Managing Inventories in Multi-echelon On-line Retail Fulfillment System with Different Response Lead Time Demands

Juan Li
Palo Alto Research Center
Webster, NY 14580
Email: juan.li@xerox.com

John A. Muckstadt
School of Operations Research and Information Engineering
Cornell University, Ithaca, NY 14853
Email: jam61@cornell.edu

***Abstract*—When designing and operating an order fulfillment system for an on-line retailer, many factors must be taken into account. In this paper, we study a multi-echlon on-line fulfillment system with different response lead time demands. We present a delayed allocation system, which is called the primary warehouse system (PWS). In this system, inventories to satisfy different response lead time demands are managed differently. Since there are many millions of items managed in the system, determining stock levels quickly is a necessity. The focus of this paper is on planning inventory levels. Specifically, our goals are to describe a model for setting stock levels for each item, to present a computationally tractable method for determining their values, and to provide numerical results that illustrate the applications of the model to the on-line retailer's environment.**

I. Introduction

When designing their multi-warehouse distribution systems, on-line retailers must ensure the system provides timely response to customer's orders while minimizing total inventory, warehousing, and transportation costs. This order fulfillment system must contain enough geographically distributed warehouse capacity to hold inventories required to meet uncertain and time varying demand for millions of items. Multiple warehouses are needed so that orders can be filled cost effectively from warehouses that are relatively close to the customer's delivery address. These "last mile delivery costs" are significant and must be controlled.

Recently we worked with a major on-line retailer to improve its fulfillment system's operations. We observed that there is a difference between a customer's order date and the customer's desired fulfillment date. Recognizing this difference changes the way inventories are managed and orders are fulfilled. By recognizing this difference, inventories and warehouse space were reduced significantly.

We will do two things in this paper. First, we will present a model and a computationally tractable method for planning procurements and allocating stocks among warehouses. Our model is based on taking advantage of the difference between the time a customer places an order and when it must be fulfilled. Second, we demonstrate the effect of increasing the percentage of demand that must be filled immediately on the total system inventories. This effect is important to understand since the on-line retailer is increasingly encouraging customers to request rapid fulfillment of their orders.

II. Background

We will now discuss some of the key attributes that exist in the on-line retailer's operating environment. The on-line retailer now operates a multi-echelon fulfillment system structured as follows. Each item is managed through a single warehouse, called the *primary warehouse* for the item. Each item has a single primary warehouse. Each warehouse serves as a primary warehouse for a collection of items. The choice of the primary warehouse for an item depends largely on the supplier's location. Balancing warehouse workloads and recognizing facility capacities also affects the number of items managed by each warehouse. Items normally ordered together also have a common primary warehouse. Once a primary warehouse is selected for an item, that warehouse becomes responsible for procuring and receiving inventory from an external supplier. Suppliers ship items only to the item's designated primary warehouse. The primary warehouse then distributes inventory to the other warehouses, which we call *regional warehouses*, on an as needed basis. We refer to this system as the primary warehouse system (PWS). Every warehouse in the PWS serves as a primary warehouse for hundreds of thousands of items and a regional warehouse for an even larger number of items.

Each primary warehouse is conceptually thought of as two entities, although there is only one physical entity. One entity performs the procurement, warehousing and allocation tasks associated with the primary warehouse. The other entity is responsible for satisfying demand that arises in the geographical region in which the warehouse is located. We call this second entity, which is a virtual warehouse, the co-located warehouse. Since this virtual co-located warehouse is physically located in the same facility as the primary warehouse, shipping to it is assumed to occur instantaneously.

Within the PWS, orders are planned to be satisfied from the regional warehouse closest to the customer's delivery address to minimize last mile delivery costs, which, as we mentioned, are a substantial component of the retailer's operating cost. Recall that when customers place orders, they also request a delivery response time. Customers may request and possibly pay extra for immediate delivery or they may request a later delivery date to reduce or waive a delivery fee. We categorize demands by their response lead time. Suppose a customer places an order that will be fulfilled from a particular regional

978-1-4799-4500-9/14/$31.00 ©2014 IEEE

warehouse. If the delivery response time specified by the customer is less than the sum of the shipping lead time from the primary warehouse to that regional warehouse and the time to ship to the customer from that warehouse, we call this a *short-response lead time demand*. Otherwise, we call it a *long-response lead time demand*.

In principle, the only reason to stock inventory for an item at a regional warehouse is to satisfy short-response lead time demand arising in nearby region. The inventory to fulfill long-response lead time demand occurred anywhere in the system is stocked at the primary warehouse. When a long-response lead time demand occurs, the primary warehouse sends inventory to the regional warehouse that is responsible for satisfying the customer's order. The inventory will be cross-docked at the regional warehouse and sent directly to the customer. The primary warehouse must also carry inventory so that it can replenish each regional warehouse's inventory. We note that for the on-line retailer we worked with, short-response lead time demand usually accounted for between $13\% - 20\%$ of the total demand for an item, and the percentage varied by item and sometimes by the time of the year. Recently, the percentage has been increasing. As mentioned a key question we are addressing in this paper is the following: how does the total inventory requirement change as this percentage increases?

There are two distinct but interrelated decisions related to setting stock levels and fulfilling customer orders. First, procurement decisions are made on an item basis. These decisions take into account fixed ordering costs, quantity discounts, holding and shortage costs, workload smoothing, the multi-echelon nature of the PWS, and, of course, the uncertainty and timing aspects of the demand process throughout a planning horizon. The model developed in this paper is focused on this planning problem. Second, fulfillment decisions are made on an order-by-order basis. Once a customer order is received, a decision needs to be made about how best to fulfill it. Determining how existing inventories should be best allocated for either replenishment or cross-docking from warehouses to fulfill a customer's order in a timely and cost effective manner is the topic of another paper [1].

The inventory levels or targets set in this planning process guide procurement decisions for individual items over time. When establishing inventory targets, we do not consider the composition of customer orders for two main reasons. First, no single customer order accounts for more than a small fraction of 1% of the demand for the items of interest. Second, the procurement lead times are normally weeks to months in length, whereas fulfilling orders must be accomplished in time intervals measured in days. Forecasting customer orders in terms of content and timing weeks or months in advance of their placement has not proven to be possible. Hence, we have separated the decision making into two segments, one for planning the target inventory levels and the other for fulfilling orders. In the remainder of this paper, we will focus our discussion on the planning model.

The organization of the remainder of the paper is as follows. In Section III, we briefly discuss literature related to our efforts. In Section IV we present an exact planning model for setting stock levels for each item at each warehouse. This model is formulated as a dynamic program that cannot be solved directly. In Section IV-B we develop an approximation approach for determining stock levels that is scalable and applicable to the planning activity in the environment that motivated this research. Numerical experiment is discussed in Section V. In Section VI we provide concluding remarks.

III. Literature Review

There are three streams of literature related to our paper's content. We will discuss them separately.

Recall we propose to operate the system with the PWS approach where we stock the inventories for long-response lead time items at the primary warehouse and the short-response lead time items at the regional warehouses. This idea is related to material found in Muckstadt et al. [2]. Our approach is similar to their "No B/C strategy" in the sense that the long-lead time demand can be viewed as their B/C type items, which are not stocked at the regional warehouses in general. The short-response lead time demand can be viewed as their A type items, which are stocked at all regional warehouses. In this paper, long-response lead time demand and short-response lead time demand may overlap. Hence, when the primary warehouse is not able to satisfy a long response lead time demand, the regional warehouse may use its on-hand inventory to satisfy the demand.

We assume the primary warehouse places orders from an external supplier under a fixed schedule. Eppen and Schrage [3] do as well. The primary focus in their paper is to demonstrate how risk and inventory requirements are reduced by operating a two echelon distribution system in a certain manner. In their system, the central warehouse places an order every period under a base stock policy and no stock is held at the central warehouse. In our system, this is not the case. Jackson [4] built both an exact cost model and a computationally tractable approximation cost model to find a ship-up-to-S allocation policy for a cyclic system that serves N warehouses. In each cycle, the central warehouse allocates inventories to each regional warehouse periodically. Jackson and Muckstadt [5] extend this idea to a two-echelon, two-period allocation problem.

Our multi-echelon inventory model is based on Clark and Scarf[6]'s echelon inventory position concept. They found that solving a distribution system type of problem is difficult due to the possible "imbalance" of inventories among the regional warehouses. Such systems are "balanced" if there is no desire to redistribute the inventories among the regional warehouses whenever inventories are allocated to regional warehouses. Clark and Scarf found that under their balance assumption that the systems can be decomposed into individual location problems which can be optimized separately. However, this balance assumption may not necessarily hold. When it is violated, the optimal allocation strategy could be to allocate negative quantities to a location. Federgruen and Zipkin [7] obtain a lower bound on a value function by relaxing the imbalance constraints. Kunnumkal and Topaloglu [8] and [9] associated Lagrangian multipliers with the balance constraints. By introducing the Lagrangian multipliers, the resulting relaxed problem can be easily solved. They discussed several approximation methods that can be used to select a good set of multiplier values. They also show that Fedegruen and Zipkin's approach is equivalent to setting the multiplier values to zero.

Hence their approach permits them to obtain tighter lower bounds on the optimal objective function value than the value achieved using Federgruen and Zipkin's methodology.

As mentioned earlier, the system provides two levels of service, which are the short-response lead time demand and long-response lead time demand. The long-response lead time demand can be thought of as having advanced demand information. The effect of the advance demand information is examined by Hariharan and Zipkin [10] in a continuous-review framework. In many papers, one of two assumptions is made. Either the advance demand must be satisfied on a fixed schedule or the system has the flexibility to satisfy the demand within some amount of time. Gallego and Özer [11] and Özer [12] make the first type of assumption and conclude that state-dependent (s, S) and base-stock policies are optimal for stochastic inventory systems having different cost structures. Wang and Toktay [13] extend this conclusion to the flexible delivery case.

IV. The Planning Model

Before we present our model, we first discuss attributes of the demand process experienced by the on-line retailer. Although there are millions of items available for purchase from the on-line retailer, most items have very low demand rates. In fact, most items have four or fewer units demanded annually. These items are sometimes stocked in a single warehouse operated by the on-line retailer. However, in most cases, these items are not stocked by the on-line retailer at all but rather in some other company's warehouse. Other items may be ordered only a few times per year; but, many units may be requested in a single order. Some textbooks are examples of such items. These items are also normally stocked in a single location. For the on-line retailer we worked with, low demand items account for over 70% of the items offered in the system.

There is another type of item, the very high demand rate items. For the on-line retailer we studied, under 2% of the items account for about 30% of the system's total sales, measured in units and in monetary terms. We will not focus on either the very low or very high demand rate items in this paper. Rather, we will focus on the roughly 27% of the items that are relatively high demand rate items and that are stocked in the multiple-warehouse system operated by the on-line retailer.

The mean and variance of the demand process varies over time for a large portion of these higher demand rate items. For many items, most of their demand occurs from mid-November through the end of December. For others, spikes in demand occur according to school calendars or perhaps due to the launch of the item into the market.

We now develop a planning model designed to establish inventory levels for the relatively high demand rate items. We first construct an exact model that cannot be used to solve realistic problems due to the size of the state space. We then construct an approximation model and a computationally tractable solution method. We begin our model development by stating our assumptions concerning the fulfillment system's operation and by introducing some nomenclature. Additional nomenclature is presented as we proceed.

We assume that decisions for each item are made on a periodic basis. Thus the model we develop is a periodic review model, a period being a day in length. Each day at each primary warehouse, two decisions must be made for each item that is managed there. The first is a procurement decision and the second is an allocation decision.

To manage both costs and workloads, for planning purposes, procurement orders for an item are placed according to a schedule. That is, we assume that for each item there is a pre-determined set of times at which orders are placed on a supplier. We call the time between the placing of successive procurement orders a cycle length. Items are ordered by the retailer using a power of two type like policy, that is, orders are placed weekly, bi-weekly, monthly, or quarterly. In practice, the exact timing of the placing of orders is done to smooth buyer and warehouse workloads. The frequency of placing orders largely depends on the economics of ordering. Determining the schedule for placing procurement orders for each item is in itself an interesting problem. Since this topic is not the focus of this paper, we assume in the following sections that the cycle lengths are known and the timing of procurement actions has been established for each item.

The second decision made each day is to determine how much inventory to allocate to each regional warehouse from the primary warehouse. The allocation decisions for an item depend on demand forecasts, costs, inventory availability at the primary warehouse, and the inventory positions at the regional warehouses.

The entire fulfillment system can be analyzed one primary warehouse at a time since there are no constraints that link decisions made for one primary warehouse to another primary warehouse. Thus our model focuses on a system consisting of one primary warehouse that manages inventories for several hundred thousand items.

Suppose there are N regional warehouses in the PWS including the co-located warehouse. Let us denote the primary warehouse as location 0. Regional warehouses numbered 1 through $N-1$ correspond to those that are not co-located with the primary warehouse. Regional warehouse N is the one co-located with the primary warehouse. Let I denote the set of items managed in the PWS.

In every period, two types of demands may arise for an item, short-response lead time and long-response lead time demands. We define $D_{it}^{n,\alpha}$ and $D_{it}^{n,\beta}$ to be random variables for the short (α) and long (β) response lead time demand at regional warehouse i in period t for item n, respectively. We also define $d_{it}^{n,\alpha}$ and $d_{it}^{n,\beta}$ to be the realizations of these random variables, and $d_{it}^{n} = d_{it}^{n,\alpha} + d_{it}^{n,\beta}$. Also, let $D_{it}^{\alpha} = \sum_{n} D_{it}^{n,\alpha}$, $D_{it}^{\beta} = \sum_{n} D_{it}^{n,\beta}$ and $D_{it} = D_{it}^{\alpha} + D_{it}^{\beta}$.

There are two types of lead times that exist in the system. The first is the customer response lead time and the second is the nominal supplier lead time or order replenishment lead time. If the system cannot respond to a customer's request in a timely manner, backorder costs will be incurred. We assume that all customer demands that have a required response lead time less than or equal to L_i^{α} periods can be satisfied at a low fulfillment cost only from stock located at regional warehouse i. These are the short-response lead time demands. When a customer's expectation for delivery is greater than L_i^α periods, the order can be satisfied at a low fulfillment cost by shipping from stock held in the primary warehouse. These are the long-response lead time demands. In this case, L_i^β measures the time following the receipt of a customer order by which the shipment must leave the regional warehouse.

Suppose the shipping lead time from the primary warehouse to regional warehouse i is L_i periods. Thus $L_i^\alpha < L_i \leq L_i^\beta$. When $L_i^\alpha = 0$, the customer order must be shipped immediately upon its receipt. Let $l_i = L_i^\beta - L_i$, the slack time between the long-response customer time window and the transportation time. We call l_i the grace period. If $l_i = 0$, the primary warehouse must immediately ship the item to the regional warehouse where it will be cross-docked and shipped to the customer. When $L_i^\alpha = 0$ and $l_i = 0$, we call this the immediate response time case. When $l_i > 0$, there is greater flexibility in the timing of shipments to the regional warehouse. This helps smooth workloads, as we discuss in the final section of the paper.

Our final assumption pertains to the sequence in which events occur in each period, which we assume occur as follows for each item. First, we observe the echelon inventory positions at all locations. Second, when appropriate, we receive a replenishment order at the primary warehouse corresponding to an order placed a procurement lead time ago. Third, we observe the demands at all regional warehouses. Fourth, when appropriate, the primary warehouse places a replenishment order on an external supplier. Fifth, based on availability, and the inventory positions at the regional warehouses, we allocate inventory on-hand at the primary warehouse to the regional warehouses. Sixth, we receive stocks at the regional warehouses that were shipped a lead time ago from a primary warehouse. These stocks can be used to satisfy the current period's short response lead time demand and the long-response lead time orders that were received from customers a transportation lead time or more ago. Seventh, we backlog the unsatisfied demands at the regional warehouses. At the end of each period, holding costs are charged based on on-hand inventories at all warehouses. Backorder costs are charged only at regional warehouses at each period's end.

A. *An Exact Model*

We now construct an exact model for determining stock levels for the fulfillment system based on the assumptions we have made. Suppose the planning horizon over which we are placing procurement orders and making inventory allocation decisions for each item is T days in length. During this horizon, there is a set of periods in which item n is permitted to be procured from the external supplier. Let P_n denote the set of days for item n. Thus q_{0t}^n, the amount ordered from the external supplier for item n in period t, can be positive only if $t \in P_n$. To simplify our discussion, we assume procurement orders can be placed only every τ^n periods for item n, which is the cycle length for item n. In reality, the cycle lengths can and do vary over time. As you will observe, this assumption can be relaxed without impacting the modeling approach we present subsequently.

Let x_{it}^n represent the echelon inventory position for item n at location i, $i \in \{0, \cdots, N\}$, at the beginning of period t before any inventory has arrived to i or has been shipped from it. Let q_{it}^n be the amount allocated from on-hand stock of item n at the primary warehouse to regional warehouse i, $i \in \{1, \cdots, N\}$, in period t. Then at regional warehouse i, $x_{i,t+1}^n = x_{it}^n + q_{it}^n - d_{it}^n$. Let $y_{it}^n = x_{it}^n + q_{it}^n$. Thus, y_{it}^n represents the echelon inventory position for regional warehouse i after the allocation is made to it but before satisfying demand in period t. Furthermore, the echelon net inventory for the primary warehouse system at the end of period $t + L_0^n$ for item n is

$$x_{0t}^n + q_{0t}^n - \sum_{k=t}^{t+L_0^n} \sum_{i=1}^{N} d_{ik}^n, \tag{1}$$

when $t \in P_n$, and L_0^n is the replenishment lead time for item n. Also, let $y_{0t}^n = x_{0t}^n + q_{0t}^n$.

There are two types of costs considered in our model, holding and backorder costs. Let h_i^n denote the per unit installation holding cost for item n at location i charged at the end of a period. Let b^n be the backorder cost for a unit of item n at a regional warehouse at the end of a period. Thus we assume the backorder cost for item n is the same across regional warehouses and time. This assumption is required to maintain convexity of the problem's formulation.

Next, we formulate the decision problem as a dynamic program. We begin by showing how to calculate the costs associated with holding inventories and incurring backorders at the end of a period.

1) Cost At A Regional Warehouse: Expected holding and backorder costs are charged in each period at each regional warehouse. An allocation of q_{it}^n units to regional warehouse i in period t results in expected costs being incurred at the end of period $t + L_i$. The net inventory at the end of period $t + L_i$ at regional warehouse i is

$$\begin{aligned} x_{it}^n + q_{it}^n - \sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} - d_{it}^{n,\alpha} - d_{it}^{n,\beta} \\ = y_{it}^n - \sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} - d_{it}^{n,\alpha} - d_{it}^{n,\beta}, \end{aligned} \tag{2}$$

when $l_i = 0$. Note we know the demand that occurred in period t prior to making the allocation decision, that is, we know $d_{it}^{n,\alpha}$ and $d_{it}^{n,\beta}$. However, the short-response lead time demands in periods $t + 1$ though $t + L_i$ are unknown at that time. The resulting expected holding and backorder costs incurred as a consequence of allocating q_{it}^n units to regional warehouse i in period t for item n, that is, ordering up to y_{it}^n, are

$$\mathbf{E}[h_i^n(y_{it}^n - \sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} - d_{it}^n)^+ + b^n(\sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} + d_{it}^n - y_{it}^n)^+], \tag{3}$$

where the expectation is taken over the short-response lead time demand random variables for periods $t+1$ through $t+L_i$. Since no inventory is held in regional warehouse N, this co-located warehouse will incur only expected backorder costs.

2) Holding Costs At the Primary Warehouse: Consider the immediate response case, that is, $L_i^{\alpha} = 0$ and $L_i^{\beta} = L_i$. In this case, backorder costs are charged against any short-response lead time demand that is unfilled at the end of a period and against any long-response lead time demand that is unfulfilled at a regional warehouse a shipping lead time L_i in the future. Since no customer demands are satisfied directly from the primary warehouse inventory, only holding costs are incurred there.

Suppose $t' \in P_n$. The echelon inventory position at the beginning of period $t' + L_0^n$ is

$$x^n_{0,t'+L_0^n} = y^n_{0,t'} - \sum_{i=1}^{N} \sum_{k=t'}^{t'+L_0^n-1} D^n_{i,k}. \tag{4}$$

For $t \in \{t' + L_0^n, \cdots, t' + L_0^n + \tau^n - 1\}$, the net inventory at the primary warehouse at the end of period t is

$$x^n_{0,t'+L_0^n} - \left\{ \sum_{i=1}^{N} [x^n_{i,t'+L_0^n} + \sum_{k=t'+L_0^n}^{t} q^n_{i,k}] \right\} \tag{5}$$

$$= y^n_{0,t'} - \sum_{k=t'}^{t'+L_0^n-1} D^n_k - \left\{ \sum_{i=1}^{N} [x^n_{i,t'+L_0^n} + \sum_{k=t'+L_0^n}^{t} q^n_{i,k}] \right\}. \tag{6}$$

Consequently, the expected holding cost incurred at the primary warehouse at the end of period t, $t \in \{t' + L_0^n, \cdots, t' + L_0^n + \tau^n - 1\}$, is

$$\mathbf{E}\left[h_0^n \left\{ (y^n_{0t'} - \sum_{k=t'}^{t'+L_0^n-1} D^n_k) - \sum_{i=1}^{N} [x^n_{i,t'+L_0^n} + \sum_{k=t'+L_0^n}^{t} q^n_{i,k}] \right\} \right] \tag{7}$$

$$= \mathbf{E}\left[h_0^n \left\{ (y^n_{0t'} - \sum_{k=t'}^{t} D^n_k) - \sum_{i=1}^{N} y^n_{it} \right\} \right]. \tag{8}$$

3) The Objective Function: The expected cost functions given in Equations (3) and (7) provide the basis for making procurement and allocation decisions. However, we do not use them directly in our decision model. Rather, we define another but equivalent set of functions, which are of the type introduced by Clark and Scarf [6] in their seminal paper and later used, for example, by Kunnumkal and Topaloglu [8], [9].

Let us first focus on each regional warehouse i, $i = 1, \cdots, N-1$, for item n. Define

$$G^n_{it}(y^n_{it}) = -h_0^n y^n_{it} + \mathbf{E}[h^n_{i,t+L_i}(y^n_{it} - \sum_{k=t+1}^{t+L_i} D^{n,\alpha}_{ik} - d^n_{it})^+] + \mathbf{E}[b^n(\sum_{k=t+1}^{t+L_i} D^{n,\alpha}_{ik} + d^n_{it} - y^n_{it})^+]. \tag{9}$$

At the beginning of time period t, we do not know the values of d^n_{it}, but we do know these values when making the allocation decision later in that period. Thus $G^n_{it}(y_{it})$ reflects the knowledge we have when making the allocation decision in period t.

For regional warehouse N,

$$G^n_{Nt}(y^n_{Nt}) = -h_0^n y^n_{Nt} + [b^n(d^n_{Nt} - y^n_{Nt})^+], \tag{10}$$

since we know d^n_{Nt} when making the allocation of q^n_{Nt} units.

Let us next turn to the primary warehouse. Let

$$G^n_{0t}(y^n_{0t}) = h_0^n y^n_{0t'}, \qquad t \in \{t' + L_0^n, \cdots, t' + L_0^n + \tau^n - 1\}, \tag{11}$$

where $y^n_{0t'}$ is the system echelon inventory position corresponding to the procurement order placed at time t'.

Observe that $G^n_{0t}(y^n_{0t}) + \sum_{i=1}^{N} G^n_{it}(y^n_{it})$ yields, except for a constant, the same period t expected costs for item n as would result from using expressions (3) and (7) to compute these costs. Thus the total expected period t cost function used in our model is

$$\sum_{n} \{G^n_{0t}(y^n_{0t}) + \sum_{i=1}^{N} G^n_{it}(y^n_{it})\}. \tag{12}$$

Expressing our cost function in this manner permits us to solve the problem more efficiently as we will observe subsequently and as was observed in [6], [8], [9] and by others.

4) A Dynamic Programming Formulation of the Planning Problem: We now construct a dynamic programming recursion that could be employed, at least theoretically, to determine the optimal procurement and allocation decisions over the T period planning horizon.

Let $V_t(\bar{x}_t)$ be the expected minimum cost over periods t through T, that could be achieved given that the system is in state $\bar{x}_t$ at the beginning of period t, where $\bar{x}_t$ is the vector of the x^n_{it} values at that time.

Three types of constraints exist when making decisions in each period. First, there is a logical constraint that implies that the quantity of an item allocated to each regional warehouse in each period cannot be negative. That is $q^n_{it} \geq 0$. This is the balance constraint. Second, procurement of each item n can take place only in period $t \in P_n$. Third, we cannot ship more than is on hand at the primary warehouse for any item in any period.

Combining the results obtained in the previous section with these constraints, we can now express $V_t(\bar{x}_t)$. Let $\bar{D}_t$ be the vector of random variables for demands arising in period t for all items at all regional warehouses and $\bar{y}_t$ be the vector of order-up-to levels, y^n_{it}. The following dynamic program is an example when $L_0^n = 0$.

$$V_t(\bar{x}_t) = \mathbf{E}_{D_t} \{ \min \sum_n \{ G_{0t}^n(y_{0t}^n) + \sum_{i=1}^N G_{it}^n(y_{it}^n) \} + V_{t+1}(\bar{y}_t - \bar{D}_t) : \quad (13)$$

$$\text{s.t.} \sum_i^N y_{it}^n \leq y_{0t}^n, \quad \forall n, \quad (14)$$

$$y_{it}^n \geq x_{it}^n, \quad \forall n, i, \quad (15)$$

$$y_{0t}^n \geq x_{0,t}^n, \quad t \in P_n, \forall n, \quad (16)$$

$$y_{0t}^n = x_{0,t}^n, \quad t \notin P_n, \forall n \ \} . \quad (17)$$

Note that we do not depend on the assumption that $L_0^n = 0$ in what follows.

Observe that this problem is separable by item. However, the size of the state space corresponding to this dynamic programming formulation for each item is still is too large to make it a useful practical approach for setting stock levels. Hence we now discuss an approximation approach for computing recommended stock levels.

B. An Approximation Approach

We begin our development of an approximate model and method for obtaining a solution to a single item problem by making some observations and assumptions.

Demand from cycle-to-cycle may be non-stationary. However, we assume the demand in any cycle is large enough so that the system echelon inventory position at the time an order is placed is not greater than the one that is desired. That is, a positive quantity will always be ordered ($q_{0t}^n > 0$, $t \in P_n$). This is virtually always the case in the environment we studied for the type of items we are considering. The experiments discussed in Section V support this assumption. By assuming so, we are able to formulate the problem as a sequence of independent problems, one for each cycle. This myopic, cycle-based approach for setting stock levels will yield optimal order-up-to levels throughout the planning horizon.

Another observation pertains to the holding and backorder costs. Since the per period backorder costs are high relative to the holding costs (over a 100 to 1 for many items), inventory levels are high enough to ensure that backorders occur only infrequently. This observation leads to the following two assumptions. First, we assume that when a procurement order arrives it is possible to make allocations so that all regional warehouses achieve their desired stock level. Second, we assume that the balance assumption will be satisfied without explicitly considering constraints which enforce the balancing of inventories. That is, in our approach we assume q_{it}^n could be negative or equivalently y_{it}^n could be less than x_{it}^n. We make this assumption for two reasons.

First, recall that inventory is held at a regional warehouse to satisfy short-response lead time demand. Recall also that these demands have historically accounted for less than 20% of the total demand each day for an item. Since most of the demand is satisfied from stock held at the primary warehouse, most of the inventory is held there. As the percentage of demand that is short-response lead time demand increases, we test the validity of the balance assumption in Section V.

Second, a cycle is always a week or longer, demand rates for items managed using the PWS are high and relatively stable from period-to-period, and our data indicate that short-response lead time demand tends to have low variance to mean ratios. Consequently, stock imbalance will likely occur, if at all, only at the end of a cycle. For example, if a cycle is a week in duration, imbalance may occur on the last day of the cycle, but is very unlikely to occur prior to that day.

1) An Approximation Model for a Single Item: Since we assume that decisions made in one cycle do not affect those made in other cycles, we will model a single cycle for some item. We drop the item identifier from our notation in this section since we focus on a single item. For ease of exposition we again assume $l_i = 0$. The effect of $l_i > 0$ on the system's performance will be addressed subsequently.

Suppose the item is ordered at time 0 and its echelon inventory position for the primary warehouse, or system, is raised to y_{00} units. The amount ordered arrives in period L_0 at which time the system echelon inventory is $x_{0,L_0} = y_{00} - \sum_{k=0}^{L_0-1} d_k$ units. Thus $P[x_{0,L_0} = w] = P[\sum_{k=0}^{L_0-1} D_k = y_{00} - w]$.

Suppose the echelon inventory position at the beginning of period L_0 is x_{0,L_0} units. By assumption, we need not consider the possibility that some of these units are not on hand at the primary warehouse at the beginning of period L_0, that is, $q_{i,L_0} \geq 0$ with probability one. Since we know d_{i,L_0} prior to making the allocation to regional warehouse i, q_{i,L_0} will depend on d_{i,L_0}.

Let y_{it} represent the order-up-to level for regional warehouse i following the allocation decision made in period t. Consequently, the expected cost incurred at regional warehouse i in period $t + L_i$ is

$$h_i \mathbf{E}[y_{it} - d_{it} - \sum_{k=t+1}^{t+L_i} D_{ik}^\alpha]^+ + b\mathbf{E}[d_{it} + \sum_{k=t+1}^{t+L_i} D_{ik}^\alpha - y_{it}]^+, \quad (18)$$

where $t \in \{L_0, \cdots, L_0 + \tau - 1\}$. Remember d_{it} includes the long-response lead time demand that arises in period t, which must be shipped to the customer in period $t + L_i$.

Similarly, the expected holding cost charged at the primary warehouse at the end of period t is $\mathbf{E}[h_0[y_{00} - \sum_i y_{it} - \sum_{k=0}^t D_k]]$ since y_{it} can be negative.

Let $G_0(y_{0t}) = h_0 y_{00}$ and $G_{it}(y_{it}) = -h_0 y_{it} + h_i \mathbf{E}[y_{it} - d_{it} - \sum_{k=t+1}^{t+L_i} D_{ik}^\alpha]^+ + b\mathbf{E}[d_{it} + \sum_{k=t+1}^{t+L_i} D_{ik}^\alpha - y_{it}]^+$.

Then $G_0(y_{0t}) + \sum_{i=1}^N G_{it}(y_{it})$ is, within a constant, the expected cost incurred resulting from the allocation decisions made in period t, $t \in \{L_0, \cdots, L_0 + \tau - 1\}$ and the procurement decision made in period 0.

Let x_{0t} be the system echelon inventory position at the beginning of period t, $t \in \{L_0, \cdots, L_0 + \tau - 1\}$, resulting only from a procurement decision made in period 0. We now construct a scalable dynamic programming recursion that can be used to find the values of y_{it} and ultimately y_{00}. Recall that we have relaxed the balance constraints. The recursion we use has a single dimensional state space and is defined as follows.

$$\begin{aligned} V_t(x_{0t}) =& \mathbf{E}_{D_t}\, [\, \min\{G_0(y_{0t}) + \sum_{i=1}^{N} G_{it}(y_{it}) + V_{t+1}(x_{0t} - D_t)\} : \\ & \sum_i y_{it} \leq x_{0t}\,] \qquad (19) \\ =& G_0(y_{0t}) + E[\min \sum_{i=1}^{N} G_{it}(y_{it}) + V_{t+1}(x_{0t} - D_t) : \\ & \sum_i y_{it} \leq x_{0t}]. \qquad (20) \end{aligned}$$

Once $V_{L_0}(x_{0,L_0})$ is computed for a range of values for x_{0,L_0}, we compute the expected value $\sum V_{L_0}(x_{0,L_0}) \cdot P[\sum_{k=0}^{L_0-1} D_k = y_{00} - x_{0,L_0}]$, which we call $F(y_{00})$. It is easy to show that $F(\cdot)$ is a convex function of y_{00} and hence it is easy to determine the optimal value of y_{00} using a line search. Since we determine the order-up-to levels for each item independent of the other items, the calculations can be executed in a parallel manner.

V. Computational Results

Our primary goal is to develop a computationally tractable method for determining stock levels for items managed at a primary warehouse. Using the approach described in the preceding section, we computed stock levels for approximately $250,000$ items for a 15 month time horizon in a system consisting of 5 regional warehouses and a primary warehouse. This is the system operated by the on-line retailer at the time we examined it. Calculating these stock levels required approximated 9.8 minutes on a single PC with an Intel Xeon Processor E5520 (2.26HHz). If calculations were performed using a parallel computing architecture and a faster processor, these stock levels could be established in well less than a minute. Thus we have demonstrated that the approach we have developed is scalable and appropriate for use in the planning process for the on-line retailer environment we examined.

Our model and algorithm were based on the assumption that stock imbalance among regional warehouses occurs rarely. Through our extensive numerical experiments, which we will now summarize, we validate the assumption that the imbalance constraints can be safely ignored in our model. We also show that as the percentage of demand that is short-response lead time demand increases, inventory requirements increase substantially.

A. Experimental Design

To test our imbalance assumption, we simulated the operation of the on-line retailer's 5 regional warehouse system. For each item, we simulated 600 cycles of operation where the cycle length ranged from one to five weeks. One primary warehouse was responsible for managing $234,564$ relatively high demand rate items in the experiment. This is approximately the number of such items managed within each of the five primary warehouses.

In the simulation experiment, we assumed that daily demand is negatively binomially distributed for each item. This assumption is based on evaluating the distribution of forecast errors in the real application. We scaled costs and demand rates to protect their true values. After scaling, daily demand rates ranged from 0.5 to 120 units per day and unit costs ranged from \$4 to \$120. Also, demand variance-to-mean ratios ranged from 1.01 to 4.10. Procurement lead times ranged from one to five weeks in length.

Shipping lead times from the primary warehouse to the regional warehouses numbered one through five were 1, 2, 3, 1 and 0 days, respectively. Holding costs were charged at a rate of 20% of the product cost on a annual basis. Thus, for example, the cost to hold an item costing \$50 for a year is \$10 or approximately \$0.0274 per day.

Recall that backorder costs are high and are most often in excess of 25 times the daily holding cost rates. For lower cost items, these costs are greater than 100 times higher. In our experiments we conservatively assumed the backorder costs were 25 times greater than the holding costs. Since the backorder costs are normally higher in practice, our computed inventory levels would be lower bounds on their true values. Thus, the results of our simulation experiments will provide an upper bound on the possibility of an imbalance situation occurring.

We began the experiment by grouping items according to their cost, demand rate, variance-to-mean ratios and procurement lead times. We found that items could be placed into 10 unit cost, 10 demand, 4 variance-to-mean ratio, and 5 procurement lead time categories, or $10 \times 10 \times 4 \times 5$ possible categories. Of these 2000 possible categories, 739 categories contained at least one item. The categories containing the most items consisted of about 3% of the items. The category containing the highest total demand had 3.1% of the items in it. We also rank the category by demand rate. The largest category accounts for about 2.3% of total demand.

For each of the 739 categories, we defined a representative item. To do so, we calculated the average unit cost, cycle length, variance-to-mean ratio, procurement lead time and demand rate over all items in the category. This representative item was simulated to obtain investment and imbalance estimates for all items within that category.

After simulating each of the 739 representative items for 600 cycles, we estimated the system inventory holding cost by multiplying the average annual holding cost for reach representative item by the number of items in the category. The simulation experiments were conducted for each of four scenarios. The percentage of the total demand that was short-response lead time demand varied in each scenario and was set at 10%, 20%, 30% or 80%.

When we first worked with the on-line retailer, inventories of all items were managed at each the five warehouses. That is, each warehouse placed and received procurement orders for each item. To illustrate the effect of adopting the PWS, we also simulated the system's performance when each warehouse

TABLE I. PERCENTAGE OF PERIODS IN WHICH IMBALANCE OCCURS

% Short Response Lead Time Demand	% Periods in Which Imbalance Occurs
10%	0.015%
20%	0.018%
30%	0.018%
80%	0.024%

TABLE II. ANNUAL HOLDING AND BACKORDER COSTS

%Short Response Lead Time Demand	Average Annual Cost
10%	$\$86,344,071$
20%	$\$92,121,685$
30%	$\$96,317,015$
80%	$\$110,281,815$

manages all items. In this system, all demands for all items that arise in a geographical region are fulfilled by the warehouse in that region. The fulfillment system has longer cycle lengths and larger saftey stocks since demand for each item was distributed among the five warehouses. We demonstrate how much the inventory costs were lowered by switching to the PWS architecture.

B. Experiment Results

The first question that we address pertains to our imbalance assumption. That is, does imbalance exist to an extent that invalidates our assumption that we can ignore imbalance constraints in our model's formulation? And how does the extent of demand imbalance change as we increase the percentage of demand that is short-response lead time demand? The data displayed in Table I indicate that our assumption pertaining to the imbalance constraint is valid. While insignificant across scenarios, we observe that the percentage of periods experiencing imbalance increases by 60% when the percentage of short-response lead time demand increases from 10% to 80%. This result is as we would expect since a greater portion of system stock is located at the regional warehouses as the percentage of short-response lead time demand increases. Thus, for the reasons discussed earlier, the on-line retailer can employ our model without being concerned about imbalance becoming an issue.

Let us now ask another important question: how do inventory holding and backorder costs change as the percentage of demand that is short-response lead time demand increases? The data displayed in Table II show that these costs are highly sensitive to this percentage. Thus average cost increases by about 31.5% as the percentage of short-response lead time demand increases from 10% to 80%. Since profit margins are low, typically well less than 1% of sales, the increased costs associated with carrying inventories could result in an operation loss.

Suppose the five warehouses operated independently and had to meet all demands in their region using only their inventories. For one of the warehouses the expected annual inventory costs would be approximately $392 million. This is over four times greater than the annual cost of operating a primary warehouse when short-response lead time demand is about 20% of the total demand. This reduction in operating costs was another key reason for adopting the PWS architecture.

Recall that the on-line retailer had about 20% of short-response lead time demand several years ago when we worked with it. In the past few years, they have encouraged customers to place short-response lead time demand by selling memberships. Qualified customers would get short-response lead time shipping with no additional cost for any amount of purchase. From our analysis, this would increase the inventory holding and backorder costs significantly. We have seen that the on-line retailer made several changes to control this cost. First, they increased the subscription fee of their membership. Second, they are offering discounts when members are willing to take the long-response lead time option. Third, inexpensive items can no longer be ordered alone to qualify for expedited shipping.

VI. CONCLUSION

In this paper we present and illustrate a computationally tractable approach for setting stock levels at a multi-echelon on-line fulfillment system. Our numerical experiments demonstrate the cost impact as the percentage of demand that is short-response lead time demand increases. The operators of on-line retailers with low profit margins need to take serious considerations of the inventory holding and backorder costs when pricing their fulfillment options.

REFERENCES

[1] J. Li and J. Muckstadt, *Fulfilling Orders in a Multi-Echelon Capacitated On-line Retail System: PART TWO, real-time purchasing and fulfillment decision making*, School of Operations Research and Information Engineering, Cornell University 14850, 2013, technical report, No. 1482.

[2] J. Muckstadt, D. Murray, and J. Rappold, *Capacitated Production Planning and Inventory Control when Demand is Unpredictable for Most Items: The No B/C Strategy*, School of Operations Research and Industrial Engineering, Cornell University 14850, 2001, technical report, No. 1306.

[3] G. Eppen and L. Schrage, “Centralized ordering policies in a multi-warehouse system with lead times and random demand,” *Management Science*, vol. 30, pp. 69–84, 1981.

[4] P. Jackson, “‘stock allocation in a two-echelon distribution system or’ what to do until your ship comes in,” *Management Science*, vol. 34, pp. 880–895, 1988.

[5] P. Jackson and J. Muckstadt, “Risk pooling in a two-period, two-echelon inventory stocking and allocation problem,” *Naval Research Logistics*, vol. 31, no. 1, pp. 1–26, 1989.

[6] A. Clark and H. Scarf, “Optimal policies for a multi-echelon inventory problem,” *Management Science*, vol. 6, pp. 475–490, 1960.

[7] A. Federgruen and P. Zipkin, “Approximations of dynamic, multiocation production and inventory programs,” *Management Science*, vol. 30, no. 1, pp. 69–84, 1984.

[8] S. Kunnumkal and H. Topaloglu, “A duality-based relaxation and decomposition approach for inventory distribution systems,” *Naval Research Logistics Quarterly*, vol. 55, no. 7, pp. 612–631, 2008.

[9] ——, “Linear programming based decomposition methods for inventory distribution systems,” *European Journal of Operational Research*, vol. 211, no. 2, pp. 282–297, 2011.

[10] R. Harihar and P. Zipkin, “Customer-order information, leadtimes, and inventories,” *Management Science*, vol. 41, pp. 1599–1607, 1995.

[11] G. Gallego and Ö. Özer, “Optimal replenishment policies for multiechelon inventory problems under advance demand information,” *Manufacturing and Service Operations Management*, vol. 5, no. 2, pp. 157–175, 2003.

[12] Ö. Özer, “Replenishment strategies for distribution systems under advance demand information,” *Management Science*, vol. 49, no. 3, pp. 255–272, 2003.

[13] T. Wang and B. Toktay, “Inventory management with advance demand information and flexible delivery,” *Management Science*, vol. 54, no. 4, pp. 716–732, April 2008.