0.07	(0.00) 3.35	(0.00) 9.44	0.03	0.13	3.35	9.44
		S_0, S_i, K_i	*2, 1, −1	**0, 1, −1	0, 1, −1	3, 0, −1	*1, 2, −1	**3, 1, −1	0, 1, 0	3, 0, −1
	0.05	C_0, C_i	9.3, 2.2	15.9, 10.7	15.9, 50.7	881.5, 12.1	4.2, 4.1	36.3, 10.3	185.8, 58.0	881.5, 12.1
0.3		$(\alpha)E(W)$	(0.04) 0.06	(0.18) 0.18	(0.18) 0.18	(0.00) 3.80	0.03	0.08	2.43	3.80
		S_0, S_i, K_i	1, 2, −1	*3, 1, −1	0, 1, −1	5, 0, −1	0, 3, −1	*1, 2, −1	3, 1, −1	5, 0, −1
	0.1	C_0, C_i	6.7, 4.3	52.7, 11.0	65.0, 52.8	1500.0, 21.4	7.6, 6.3	44.8, 21.5	231.6, 53.5	1500.0, 21.4
		$(\alpha)E(W)$	(0.5) 0.03	(0.08) 0.13	(0.35) 0.38	(0.00) 0.35	0.05	0.24	0.56	3.35
		S_0, S_i, K_i	*6, 2, −1	0, 2, −1	4, 1, −1	2, 1, −1	*1, 4, −1	1, 3, 0	1, 2, 0	2, 1, 0
	0.2	C_0, C_i	17.8, 4.2	44.2, 21.9	338.7, 55.9	725.1, 213.9	9.0, 8.3	64.6, 32.3	390.1, 114.6	1649.4, 253.5
		$(\alpha)E(W)$	(0.01) 0.02	(0.17) 0.12	(0.28) 0.41	(47) 1.04	0.02	0.16	1.11	4.24
		S_0, S_i, K_i	*0, 1, −1	0, 1, −1	0, 1, −1	1, 0, −1	*0, 2, −1	1, 1, −1	0, 1, −1	1, 0, −1
	0.02	C_0, C_i	1.1, 2.3	1.1, 10.3	1.1, 50.3	316.0, 27.1	0.3, 4.1	11.7, 10.4	14.0, 53.3	316.0, 27.1
		$(\alpha)E(W)$	(0.07) 0.07	(0.07) 0.07	(0.07) 0.07	(0.00) 9.44	0.03	0.13	1.15	9.44
		S_0, S_i, K_i	*1, 1, −1	*0, 1, −1	0, 1, −1	0, 1, −1	*1, 2, −1	*0, 2, −1	1, 1, −1	0, 1, 0
	0.05	C_0, C_i	6.0, 2.9	6.8, 11.6	6.8, 51.6	6.8, 201.6	2.9, 4.2	5.0, 21.2	75.9, 56.0	79.6, 218.6
0.7		$(\alpha)E(W)$	(0.11) 0.11	(0.18) 0.18	(0.18) 0.18	(0.18) 0.18	0.03	0.16	0.84	2.43
		S_0, S_i, K_i	0, 2, −1	*0, 1, −1	0, 1, −1	0, 1, −1	1, 3, −1	*1, 2, −1	0, 2, 0	0, 1, 0
	0.1	C_0, C_i	2.9, 4.7	27.9, 16.5	27.9, 56.5	27.9, 206.5	3.0, 6.2	24.9, 23.5	35.0, 108.2	290.4, 267.8
		$(\alpha)E(W)$	(0.05) 0.03	(0.35) 0.38	(0.35) 0.38	(0.35) 0.38	0.02	0.24	0.52	4.61
		S_0, S_i, K_i	0, 3, −1	0, 2, −1	0, 2, −1	0, 1, −1	1, 4, −1	2, 3, −1	2, 2, 0	2, 1, 0
	0.2	C_0, C_i	3.2, 6.7	19.0, 24.4	19.0, 104.4	399.2, 293.1	5.0, 8.7	32.7, 33.0	189.9, 121.0	935.5, 324.9
		$(\alpha)E(W)$	(0.04) 0.02	(0.17) 0.12	(0.17) 0.12	(0.54) 3.17	0.02	0.10	0.66	4.24

Note. Customer waiting cost is $c_W = 2{,}000$, shipping times are $L = 30$, $T = 7$, $F = 2$, and shipping costs are $C_F = 500$ and $C_T = 50$. The retailer absorbs fraction R of the customer waiting cost and the cost of emergency shipping in the system. Single asterisks indicate that the distribution center is worse-off, and double asterisks that the retailers are worse-off when transshipment is allowed.

In contrast, the retailers' stocking levels are higher than or equal to those in the centralized supply chain. In other words, even if the retailers absorb a relatively small share of 30% of the customer waiting and emergency shipping costs, they are subject to significant free riding by the distributor.

It is interesting to note that the problem of free riding by the distributor is more severe in the supply chain without transshipment. The intuition is that, when transshipment is allowed, the retailers can assist each other and are thus less vulnerable to reductions in the distributor's stocking level. The effect of mutual assistance among retailers is reinforced when high values of R further sharpen their incentives to hold inventory. Comparing the left and the right segments of Table 4, we observe that inventory sharing makes the retailers worse-off only in the two cases for $R = 0.3$, marked with double asterisks. The distributor is made worse-off in the seven cases marked with single asterisks. Importantly, inventory sharing is efficient in all other cases in the sense that it makes at least one party strictly better-off while making no one else worse-off. Overall, the results imply that sharing of inventory is usually in the best interest of all the participants in the supply chain. However, as the retailers absorb more of the emergency shipping and customer waiting costs, they tend to become relatively more interested, and the distributor relatively less interested, in instituting inventory-sharing arrangements.

Table 5 shows results for the decentralized supply chain with a mix of small and large retailers. Here, the number of equilibrium inventory policies is large, and results depend on the assumptions about

Table 5 **Locally Optimal Policies in a Decentralized Supply Chain with Three Big (Subscript *b*) and Seven Small (Subscript *s*) Retailers Accounting for 60% and 40% of the Overall Demand, Respectively**

		With Transshipment				Without Transshipment			
λ_0		$h = 2$	$h = 10$	$h = 50$	$h = 200$	$h = 2$	$h = 10$	$h = 50$	$h = 200$
	S_0, S_b, S_s	0, 1, 1	0, 0, 1	0, 1, 0	1, 0, 0	2, 1, 1	0, 1, 1	2, 0, 0	1, 0, 0
0.02	$C_0(\alpha_b, \alpha_s)$	4.1 (0.13, 0.05)	43.3 (1.0, 0.10)	81.1 (0.21, 0.98)	470.6 (0.00, 0.00)	5.2	44.3	200.0	470.6
	C_b, K_b, W_b	*2.5, −1, 0.13	5.9, −1, 2.00	*51.8, −1, 0.64	23.2, −1, 9.44	*2.1, −1, 0.05	15.3, 0, 1.96	*8.6, −1, 3.35	23.2, −1, 9.44
	C_s, K_s, W_s	2.0, −1, 0.05	10.1, −1, 0.10	*4.2, −1, 5.66	6.6, −1, 9.44	2.0, −1, 0.02	10.5, −1, 0.66	*2.4, −1, 3.35	6.6, −1, 9.44
	S_0, S_b, S_s	0, 2, 1	0, 1, 1	0, 0, 1	3, 0, 0	0, 3, 2	2, 1, 1	0, 1, 1	3, 0, 0
0.05	$C_0(\alpha_b, \alpha_s)$	*6.8 (0.05, 0.11)	23.7 (0.29, 0.13)	120.5 (1.0, 0.25)	881.5, (0.00, 0.00)	*3.9	43.9	249.2	881.5
	C_b, K_b, W_b	4.3, −1, 0.03	12.6, −1, 0.29	15.6, −1, 2.13	24.1, −1, 3.80	6.3, −1, 0.05	12.8, −1, 0.45	79.0, 0, 4.61	24.1, −1, 3.80
	C_s, K_s, W_s	2.3, −1, 0.11	*10.3, −1, 0.13	50.7, −1, 0.27	6.9, −1, 3.80	4.1, −1, 0.06	*10.2, −1, 0.13	52.8, −1, 1.62	6.9, −1, 3.80
	S_0, S_b, S_s	4, 2, 1	3, 1, 1	1, 0, 1	5, 0, 0	2, 3, 2	3, 2, 1	0, 2, 1	5, 0, 0
0.1	$C_0(\alpha_b, \alpha_s)$	*13.6 (0.01, 0.02)	*62.9 (0.14, 0.05)	*369.0 (0.94, 0.35)	1500.0 (0.00, 0.00)	*8.1	*57.5	*320.4	1500.0
	C_b, K_b, W_b	4.2, −1, 0.01	13.8, −1, 0.23	40.1, −1, 2.89	42.9, −1, 3.35	6.3, −1, 0.03	21.2, −1, 0.09	121.9, +1, 1.60	42.9, −1, 3.35
	C_s, K_s, W_s	2.3, −1, 0.06	10.4, −1, 0.09	52.3, −1, 0.51	12.2, −1, 3.35	4.1, −1, 0.03	11.2, −1, 0.33	60.2, 0, 2.76	12.2, −1, 3.35
	S_0, S_b, S_s	0, 4, 2	5, 2, 1	0, 2, 1	9, 0, 0	0, 6, 3	2, 4, 2	4, 2, 1	9, 0, 0
0.2	$C_0(\alpha_b, \alpha_s)$	*16.3 (0.05, 0.11)	*95.6 (0.08, 0.12)	220.7 (0.39, 0.48)	2648.8 (0.00, 0.00)	*12.5	*73.5	447.7	2648.8
	C_b, K_b, W_b	9.1, −1, 0.02	22.8, −1, 0.07	115.1, −1, 0.44	72.8, −1, 2.81	12.5, −1, 0.02	42.8, 0, 0.10	116.7, 1, 0.47	72.8, −1, 2.81
	C_s, K_s, W_s	4.5, −1, 0.05	11.6, −1, 0.17	57.0, −1, 0.80	20.8, −1, 2.81	6.5, −1, 0.08	22.1, −1, 0.29	58.0, 0, 0.94	20.8, −1, 2.81

Note. Customer waiting cost is $c_W = 2{,}000$, shipping times are $L = 30$, $T = 7$, and $F = 2$, and shipping costs are $C_F = 500$ and $C_T = 50$. The retailers absorb a fraction $R = 0.3$ of the customer waiting costs and the cost of emergency shipping in the system. Asterisks indicate the cases in which the respective participant is worse-off when transshipment is allowed.

the subgame played among the retailers for any fixed stocking level S_0 at the distribution center. For illustrative purposes, we assume that the large retailers act as Stackelberg leaders in the subgame with the small ones. Put differently, big retailers are assumed to be better informed and to exercise power over the smaller parties in the supply chain.

Table 5 confirms that transshipment of inventory is frequently, but not always, efficient. As marked by asterisks, any of the three classes of parties (distributor, big retailer, small retailer) can end up worse-off once transshipment is allowed. Hence, depending on the particular set of parameter values, inventory sharing may work both in favor of and against any of the participants, regardless of their relative power and leadership position in the supply chain. This result indicates that sharing arrangements in the decentralized supply chain frequently need to be accompanied by mutual monitoring and enforcement mechanisms and possibly prenegotiated cash payments among the parties.

5.3. Sharing and Transshipment of Inventory and Lost Sales Resulting from Stockouts

The model we have presented so far is easily adaptable to the case in which some customers refuse to wait, which in turn results in lost sales in the supply chain. The effective demand rate at retailer i is state dependent in this case because some customers are lost when the retailer is out of stock. Denoting the fraction of impatient customers P_L, we observe that the effective demand rate at retailer i when its net stock is j can be written as $\lambda_i - A^j P_L \lambda_i$. It is now straightforward to rewrite Equations (9) and (10) and obtain the following modified result:

$$\pi_i^{S_i-k} = \pi_i^{S_i} \frac{\prod_{l=0}^{k-1}[\lambda_i - A^{S_i-l}P_L\lambda_i + (1-A^{S_i-l}E_i^{S_i-l-1})\delta_i]}{k!\prod_{l=1}^{k}[(1-E_i^{S_i-l})\mu_i + E_i^{S_i-l}\eta_i]},$$
$$k = 1, 2, \ldots. \qquad (11')$$

Once we know stationary probabilities π_i^j, lost sales per time period at retailer i are found using

$$P_L\lambda_i \sum_{k=S_i}^{\infty} \pi_i^{S_i-k}. \qquad (18)$$

The above expressions for each retailer i can now be added to the performance measure (14) with an appropriate penalty for each lost customer. As could be expected, imposing high penalties for each lost customer makes retailers more likely to hold very expensive items and switch to the emergency resupply mode before experiencing a stockout. Because detailed results are intuitive and consistent with those presented so far, we omit them here to save space.

6. Conclusions

In this paper, we have developed a solution methodology for analyzing a supply chain with expensive, low-demand items, and the particular pattern of transshipment within the lowest echelon. Our methodology closely approximates the values obtained using simulation. Moreover, the methodology is computationally efficient and can be used to analyze a variety of problems that involve numerous, variously sized retailers.

We have used the developed methodology to study the benefits of sharing and transshipment of inventory in a particular class of supply chains—specifically, those in which proprietary technology or the retailer's exclusive relationship with the manufacturer forces the customer to wait for the desired item. Our results confirm the intuition and anecdotal evidence presented by Narus and Anderson (1996) that the ability to quickly move inventory within the lowest echelon can reduce the overall cost by up to approximately 20%. However, we reach a somewhat counterintuitive conclusion that this savings is not always accompanied by a reduction in the overall inventory in the supply chain. Furthermore, any reductions in inventory occur at the distribution center, while retailers experience stable or even increasing inventory levels.

These opposing trends can cause problems in decentralized supply chains in which some participants may need extra incentives and assurances to join the inventory-sharing and transshipment arrangement. While we can make no definitive predictions about optimal policies in the decentralized supply chain, we do find that as retailers absorb more of the customer waiting and emergency shipping costs, they tend to be more motivated to introduce sharing and transshipment of inventory, whereas the distributor becomes more likely to oppose it. Finally, the developed methodology can be used as a tool to detect potential problems and to decide whether, and what kinds of, side payments or safeguards can be used to address these problems.

Acknowledgments

This research was generously supported by the A. B. Freeman School of Business at Tulane University. The authors thank the departmental editor, the associate editors, and the two anonymous referees for their helpful comments on two earlier versions of this paper.

References

Aggarwal, P. K., K. Moinzadeh. 1994. Order execution in multi-echelon production distribution systems. *IIE Trans.* **26** 86–96.

Axsäter, S. 1990. Modeling emergency lateral transshipments in inventory systems. *Management Sci.* **36** 1329–1338.

Bowman, J: P. 1986. A multi-echelon inventory stocking algorithm for repairable items with lateral resupply. Working paper.

Cohen, M. A., P. R. Kleindorfer, H. L. Lee. 1986. Optimal stocking policies for low usage items in multi-echelon inventory systems. *Naval Res. Logistics. Quart.* **33** 17–38.

Dada, M. 1984. Inventory systems for spare parts. Unpublished Ph.D. dissertation, Sloan School of Management, MIT, Cambridge, MA.

Feeney, G. J., C. C. Sherbrooke. 1966. The (S-1, S) inventory policy under compound Poisson demand. *Management Sci.* **12** 351–411.

Graves, S. C. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Sci.* **31** 1247–1256.

Hausman, H. H., N. K. Erkip. 1994. Multi-echelon vs. single-echelon inventory control policies for low-demand items. *Management Sci.* **40** 597–602.

Hoadley, B., D. P. Heyman. 1977. A two-echelon inventory model with purchases, dispositions, shipments returns, and transshipments. *Naval Res. Logistics Quart.* **24** 1–19.

Karmarkar, U. S., N. R. Patel. 1977. The one-period, N-location distribution problem. *Naval Res. Logistics Quart.* **24** 559–575.

Lee, H. L. 1987. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Sci.* **33** 1302–1316.

Moinzadeh, K., P. K. Aggarwal. 1997. An information based multiechelon inventory system with emergency orders. *Oper. Res.* **45** 694–701.

——, C. P. Schmidt. 1991. An (S-1, S) inventory system with emergency orders. *Oper. Res.* **39** 308–321.

Muckstadt, J. A., L. J. Thomas. 1980. Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? *Management Sci.* **26** 483–494.

Narus, J. A., J. C. Anderson. 1996. Rethinking distribution. *Harvard Bus. Rev.* **74** 112–120.

Palm, C. 1938. Analysis of the Erlang traffic formula for busy-signal arrangements. *Ericsson Technics* **5** 39–58.

Shanker, K. 1981. Exact analysis of a two-echelon inventory system for recoverable items under batch inspection policy. *Naval Res. Logistics. Quart.* **28** 579–601.

Sherbrooke, C. C. 1968. METRIC: A multi-echelon technique for recoverable item control. *Oper. Res.* **16** 122–141.

Simon, R. M. 1971. Stationary properties of a two-echelon inventory model for low demand items. *Oper. Res.* **19** 761–773.

Slay, F. M. 1986. Lateral resupply in a multi-echelon inventory system. Working note, Logistics Management Institute, Washington, D.C.

Svoronos, A., P. Zipkin. 1991. Evaluation of one-for-one replenishment policies for multi-echelon inventory systems. *Management Sci.* **37** 68–83.

Accepted by Christopher S. Tang; received January 2, 1999. This paper was with the authors 9 months for 2 revisions.