Chapter 12
Multi-location Inventory Models for Retail Supply Chain Management

A Review of Recent Research

Narendra Agrawal and Stephen A. Smith

1 Introduction

Research on multi-level inventory systems is critical to retail supply chain management. Multi-level systems are commonly observed in most retail environments, where regional distributions centers (warehouses) stock products to replenish inventory at the retail stores. There is a rich and vast literature in the field of operations management that focuses on the design and management of multi-echelon inventory systems, which can be applied to retailing. Even so, a variety of open problems remain, and this continues to be a fruitful area for researchers. While more than two echelons are also observed in practice, most retailers now prefer to move toward the simpler, two-echelon systems. Such structures are common even in pure play "E-tailers," such as Amazon.com. Amazon.com started with the idea of owning no distribution centers at all, and relying on direct shipments of books from publishers to customers for demand fulfillment. However they now manage a small number of distribution centers, and use a combination of direct shipments from vendors and shipments from their warehouses for demand fulfillment. Traditional "bricks and mortar" retailers today also face the problem of designing inventory management systems for items that are purchased through their Internet sales channels, in combination with normal store replenishment.

This review paper covers a subset of the research on this topic. Because of the vastness of the literature on multi-level inventory systems, we felt it was important to limit the scope of our survey in a meaningful way. First, we restrict our attention to papers after 1993, and refer the reader to the reviews in other papers for articles prior to 1993. For example, Axsater (1993a), Federgruen (1993), and Nahmias and

N. Agrawal (✉) • S.A. Smith
Department of Operations Management and Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
e-mail: nagrawal@scu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_12

Smith (1993) contain excellent reviews of the work up to that point. We discuss some of the earlier articles that provide foundations for results that we are presenting, or were not included in the reviews listed above. Second, we omit papers on certain model formulations that are not typical of retail inventory management. For example, we exclude the literature on serial systems, since they are not representative of typical retail chains, and are a special case of general multi-location multi-echelon systems. Also excluded are papers that assume deterministic demand, since demand uncertainty is a key aspect of most retail systems.

Finally, we focus our attention primarily on periodic review systems. Most retail chains today employ technologies such as point-of-sale (POS) scanner systems that provide real time access to sales and inventory data. Consequently, in principle, continuous review models could be an appropriate construct for these retail systems. However, two issues limit the practical applicability of this assumption. First, due to contracts with vendors and shipping companies, shipments occur primarily on a pre-specified schedule, and often a variety of items are delivered simultaneously. Second, despite the real time access to sales information, the ERP databases and inventory allocation algorithms are typically updated periodically. Thus, strictly speaking, inventory decisions must be made by planners according to predefined cycles. Consequently, periodic review systems are a better representation of the inventory management systems used by most retailers. For the sake of completeness, in the appendix we briefly present the formulation of some continuous review models along with a few key references.

The rest of the paper is organized as follows: We begin by discussing the key modeling issues in Sect. 2. In Sect. 3, we present the general formulation for periodic review inventory model, and review the relevant literature. Key conclusions and opportunities for further research are discussed in Sect. 4. The continuous review model is discussed briefly in the Appendix.

2 Modeling Issues

2.1 The Key Decision

The fundamental decision to be made in two-echelon retail inventory systems is the appropriate division of inventory between the central (warehouse) location, and each of the retail stores.[1] Clearly, more inventory at the retail stores provides a higher service level to customer demand, but this also increases costs associated with carrying the inventory. The holding cost is higher at stores, due to increased shrinkage and because space in retail stores is typically more costly than warehouse space.

[1] Earlier papers used the term "retailers" to refer to individual retail locations, while more recent papers have used the term "stores." In this paper, we will use the term stores or retail stores for the lowest echelon level in the inventory system.

Higher costs also result from transporting additional items to stores, which increases the product's value. Also, immediate distribution of a large proportion of the inventory to stores makes it difficult to address subsequent inventory imbalances across stores, because lateral shipments between stores are not part of normal replenishment. That is, keeping additional inventory at the warehouse offers the advantage of risk pooling, since inventory can be directed to those stores that need it most. This can potentially reduce overall inventory investments and costs. However, the resulting shipment delays may adversely affect customer service levels. This type of risk pooling has been referred to as the *depot effect*. The other advantage of having a warehouse is the possibility of risk pooling over the length of the replenishment lead time from the external supplier. This is sometimes referred to as the *joint replenishment effect*. In other words, while replenishment orders placed by the warehouse take into account actual demands at the retail stores, the actual decision to allocate this inventory to stores can be delayed until the replenishment order is received. The additional demand information gained during this lead time can be used to make more efficient inventory decisions. Note that this benefit can be realized even if the warehouse holds no inventory.

2.2 Modeling Demand

The Poisson distribution is often used to model retail store demand, using a probability function of the form

$$P\{\text{Demand} = k\} = e^{-\lambda}\lambda^k/k! \qquad k = 0, 1, 2, \ldots$$

with mean = variance = λ. The Poisson distribution is a particularly attractive assumption for modeling demand in continuous review systems because it requires only a single parameter (λ), and the resulting analysis is more tractable.

When mean demand per period is large, the normal distribution can be used to approximate the Poisson. To model discrete demand, the discrete probabilities can be approximated by

$$P\{\text{Demand} = k\} = \Phi\left(k + 0.5\middle|\mu, \sigma\right) - \Phi\left(k - 0.5\middle|\mu, \sigma\right) \qquad k = 0, 1, 2, \ldots$$

where $\Phi(x|\mu,\sigma)$ = normal cumulative distribution with mean μ and variance σ^2.

Some empirical studies of retail data (e.g., Agrawal and Smith 1996) have found that retail demands are more variable than the Poisson distribution, which has a fixed variance to mean ratio of one. There are some practical reasons why actual demand may have higher variance than would be predicted by a Poisson distribution. Random variations may occur in the underlying Poisson arrival rate due to the weather, competitors' promotions, or special events that are not captured by the inventory system's forecasts. Second, customers whose purchases are Poisson arrivals may introduce additional variability by purchasing multiple items of the

same kind. The normal distribution can accommodate more variation, by selecting a larger variance, but the empirical analysis mentioned above found that the normal distribution fit low demand items poorly because it assigns probability to negative values and because it is symmetric about its mean.

This suggests that a compound Poisson distribution or a negative binomial distribution may provide a better choice for modeling retail store demand. In particular, the negative binomial can be generated either from a Poisson distribution whose parameter λ has a gamma distribution, or from a compound Poisson with a geometrically distributed purchase quantity. Agrawal and Smith (1996) found that the negative binomial fit the store level demand data better than either the Poisson or normal distributions. The negative binomial distribution with parameters N and p has the following discrete probability function:

$$P\left(D = k \middle| N, p\right) = f_k(N,p) = \binom{N+k-1}{N-1} p^N (1-p)^k,$$
$$0 < p < 1, \quad N > 0, \quad k = 0, 1, \ldots$$

where the cumulative probability distribution is

$$F_k(N,p) = \sum_{j=0}^{k} \binom{N+j-1}{N-1} p^N (1-p)^j.$$

The mean and variance are

$$\mu = N\left(\frac{1}{p} - 1\right), \text{ and } \sigma^2 = N\left(\frac{1-p}{p^2}\right).$$

The ratio of the variance to the mean is 1/p, which is greater than one and can be arbitrarily large. This makes the negative binomial distribution particularly attractive for retailing applications that have high demand variability.

Other assumptions for modeling retail demand include the Gamma (Bradford and Sugrue 1990), Gumbel (Lariviere and Porteus 1999), and the general exponential family of distributions (Agrawal and Smith 2012).

We also note that the majority of papers assume that demand at different locations is independently distributed. There are a few exceptions that allow correlations across stores or across time, which are described later in this chapter.

Finally, in any store level model, it is important to specify assumptions regarding the treatment of excess demand at the stores. Primarily for analytical tractability, most papers assume that unmet demand is backordered, not lost. While backordering is common for some classes of expensive retail items, excess demands for most department store and grocery items result in lost sales to another retailer, or possibly substitution of another item in the store. Backordering can serve as a good approximation to the lost sales case, provided that the inventory service level at the store is sufficiently high.

A few researchers have assumed lost sales for unmet store demands. Because of the complexity of modeling lost sales, these papers generally assume that the latest store demand information is available with zero delay prior to store replenishment. This zero delay assumption is generally correct in today's retail environment, since electronic data interchange (EDI) can provide essentially continuous communication of demand information across locations, and stores are typically replenished after hours, when no sales are occurring. But the lost sales case is significantly more complex analytically than the backorder case. With lost sales, the inventory level at any time t depends on all the individual demands and replenishments that have occurred previously, while in the backorder case, computing the inventory level requires knowledge of only the total demand over the previous periods. That is, in the backorder case, the inventory level at time t ($IL(t)$) follows from the well known relationship between inventory position (IP(t)) and total demand during the lead time ($D(t-L,\ t)$), where $IL(t) = IP(t-L) - D(t-L,\ t)$. Therefore, knowledge of the actual demand or order placed in every period is not needed to determine the inventory level in a given period. This does not hold for lost sales, adding significant complexity to the analysis.

2.3 Lead Times

Two types of lead times are relevant in such systems. The first is the replenishment lead time at the warehouse for orders placed with external suppliers. Since most researchers assume no capacity constraints on the supplier, these lead times may be assumed to be constant. Exceptions are papers that explicitly model production capacity constraints. We briefly mention this literature later. The second lead time is for orders placed by retail stores at the warehouse. This consists of two components—the shipment time, which is generally assumed to be constant (but may vary across locations), and the lead time due to shortage delays at the warehouse, which is random. Consequently, the effective lead time at the stores, i.e., the sum of the two components is always stochastic due to the possibility of stockouts at the warehouse. It is also a function of the specific allocation rules at the warehouse when shortage occurs. Thus, determining the store lead time distribution is a key analytical challenge.

2.4 Allocation Policies Used at the Warehouse

How the warehouse allocates inventory among competing store demands in shortage situations is a critical determinant of the complexity of multi-location inventory models. It also affects the service level and the cost structure for the retail stores. Conceptually, researchers have considered four different policies for what the warehouse does with the inventory it receives from the external supplier (McGavin et al. 1993). The first policy is essentially a "pass-through," where the warehouse holds no stock, but allocates and ships it to the stores as soon as stock is

received from the supplier(s). This is similar to the cross-docking policy that is practiced at many retail warehouses today. The second policy, called the equal interval policy, attempts to balance the stores' inventory at regular intervals. The third policy is called a two-interval policy, where the warehouse makes two shipments during the period between consecutive replenishments from the supplier. The final policy is called as the virtual allocation policy, where units of inventory at the warehouse are reserved for specific demands as they occur at the retail stores. This essentially imposes a first come first served discipline on demand fulfillment. We will discuss the modeling implications of each of these policies in the next section.

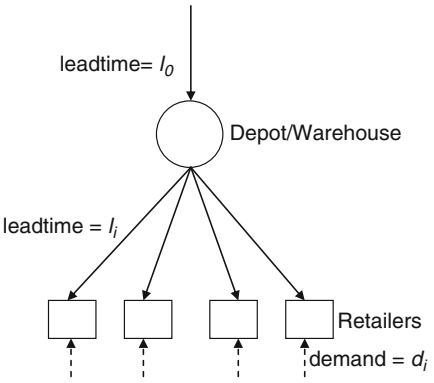
3 The General Periodic Review Inventory Model

Consider a single-item, discrete-time, two-echelon system, where the top echelon consists of a depot (also referred to as the warehouse) which supplies a collection of N retail stores, numbered $1,..,N$, with l_0 and l_i corresponding to the lead times for the depot and the retail outlet i respectively. Random demand occurs in each period at each retail store, with

$D_i(t, t+s) =$ the total demand at location i during periods $t, \ldots, t+s$, and

$$D_0(t, t+s) = \sum_{i=1}^{N} D_i(t, t+s)$$

is the system wide demand during the same period. We let $D_i^{(l)}$ and $D_0^{(l)}$ be the l-period demand at retailer i and the warehouse with cumulative distribution functions $F_i^{(l)}$ and $F_0^{(l)}$ respectively. Unmet demand is backlogged at the retailer, with a penalty cost of p_i per unit backordered and h_0 and $(h_0 + h_i)$ are the inventory holding costs assessed on ending inventory at the depot and the retailer i, respectively.



In each period, we define the following sequence of events:

1. Current period's ordering and shipment decisions are made.
2. Shipments are received.
3. Demand occurs.
4. Holding and penalty costs are assessed based on ending inventory levels.

Define $I_i(t)$ as the echelon stock (stock on hand plus in transit to and on hand at successor points minus backorders from external customers) at location i at the beginning of any period t just after the receipt of a shipment, and $\hat{I}_i(t)$ as the corresponding value at the end of the period t. Define $\hat{I}_i(t) = \hat{I}_i^+(t) - \hat{I}_i^-(t)$. Then $\hat{IP}_i(t)$ and $IP_i(t)$ are the echelon inventory positions just before and after ordering (at the depot) or shipment (if i is a retailer), where echelon inventory position is the echelon stock level plus all orders in transit to that location.

At the end of any period t, the total cost for the whole system, which includes holding and penalty costs, can be expressed as

$$\begin{aligned} & h_0\left(\hat{I}_0(t) - \sum_j \hat{I}_i(t)\right) + \sum_j (h_0 + h_i)\hat{I}_i^+(t) + \sum_j p_i \hat{I}_i^-(t) \\ & \quad = h_0\hat{I}_0(t) + \sum_i \left(h_i\hat{I}_i(t) + (h_0 + h_i + p_i)\hat{I}_i^-(t)\right). \end{aligned}$$

Then, using the notation

$$C_0(t) = h_0\hat{I}_0(t), \text{ and } C_i(t) = h_i\hat{I}_i(t) + (h_0 + h_i + p_i)\hat{I}_i^-(t).$$

The total cost is equal to:

$$C_0(t) + \sum_{i=1}^{N} C_i(t).$$

The expected system costs then depend on the *ordering decision* at the warehouse (which raises the inventory position $IP_0(t)$ of the system to, say, y_0), and on how shipment quantities for retail stores are determined, i.e., the *allocation decision*. Let the corresponding inventory positions at the retailers be denoted by $y_1, \ldots, y_N$. The first decision determines the expected cost at a warehouse at the end of period $(t + l_0)$, and limits the amount to which the aggregate echelon inventory positions of the retail stores can be raised in period $(t + l_0)$. The later decision is particularly relevant in case of shortage situations. These decisions are not independent, which makes the overall optimization problem challenging. So, the upper limit on the aggregate echelon inventory position of the stores can be specified as

$$\sum_{i=1}^{N} IP_i(t + l_0) \leq y_0 - D_0(t, t + l_0 - 1).$$

Obviously, these decisions influence the cost at echelon i at the end of period $(t + l_0 + l_i)$. Therefore, the effect of decisions made in period t, $C(t)$, is

$$C(t) = C_0(t + l_0) + \sum_{i=1}^{N} C_i(t + l_0 + l_i).$$

Thus, for any given ordering policy, the expected long-run average cost is given as

$$\lim_{T\to\infty}\frac{1}{T}E\left[\sum_{t=0}^{T-1}\sum_{i=0}^{N}C_i(t)\right]=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}E[C(t)].$$

Minimization of the long run average expected value of this function is the overall objective in the two echelon system.

3.1 Solution Methodologies

Determining optimal strategies for general two echelon systems remains difficult. Consequently, most papers use approximations. While some papers make use of relaxation techniques to obtain bounds on the true costs or profits, others impose specific restrictions on the class of inventory policies and then determine the optimal policy within that class. In all cases, the issue of inventory allocation must be addressed carefully.

The form of the optimal solution can be characterized in special cases. One way of rationing, called as the *myopic allocation method*, allocates the echelon stock of the warehouse at the beginning of period $(t + l_0)$ such that the sum of the expected costs at the stores in period $(t + l_0 + l_i)$ is minimized, without regard to later periods. A relaxation of this problem allows the quantities allocated to stores to be negative (by ignoring the constraint that the retail stores' inventory positions must be greater than at the beginning of period $t + l_0$). This is called as the *balance assumption*. The key advantage of the balance assumption is that the echelon stock (sum of the total inventory in the system) suffices to determine the warehouse ordering decision. Further, it also makes the myopic allocation policy optimal. The drawback is that this approach gives up the risk pooling advantage associated with holding stock back at the warehouse. In any case, the balance assumption underestimates the total costs since it is a relaxation. However, absent these assumptions, it turns out that base stock policies are not optimal for such systems (Clark and Scarf 1960). Van Donselaar and Wijngaard (1987), Eppen and Schrage (1981) and Federgruen and Zipkin (1984a) discuss the consequences of making this assumption in detail. These early papers consider special cases of the problem: for example, Eppen and Schrage (1981) consider a two echelon model with identical retailers and a depot that doesn't carry any stock. Jackson (1988) extends the Eppen and Schrage model to allow the warehouse to carry stock, while Jackson and Muckstadt (1989) allow non-identical retailers, but with identical cost parameters. Federgruen and Zipkin extend the Eppen and Schrage model to include non-identical retailers, non-stationary demand, and (*s*,*S*) ordering at the warehouse, but they determine their allocation policies under the assumption that the warehouse is stock-less. Jonsson and Silver (1987) also assume that the warehouse is stock-less, but extend the Eppen and Schrage model to include the possibility of a single, complete

redistribution of inventory between the retailers in the period before the end of any review cycle for the warehouse. Erkip et al. (1990) consider a model like Eppen and Schrage (1981) but allow demand correlation across retailers as well as time. Chen and Zheng (1994) develop lower bounds for costs, based on a cost allocation mechanism, for serial, assembly and distribution systems. Our system is an example of their distribution system.

McGavin et al. (1993) model a system with identical retailers, zero lead times for shipments from the warehouse to each retailer, centralized control and periodic replenishment at the warehouse. The overall stock allocation consists of four decisions: the number of withdrawals from the warehouse stock (which is an opportunity to allocate inventory to retailers), the time between these withdrawals, the quantity withdrawn, and the division of the withdrawn stock to each retailer. The first three decisions are set when the warehouse is replenished and the last one depends on retailer inventories. In particular, they model two opportunities for allocating stock from the warehouse to the retailers, which need not be equally spaced between warehouse replenishments. They seek to determine the effective timing of these two instances and the allocated quantities, so as to minimize lost sales per retailer. This assumption of lost sales makes this paper's contribution a significant departure from the majority of the literature in this stream of work. However, as noted before, this requires the retailer lead time to be zero. They show that the best allocation policy is one that balances retailer inventories (i.e., maximizes the minimum retailer inventory). Heuristic policies are developed assuming that the number of retailers is infinitely large, and are numerically tested in the finite retailer case. In particular, they test the 50/25 heuristic, where the first interval is 50 % of the replenishment cycle and the second withdrawal quantity is 25 % of the replenishment cycle's mean demand. The resulting analysis suggests the insight that the choice of the withdrawal quantity and division of inventory may matter more than the number of withdrawals.

Ahire and Schmidt (1996) consider a mixed continuous and periodic review system with one warehouse and multiple, non-identical retailers. While the retailers follow a continuous review (r, Q) policy, the warehouse follows a periodic review policy (with review period T). At the warehouse, the review period is divided into equally spaced intervals, where at each such point, a group of identical retailers (say, within a geographic zone) are reviewed. Each such zone, however, is reviewed only once per review cycle. The implication of this setup is that the retailer system is equivalent to a (nQ, r, T) system. The lead time consists of a deterministic component, the shipping lead time from the warehouse, and a stochastic component, due to possible shortages at the warehouse (however, order splitting is not allowed), and due to the fact that their orders are only reviewed periodically. Thus, an order may have to wait for anywhere from 1 to T periods before it is even reviewed by the warehouse. Results from Little's Law are used to approximate the shortage delays. Retailer demand is assumed to be Poisson, while the warehouse demand is approximated by a normal distribution, whose parameters are computed. The resulting approximations for financial and operational performance metrics compare well to those obtained through simulation.

Graves (1996) considers a general distribution network following a periodic review, order up to policy at each location. Under the assumption that each location orders at pre-set and known times, he specifies a *virtual allocation* policy where a unit at the supply location is committed/reserved for each unit demanded at the time of the occurrence of the demand. This assumes that the warehouse has real time visibility into the retail demand. Shipments, however, occur only at the next appropriate time after order receipt. The committed units can not be used to satisfy any other order. Unmet demand at the warehouse is backordered and satisfied in a first-come-first-served manner. Independent demand occurs at each retail location following a Poisson process, and excess demand is backordered. Since the order interval is present and excess demand is backordered, each location orders an amount that equals the total demand since the last order. The analysis requires the characterization of the run-out time, the time at which the warehouse runs out of inventory to allocate to the retail sites. The demand at the warehouse is approximated with a Negative Binomial distribution, whose moments can be determined. Various performance metrics can then be quantified using this approximation. Diks and de Kok (1998) model a general N-echelon divergent system where every location can hold stock, and determine policies that minimize long run average costs.

This idea of pre-set, staggered schedules for ordering is also considered in Chen and Samroengraja (2000). In a one-warehouse, multi-retailer model, where retailers are identical, and face i.i.d. demands, they assume that the warehouse follows a periodic review (s, S) policy to receive shipments from a source of unlimited supply with lead time L. The warehouse orders are based on its local inventory position. Between consecutive warehouse ordering epochs, the retailers, whose ordering points are pre-set and equally staggered with groups of retailers ordering at each such epoch, place orders, following base-stock policies with a common order up to level. Two different allocation policies are evaluated. The first, called past priority allocation (PPA) backlogs the unmet demand from a retailer, and fills it in a first-come-first-served manner from the inventory at the warehouse. However, actual shipment occurs only at the next epoch when the retailer places an order with shipment lead time l. The second policy, called current priority allocation (CPA) gives priority to the current order and backorders for the retailer designated to order in a given period. Thus, under PPA, the warehouse may carry inventory earmarked for a retailer while it denies inventory to orders from other retailers. In the second case, some retailers may be backlogged for several consecutive periods while others get replenished. The PPA model lends itself better to exact analysis. Solutions for this formulation are obtained through an approximation procedure. The CPA model is harder to evaluate exactly, but simulation studies indicate that the optimal policies are close to those under the PPA regime. Unlike in the Graves (1996) paper where inventory at the warehouse is committed to demands as they occur, here, the allocation decision is delayed until the retailer actually places an order. Their derivation of the exact cost function in the PPA case is based on a different accounting scheme. Warehouse holding costs occurring in period $(t+L)$ are charged to period t. For retailers, in period t, they charge the total holding and

backorder cost over the next N periods (N is the number of ordering epochs within each warehouse cycle) for the retailer designated to order in that period. The exact calculation under the CPA method is difficult since the distribution of a retailer's inventory position at any time depends not only on the inventory position L periods ago, but also on the exact pattern of deliveries from the outside supplier.

Continuing in the spirit of generalization, Axsater et al. (2002) allow the retailers to be non-identical. The warehouse holds stock and orders from an external supplier in multiples of a given batch size, receiving shipments after a fixed lead time. Lead times for shipping to retailers is constant, but can vary by retailer. Instead of the balance assumption, they consider the virtual assignment rule, where the inventory ordering decision at the warehouse accounts for all retailer inventory positions and assigns inventory to retailers as soon as orders are placed. The final inventory allocation, however, is made only upon the arrival of the replenishment. This is a more restrictive policy that overstates costs. Instead of the myopic allocation policy, they consider a two-step allocation policy, which allows some inventory to be retained at the warehouse. Essentially, at the beginning of each period, the remaining time until the next ordering opportunity is assumed to consist of two intervals, the second one being a single period, at which point reallocation can be done again. An optimization methodology is developed under these assumptions and the results are found to compare very favorably with the case of balance assumption and myopic allocation.

Under the balance assumption, Dogru et al. (2013) establish the convexity of the cost function for the infinite horizon case and discrete demand case, which implies the existence of optimal policies that are base stock policies. They also characterize newsvendor inequalities that must be satisfied by the optimal solutions. For example, for the special case of identical retailer holding and penalty costs at the retailers, and under the myopic allocation and balance assumptions, the well known *critical fractile* solution yields the optimal stocking policy for each location.

3.2 Batch Ordering

The use of batch ordering policies imposes additional complexities on the model since the demand at the warehouse is no longer a simple convolution of the retailers' individual demands. Further, if the retailers follow a periodic review policy, a retailer's order consisting of multiple batches may have to be split across multiple shipments. Of course, the issue of allocation of scarce warehouse inventory remains. Analytically, the key challenge is to determine the distribution of the retailers' replenishment lead time, which consists of both the shipping time (constant) and additional delays due to shortages at the warehouse. Two approaches have been used in the literature for this purpose. One is to evaluate when a batch is ordered by the retailer relative to when the warehouse orders it (as in Svoronos and Zipkin 1988). The second is to evaluate when a batch is ordered by the warehouse relative to when the retailer orders it. In cases with a single warehouse, the later approach is more tractable. This is the approach used in the following two papers.

Cachon (1999) considers a one warehouse N (non-identical) retailer model where the retailers as well as warehouse follow (R, nQ) policies. Retailers follow a periodic review policy with period T, but the ordering process is balanced in the sense that a fixed number N/T of retailers order every period. Unmet demand is backordered, and partial fulfillment is allowed. Retailer orders are randomly shuffled upon receipt, and fulfilled in a first-come-first-serve manner. Exact expressions are derived for costs, as well as demand variability at the warehouse. The key result is that the warehouse demand variability decreases due to balancing (rather than synchronizing retailer orders, where all retailers order simultaneously). Further, under a balanced system, increasing the length of the review period T and decreasing the order batch size also helps lower the supplier's demand variability. However, these strategies may not necessarily decrease total supply chain costs, since they might actually increase the retailers' ordering or inventory costs.

Cachon (2001a) considers a similar model but with identical retailers, and where each location reviews and orders in each period. All locations follow a batch ordering policy. Demand is stochastic and discrete. Average inventory and backorder levels and fill rates are evaluated exactly at each location. Safety stock requirements are determined exactly at the retailers, but approximately at the warehouse.

3.3 Lost Sales

All papers described thus far assume that unmet demand is backordered, McGavin et al. 1993. Another exception is Nahmias and Smith (1994), which focuses on a one warehouse multi retailer system, and assumes that a given fraction of unmet retailer demand is lost. Order up-to policies are used at the retailers, and the replenishment lead time from the warehouse is assumed zero. The warehouse also uses an order up to policy with zero lead times. The length of the review period at the warehouse is a multiple of the retailer's review period, and the stock levels are such that shortages only occur in the mth period within any cycle. This assumption, along with that of zero lead times, is necessary to lend tractability to the model.

In contrast to most other papers, they assume that the demand at the retailers follows a negative binomial distribution, which has been shown to fit retail data well (Agrawal and Smith 1996) because the variance to mean ratio is often larger than one. Since the warehouse supports many stores, the warehouse demand can be approximated by a normal distribution. Exact expressions are derived for the average inventory level and lost sales at stores and the warehouse. Representative retail data is used to illustrate the results and generate managerial insights. For example, they show that the benefits of holding stock at the warehouse depend upon item characteristics—items with low optimal service levels at stores derive the most benefit by holding the majority of the stock at the warehouse. Increasing the

frequency of store delivery can also reduce costs, especially for items that require high optimal service levels at stores.

Anupindi and Bassok (1999) quantify the benefit of centralizing stocks in a single warehouse, two-retailer setting, where a fixed fraction, $1-\alpha$, of unmet demand at the retailers is lost. The remaining customers look for the product at the other retailer. They too assume zero lead times for shipments to retailers. Each retailer faces an independent demand (with known distribution), buys from the warehouse at a unit cost w and sells it to their customers for a price p. Since they consider a stationary, infinite horizon model, the problem boils down to a single period newsvendor-type problem. In the simplest case where $\alpha = 0$, i.e., all unmet demand is lost, they show that centralization does not necessarily increase sales. This depends upon the nature of the demand distribution, as well as the value of the critical fractile. For example, for demand with a normal or exponential distribution, centralization leads to higher sales, while for a Uniform distribution, this happens only if the critical fractile has a value less than 0.77.

In the general case when $\alpha > 0$, the solution corresponds to a Nash equilibrium. They find that the expected total profits for the retailers are greater when stocks are centralized. However, the total sales are greater in the centralized case only if α is smaller than a certain threshold. The manufacturer/warehouse will prefer the centralized case only if α is smaller than a threshold (one interpretation for α in their model is the fraction of customers that, when unsatisfied at a local retailer due to stockouts, search for the goods at other retailers). Interestingly, even the total supply chain profit may decrease due to centralization in some cases. This happens when α is larger than some threshold value, which in turn is a function of the wholesale price w. These insights apply even when coordinating contracts are used. Thus, the main insight from this analysis is that while conventional wisdom dictates that costs decrease (and profits increase) under centralized systems due to risk pooling benefits, this benefit may not result for all parties in the supply chain.

3.4 Decentralized Environments (Quantifying the Value of Information Sharing)

The discussion thus far assumed that the entire supply chain was under central control, and information about all locations was available to the central decision maker. This assumption is not appropriate when the entities at the different echelons operate independently. When decisions are made so as to optimize local incentives, the overall supply chain performance may not be optimal. The consequences of the resulting actions by the supply chain participants include the well known bullwhip effect, as discussed in Lee et al. (1997a, b).

In an early paper, Eppen (1979) showed that in a multi-location model with normal and correlated demand, the total holding and penalty costs are lower in a centralized system than in a decentralized system. This result was later generalized

for other distributions in Chen and Lin (1989) and Stulman (1987), and to include inter-node transportation costs in Chang and Lin (1991).

Recently, however, spurred by the advances in information technology and software solutions, explicitly quantifying the potential value of information sharing in supply chains has been the subject of a number of papers. For example, Cachon and Fisher (2000) quantify this value in the case of a single warehouse multi-retailer environment. The retailers are identical, and use periodic review batch ordering policies. Retailers order periodically, in batches of a given size Q, and receive shipments after a fixed lead time. The warehouse also orders in multiples of Q, and receives its orders from an external supplier after a constant lead time. Inventory is allocated using a batch priority rule, where each batch order is assigned a priority, and shipments are done in the order of priority. By comparing the total supply chain costs with and without information sharing, they conclude that the value of information sharing is rather limited, 2.2 % on average. However, the benefit from shorter lead times and smaller batch sizes was nearly 20 % each. The explanation they offer is that demand information only matters when the retailer inventory levels are very low, since otherwise, they don't need to place orders. However, this is precisely when retailers actually place orders, so essentially, the demand data is already captured in the order information.

Lee et al. (2000) quantify the value of information sharing, albeit in a one warehouse one retailer supply chain. In contrast to the earlier papers which assume the demand is independent and identical across time, they assume that demand at the retailer is auto-correlated [AR(1)], such that

$$D_t = d + \rho D_{t-1} + \varepsilon_t,$$

where $d > 0$, $-1 < \rho < 1$, and ε_t is normally distributed with mean zero and standard deviation of σ. Both locations order every period in a periodic review system, with fixed lead times for shipments to each location. Unmet demand at the retailer is backordered, while at the warehouse excess demand is met with a special order placed at an external supplier at an additional cost. They assume that the manufacturer bears the full cost of guaranteeing supply to the retailer. They characterize the retailer's ordering process, which becomes the demand process for the manufacturer. In the case of no information sharing, the manufacturer only receives the retailer's orders. In the case of information sharing, the manufacturer also receives information about actual demand, which allows him to obtain the value of the error term ε_t, thereby lower demand variability. Since the manufacturer bears the full cost of assuring supply, the retailer's inventory costs remain unchanged with information sharing. However, information sharing leads to lower inventory levels as well as lower costs for the manufacturer. Further, they show that the benefit of information sharing is greater when the auto-correlation or demand variance is high. This analysis is complicated by the fact that when demand is auto-correlated, exact expressions for average inventory levels cannot be derived. Consequently, they make use of approximations for the retailer's and manufacturer's inventory levels.

Chen (1998) also quantifies value of information, but in a serial system with continuous review policies. They report cost benefits in the range of 2–9 %. Gavirneni et al. (1999) also consider a serial system (one warehouse, one retailer), but extend the model to the case where the manufacturer's capacity is limited. By comparing the base case to one in which the manufacturer obtains information about the retailers' demand distribution and inventory policy parameters, they are able to quantify the value of information. They find that the value of information is more compelling when end item demand is not very variable, when the retailer's $(S - s)$ is not very large or very small, or, when supplier's capacity is large. Aviv and Federgruen (1998) also consider the benefits resulting from sharing demand forecasts, also with limited supplier capacity.

3.5 Lateral Pooling

There is a large body of research that focuses on the issue of lateral pooling, also referred to as transshipments. In practice, this is rarely done for low-ticket items, since the cost and time involved in repackaging leftover inventory, shipping it to another location, and unpacking it again can easily wipe out the margins. However, for bigger ticket items, like electronics, expensive jackets and suits, and automobiles, this practice is common. Obviously, the presence of an information technology solution that provides information about inventory levels is a prerequisite for this system. One stream of research on transshipments addresses the problem in the context of repairable items. In the interest of staying focused on the retail environment, we will not review this literature, but instead direct the interested reader to Cohen et al. (2006), Muckstadt (2004), Axsater (1990) and Lee (1987), and the references contained therein. A more recent review of the literature can be found in Paterson et al. (2011).

Since the other locations serve as a backup location from which to fill unmet demand, albeit at some cost, this alters the penalty incurred due to shortages. Similarly, since there is the possibility of selling excess inventory to other locations, it alters the salvage value. Depending upon the cost of transshipment and the terms of the exchange, a retail location may, in some conditions, find it profitable to transfer its inventory to another location even when it has its own demand to meet. Clearly, each location will need to determine rules for when is it appropriate to give up its inventory. In any case, the inventory stocking policy must be modified. A second factor to consider is whether the stocking decisions are made centrally, or in a decentralized manner. In the later case, a game theoretic formulation is necessary to determine the optimal inventory ordering and allocation rules to appropriately model the incentives for each party. This results from the externality created due to decentralized decision making—larger inventory carried by one location could lower the stockout cost for others. Similarly, lower inventory levels at one location make it more economical for another location to dispose of its excess inventory. An important source of

distinction between papers on this topic is whether the redistribution of stock occurs *after* or *before* demand is realized.

We begin with the former category first. Early works on this topic include Krishnan and Rao (1965) and Karmarkar and Patel (1977). Both assume identical costs at retailers, an assumption later generalized by Tagaras (1989). Robinson (1990) formulates the problem for an arbitrary number of non-identical retailers, and shows the optimality of order up to policies. However, analytical solutions can be determined only for the case of identical retailers, or when there are only two retailers. Consequently, Monte Carlo simulation has been used to solve the general case. All these papers assume zero replenishment and shipment lead times. This assumption leads to the result of "complete pooling" (Tagaras 1989), which implies that if transshipment is economically viable, then it is optimal for each location to make its excess inventory available for lateral shipments, *i.e.,* there is no reason for holding inventory back at any location. This logic, *a priori*, may not hold if the replenishment lead times are non-zero. This factor is the focus of Tagaras and Cohen (1992), which we discuss next due to its generality.

Tagaras and Cohen (1992) model a multi-period, one-warehouse, two-retailer locations system, where demands occur independently at the retail locations. Shipments from the warehouse to retailer i arrive after L_i periods. Order-up-to policies are followed by each retailer, who faces a unit holding cost c_{hi} on the ending inventory OH_i as well as shortage cost c_{pi} on the backorders BO_i. Additionally, there is a unit lateral shipment cost c_{ij} incurred for the X_{ij} units shipped from i to j. The transshipment policy is determined by whether the inventory level (or inventory position) at the shipping location i is above a threshold level r_i, and target inventory level t_j, (or inventory position) at the receiving location j, which must not be exceeded after transshipment. Four transshipment policies are thus generated. The first two involve on-hand inventory level as the criteria. In the first case, transshipment occurs only if a location faces a shortage (i.e., $t_i = t_j = 0$). Under the second policy, transshipment can take place even if there are no shortages (i.e., $t_i = r_i = 0, i = 1,2$). Obviously, $r_i = r_j = 0$ implies complete pooling in this case. The third and fourth policies are similar to these two, except that the triggers are inventory positions. The objective is to determine order quantities Q_i that minimize total expected costs, as given by:

$$E(C) = \sum_{1}^{2} \left\{ c_i E(Q_i) + c_{hi} E(OH_i) + c_{pi} E(BO_i) + \sum_{j=1, j\neq i}^{2} c_{ij} E\left(X_{ij}\right) \right\}.$$

Exact analysis of this formulation is mathematically intractable. Consequently, search procedures are used to determine optimal solutions. They also derive heuristics based on the assumption of zero lead times. The key finding is that the complete pooling policy always dominates, as was the case when lead times are zero. In other words, hedging, by holding back inventory, or transshipping in anticipation of shortages is not optimal. Also, the heuristics were found to be near-optimal. These results are extended to the case where the transshipment lead times are non-zero in Tagaras and Vlachos (2002).

Archibald et al. (1997) also consider a two-location model, but assume that unmet demand at a location can be met either through transshipment from the other location, or through an emergency shipment from the supplier (no warehouse is assumed). The demand distribution is assumed to be Poisson. A Markov chain formulation is developed to characterize the optimal policies, which are shown to be of the order up to type. The model is then extended to the case of multiple items with constraints on the amount of inventory that can be carried at any location.

Herer et al. (2006) generalize Robinson (1990) to include more general cost structures, and develop an optimization approach that is guaranteed to converge, as compared to Robinson's heuristic which does not provide such a guarantee. They too assume zero lead times, show optimality of order up to policies, which are computed using Infinitesimal Perturbation Analysis. The transshipment quantities are determined by solving a linear programming formulation.

Bertrand and Bookbinder (1998), on the other hand, consider a general, periodic review model for the case where the redistribution decision is made *before* demand realization. They consider a model with multiple non-identical retailers that are supplied by a warehouse. The warehouse does not carry any stock, but allocates it to stores on the basis of their inventory levels so as to minimize total costs. In the period immediately before the end of the cycle (after which the warehouse orders again), inventory can be redistributed so as to minimize shortage in the last period. The assumption is that shortages primarily occur in the last period in any cycle. The redistribution decision is determined using a greedy heuristic. The optimal policies, and the corresponding costs and service level are determined using simulation, since any analytical treatment is intractable. Similar assumptions were made earlier in Jonsson and Silver (1987), but the objective was to minimize the total number of stockouts.

Anupindi and Bassok (1999), which was discussed earlier, model interactions between retailers when transshipments are possible. Similarly, Rudi et al. (2001) consider interactions between retailers in a game-theoretic setting, although their work is based on ideas contained in earlier papers by Parlar (1988) and Lippman and McCardle (1997). In the later two papers, in case of stockouts, it is the customer demand that is directed to the other location. This is different from the currently assumed scenario more relevant to us where products are transferred (albeit at a cost). Nonetheless, the modeling mechanics are similar. Rudi et al. (2001) consider the interactions between two firms, each modeled as a newsvendor within a single period framework. They assume that transshipment occurs *after* demand is realized, and the number of units exchanged from location i to location j is

$$T_{ij} = \min\left\{ (D_j - Q_j)^+, (Q_i - D_i)^+ \right\}.$$

A unit cost is incurred for each unit shipped, and a unit price is charged that varies by shipping location. The resulting profit functions follow in a straightforward manner from the newsvendor methodology. They characterize the optimal decision in the centralized as well as the decentralized cases by solving for the Nash

equilibrium. The pricing decision is also evaluated. Extending this approach to the case of more than two locations is complicated by the specific construction of the schedule of transshipment prices and costs.

Anupindi et al. (2001) develop a more generalized framework for the analysis of decentralized distribution systems. They assume N retailers who face stochastic demands and hold stocks locally and/or at one or more central locations. An exogenously specified fraction of any unsatisfied demand at a retailer could be satisfied using excess stocks at other retailers and/or stocks held at a central location. The operational decisions of ordering inventory and allocation of stocks and the financial decision of allocation of revenues/costs must be made in a way consistent with the individual incentives of the various independent retailers. They develop a "coopetitive" framework for the sequential inventory and allocation decisions. They define *claims* that allow them to separate the ownership and the location of inventories in the system. For the cooperative shipping and allocation decision, they develop sufficient conditions for the existence of the core of the game. For the inventory decision, they develop conditions for the existence of a pure strategy Nash Equilibrium. They show that there exists an allocation mechanism that achieves the first-best solution for inventory deployment and allocation, and develop conditions under which the first best equilibrium will be unique.

Dong and Rudi (2004) include the consequences of lateral shipments between retailers on the warehouse/manufacturer in their study. However, they do so in a single period setting with identical retailers. Recall that Anupindi and Bassok (1999) solved only the two retailer case. They analyze the case where the manufacturer is a price taker as well as one where he is a price setter (i.e., a Stackelberg leader). Following an analysis in a newsvendor type setting, they find that the benefit of transshipment is no longer guaranteed, rather it depends upon the parameters of the problem.

In an interesting paper, Zhao et al. (2005) formulate the problem faced by a network of decentralized retailers who stock inventory of a common item (they consider this problem in the context of a spare parts dealer network). Each location follow an (S, K) type policy. S denotes the order up to level while K denotes a threshold rationing level such that inventory will be shared with the other dealer only if the inventory level exceeds the threshold. Higher values of K imply that smaller portions of inventory are available for sharing. While demand occurs independently at each location, this possibility of inventory sharing changes the cost structure. Thus, each location needs an incentive to share inventory. Otherwise, it might find it profitable to retain inventory to satisfy future demand (understandably, the complete pooling result does not always hold in the decentralized setting). This manufacturer can either provide incentives for sharing, or subsidize the cost of sharing the inventory. The consequences differ. The first incentive induces the locations to lower their threshold rationing levels instead of increasing their stocking levels. The second induces them to lower their stocking levels, which results in lower service levels. Thus, from the manufacturer's point of view, a combination of such incentives may be best.

3.6 Fashion Products

The majority of the papers discussed thus far model environments in which the product being managed is a basic, replenishable item. In contrast, there is a smaller literature that explores issues relevant to the management of fashion products in large, multi-echelon retail chains. Fashion products tend to have very short selling seasons, with replenishment lead times that may be substantially longer than the length of the selling season. Consequently, these environments differ in that the retailer may have a very limited number of opportunities (often one or two) to place inventory in stores, and demand uncertainty tends to be large. At the same time, for many fashion forward retailers, sales from such products form the bulk of revenues.

For single retail location environments, the problem can be modeled in a straightforward manner using the well known newsvendor formulation. Extensions to the case of multiple locations, but with only a single opportunity to position the retail inventory, are fairly straightforward too. However, the problem is more complicated when there are multiple locations, limited inventory on hand, and more than one opportunity to stock stores. Multiple stocking opportunities also offer the possibility of forecast updates based on observed sales.

Fisher and Rajaram (2000) consider a demand model, with different store types. They consider the problem of determining the optimal set of test stores to stock prior to the beginning of the selling season. Using sales histories of comparable products from a prior season, they cluster the stores in the chain deterministically using a store similarity measure and then choose one test store from each cluster. Then, in the test period, inventory is placed in the test stores so that demand can be observed, from which, regression is used to estimate sales for the season. They use linear regression to estimate forecasts for season sales. Test stores are obtained deterministically by considering only the prior season sales.

Agrawal and Smith (2012) develop a two period inventory decision model for seasonal items at a retail chain with non-identical stores. As is typical in such scenarios, they assume that store demands can be correlated across the chain, and across the two time periods. At the beginning of the second period, demand forecasts and inventory policies can be revised, based on the observed demands in the first period. They develop a generalized Bayesian inference model assuming that the store demand distributions share a common unknown parameter. They also develop a two stage optimization methodology to determine the total order quantity, as well as the initial and revised store stocking policies for the two periods, taking into account the fact that store stocking policies in the first period affect the demand information that is collected. If many stores are stocked in the first period, better information about demand may be possible, but fixed costs associated with stocking stores, especially at low-volume ones, can lower profits. Additionally, ordering and inventory allocation decisions made in the first period also affect the amount of inventory that will be available for stores in the second period. To reduce the state space of this problem, they develop a normal approximation for the excess inventory left over at the end of the first period, which greatly simplifies the analysis.

By comparing the performance of the system under different supply chain flexibility arrangements, they develop counterintuitive insights regarding the magnitude of benefits resulting from (1) using updated demand information to modify store inventory levels and the set of stores that are stocked in mid-season (internal flexibility), and (2) flexible supply arrangements that allow the total replenishment quantity to be adjusted in mid-season (external flexibility). They find that the value from store adjustment can be significant even without learning (i.e., the ability to update demand forecasts based on observed sales) or external flexibility. The incremental value of external flexibility can also be significant, but only if it is accompanied by learning. On the other hand, the value of learning alone was small without either external flexibility or store adjustment capability. Thus, internal flexibility (store adjustment and learning) increases the value of external ordering flexibility.

3.7 Transportation Issues

A closely related problem in multi-location systems is that of determining optimal policies and routes for scheduling vehicles to deliver products to the various retailers in the network. The well known joint replenishment problem is also a part of this stream of work. This area represents a substantial body of research, and we will not review it in this paper. However, we will briefly point to some of the papers, and encourage the interested readers to follow the references therein.

Papers that focus on the joint replenishment problem when demand is deterministic include Jackson et al. (1985), Anily and Federgruen (1991), Federgruen and Zheng (1992), Vishwanathan and Mathur (1997), Speranza and Ukovich (1994) and Bramel and Simchi-Levi (1995). Papers that consider stochastic demands include Balintfy (1964) (can order, must order, order up to levels in a continuous review setting); Silver (1981) and Federgruen et al. (1984) (determining can-order policies); Atkins and Iyogun (1988) (periodic review policies for coordinated replenishments); Pantumsinchai (1992) (heuristics for Q, S policies for multiple items); Viswanathan (1997) ((T, s, S) policies); Pryor et al. (1999) (single item with transportation set up costs), and Cachon (2001b) (single store but multiple items, capacitated vehicles).

There are also many papers that consider vehicle routing along with inventory costs, but the few among these that allow for stochastic demand include Federgruen and Zipkin (1984b), McGavin, et al. (1993), Adelman and Kleywegt (1999) and Reinman et al. (1999).

3.8 Additional Issues

While the focus of the papers discussed thus far was primarily on cost minimization, another approach to system design may be driven by service level targets. For this type of problem, de Kok (1990) assumes that the depot does not carry any stock

and imposes a service level target at the retail locations. This model is extended in Verrijdt and de Kok (1995) for more general N-echelon networks, and in de Kok et al. (1994) to allow the depot to hold stock as well. Diks and de Kok (1998) derive newsvendor equalities for such systems under continuous demand.

In an interesting paper, Erkip et al. (1990) consider a multi-echelon model with multiple retail outlets whose demands may be correlated with each other and also across time, but do not consider forecast revision as demand data becomes available. They model demand at retailer j in period t as

$$d_{jt} = R_j \hat{D}_t L_t + \varepsilon_{jt},$$

where R_j is the average fraction of chain-wide demand at store j, $\hat{D}_t$ is the forecasted chain-wide demand, L_t is the normally distributed (with unit mean) *index variable* for period t, and ε_{jt} is the normally distributed (with zero mean) random forecast error at store j. The index variable parameter, common to all stores, is assumed to be an autoregressive process of order one. This is what induces correlation across stores and time. To lend tractability to their analysis, they need to assume that the coefficient of variation of demand at each store is equal. This assumption, along with the allocation assumption at the warehouse allows them to derive newsvendor type cost minimizing solutions for the problem.

While allocation policies are clearly important in the papers discussed above, this issue is also the subject of other papers developed in the context of assembly/production systems. In this case, when multiple products require the same common component, the available stock of components needs to be allocated in shortage situations. Similarly, in single location problems where there are multiple "classes" of demand, some allocation mechanism must be designed. Comparing these settings to distribution systems, it is clear that in both these cases, the inventory dynamics at the retail locations are not relevant, but the problem of inventory allocation is similar to that faced by the warehouse in our model. Without reviewing in detail, we list some of the papers in this category for the sake of completeness: Collier (1982), Baker et al. (1986), Gerchak and Henig (1986), Gerchak et al. (1988), Ha (1997) and Agrawal and Cohen (2001).

In papers discussed thus far, the locations of the various facilities was given. However, this may very well be a decision if the objective is to design (or redesign) a firm's supply chain network. This is the subject of investigation in Berman et al. (2012). They consider the joint problem of choosing the location of the DCs, assignment of retailers to DCs, and setting inventory policies at the retail locations. Using approximations for the cost average functions at the retail locations, the problem is formulated as a non-linear integer program, and a Lagrangian relaxation method is developed and tested to solve the problem.

Finally, for versions of our problem that include capacity constraints, i.e., capacitated production/distribution systems, see Glasserman and Tayur (1994) and Rappold and Muckstadt (2000), and the references therein.

4 Conclusions

The reviews presented in this paper as well as earlier ones clearly show that much has been accomplished in the area of designing and managing multi-location retail supply chain structures. However, our collaboration with a number of prominent retail chains has identified several of practical issues that have yet to be examined in any detail. The brief description of these issues that follows here is by no means an exhaustive list, and the interested reader should append this list to the other open questions discussed in many of the papers that we have reviewed here.

The trend towards micro-merchandising presents the first set of opportunities. Since local consumer preferences vary by location, retailers are attempting to customize their product assortments and model stocks to such local needs. However, this requires investing in mechanisms and methodologies that can allow retailers to determine what such differences are, and how best to let inventory policies be influenced by such information. Correlations between demand across stores and across time add additional complexity to such decisions in general. Agrawal and Smith (2012) present one approach for addressing this problem. This work can be generalized to include multiple products, multiple planning periods, and the potential to use pricing as yet another instrument for supply chain flexibility.

As we move from planning of one product to multiple products that form an assortment, practical considerations relating to product packaging become important. Products often move in supply chains in the form of pre-packs. For example, for an apparel retailer, a pre-pack might consist of one red, two black, and one grey t-shirt. Such pre-packs may also contain products corresponding to different sizes. Designing such pre-packs is critical to supply chain efficiency. Obviously, smaller pre-packs maximize the ability of stores to match supply and demand cost effectively. However, larger pre-packs minimize packaging and material handling costs throughout the supply chain. They also result in the possibility of shipping more units than are really needed at stores. When retail stores vary greatly in their sales rates, the problem of pre-pack design assumes even greater complexity.

While the mathematical models described in this paper have the ability to make unique inventory decisions at the store level, in practice, for large chains with thousands of stores, managing such a large number of policies is prohibitive. Consequently, stores are often grouped into a manageable number of categories (e.g., 4–10), such that the same policy can be implemented within a category. While mathematically suboptimal, the practical advantages are substantial. However, this raises the interesting question of how best to specify such categories, particularly considering store differences across geographies and product categories.

Pricing and markdown strategies in retail chains are yet another rich area of research. The majority of papers we have discussed here ignore the pricing decision. Most pricing papers that we are aware of are single location models. How best to determine pricing and inventory policies simultaneously across chains is an important research topic for retailing.

Finally, no discussion of the retail industry can be complete without recognizing the tremendous opportunities afforded by multi-channel formats, where retailers attempt to access customers using the traditional store, plus the Internet and catalog channels. Retailers vary greatly in their capabilities to deliver their products and services in this manner, and few appear to have realized any potential supply chain synergies from jointly optimizing such formats. This, we hope, will be a topic that researchers in the area of supply chain management will explore in the coming years.

Appendix: Continuous Review Inventory Systems

Many of the results in this research area, particularly for centrally controlled continuous review systems, grew out of the METRIC approximation derived in the seminal work done by Sherbrooke (1968). Consider a one-warehouse multi-retailer system where inventory is managed using a one-for-one $(S-1, S)$ inventory policy. Further, let the demand distribution at each retailer i be independent and Poisson (λ_i). Then, it follows that the demand faced by the warehouse is Poisson $(\lambda_0 = \Sigma_{i=1..N}\ \lambda_i\ )$. Using Palm's theorem, it then follows that the number of outstanding orders at the warehouse has a Poisson distribution with mean $\lambda_0\ L_0$, where L_0 is the replenishment lead time at the warehouse. Then, for a given order up to level S_0, expressions for expected backorders (B_0), waiting time (W_0) as well as inventory levels (I_0) can be derived as follows:

$$E(B_0) = \sum_{j=S_0+1}^{\infty} (j - S_0)\frac{(\lambda_0 L_0)^j}{j!}\exp\big(-\lambda_0 L_0\big),$$

$$E(W_0) = E(B_0)/\lambda_0,$$

$$E(I_0) = \sum_{j=0}^{S_0-1} (S_0 - j)\frac{(\lambda_0 L_0)^j}{j!}\exp\big(-\lambda_0 L_0\big).$$

While the actual lead time is random, the average lead time for retailer orders now equals the shipping lead time plus the average delay time due to shortages at the warehouse. The problem is that the random replenishment lead times for retailers are not independent, since they all depend upon the inventory situation at the warehouse. The METRIC approximation ignores this correlation, and replaces the random lead time with its expected value. This allows results similar to the ones for the warehouse to be derived for the retailers as well. Thus, cost expressions can be derived and optimized.

Exact expressions can be obtained by characterizing the steady state distributions of inventory levels. While the previous papers focused on characterizing the distribution of the retailer lead times, an alternate approach was taken by Axsater (1990) to develop an exact evaluation methodology for the costs directly. In particular, he observed that any unit ordered by facility i will be used to fill the

S_i-th unit of demand at this facility following that particular order, where S_i is the order up to level. Therefore, the distribution of the time elapsed between an order and the occurrence of the unit of demand that it will satisfy will have an Erlang (λ_i, S_i) distribution, with the following density function:

$$g_i^{S_i}(t) = \frac{\left(\lambda_i^{S_i} t^{S_i - 1}\right)^j}{(S_i - 1)!} \exp(-\lambda_i t).$$

Now, conditioning on the delay at the warehouse (which also has an Erlang distribution similar to the one above), cost expressions for that unit can be derived (consisting of holding and backordering costs). Axsater derived a recursive procedure for evaluating the resulting costs. Thus, this method primarily focuses on keeping track of costs associated with arbitrary supply units.

Such procedures and results become ineffective when we consider general systems where one-for-one policies are replaced by batch ordering policies (R, Q) due to fixed ordering costs. In this case, the demand arising from retailers is no longer Poisson, but Erlang instead. Consequently, the demand process at the warehouse is the sum of N Erlang processes, which is more complicated to analyze.

This generalization is considered in Axsater (1993b), where the author considers a one warehouse multi-retailer inventory system, with N *identical* retailers facing independent Poisson demand. However, all locations are allowed to order in batches using a *(R, Q)* policy, and the policies at the warehouse are defined in terms of retailer batches. Lead times are assumed to be constant. Unmet demand is assumed to be backordered, and costs include proportional holding as well as backordering costs. The basic idea stems from a similar observation in Axsater (1990). In this case, a sub-batch ordered at the warehouse will fill the $(R_w + 1)$th subsequent order for a retailer batch at the warehouse. Of course, this will happen after a random number of system demands. The costs are then derived by conditioning on which subsequent demand triggers an order. Exact as well as approximate evaluation procedures are derived.

Following a similar logic, in Axsater (1997), the results are further generalized to a two-level inventory system with one warehouse N retailers and constant lead times (transportation times), but where the retailers face *different compound Poisson* demand processes. All facilities apply continuous review ***echelon*** stock (R, Q) policies and backorder unmet demands. They provide a method for exact evaluation. Note however that echelon stock based policies may not always dominate installation stock based policies.

The third approach to solving such problems is based on characterizing the steady state distribution of inventory levels. For example, Graves (1985) fitted a two parameter Negative Binomial distribution to the number of outstanding orders for the basic METRIC model. In a similar manner, Chen and Zheng (1997) consider a one warehouse N retailer system where the retailers face different but independent compound Poisson demands, lead times are fixed, and orders are restricted to be batches of some specified lot size. They too assume *installation* stock based replenishment policies. For the case of simple Poisson demands, exact results are

possible. The inventory level at the warehouse can be determined easily, since its echelon inventory position has a uniform distribution. The distribution of the inventory level at the retailer locations is more complicated, for which the authors determine an exact procedure. For the case of compound Poisson demand, approximate evaluation methods are derived.

References

Adelman, D., & Kleywegt, A. (1999). *Price directed inventory routing*. Working paper, University of Chicago, Chicago, IL.

Agrawal, N., & Cohen, M. A. (2001). Optimal material control in an assembly system with component commonality. *Naval Research Logistics, 48*, 409–429.

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839–861.

Agrawal, N., & Smith, S. A. (2012). Optimal inventory management for a retail chain with diverse store demands. *European Journal of Operational Research, 225*, 393–403.

Ahire, S. L., & Schmidt, C. P. (1996). A model for a mixed continuous-periodic review one warehouse, N retailer inventory system. *European Journal of Operational Research, 92*, 69–82.

Anily, S., & Federgruen, A. (1991). Capacitated two-stage multi-item production/inventory model with joint set up costs. *Operations Research, 39*(3), 443–455.

Anupindi, R., & Bassok, Y. (1999). Centralization of stocks: Retailers vs. manufacturer. *Management Science, 45*(1), 178–191.

Anupindi, R., Bassok, Y., & Zemel, E. (2001). A general framework for the study of decentralized distribution systems. *Manufacturing and Service Operations Management, 3*(4), 349–368.

Archibald, T. W., Sassesn, S. A. E., & Thomas, L. C. (1997). An optimal policy for a two depot inventory problem with stock transfer. *Management Science, 43*(2), 173–183.

Atkins, D., & Iyogun, P. (1988). Periodic versus "can-order" policies for coordinated multi-item inventory systems. *Management Science, 34*(6), 791–796.

Aviv, Y., & Federgruen, A. (1998). *The operational benefits of information sharing and vendor managed inventory (VMI) programs*. Working paper, Washington University, St. Louis, MO.

Axsater, S. (1990). Modeling emergency lateral transshipments in inventory systems. *Management Science, 36*, 1329–1338.

Axsater, S. (1993a). Continuous review policies for multi-level inventory systems with stochastic demand. In S. C. Graves, A. H. G. Rinnooy Kan, & P. H. Zipkin (Eds.), *Handbooks in operations research and management science* (Logistics of production and inventory, Vol. 4, pp. 175–197). Amsterdam: Elsevier Science Publishing Company B.V.

Axsater, S. (1993b). Exact and approximate evaluation of batch ordering policies for two-level inventory systems. *Operations Research, 41*(4), 777–785.

Axsater, S. (1997). Simple evaluation of echelon stock (R, Q) policies for two-level inventory systems. *IIE Transactions, 29*, 661–669.

Axsater, S., Marklund, J., & Silver, E. A. (2002). Heuristic methods for centralized control of one-warehouse, N-retailer inventory systems. *Manufacturing and Service Operations Management, 4*(1), 75–97.

Baker, K. R., Magazine, M. J., & Nuttle, H. L. (1986). The effect of commonality on safety stock in a simple inventory model. *Management Science, 32*, 982–988.

Balintfy, J. (1964). On a basic class of multi-item inventory problems. *Management Science, 10*, 287–297.

Berman, O., Krass, D., & Tajbaksh, M. M. (2012). A coordinated location-inventory model. *European Journal of Operational Research, 217*, 500–508.

Bertrand, L. P., & Bookbinder, J. H. (1998). Stock redistribution in two-echelon logistics systems. *Journal of the Operations Research Society, 49*, 966–975.

Bradford, J. W., & Sugrue, P. K. (1990). A Bayesian approach to the two-period style-goods inventory problem with single replenishment and heterogeneous Poisson demands. *The Journal of the Operational Research Society, 41*(3), 211–218.

Bramel, J., & Simchi-Levi, D. (1995). A location based heuristic for general routing problems. *Operations Research, 43*, 649–660.

Cachon, G. P. (1999). Managing supply chain demand variability with scheduled ordering policies. *Management Science, 45*(3), 843–856.

Cachon, G. P. (2001a). Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review. *Operations Research, 49*(1), 79–98.

Cachon, G. (2001b). Managing a retailer's shelf space, inventory and transportation. *Manufacturing and Service Operations, 3*(3), 211–229.

Cachon, G. P., & Fisher, M. L. (2000). Supply chain inventory management and the value of shared information. *Management Science, 46*(8), 1032–1048.

Chang, P. L., & Lin, C. T. (1991). On the effects of centralization on expected costs in a multi-location newsboy problem. *Journal of Operational Research Society, 42*, 1025–1030.

Chen, F. (1998). Echelon reorder points, installation reorder points, and the value of centralized demand information. *Management Science, 44*(12), S221–S234.

Chen, M. S., & Lin, C. T. (1989). Effects of centralization on expected costs in a multi-location newsboy problem. *Journal of Operational Research Society., 40*, 597–602.

Chen, F., & Samroengraja, R. (2000). A staggered ordering policy for one-warehouse multi-retailer systems. *Operations Research, 48*(2), 281–293.

Chen, F., & Zheng, Y. S. (1994). Lower bounds for multi-echelon stochastic inventory systems. *Management Science, 40*(11), 1426–1443.

Chen, F., & Zheng, Y. S. (1997). One-warehouse multiretailer systems with centralized stock information. *Operations Research, 45*(2), 275–287.

Clark, A. J., & Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management Science, 40*, 1426–1443.

Cohen, M. A., Agrawal, N., & Agrawal, V. (2006). Achieving breakthrough service delivery through dynamic asset deployment strategies. *Interfaces, 36*(3), 259–271.

Collier, D. A. (1982). Aggregate safety stock levels and component part commonality. *Management Science, 28*, 1296–1303.

de Kok, A. G. (1990). Hierarchical production planning for consumer goods. *European Journal of Operations Research, 45*, 55–69.

de Kok, A. G., Lagodimos, A. G., & Siedel, H. P. (1994). *Stock allocation in a two-echelon distribution network under service constraints*. Working Paper. Department of Industrial Engineering and Management Science, Eindhoven University of Technology, EUT 94-03.

Diks, E. B., & de Kok, A. G. (1998). Optimal control of a divergent multi-echelon inventory system. *European Journal of Operations Research, 111*, 75–97.

Dogru, M. K., de Kok, A. G., & van Houtum, G. J. (2013). Newsvendor characterizations for one-warehouse multi-retailer systems with discrete demand under the balance assumption. *Central European Journal of Operations Research, 21*, 541–559.

Dong, L., & Rudi, N. (2004). Who benefits from transshipment? Exogenous vs. endogenous wholesale prices. *Management Science, 50*(5), 645–657.

Eppen, G. D. (1979). Effect of centralization on expected costs in a multi-location newsboy problem. *Management Science, 25*, 498–501.

Eppen, G., & Schrage, L. (1981). Centralized ordering policies in a multi-warehouse system with lead times and random demands. In L. B. Schwarz (Ed.), *Multi-level production/inventory control systems: Theory and practice* (pp. 51–67). Amsterdam: North-Holland.

Erkip, N., Hausman, W., & Nahmias, S. (1990). Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Science, 36*(3), 381–392.

Federgruen, A. (1993). Centralized planning models for multi-echelon inventory systems under uncertainty. In A. H. G. Rinnooy Kan & P. H. Zipkin (Eds.), *Logistics of production and inventory. Handbooks in operations research and management science* (Vol. 4, chap. 3, pp. 133–173). Amsterdam: Elsevier.

Federgruen, A., Groenevelt, H., & Tijms, H. (1984). Coordinated replenishments in a multi-item inventory system with compound Poisson demands and constant lead times. *Management Science, 30*, 344–357.

Federgruen, A., & Zheng, Y. S. (1992). The join replenishment problem with general joint cost structures. *Operations Research, 40*, 384–403.

Federgruen, A., & Zipkin, P. H. (1984a). Approximation of dynamic, multi-location production and inventory problems. *Management Science, 30*, 69–84.

Federgruen, A., & Zipkin, P. (1984b). A combined vehicle routing and inventory allocation problem. *Operations Research, 32*(5), 1019–1037.

Fisher, M. L., & Rajaram, K. (2000). Accurate retail testing of fashion merchandise: Methodology and application. *Marketing Science, 19*(3), 266–278.

Gavirneni, S., Kapuscinski, R., & Tayur, S. (1999). Value of information in capacitated supply chains. *Management Science, 45*(1), 16–24.

Gerchak, Y., & Henig, M. (1986). An inventory model with component commonality. *Operations Research Letters, 5*, 157–160.

Gerchak, Y., Magazine, M. J., & Gamble, A. B. (1988). Component commonality with service level requirements. *Management Science, 34*, 753–760.

Glasserman, P., & Tayur, S. (1994). The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Operations Research, 42*(5), 913–925.

Graves, S. C. (1985). A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science, 31*(10), 1247–1256.

Graves, S. C. (1996). A multi echelon inventory model with fixed replenishment intervals. *Management Science, 42*(1), 1–18.

Ha, A. (1997). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science, 43*(8), 1093–1103.

Herer, Y. T., Tzur, M., & Yucesan, E. (2006). The multilocation transshipment problem. *IIE Transactions, 38*, 185–200.

Jackson, P. L. (1988). Stock allocation in a two-echelon inventory system or what to do until your ship comes in. *Management Science, 34*, 880–895.

Jackson, P., Maxwell, W., & Muckstadt, J. (1985). The joint replenishment problem with power-of-two intervals. *IIE Transactions, 17*, 25–32.

Jackson, P. L., & Muckstadt, J. A. (1989). Risk pooling in a two-period multi-echelon inventory stocking and allocation problem. *Naval Research Logistics, 36*, 1–26.

Jonsson, H., & Silver, E. A. (1987). Analysis of a two-echelon inventory control system with complete redistribution. *Management Science, 33*(2), 215–227.

Karmarkar, U. S., & Patel, N. R. (1977). The one-period, N-location distribution problem. *Naval Research Logistics Quarterly, 24*, 559–575.

Krishnan, K. S., & Rao, V. R. K. (1965). Inventory control in N warehouse. *Journal of Industrial Engineering, 16*, 212–215.

Lariviere, M. A., & Porteus, E. L. (1999). Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science, 45*(3), 346–363.

Lee, H. L. (1987). A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science, 45*(5), 633–640.

Lee, H. L., Padmanabhan, P., & Whang, S. (1997a). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*, 546–558.

Lee, H. L., Padmanabhan, P., & Whang, S. (1997b). Bullwhip effect in a supply chain. *Sloan Management Review, 38*, 93–102.

Lee, H. L., So, C., & Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science, 46*(5), 626–643.

Lippman, S. A., & McCardle, K. F. (1997). The competitive newsboy. *Operations Research, 45*(1), 54–65.

McGavin, E. J., Schwarz, L. B., & Ward, J. E. (1993). Two-interval inventory allocation policies in a one-warehouse N-identical retailer distribution system. *Management Science, 39*(9), 1092–1107.

Muckstadt, J. A. (2004). *Analysis and algorithms for service parts supply chains* (Springer series in operations research and financial). Berlin: Springer Verlag.

Nahmias, S., & Smith, S. A. (1993). Mathematical models of retailer inventory systems: A review. In R. K. Sarin (Ed.), *Perspectives in operations management* (pp. 249–278). MA: Kluwer Academic Publishers.

Nahmias, S., & Smith, S. A. (1994). Optimizing inventory levels in a two-echelon retailer system with partial lost sales. *Management Science, 40*(5), 582–596.

Pantumsinchai, P. (1992). A comparison of three joint ordering inventory policies. *Decision Science, 23*, 111–127.

Parlar, M. (1988). Game theoretic analysis of the substitutable product inventory problem with random demands. *Naval Research Logistics, 35*, 397–409.

Paterson, C. G., Kiesmuller, R., & Teunter, K. G. (2011). A comparison of three joint ordering inventory policies. *European Journal of Operational Research, 210*, 125–136.

Pryor, K., Kapuscinski, R., & White, C. (1999). *A single item inventory problem with multiple setup costs assigned to delivery vehicles*. Working paper, University of Michigan, Ann Arbor.

Rappold, J. A., & Muckstadt, J. A. (2000). A computationally efficient approach for determining inventory levels in a capacitated multi-echelon production-distribution system. *Naval Research Logistics, 47*, 377–398.

Reinman, M., Rubio, R., & Wein, L. (1999). Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transportation Science, 33*(4), 361–380.

Robinson, L. W. (1990). Optimal and approximate policies in multi-period, multi-location inventory models with transshipments. *Operations Management, 38*(2), 278–295.

Rudi, N., Kapur, S., & Pyke, D. F. (2001). A two location inventory model with lateral shipment and local decision making. *Management Science, 47*(12), 1668–1680.

Sherbrooke, S. C. (1968). Metric: A multi-echelon technique for recoverable item control. *Operations Research, 16*(1), 122–141.

Silver, E. (1981). Establishing reorder points in the (S, c, s) coordinated control system under compound Poisson demand. *International Journal of Production Research, 19*, 743–750.

Speranza, M. G. W., & Ukovich, W. (1994). Minimizing transportation and inventory costs for several products on a single link. *Operations Research, 42*(5), 879–896.

Stulman, A. (1987). Benefits of centralized stocking for the multi-center newsboy problem with first-come-first-serve allocation. *Journal of Operational Research Society, 38*, 827–832.

Svoronos, A., & Zipkin, P. (1988). Estimating the performance of multi-level inventory systems. *Operations Research, 36*(1), 57–72.

Tagaras, G. (1989). Effects of pooling on the optimization and service levels of two-location inventory systems. *IIE Transactions, 21*(3), 250–257.

Tagaras, G., & Cohen, M. A. (1992). Pooling in two-location inventory systems with non-negligible replenishment lead times. *Management Science, 38*(8), 1067–1083.

Tagaras, G., & Vlachos, D. (2002). Effectiveness of stock transshipment under various demand distributions and nonneglibible transshipment times. *Production and Operations Management, 11*(2), 183–198.

Van Donselaar, K., & Wijngaard, J. (1987). Commonality and safety stocks. *Engineering Costs and Production Economics, 12*, 197–204.

Verrijdt, J. H. C. M., & de Kok, A. G. (1995). Distribution planning for a divergent N-echelon network without intermediate stock under service restrictions. *International Journal of Production Economics, 38*, 225–243.

Vishwanathan, S., & Mathur, K. (1997). Integrating routing and inventory decisions in a one-warehouse multi-retailer multi-product distribution system. *Management Science, 43*(3), 294–312.

Viswanathan, S. (1997). Periodic review (s, S) policies for joint replenishment inventory systems. *Management Science, 43*(10), 1447–1453.

Zhao, H., Deshpande, V., & Ryan, J. K. (2005). Inventory sharing and rationing in decentralized dealer networks. *Management Science, 51*(4), 531–547.