Fulfilling Orders in a Multi-Echelon Capacitated On-line Retail System: PART TWO, real-time purchasing and fulfillment decision making

Juan Li
Xerox Research Center Webster, Xerox Innovation Group, juan.li@xerox.com,

John A. Muckstadt
School of Operations Research and Information Engineering, Cornell University, jam61@cornell.edu,

When fulfilling customer orders, on-line retailers must operate their multi-warehouse systems with great care to ensure that these orders are satisfied in a timely and cost effective manner. We worked closely with a major on-line retailer to design an effective and efficient fulfillment system. This included establishing policies and procedures for ordering, receiving, storing and shipping of goods. Internal warehousing and transportation practices were addressed, and new approaches for managing inventories were established. In this paper we focus on one type of inventory management problem faced by the company when making daily purchasing and allocation decisions. These decisions are of two types, the positioning of inventories in their multi-echelon system and the detailed manner in which they use inventories to fulfill specific customer orders. After reviewing some of the key attributes of models that address the two types of decision problems, we present a computationally tractable approach for solving these problems for a system that must fulfill many hundreds of thousands of orders daily.

Key words: Multi-echelon, Capacitated, Multi-item, level, Inventories

1. Introduction

On-line retailers have constructed distribution systems in the United States to fulfill customer orders in a timely and cost effective manner. These systems contain multiple warehouses in which millions of different items are stocked. Determining what items to stock in what quantities in each warehouse and how to fulfill each customer's order must be done carefully to ensure that the retailer is financially and operationally successful.

In recent years, we worked closely with a major on-line retailer to address a broad range of issues related to the design and operation of its fulfillment system. A revised approach to procuring, receiving, stocking, and shipping of goods was developed, which is discussed in greater detail in our companion paper, Li and Muckstadt (2013a). Other aspects of our interaction with the company are also discussed in that paper. The organizational and operational structure described in that paper is called the Primary Warehouse System (PWS). Given this structure, we developed a model and an algorithm for setting target inventory levels for each location in the PWS. That model, which is a two-echelon, multi-item, multi-period model, is intended to be used by the retailer in its sales and

operations planning activities. This model focuses on setting target inventory levels for each item at each location over a planning horizon, which is typically 3 to 15 months in length. It does not directly focus on fulfilling customer orders. These orders are often for multiple items (about 60% for the retailer we studied). The order quantity for a single item type may be greater than one as well (approximately 2% for the retailer we studied). Customer orders also have due dates associated with them which differ among orders. Allocation and fulfillment decisions require coordinating shipping activities among all warehouses. This coordination must be executed carefully to ensure the timely fulfillment of customer orders while minimizing costs. Hence, an approach that differs from the one used for planning activities is needed to make daily procurement, allocation and fulfillment decisions in the retailer's highly dynamic environment. Our objective in this paper is to describe an approach for making execution decisions for hundreds of thousands of orders and items that can be executed in close to real time.

The remainder of this paper is organized as follows. In Section 2 we review the PWS's structure and other system operating characteristics. In Sections 3 and 4 we discuss our modeling assumptions and present nomenclature used throughout the paper. In Sections 5, 6 and 7 we present our execution models and computational methods for making daily procurement, allocation and fulfillment decisions. We conclude with some final comments.

2. Background

We now summarize some of the important attributes of the on-line fulfillment system we examined.

The fulfillment system contains warehouses located across the United States. Each warehouse is conceptually considered as two entities. The first entity directly fulfills customer orders that arise in the geographical region in which the warehouse is located. We call this part of the warehouse the *regional warehouse*. The on-line retailer had five such regional warehouses in the United States at the time we worked with it. Customer orders are generally satisfied from the warehouse that is located closest to the shipping address to reduce the "last-mile delivery costs," which are significant. Normally the last-mile-delivery cost is a concave function of the volume or weight associated with an order and the distance to the customer from the shipping location. Hence it is desirable to have all items needed to satisfy an order on hand at the regional warehouse closest to the customer. The second conceptual entity makes procurement and allocation decisions. The external supplier for an item interacts with this entity and sends inventories to it. We call this second entity the *primary warehouse* for that item. That means each physical warehouse conceptually serves as a primary warehouse for a collection of items and serves as a regional warehouse for all items needed to fulfill orders received from nearby customers.

Depending on the item, a regional warehouse can be categorized as a co-located regional warehouse or a non-co-located regional warehouse. A warehouse is a *co-located regional warehouse* for items whose primary warehouse is at the same physical location. If a warehouse is the primary warehouse for an item, the co-located warehouse does not conceptually carry any stock of the item. All stock is assumed to be located in the primary warehouse for that item. A regional warehouse is a *non-co-located regional warehouse* for an item when its primary warehouse is elsewhere.

As mentioned earlier, customers request a response time for fulfilling their orders. Some customers are willing to pay extra for immediate delivery while others prefer to delay the shipment to receive a discounted shipping cost. We categorize the orders into *short response lead time orders* and *long response lead time orders* depending on the customer's desired response time. If an order's required delivery response time is less than the shipping lead time from the primary warehouse to the regional warehouse plus the shipping time from the regional warehouse to the customer, then we call it a short response lead time demand. Otherwise, we call it a long response lead time demand. For the system we studied, the short response lead time demand accounted for about $13\% - 20\%$ of the total demand. The percentage varied by item and by time of the year.

Recall that an order should be satisfied from the regional warehouse that is closest to the customer's shipping address to minimize the last-mile-delivery cost. A regional warehouse stocks inventory to satisfy short response lead time demand that arises in its region. The regional warehouses are replenished by the primary warehouse. Hence, the primary warehouse carries inventory for two reasons. First, all the inventory required to satisfy all long response lead time demand for an item is stocked at its primary warehouse. When a customer places a long response lead time order, the inventory will be sent from the primary warehouse to the designated regional warehouse where it will be cross-docked and sent to the customer by a third-party logistics provider. Second, the primary warehouse also carries inventory to replenish regional warehouse stocks on a daily basis. Thus a primary warehouse sends inventory to a regional warehouse to replenish regional warehouse stock and to fulfill long response lead time demands.

In practice, labor and equipment availability constrain the amount of inventory that can be shipped every day. These constraints exist so that workloads are relatively smooth over time. These constraints are not "hard" ones. A moderate amount of overshoot above the target maximum workload level (less than 1% of total capacity) is permitted.

As we mentioned, when customers place orders, they sometimes request multiple items in one order. These items may or may not be managed by the same primary warehouse. Suppose a long response lead time order consists of items that are managed by different primary warehouses.

Then each item in the order should arrive at the designated order fulfilling regional warehouse at the same time. Careful coordination is needed to minimize incremental warehousing costs (labor and capacity) and transportation costs (split shipments). Establishing the warehouse processes and information system mechanisms required to make this coordination work properly were very important components of our system design efforts. More will be said about our design efforts in the final section. In this paper we present the models we designed to coordinate the shipping activities across warehouses.

Although there are many millions of items available for purchase from an on-line retailer, most have demand rates of four or fewer units per year. About 70% of the items offered by the on-line retailer we studied were low demand rate items. Most of these items are not stocked by the on-line retailer but rather are held in another company's warehouse. The low demand rate items stocked by the on-line retailer were stocked in one centrally located warehouse in the United States. Since our objective in this paper is to develop an approach for coordinating activities across regional warehouses, we will focus on the items stocked in multiple regional warehouses.

The series of execution models we now develop are designed to assist in making day-to-day procurement, allocation, and order fulfillment decisions. These models are based on the attributes we have discussed and the results presented in our companion paper Li and Muckstadt (2013a).

A review of the relevant literature can be found in Li and Muckstadt (2013a). and will not be repeated here. We, however, do include a set of references.

3. Assumptions Underlying The System's Operation

We begin by stating some assumptions about our model and the fulfillment system's operation.

Our execution models are periodic review models, a period being one day in length. There are three types of decisions of interest. The first is a procurement decision at the primary warehouse. This decision determines whether or not to order an item and how much to order. The second is the daily allocation decision for every item. The last decision establishes which orders to satisfy either completely or partially.

In Li and Muckstadt (2013a), an echelon stock based order-up-to policy is followed when making procurement decisions. An order-up-to policy remains in use in the execution stage and is based on the one generated when solving the planning problem, as we will see shortly.

The primary warehouses allocate their inventories to regional warehouses on a daily basis. As mentioned, we assume the shipping capacity from one primary warehouse to a non-co-located regional warehouse is limited. Recall that this shipping constraint is a workload smoothing mechanism.

The allocation includes three types of inventories, and we have to assign priorities to each type. First, shipping capacity is allocated to satisfy the long response lead time items that must be sent from the primary warehouse to the regional warehouse where the order is to be fulfilled. By must, we mean that the order will be backordered otherwise. Second, the remaining capacity is used, perhaps partially, to replenish the regional warehouse stocks. Remember, the regional warehouse's stocks are used to satisfy the unknown short response lead time demand that may arise over the primary warehouse to regional warehouse shipping lead time. Third, the remaining inventories and capacity are then allocated to fulfill certain long response lead time orders in advance, that is, prior to their due date. As mentioned previously, the problem is to coordinate the timing of these allocations to regional warehouses from the primary warehouses to fulfill the orders both known and unknown.

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, we place a replenishment order by the primary warehouse on an external supplier. Fifth, based on availability, the inventory positions at the regional warehouses, and the shipping capacity, we allocate inventory on-hand at the primary warehouse to the regional warehouses. Sixth, we receive replenishment 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 the regional warehouses at each period's end.

4. Nomenclature

In this section, we introduce some nomenclature that will be used throughout the paper. Other nomenclature specific to individual models will be introduced later.

Let i denote an order and j denote an item. Let $m(j)$ denote the primary warehouse for item j and let $n(i)$ denote the regional warehouse n that is planned to satisfy order i.

As mentioned, an order may contain multiple items and more than one unit may be requested for any item. Let $a_i^j \geq 0$ denote the number of units of item j requested in order i.

Let $\kappa(i)$ denote the time period during which customer order i is received. Time $\tau(i)$ is the period by which order i must be sent from regional warehouse $n(i)$ to the customer. Hence the required response time along with each order is $L_i = \tau(i) - \kappa(i)$. For each order i, we define an order fulfillment time $\bar{L}_i$, which is the number of days before $\tau(i)$ by which a primary warehouse must ship items to fulfill a long response lead time order at regional warehouse $n(i)$.

Let L be the time required to ship between two distinct warehouses. We assume shipments from the primary warehouse to its co-located regional warehouse occur instantaneously. Then

$$\bar{L}_i = \begin{cases} L, & \text{if there is an item } j \text{ in order } i \text{ for which } n(i) \neq m(j) \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

The *maximum grace period* for a long response lead time order i is defined to be $l_i = L_i - \bar{L}_i$. *Current grace period* for an outstanding long response lead time order i in time period t is $l'_i = (\tau(i) - t - \bar{L}_i)^+$.

Next let $d_{nt}^{j\alpha}$ and $d_{nt}^{j\beta}$ denote the observed short response lead time demands for item j at regional warehouse n in time period t, and the remaining unfilled known long response lead time demand for item j that is due to be shipped in period t at regional warehouse n, respectively. $D_{nt}^{j\alpha}$ and $D_{nt}^{j\beta}$, random variables, denote the short and long response lead time demands in some future period t for item j at regional warehouse n, respectively.

Suppose in time period t that there is a long response lead time order i that must be shipped from regional warehouse $n(i)$ to customers by period $\tau(i)$. If $\tau(i) - L < t$, then any item j for which $m(j) \neq n(i)$ in that order that has not been shipped previously cannot be sent to the customer on time using stock located at the primary warehouse. Therefore, fulfilling a long response lead time order i at regional warehouse $n(i)$ by time period $\tau(i)$ can only be accomplished using stock on hand at the regional warehouse $n(i)$ when $\tau(i) - L < t$ for item j for which $a_i^j > 0$ and $m(j) \neq n(i)$.

We let C_{mnt} denote the shipping capacity from primary warehouse m to regional warehouse n in time period t. Normally, $C_{mnt} > \mathbb{E}[\sum_{j:m(j)=m} D_{nt}^j]$, where D_{nt}^j is a random variable for both short and long response lead time demand for item j that will be fulfilled at regional warehouse n that arises on day t.

5. A Procurement Model

In Li and Muckstadt's (2013a) planning model, a primary warehouse places orders according to a fixed schedule for each item. In reality, the primary warehouse does not strictly adhere to the planned schedule. Whenever much more demand occurs than expected for an item, the primary warehouse will place an order to reduce the probability of incurring a stockout. Similarly, if fewer

customer orders are received than expected, the primary warehouse may place the next procurement order beyond the planned time. Even though the length of the cycle is not fixed in the execution of the fulfillment system, we should expect that the interval between two procurement decisions will be close to the fixed cycle-length used in the planning model. If there is a large discrepancy between the planned cycle length and the actual cycle length, the planning model is no longer accurate, and would be executed again. The planning model would be executed at least monthly to ensure the inventory, warehousing and transportation tactics are aligned properly. In this section, we introduce a procurement model that can be used to determine whether the primary warehouse should place an order for an item in the current time period and what the order quantity should be. The procurement decision is made based on the current echelon inventory position and the target echelon inventory level computed in the planning model.

5.1. Procurement Model Formulation

In practice, procurement decisions are decoupled from allocation and order fulfillment decisions due to the lengthy procurement lead time. The procurement lead times are normally several weeks or longer in length. In some cases, they are measured in months. Allocation decisions are made to reflect the dynamics of the fulfillment system over a horizon measured in days.

When constructing the procurement model, we use the target echelon inventory position computed through the use of the planning model to guide the execution strategy. From the target echelon inventory position, we are able to compute the planned in-cycle service level. The goal of the procurement model is to maintain this service level during the execution stage. As mentioned, an order-up-to policy is used in the execution model.

Let y^*_{jk} denote the target system echelon inventory position for item j at the beginning of cycle k as determined in the planning model. Let $D^j_{[a,b]}$ denote the random variable for total system demand for item j over the interval $[a,b]$. Also $\tilde{L}_j$ is the procurement lead time for item j and δ_j is the planned cycle length. The desired in-cycle service level of item j at primary warehouse $m = m(j)$ is

$$p^*_{jk} = P[D^j_{[0,\tilde{L}_j+\delta-j-1]} \leq y^*_{jk}], \tag{2}$$

which is the probability that total system demand over the time horizon of length $\tilde{L}_j + \delta_j$ days does not exceed y^*_{jk}. Thus p^*_{jk} is the planned probability of not running out of stock in cycle k.

Suppose t is some period in a cycle. Let y^j_t represent the system echelon stock after demand is observed in period t. Calculate

$$p^j_t = P[D^j_{[t+1,t+\tilde{L}_j]} \leq y^j_t], \tag{3}$$

which is the probability that demand does not exceed the current system echelon inventory position over the subsequent $\tilde{L}_j$ periods.

If $p_t^j \leq p_{jk}^*$, then the primary warehouse will place an order to raise the system echelon inventory position of item j up to the target $y_{j,k+1}^*$.

Thus procurement decisions are made item by item. Using a parallel computing environment, all computations associated with these decisions for each primary warehouse can be executed in well less than a minute.

6. Inventory Allocation

In this section we present a model that can be used to allocate inventories from a primary warehouse to the regional warehouses when flexible delivery is allowed, that is, when long response lead time orders can be shipped to the regional warehouses any time within the grace period.

6.1. Assumptions and Nomenclature

The objective of our allocation model is to determine the amount of each item to ship to regional warehouses so as to minimize the expected holding, backorder and delivery costs over a horizon whose length is several days in duration.

Note that we will only determine what quantity of each item to allocate in the current period. For the purpose of smoothing daily allocations, we will allocate inventories to fulfill long response lead time orders only after the order's content is known. Hence, the planning horizon is at least as long as the number of days over which long response lead time shipping requirements are known. More will be said about the length of the planning horizon in Section 6.2.2.

Another important observation pertains to how the backlogged items are filled in the system. At the end of a period, we charge a backorder cost based on the amount of unfilled demand for order i that is not sent from $n(i)$ by time $\tau(i)$. In the actual operation of the fulfillment system we worked on, when an item is backlogged, the primary warehouse inventories are used to satisfy those backorders directly when inventories become available. As a consequence, the fulfillment of backorders does not consume the shipping capacity from the primary warehouse to regional warehouses. We will address this assumption again when we introduce the constraints and models.

There are three types of decisions to make in this stage. The inventories and capacities are first allocated to regional warehouse $n(i)$ to satisfy order i that must be sent from the primary warehouse today, that is, $\tau(i) = t + \bar{L}_i$. The decisions are denoted by u_i^j, where u_i^j is the number of units of item j shipped to satisfy order i. Clearly, $u_i^j \leq a_i^j$. When there is not enough stock at the primary warehouse to fulfill an order completely, inventory on-hand at regional warehouse $n(i)$ may

be used to fill the order. Let w_i^j denote the units of inventory of regional warehouse $n(i)$ on-hand stock planned to be used to fill order i's requirements for item j, where $n(i)$ is not a co-located warehouse. Since the co-located regional warehouse does not hold any inventory, $w_i^j = 0$ there. If order i that is due in period $t + \bar{L}_i$ is expected to be filled completely in period t, then $u_i^j + w_i^j = a_i^j$.

The second decision pertains to the replenishment of regional warehouse stock. Let y_{nt}^j represent the number of units of item j allocated from the primary warehouse stock to replenish the stock of item j at regional warehouse n in time period t. The goal is to raise the inventory level at regional warehouses to satisfy future unknown short response lead time demands.

Following the allocation of inventory to a regional warehouse to meet today's long response lead time requirements and to replenish the regional warehouse inventories, some shipping capacity may remain unused. We may use some of this remaining capacity to ship inventories to satisfy long response lead time orders for which $\tau(i) > t + \bar{L}_i$. Thus a third decision is to determine whether or not to fulfill such orders. This decision is denoted by a binary variable x_{it} for each order i, where

$$x_{it} = \begin{cases} 1, \text{ if inventory and capacity are allocated in period } t \text{ to fulfill order } i \text{ completely,} \\ 0, \text{ otherwise.} \end{cases} \tag{4}$$

These allocations are made only if an existing order is to be fulfilled totally.

Let z_i be a binary variable that assumes a value of 1 when order i is not planned to be filled completely by period $\tau(i)$. Let the binary variable $\tilde{z}_i = 1$ if order i is fulfilled completely in period $\tau(i)$. We have

$$\sum_{k=1}^{\tau(i)-\bar{L}_i-1} x_{ik} + \tilde{z}_i + z_i = \tilde{x}_i + \tilde{z}_i + z_i = 1, \tag{5}$$

where $\tilde{x}_i = \sum_{k=1}^{\tau(i)-\bar{L}_i-1} x_{ik}$, which indicates whether or not order i is planned to be fulfilled before period $\tau(i)$. When order i is not expected to be satisfied completely on time, a penalty cost Q_i is charged, where Q_i measures the expected incremental cost incurred due to the partial fulfillment of order i. More will be said about this penalty cost subsequently. Note that we say planned to not be fulfilled. It still may be fulfilled even though $z_i = 1$. This can occur when short response lead time demand at regional warehouse $n(i)$ is less than the on-hand stock during the lead time of length $\bar{L}_i$ thereby resulting in inventory becoming available at $n(i)$ to satisfy these long response lead time demands in time period $\tau(i)$.

Recall that $d_{nt}^{j\beta}$ represents the outstanding long response lead time demand for item j due in period t at regional warehouse n following the allocation of u_i^j units. Hence, $d_{nt}^{j\beta} = \sum_{i:\tilde{x}_i=0,\tau(i)=t} (a_i^j - u_i^j)$. At a co-located regional warehouse, the allocation decisions are made following the receipt of

that day's short response lead time demand. Therefore, the allocations to the co-located regional warehouse are made to fill orders completely whenever possible. Let $d^j_{N_j t} = \sum_{i:n(i)=N_j, \tau(i)=t, \tilde{x}_i=0} (a^j_i - u^j_i)$, where N_j is the co-located warehouse for item j. Note that $d^j_{N_j t}$ includes both the short response lead time demand as well as the remaining long response lead time demand at N_j unfilled by the end of period t.

There are two goals we shall keep in mind in the allocation stage. First, we would like to minimize backorders and second we would like to maximize th number of orders filled completely.

In the next section, we describe a dynamic program that could conceptually be used to determine the optimal allocation decisions.

6.2. Single Primary Warehouse System

Recall that each warehouse serves as a primary warehouse for some items as well as a regional warehouse for the other items. As a result, the allocation model is complicated since the decision to fulfill an order may trigger inventory shipments from multiple warehouses. Hence, when there are N warehouses in the fulfillment system, the model must include N primary warehouses and their interactions when making allocation decisions. For ease of notation and discussion, let us start with a simple case in which there is only one primary warehouse in the fulfillment system. This means that all items in the system share the same primary warehouse. As we will see, the extension to the N primary warehouse case is straight-forward. In this section we let the warehouses be numbered so that regional warehouse N is the primary warehouse for all items.

6.2.1. Dynamic Program We now outline a dynamic program that could be used on a daily basis to make the three allocation decisions. When making these allocations, three types of costs would be considered. At the primary warehouse, a holding cost h^j_0 would be charged at the end of each period proportional to its on-hand inventories. At regional warehouse n, we would charge a holding cost h^j_n for each unit of on-hand inventory of item j held at the end of a period and a backorder penalty cost b^j for each backordered unit of item j. We assume $h^j_0 < h^j_n$, $n = 1, \cdots, N-1$. As noted, when a long response lead time order i is not fully fulfilled by time $\tau(i)$, an incremental cost Q_i is charged to order i. The magnitude of Q_i would reflect a combination of the order's priority and its content (large/heavy orders have priority due to the concave nature of the transportation cost functions). This penalty cost is also charged to the short response lead time orders unfilled at the co-located warehouse.

The Model

To formulate the problem as a dynamic program, one would first construct a one-period cost function that includes the holding costs, backorder costs and penalty shipping costs at all warehouses. The dynamic program's objective would be expressed as the one-period cost function plus future expected costs over the planning horizon.

Suppose I_{0t}^j represents the net inventory of item j at the primary warehouse at the end of period t and q_{0t}^j denotes the replenishment order for item j placed on the outside supplier in period t. Recall that the replenishment lead time for item j from the supplier to the primary warehouse is $\tilde{L}_j$ and that y_{nt}^j would be the replenishment allocation to regional warehouse n for item j in period t. At the end of period t the net inventory of item j at the primary warehouse would be the net inventory at the beginning of the period plus the replenishment order that is scheduled to arrive in that period, if any, minus the total amount of inventory allocated in period t, which includes all three types of allocations. That is

$$I_{0t}^j = I_{0t-1}^j + q_{0,t-L_j'}^j - \sum_{i:\tau(i)=t+\bar{L}_i,\tilde{x}_i=0} u_i^j - \sum_{n=1}^{N-1} y_{nt}^j - \sum_{i:\tau(i)>t+\bar{L}_i} x_{it} a_i^j. \tag{6}$$

When allocating inventories to fill long response lead time demands, we would plan to minimize the number of packages used to fulfill one order. Therefore, when there is not enough primary warehouse inventory or shipping capacity to fill all orders, we would plan to use regional warehouse stock to fill this order completely if possible. Recall that w_i^j denotes the amount of regional warehouse stock of item j planned to be used to satisfy order i . This would be a planned fulfillment rather than an actual fulfillment of the order as discussed more fully below.

The net inventory level at the end of period t at regional warehouse $n \in \{1, 2, \cdots, N-1\}$ would be the net inventory level at the beginning of the period plus the replenishment stock $y_{n,t-L}^j$ shipped to regional warehouse n L periods ago minus the amount of stock allocated to satisfy long response lead time demands that have not previously been satisfied from the stock coming from the primary warehouse, which we have denoted by $d_{nt}^{j\beta}$, and the short response lead time demands that would be satisfied from the regional warehouse n stock, which we have denoted by $d_{nt}^{j\alpha}$.

Then the net inventory of item j at the end of period t at regional warehouse n would be

$$I_{nt}^j = I_{nt-1}^j + y_{nt-L}^j - d_{nt}^{j\alpha} - d_{nt}^{j\beta} \tag{7}$$

$$= I_{nt-1}^j + y_{nt-L}^j - d_{nt}^{j\alpha} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a_i^j - u_i^j). \tag{8}$$

The dynamic programming recursion in period t would represent the costs incurred in that period as a function of the incoming state of the system, that is, the values of $I_{0,t-1}^j$ and $I_{n,t-1}^j$,

the demands ($d_{nt}^{j\alpha}$, $d_{nt}^{j\beta}$), the procurement quantities ($q_{0,t-\bar{L}_j}$) and allocation decisions (u_i^j, $y_{n,t-L}^j$), plus expected future costs, which would depend on I_{0t}^j and I_{nt}^j. The allocation decisions in period t would be made considering several constraints. For example, the primary warehouse could not allocate more of an item to the regional warehouses than the amount of stock of the item it has on-hand; the amount shipped to a non-co-located warehouse from the primary warehouse could not exceed its shipping capacity; and regional warehouses could not plan to allocate more stock to fill orders on their due date (the w_i^j variables) than it would have on-hand. There would also be logical constraints that would indicate whether or not an order is fulfilled fully by its due date (z_i, $\bar{z}_i$).

The size of the state space associated with this dynamic program would be extremely large considering the number of items and locations in the system. Thus, such a dynamic programming formulation would not be able to be solved in a timely manner for practical problems of interest. Hence, we now turn our attention to developing a sequence of approximate models that will generate the desired allocation and fulfillment quantities.

6.2.2. An Approach for Making Allocation and Order Fulfillment Decisions We now present a computationally tractable approach for making allocation and fulfillment decisions. Recall that there are four distinct decisions that must be made each day. We address them in sequence. In the first sub-model we allocate inventory to satisfy existing backorders as completely as possible. In the second sub-model, we determine which outstanding long response lead time orders i for which $n(i) \neq N$ we expect to fill that are due $\bar{L}_i$ periods in the future, and all outstanding orders i due today at N. In the third sub-model, we determine how to allocate inventory of each item to replenish regional warehouse stocks. In the last sub-model we determine which long response lead time orders to fulfill completely prior to their due dates.

Sub-Model 1: Satisfy Unfilled Backorders At the Co-located Regional Warehouse

We first allocate on-hand inventory at the primary warehouse to satisfy as much of the backlogged demand as possible. Let y_{n0}^j denote the amount of item j allocated to satisfy backorders that exist at regional warehouse n.

The available on-hand inventory of item j at the primary warehouse is I_{00}^j at the beginning of the time horizon, which we denote as day 1. Let q_{0t}^j denote the quantity of item j ordered on day t. After fulfilling existing backorders, the effective inventory level there for item j at the beginning of day 1 is $I_{00}^{j\prime} = (I_{00}^j - \sum_{n=1}^{N} y_{n0}^j + q_{0,t-\tilde{L}_j}^j)$. The echelon inventory of item j at regional warehouse n increases by y_{n0}^j. Hence, $I_{n0}^{j\prime} = I_{n0}^j + y_{n0}^j$. Note that $\sum_n y_{n0}^j \leq [I_{00}^j + q_{0,t-\tilde{L}_j}^j]^+$. In practice, the

allocation y_{n0}^j would be sent directly to a customer and the value of the net inventory at regional warehouse n would be adjusted accordingly. Thus there would not be a physical transfer of stock but rather an accounting transfer of stock from the primary warehouse to the regional warehouses.

Sub-Model 2: Fulfill Orders That Must be Sent Out Today

We next determine how inventories at the primary warehouse and regional warehouses should be allocated to satisfy long response lead time orders that must be filled at the regional warehouses $n \neq N$ a lead time $\bar{L}_i$ in the future and short and long response lead time orders due in period 1 at regional warehouse N. Thus, we only allocate inventory to regional warehouse $n \neq N$ that are due on day $1 + \bar{L}_i$ and to satisfy all orders due on day 1 at N.

To begin, we first identify those outstanding long response lead time orders for which $\tau(i) = 1 + \bar{L}_i$ that can be completely satisfied from the primary warehouse stock. We also include short response lead time orders due at the co-located regional warehouse N in this set of orders. For each item j, determine if $I_{00}^{j\prime} \geq \sum_{i:\tau(i)=1+\bar{L}_i, \tilde{x}_i=0} a_i^j$. Let J be the set of items j for which the above inequality holds. Next, let O_J be the set of outstanding orders i such that $\tau(i) = 1 + \bar{L}_i$ and $a_i^j > 0$ but only for items $j \in J$. Allocate stocks to these orders and decrement stocks accordingly. For all $i \in O_J$, set $x_i = 1$. Then $j \notin J$ implies there is not enough stock of item j on hand at the primary warehouse to satisfy all orders for which $a_i^j > 0$.

We have assumed that shipping capacity is available to execute these allocations. As we will see when we consider Sub-Model 4, a large proportion of the long response lead time orders due on day 1 would have been allocated previously. Hence the assumption is not an unreasonable one to make. However, if shipping capacity is not adequate to allocate all the required inventory to satisfy the demand for these long response lead time orders $i \in O_J$, then we will determine the allocation using the following model and algorithm.

Rather than maximizing the number of orders that would be filled on time, we will minimize

$$\sum_{i \in O_J} Q_i(1 - x_i). \tag{9}$$

Equivalently, we can maximize

$$\sum_{i \in O_J} Q_i x_i \tag{10}$$

$$\text{s.t.} \quad \sum_{i \in O_J, n(i)=n} x_i \{ \sum_{j \in J} a_i^j \} \leq C_n \tag{11}$$

$$x_i = 0, 1, \tag{12}$$

where C_n is the shipping capacity from the primary warehouse to regional warehouse n, $n \neq N$, and Q_i is the incremental cost of shipping the entire order i in an expedited manner from the primary warehouse. Since the problem is separable by regional warehouse, we solve

$$\max \quad \sum_{i \in O_J^n} Q_i x_i \tag{13}$$

$$\text{s.t.} \quad \sum_{i \in O_J^n} x_i \{ \sum_{j \in J} a_i^j \} \leq C_n \tag{14}$$

$$x_i = 0, 1, \tag{15}$$

where O_J^n is the set of orders that will be ideally delivered to customers from regional warehouse n. This problem can be solved, using a greedy algorithm, separately for each $n \neq N$, and hence in a parallel fashion. The greedy algorithm is as follows.

Rank order the $i \in O_J^n$ according to the values $\frac{Q_i}{\sum_{j \in J} a_i^j}$, from largest to smallest. Suppose the orders are numbered according to this ranking. Then we use the following algorithm for each n for which capacity constraints are tight.

Algorithm 1

Step 0 : Set $C = 0$ and $i = 1$. Go to Step 1.

Step 1 : If $C < C_n$, set $x_i = 1$ and $C = C + \sum_{j \in J} a_i^j$.

Step 2 : Increment i and return to step 1.

Note that this algorithm may result in an allocation that violates the shipping capacity constraint but by a very small amount relative to the total shipping capacity. Since shipping capacity constraint is not a "hard" one, a relatively small overshoot is acceptable in practice.

Then for each n, reduce the amount of shipping capacity and inventory levels commensurate with the allocations made so far in this sub-model, that is, set

$$C_n \leftarrow C_n - \sum_{i \in O_J^n} \sum_{j \in J} a_i^j x_i, \text{ and} \tag{16}$$

$$I_{00}^{j'} \leftarrow I_{00}^{j'} - \sum_{i \in O_J} a_i^j x_i. \tag{17}$$

Next, we allocate available primary warehouse inventory to orders $i \notin O_J$ for which $j \notin J$, $I_{00}^{j'} > 0$, $a_i^j = 1$ for only one j , $\tau(i) = 1 + \bar{L}_i$, and $C_{n(i)} > 0$. Since $I_{00}^{j'} < \sum_{i \notin O_J : \tau(i) = 1 + \bar{L}_i, \bar{x}_i = 0} a_i^j$, it may not be possible to satisfy all such orders even if there is enough shipping capacity. We use the following algorithm to determine the allocation. We begin by ranking the orders in descending (non-increasing) values of $Q_{i.}$ Renumber the orders, beginning with 1, corresponding to this ranking.

Algorithm 2

Step 0: Set $i = 1$

Step 1: For j such that $a_i^j = 1$, if $I_{00}^{j'} > 0$ and $C_{n(i)} > 0$, then set $x_i = 1$, $I_{00}^{j'} \leftarrow I_{00}^{j'} - 1, C_{n(i)} \leftarrow C_{n(i)} - 1$; otherwise, set $x_i = 0$.

Step 2: Increment i and return to Step 1.

It is important to recall that almost all orders will be completely satisfied in this first phase of Sub-Model 2. This is the case since about 40% of the orders are for a single unit of a single item, most items have very high fill rates, and capacity will likely be available to ship the items.

The second part of this sub-model is concerned with making allocations of inventories to orders that have not yet been satisfied. The objective is to allocate remaining inventories to those unsatisfied orders so as to minimize the incremental economic consequences. Define O to be the set of remaining unfilled long response lead time orders that are due at time $\tau(i) = 1 + L$ at regional warehouses $n \neq N$, and the remaining unfilled short and long response lead time orders that are due at the co-located regional warehouse N for which $\tau(i) = 1$.

Our objective now is to minimize the sum of the penalty costs, where we define the penalty cost for order i to be

$$Q_i \cdot \left(\frac{\sum_j (a_i^j - u_i^j - w_i^j)}{\sum_j a_i^j} \right). \tag{18}$$

This is the weighted fraction of the total incremental transportation and handling costs (Q_i). The greater the fraction of the order that is shipped via the normal delivery plan, the smaller the penalty cost. Hence our objective is to minimize the total incremental shipping costs. As an approximation, we choose to

$$\min \sum_{i \in O} Q_i \cdot \left(\frac{\sum_j (a_i^j - u_i^j - w_i^j)}{\sum_j a_i^j} \right). \tag{19}$$

When making the allocation decisions, we must consider several conditions. First, the sum of u_i^j and w_i^j must be no larger than a_i^j. The next two constraints limit the quantity of item j that can be allocated to order i. These constraints ensure that allocations made at the primary warehouse and at the regional warehouses do not exceed the inventories on hand at the respective locations. The final constraint ensures that shipping capacity is available.

To ensure that we allocate inventory from the primary warehouse to satisfy orders (u_i^j) before allocating regional warehouse stock (w_i^j), we include the term $\epsilon \sum_{i \in O} \sum_j w_i^j$ in the objective function, where ϵ is a small, positive number.

Hence, the formulation of the model is

$$\min \quad \sum_{i\in O}\left\{Q_i\cdot\left(\frac{\sum_j(a_i^j-u_i^j-w_i^j)}{\sum_j a_i^j}\right)+\epsilon\sum_j w_i^j\right\} \tag{20}$$

$$\text{s.t.} \quad u_i^j+w_i^j\le a_i^j, \qquad \forall\ i\in O, j, \tag{21}$$

$$\sum_{i\in O}u_i^j\le I_{00}^{j\prime}, \qquad \forall\ j, j\notin J \tag{22}$$

$$\sum_{i\in O:n(i)=n} w_i^j\le\left[\left[I_{n0}^{j\prime}+\sum_{k=1}^{L}S_{nk}^j-d_{n1}^{j\alpha}-\sum_{k=2}^{L}E[D_{nk}^{j\alpha}]\right]\right]^+, \qquad \forall\ j,\ n(i)\neq N, \tag{23}$$

$$w_i^j=0, \qquad \forall\ n(i)=N, \tag{24}$$

$$\sum_{i\in O:n(i)=n} u_i^j\le C_n, \qquad \forall n(i)\neq N, \tag{25}$$

$$u_i^j, w_i^j\ge 0 \text{ and integer}, \tag{26}$$

where S_{nk}^j is the number of units of item j received at regional warehouse n in period k corresponding to replenishment shipments made from the primary warehouse in period $k-L\le 0$ when $n\neq N$. Thus the S_{nk}^j values correspond to decisions made in prior periods. Note that we consider the future expected short lead time demand when setting the value of w_i^j.

Observe that minimizing $\sum_{i\in O}\left\{Q_i\cdot\left(\frac{\sum_j(a_i^j-u_i^j-w_i^j)}{\sum_j a_i^j}\right)+\epsilon\sum_j w_i^j\right\}$ is equivalent to maximizing

$$\sum_{i\in O}\sum_j\frac{Q_i}{\sum_k a_i^k}u_i^j+\sum_{i\in O}\sum_j\left(\frac{Q_i}{\sum_k a_i^k}-\epsilon\right)w_i^j. \tag{27}$$

Note that constraints (21) through (24) pertain to only one item. If there is adequate shipping capacity, that is, constraints (25) are inactive, then the problem can be solved one item at a time and in a parallel fashion. The algorithm we will discuss holds for situations when the capacity constraints are active or inactive.

Let $\tilde{Q}_i=\frac{Q_i}{\sum_j a_i^j}$. To maximize expression (27) we would prefer to have u_i^j be as large as possible before permitting w_i^j to be positive. Let

$$W^j=[\left[I_{n0}^j+\sum_{k=1}^{L}S_{nk}^j-d_{n1}^{j\alpha}-\sum_{k=2}^{L}E[D_{nk}^{j\alpha}]\right]]^+.$$

Then we can employ the following algorithm to obtain the desired allocation.

Algorithm 3

Step 1: For $i\in O$ and $C_{n(i)}>0$, rank order the $\tilde{Q}_i$ values from largest to smallest and renumber the items accordingly. Set $i=1$.

Step 2: For j such that $a_i^j > 0$, set $u_i^j = \min\{a_i^j, I_{00}^{j\prime}, C_{n(i)}\}$, decrement $I_{00}^{j\prime}$ by u_i^j and decrement $C_{n(i)}$ by u_i^j. Increment i and repeat this step until $i = |O| + 1$.

Step 3: Set $i = 1$.

Step 4: For each j for which $a_i^j > 0$, set $w_i^j = \min(a_i^j - u_i^j, W^j)$ and decrement W^j. Increment i and repeat Step 4 until $i = |O| + 1$ or until $W^j = 0$ for all j for which $a_k^j > 0$ for some $k > i$.

Sub-Model 3: Allocate Replenishment Stock to Regional Warehouses

After making allocations u_i^j and w_i^j that approximately minimize the number of potentially unsatisfied long response lead time orders and incremental shipping costs, we next allocate stocks to replenish regional warehouse inventories. Thus, the goal of the third sub model is to determine the amount of replenishment stock to allocate of each item type to each regional warehouse. Recall that the majority of the orders are long response lead time orders. Therefore, the replenishment decision should be made without jeopardizing the inventory that may be needed to fulfill long response lead time orders.

The supply at the primary warehouse available for replenishing regional warehouse inventory for item j following the allocations made when executing the previous sub models is denoted by $I_{00}^{j\prime}$.

The decision variable value that we next determine is y_{n1}^j, which is the amount of item type j replenishment stock to ship to regional warehouse n in period 1. Therefore, the net inventory of item j at the primary warehouse at the end of period 1 is

$$I_{00}^{j\prime} - \sum_n y_{n1}^j. \tag{28}$$

Let Γ_j denote the period before the next procurement shipment of item j is due to arrive at the primary warehouse. For each item j, let $R_j^{\min}$ and $R_j^{\max}$ denote $\min\{L_i : a_i^j > 0\}$ and $\max\{L_i : a_i^j > 0\}$, respectively. The inventory on hand at the primary warehouse at the end of period 1 would have to meet all short and long response lead time demand occurring at the co-located regional warehouse N through period Γ_j and all known long response lead time demand occurring at regional warehouse n, $n \neq N$, from period 2 through period $\min(\Gamma_j + L, R_j^{\max} + 1)$ that remains unsatisfied at the end of period 1, that is, $\sum_{k=2}^{L+1} \sum_{i:n(i)\neq N, \tau(i)=k, \tilde{x}_i=0} (a_i^j - u_i^j - w_i^j) + \sum_{n\neq N} \sum_{k=L+2}^{\min(\Gamma_j+L, R_j^{\max}+1)} d_{nk}^{j\beta}$, plus the unknown long response lead time demand that will occur in periods $R_j^{\min} + 2$ through $\Gamma_j + L$, which is $\sum_{n\neq N} \sum_{k=R_j^{\min}+2}^{\Gamma_j+L} D_{nk}^{j\beta}$ when $\Gamma_j + L \geq R_j^{\min} + 2$. Note that the values of the u_i^j and w_i^j variables for each order i for which $\tau(i) \leq L + 1$ are known at this point. For $n = N$, both the short and long response lead time demand during periods 2 through Γ_j should

be met from stock on-hand at the primary warehouse. Thus, the net inventory at the primary warehouse at the end of period Γ_j would be

$$I_{00}^{j\prime} - \sum_{n\neq N} y_{n1}^{j} - \sum_{k=2}^{L+1} \sum_{i:n(i)\neq N,\tau(i)=k,\tilde{x}_i=0} (a_i^j - u_i^j - w_i^j) - \sum_{n\neq N} \sum_{k=L+2}^{\min(\Gamma_j+L,R_j^{\max}+1)} d_{nk}^{j\beta} - \sum_{n\neq N} \sum_{k=R_j^{\min}+2}^{\Gamma_j+L} D_{nk}^{j\beta}$$
$$- \sum_{k=2}^{\Gamma_j} D_{Nk}^{j\alpha} - \sum_{n\neq N} \sum_{k=L+2}^{\Gamma_j} D_{nk}^{j\alpha} - \sum_{k=2}^{\min(\Gamma_j,R_j^{\max}+1)} d_{Nk}^{j\beta} - \sum_{k=R_j^{\min}+2}^{\Gamma_j} D_{Nk}^{j\beta}, \tag{29}$$

assuming no further replenishment stocks are directly allocated to non-co-located regional warehouses until the next shipment of item j is received at N in period $\Gamma_j + 1$ and assuming all short response lead time demand occurring at $n \neq N$ during periods $L+2$ through Γ_j are satisfied directly from the primary warehouse stock. Let

$$\bar{I}_{00}^{j} = I_{00}^{j\prime} - \sum_{k=2}^{L+1} \sum_{i:n(i)\neq N,\tau(i)=k,\tilde{x}_i=0} (a_i^j - u_i^j - w_i^j) - \sum_{n\neq N} \sum_{k=L+2}^{\min(\Gamma_j+L,R_j^{\max}+1)} d_{nk}^{j\beta} - \sum_{k=2}^{\min(\Gamma_j,R_j^{\max}+1)} d_{Nk}^{j\beta}. \tag{30}$$

The net inventory of item j at regional warehouse $n \neq N$ at the end of period $L+1$ is The net inventory of item j at regional warehouse $n \neq N$ at the end of period $L+1$ is

$$I_{n0}^{j\prime} + \sum_{k=1}^{L} S_{nk}^{j} + y_{n1}^{j} - d_{n1}^{j\alpha} - \sum_{k=2}^{L+1} D_{nk}^{j\alpha} - \sum_{k=1}^{L+1} \sum_{i:\tau(i)=k,\tilde{x}_i=0} (a_i^j - u_i^j), \tag{31}$$

where $\tilde{x}_i = 0$ implies that order i was not shipped prior to day 1. In this expression, the only random variables are the one period short response lead time demand random variables, $D_{nk}^{j\alpha}$. To simplify the notation, let

$$\bar{I}_{n0}^{j} = I_{n0}^{j\prime} + \sum_{k=1}^{L} S_{nk}^{j} - d_{n1}^{j\alpha} - \sum_{k=1}^{L+1} \sum_{i:\tau(i)=k,\tilde{x}_i=0} (a_i^j - u_i^j)). \tag{32}$$

Hence the net inventory level of item j at regional warehouse n at the end of time period $L+1$ can be rewritten as $\bar{I}_{n,L+1}^{j} = \bar{I}_{n0}^{j} + y_{n1}^{j} - \sum_{k=2}^{L+1} D_{nk}^{j\alpha}$, $n \neq N$.

Finally, the net inventory at the end of period 1 at regional warehouse N is

$$I_{N0}^{j\prime} + \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} u_i^j - \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} a_i^j. \tag{33}$$

Let

$$\bar{I}_{N0}^{j} = I_{N0}^{j\prime} + \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} u_i^j - \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} a_i^j. \tag{34}$$

In a dynamic program that produces an optimal allocation policy, the consequence of today's allocation decisions on the future allocation decisions is accounted for directly. In our sub-model, we

approximate the effect of the current decision on future costs and decisions. We do this as follows. In addition to calculating the current period's holding and backorder costs, we also approximate the expected future holding and backorder costs that would be incurred before the next procurement arrives.

The approximate cost model we use to determine y_{n1}^j for all j is

$$\begin{aligned}
&\sum_j h_0^j \mathbf{E}[\bar{I}_{00}^j - \sum_{n\neq N} y_{n1}^j - \sum_{n\neq N} \sum_{k=R_j^{\min}+2}^{\Gamma_j+L} D_{nk}^{j\beta} - \sum_{k=\ 2}^{\Gamma_j} D_{Nk}^{j\alpha} - \sum_{n\neq N} \sum_{k=L+2}^{\Gamma_j} D_{nk}^{j\alpha} - \sum_{k=R_j^{\min}+2}^{\Gamma_j} D_{Nk}^{j\beta}]^+ \\
&+ \sum_j b^j \mathbf{E}[-\bar{I}_{00}^j + \sum_{n\neq N} y_{n1}^j + \sum_{n\neq N} \sum_{k=R_j^{\min}+2}^{\Gamma_j+L} D_{nk}^{j\beta} + \sum_{k=\ 2}^{\Gamma_j} D_{Nk}^{j\alpha} + \sum_{n\neq N} \sum_{k=L+2}^{\Gamma_j} D_{nk}^{j\alpha} + \sum_{k=R_j^{\min}+2}^{\Gamma_j} D_{Nk}^{j\beta}]^+ \\
&+ \sum_{n\neq N} \sum_j h_n^j \mathbf{E}[\bar{I}_{n0}^j + y_{n1}^j - \sum_{k=2}^{L+1} D_{nk}^{j\alpha}]^+ + \sum_{n\neq N} \sum_j b^j \mathbf{E}[-\bar{I}_{n0}^j - y_{n1}^j + \sum_{k=2}^{L+1} D_{nk}^{j\alpha}]^+ + \sum_j b^j (-\bar{I}_{N0}^j)^+,
\end{aligned}$$

where we assume in this approximation that all short response lead time demand occurring in periods $L+2$ through Γ_j at non-co-located regional warehouses are satisfied directly from the primary warehouse's inventory.

The replenishment quantities are constrained by the remaining available inventory at the primary warehouse and the remaining shipping capacity from the primary warehouse to each non-co-located regional warehouse. That is,

$$\sum_{n=1}^{N-1} y_{n1}^j \leq I_{00}^{j\prime}, \qquad \forall j \tag{35}$$

and

$$\sum_j y_{n1}^j \leq C_n, \qquad \forall n \neq N. \tag{36}$$

where C_n is the shipping capacity remaining to regional warehouse n after the allocations that were made in the previous steps.

Let $Y^j = \sum_{n\neq N} y_{n1}^j$,

$$\begin{aligned}
F_{j0}(Y^j) = \quad & h_0^j \mathbf{E}[\bar{I}_{00}^j - \sum_{n\neq N} y_{n1}^j - \sum_{n\neq N} \sum_{k=R_j^{\max}+2}^{\Gamma_j+L} D_{nk}^{j\beta} - \sum_{k=\ 2}^{\Gamma_j} D_{Nk}^{j\alpha} - \sum_{n\neq N} \sum_{k=L+2}^{\Gamma_j} D_{nk}^{j\alpha} - \sum_{k=R_j^{\min}+2}^{\Gamma_j} D_{Nk}^{j\beta}]^+ \\
& + \sum_j b^j \mathbf{E}[-\bar{I}_{00}^j + \sum_{n\neq N} y_{n1}^j + \sum_{n\neq N} \sum_{k=R_j^{\max}+2}^{\Gamma_j+L} D_{nk}^{j\beta} + \sum_{k=\ 2}^{\Gamma_j} D_{Nk}^{j\alpha} + \sum_{n\neq N} \sum_{k=L+2}^{\Gamma_j} D_{nk}^{j\alpha} + \sum_{k=R_j^{\min}+2}^{\Gamma_j} D_{Nk}^{j\beta}]^+
\end{aligned} \tag{37}$$

and

$$F_{jn}(y_{n1}) = h_n^j \mathbf{E}[\bar{I}_{n0}^j + y_{n1}^j - \sum_{k=2}^{L+1} D_{nk}^{j\alpha}]^+ + b^j \mathbf{E}[-\bar{I}_{n0}^j - y_{n1}^j + \sum_{k=2}^{L+1} D_{nk}^{j\alpha}]^+. \tag{38}$$

Then the allocation optimization problem can be stated as

$$\min \quad \sum_j F_{j0}(Y^j) + \sum_{n \neq N} \sum_j F_{jn}(y_{n1}^j) \tag{39}$$

$$\text{s.t.} \quad \sum_{n=1}^{N-1} y_{n1}^j \leq I_{00}^{j\prime}, \qquad \forall j, \tag{40}$$

$$\sum_j y_{n1}^j \leq C_n, \qquad \forall n \neq N, \tag{41}$$

$$y_{n1}^j \geq 0. \tag{42}$$

Observe that the objective function is separable by item. The functions $F_{j0}(Y^j)$ and $F_{jn}(Y_n^j)$ are convex in each y_{n1}^j variable separately and jointly. Note that by relaxing the $N-1$ shipping constraints, the optimal solution could be determined by solving independent sub-problems, one for each item. Given that the objective function is convex and that the demand random variables assume non-negative integer values, each individual item subproblem can be solved efficiently by employing a straightforward marginal analysis algorithm. Thus, the optimal solution to the relaxed problem can be found by making use of a parallel computing architecture, which is of practical significance.

Let $\tilde{y}_{n1}^j$ be the optimal allocation for item j to regional warehouse n when solving the relaxed problem. Note that the $\tilde{y}_{n1}^j$ values, as well as the optimal allocation quantities for problem (39)-(42), will typically be for 10 or fewer units, and often for just one or two units. This is the case since over 80% of the orders have long response lead times. For example, when there is a system demand for 200 units of an item on a day, then there might be only 30 to 40 units requested in short response lead time orders. When spread over 5 regions, there will only be 6 to 8 units of the item that will be part of the short response lead time orders fulfilled by a regional warehouse. Unless target inventory levels are altered for a regional warehouse, the replenishment quantities will be about the same value as the short response lead time demand. Since the values of the $\tilde{y}_{n1}^j$ values are typically low, these item allocations can be determined very quickly.

Suppose we have solved the relaxed problem. If the shipping constraints are not violated, that is, if

$$\sum_j \tilde{y}_{n1}^j \leq C_n, \quad \forall n \neq N, \tag{43}$$

then $\tilde{y}^j_{n1}$ equals y^{j*}_{n1}, the optimal allocations for problem (39)-(42).

Suppose that condition (43) is not satisfied. We will consider two cases. In the first case, suppose all $N-1$ shipping constraints are violated by the solution to the relaxed problem. We find the optimal allocation in this case as follows.

For each $n \neq N$, find the item j that increases the expected cost by smallest amount when reducing its allocation by one unit. To do this, we would measure the expected incremental costs at both the primary and regional warehouses, using the approach discussed below. Next, determine the item, regional warehouse combination that increases the expected cost by the smallest amount. Let $j^*(n)$ be this item and regional warehouse combination. Set $\tilde{y}^{j^*(n)}_{n1} \leftarrow \tilde{y}^{j^*(n)}_{n1} - 1$. Repeat this process for each violated regional warehouse constraint until all the shipping constraints are satisfied. Note, the values required to execute this step were determined when solving the relaxed problem.

Now, in the second case, suppose that some, but not all, of the shipping constraints are violated by the solution to the relaxed problem. We find the optimal allocation by employing the following procedure.

The algorithm is somewhat more complicated when only a subset of the shipping constraints are violated. In particular, calculating the net impact of reducing the value of y^j_{n1} by one unit for some item j may require additional trade-offs to be considered.

For each violated shipping constraint, n, do the following. Suppose $\bar{y}^j_{n1}$ is the current value of y^j_{n1}.

Then for each item j for which $\bar{y}^j_{n1} > 0$, determine the net impact of reducing its value by one unit as follows.

First suppose for item j that

$$\sum_{n \neq N} \bar{y}^j_{n1} < I^{j\prime}_{00}. \tag{44}$$

When reducing $\bar{y}^j_{n1}$ by one unit, the primary warehouse stock of item j increases by one unit. Then $\Delta F_{j0}(Y^j) = F_{j0}(Y^j - 1) - F_{j0}(Y^j)$ measures the incremental expected primary warehouse cost. Compute $\Delta F_{jn}(\bar{y}^j_{n1})$, where $\Delta F_{jn}(\bar{y}^j_{n1}) = F_{jn}(\bar{y}^j_{n1} - 1) - F_{jn}(\bar{y}^j_{n1}) \geq 0$.

Let $j^*_1(n) = \arg\min_j \Delta F_{jn}(\bar{y}^j_{n1}) + \Delta F_{j0}(Y^j)$. Then $G^1_n = \Delta F_{j0}(Y^{j^*_1(n)}) + \Delta F_{j^*_1(n),n}(\bar{y}^{j^*_1(n)}_{n1})$ is the net impact of reducing $\bar{y}^{j^*_1(n)}_{n1}$ by one unit at regional warehouse n. Remember we only consider item j for which (44) is not tight.

Next consider items for which the inventory constraint (44) is tight. There are now two possibilities. If $\bar{y}^j_{n1}$ is reduced by one unit, then there may be another regional warehouse n' for

which the shipping constraint is not tight and $\Delta F_{jn'}(\bar{y}^j_{n'1}+1)<0$. If there exists an n' such that $\Delta F_{jn'}(\bar{y}^j_{n'1}+1)<0$, then let $n^*(j)=\arg\min_{n':C_{n'}>0}\Delta F_{jn'}(\bar{y}^j_{n'1}+1)$. That is, $n^*(j)$ is the regional warehouse that would most benefit from having its allocation increased by one unit of item j. The net impact for item j when there exists at least one n' such that $\Delta F_{jn'}(\bar{y}^j_{n'1}+1)<0$ is

$$\Delta F_{jn}(\bar{y}^j_{n1})+\Delta F_{jn^*(j)}(\bar{y}^j_{n^*1}+1).$$

The second possibility is that there is no n' for which $\Delta_{jn'}F_{jn'}(\bar{y}^j_{n'1}+1)<0$.

In this case the net impact of reducing the shipment of item j to regional warehouse n would be $\Delta_j(\bar{y}^j_{n1})=\Delta F_{j0}(Y^j)+\Delta F_{jn}(\bar{y}^j_{n1})$.

Let $j^*_2(n)=\arg\min_j\Delta_j(\bar{y}^j_{n1})$ and let $G^2_n=\Delta F_{j0}(Y^{j^*_2(n)})+\Delta F_{j^*_2(n),n}(\bar{y}^{j^*_2(n)}_{n1})$.

Then $G_n=\min(G^1_{n,}G^2_n)$, that is, we would select the item j that would have its allocation to regional warehouse n reduced by one unit that yields the overall smallest expected incremental cost.

Among the constraints n that remain violated, select n^* such that $n^*=\arg\min_n G_n$. Then at regional warehouse n^* we would decrement item $j^*(n^*)$'s allocation by one unit. We would reduce the constraint violation at n^* by one unit as well.

The above process would be repeated until no violated shipping constraints remain. Note that in each iteration of the algorithm calculations for only one item need to be updated. Virtually all of the calculations required at each step to update the Δ values for item j^* were made when the values of $\tilde{y}^j_{n1}$ were calculated.

Note also that once the updated value for item j^* is calculated, the comparison step corresponds to inserting this newly computed value into a list, which has a complexity of $log\ m$, where m is the number of items being considered. If $M=\sum_{n\neq N}[\sum_j\tilde{y}^j_{n1}-C_n]^+$ then the complexity of the algorithm is $Mlog\ m$.

Sub-Model 4: Allocate Inventory to Fill Orders Due in the Future

In this step we determine which, if any, long response lead time orders to fulfill in advance with the remaining inventories and shipping capacities. Let γ^j_1 be the amount of inventory the primary warehouse should hold in reserve to replenish regional warehouse stocks through period Γ_j. This amount of inventory is computed by minimizing the future expected holding and backorder costs at all regional warehouses.

Let γ^j_{n1} denote the inventory that will be held at the primary warehouse to possibly replenish regional warehouse n for item type j through period Γ_j. We find γ^{j*}_{n1} by solving the following newsvendor type of problem.

$$\min \sum_n h_n^j \mathbf{E}\left[\bar{I}_{n0} + y_{n1}^j + \gamma_{n1}^j - \sum_{k=2}^{\Gamma_j+L} D_{nk}^{j\alpha}\right]^+ + \sum_n b^j \mathbf{E}\left[-\bar{I}_{n0} - y_{n1}^j - \gamma_{n1}^j + \sum_{k=2}^{\Gamma_j+L} D_{nk}^{j\alpha}\right]^{\cdot +}.$$

Let $\gamma_1^j = \sum_n \gamma_{n1}^{j*}$ be the minimum amount of inventory to hold of item j at the primary warehouse to meet future short response lead time demand at all regional warehouses, $n = 1, ...N$.

Hence, the inventory available to allocate for long response lead time orders that are not yet due is $\bar{I}_{02}^j = [I_{00}^{j\prime} - \sum_n y_{n1}^j - \gamma_1^j]^+$.

Let O_t denote the set of unfilled long response lead time orders that are due to be shipped to a customer in time period $\bar{L}_i + t$. Sub-model 2 was designed to determine which of the orders in O_1 should be shipped to the corresponding regional warehouses in their entirety in period 1. Let us start with period 2, where C_n is the remaining shipping capacity from the primary warehouse to regional warehouse n.

In this sub-model, our goal is to allocate inventory to regional warehouses to satisfy only those orders $i \in O_2$ for which there exists an ample supply at the primary warehouse. Let J be the set of items j for which $\bar{I}_{02}^j \geq \sum_{i \in O_2} a_i^j$. Let $\tilde{O}_2 \subset O_2$ be the set of orders for which $a_i^j > 0$ but only for items $j \in J$, and let $\tilde{O}_2^n \subset \tilde{O}_2$ be the orders that should be shipped to customers directly from regional warehouse n.

Then to determine which orders $i \in \tilde{O}_2$ to fulfill prior to the due date we propose solving for each n for which $C_n > 0$

$$\begin{aligned} \max \quad & \sum_{i \in \tilde{O}_2^n} x_i \\ \text{s.t.} \quad & \sum_{i \in \tilde{O}_2^n} x_i \left\{ \sum_j a_i^j \right\} \leq C_n \\ & x_i = 0, 1 \end{aligned}$$

where, as before, C_n is the remaining shipping capacity. $C_N = \infty$.

These problems have the same form as Problem (13)-(15) discussed in the section on Sub-Model 2 and hence can be solved in the same manner.

After solving this series of problems for $t = 2$, update $\bar{I}_{00}^j \leftarrow \bar{I}_{00}^j - \sum_{i \in O_2} a_i^j x_i$ and $C_n \leftarrow C_n - \sum_{i:n(i)=n, x_i=1} a_i^j$. Solve the problems for $t = t + 1$ and continue until capacity or inventories run out or the end of the maximum grace period is reached.

6.3. The N Primary Warehouse Case

We now extend our results from the one primary warehouse system to one where each warehouse serves as a primary warehouse for some items and as a regional warehouse for all items. Since a customer may order multiple items in the same order, we want to send all ordered items to the customer in the same shipment, if possible. Remember if items in the same order are managed by different warehouses, then the allocation decisions made at one primary warehouse may affect the decisions made at another primary warehouse. It is possible to formulate an exact dynamic program for this coordination problem. However, as in the single primary warehouse case, it is not possible to solve practical problems using this exact formulation. We propose to make allocation decisions using a modification of the approximation approach presented above. The sequence of the approximation methods for the N primary warehouse case is the same as in the single warehouse case. Hence, we will discuss the four sub-models for satisfying backorders, long response lead time orders that must be shipped to desired regional warehouses, replenishment orders and advance long response lead time orders.

First, we divide long response lead time orders into three categories.

1. Let θ_1 denote the set of these orders i for which $m(j) = n(i)$ for all items, that is, $\theta_1 = \{i : m(j) = n(i), \forall j \ni a_i^j > 0\}$. In this case, regional warehouse $n(i)$ serves as the co-located regional warehouse for all items requested in order i.

2. Let θ_2 denote the set of these orders i for which $m(j) \neq n(i)$ for all items, that is, $\theta_2 = \{i : m(j) \neq n(i), \forall j \ni a_i^j > 0\}$. Each item j required to fulfill order i has a primary warehouse, $m(j)$, that differs from the regional warehouse that will fulfill order i, that is $n(i)$.

3. Let θ_3 denote the set of these orders for which $m(j) = n(i)$ for some j and $m(j) \neq n(i)$ for some other j. Thus $\theta_3 = \{i : \exists j \ni a_i^j > 0$ and $m(j) = n(i)$ and $\exists j \ni a_i^j > 0$ and $m(j) \neq n(i)\}$. In this case the fulfilling regional warehouse $n(i)$ is a co-located regional warehouse for some items in order i and receives stocks of the other items from other primary warehouses.

Second, we define order fulfillment time $\bar{L}_i$ for each category of orders. Clearly $\bar{L}_i = 0$ for orders $i \in \theta_1$ and $\bar{L}_i = L$ for orders $i \in \theta_2$. Now, suppose i in θ_3. In this case, we assume that if a primary warehouse allocates units of item j in time period t to satisfy order i and $m(j) \neq n(i)$, then the co-located regional warehouse will set aside a_i^j units for all items j for which $m(j) = n(i)$ in period t as well. This will guarantee that the order will be filled completely by its due date. Hence, for a long response lead time order $i \in \theta_3$, the order fulfillment time is $\bar{L}_i = L$ for all primary warehouses, including the one that also serves as the co-located regional warehouse for items j

for which $m(j) = n(i)$. That is, all relevant primary warehouses must either send out or set aside inventory to satisfy order i in period $\tau(i) - L$.

By employing this strategy, we can address the allocations of inventory and capacity for all orders together using the approach described below.

Since backorders are costly from both financial and customer satisfaction perspectives, inventory is first used to satisfy any outstanding backorders. Sub-Model 1 can be directly used to allocate inventories in the N primary warehouse case.

After making allocations to fulfill backorders, the next priority is to fulfill long response lead time orders that are due to be shipped from the desired regional warehouse upon receipt from the primary warehouses. Some minor modifications may need to be made to Sub-Model 2. When adequate shipping capacity is available, the coordination of the allocation of the units for the items required to satisfy order i is obvious based on the fulfillment time $\bar{L}_i$ for each order. All items j needed to satisfy order i will be sent at the same time from the primary warehouses $m(j)$ to the regional warehouse $n(i)$. Thus Sub-Model 2 can be applied directly in this case. Remember, about 40% of the orders are for a single item type and hence no coordination of allocations is required. As pointed out previously, item j in long response lead time orders i are often shipped to regional warehouse $n(i)$ prior to day $\tau(i) - \bar{L}_i$. Thus there almost always is adequate capacity available on day $\tau(i) - \bar{L}_i$ to ship items of type j to regional warehouse $n(i)$ from primary warehouse $m(j)$ to fulfill orders i for which $a_i^j > 0$ and the delivery due date is $\tau(i)$

When there is not enough capacity available to fulfill all orders $i \in O_J$, where O_J is defined in the discussion of Sub-Model 2, an alternative approach is needed to obtain the desired allocation.

The allocation problem can be stated on

$$\max \quad \sum_{i \in O_J} Q_i x_i \tag{45}$$

$$\text{s.t.} \quad \sum_{i \in O_J^n} x_i \left\{ \sum_{j \in J: m(j)=k} a_i^j \right\} \leq C_{kn}, \tag{46}$$

$$x_i = 0, 1, \tag{47}$$

where C_{kn} is the amount of capacity that exists for shipping items from a primary warehouse k to a regional warehouse n, and x_i is 1 if order i is to be filled completely and 0 otherwise. There is one constraint in this model for each k n combination for which

$$\sum_{i \in O_J^n} \sum_{j \in J: m(j)=k} a_i^j > C_{kn}. \tag{48}$$

In the practical application we studied, a maximum of 20 such constraints could exist.

We can construct a greedy algorithm that can be used to determine the allocation plan. Algorithm 1 as stated in the discussion of Sub-Model 2 would be restated as follows. Let $M(k)$ denote the set of items for which location k is the primary warehouse.

Algorithm 4

Step 0: Rank order $i \in O_J$ according to the values of $Q_i / \sum_j a_i^j$, largest to smallest, and renumber them beginning with 1, and set $i = 1$.

Step 1: For each j for which $a_i^j > 0$, if

$$a_i^j \leq C_{m(j),n(i)} \ \forall j, \text{ set } x_i = 1 \text{ and decrement } C_{k,n(i)} \text{ by } \sum_{j \in M(k)} a_k^j.$$

Step 2: Increment i and return to step 1.

As we noted in our discussion of Algorithm 1, the allocation generated using this greedy method may result in using more capacity than is available. However, since these capacities are not "hard" ones, a small overshoot is acceptable in the real world setting.

Algorithm 2 of Sub-Model 2 can be used in the N primary warehouse environment to allocate items to orders for which there is demand for only a single unit of a single item.

To allocate the remaining inventories to long response lead time orders i that are due on day $\tau(i) = 1 + L$ at a regional warehouse $n(i) \neq m(j)$ when $a_i^j > 0$ and the short and long response lead time orders due on day 1 when $m(j) = n(i)$ and $a_i^j > 0$, we can apply Algorithm 3 presented in our discussion of Sub-Model 2 with one minor change. In Step 1, we would let $u_i^j = \min \left\{ a_i^j, I_{m(j),0}^{j\prime}, C_{m(j),n(i)} \right\}$,where $I_{m(j),0}^{j\prime}$ is the remaining initial inventory at primary warehouse $m(j)$ for item j and $C_{m(j),n(i)}$ is the remaining shipping capacity from $m(j)$ to $n(i)$. Other obvious subscript changes are required as well in Algorithm 3.

Next, we allocate replenishment stocks to non-co-located regional warehouses. This allocation is made for each item j considering the remaining initial stocks at $m(j)$ and the remaining shipping capacities to each regional warehouse. Except for minor changes to the subscripts, the approach taken in our discussion of Sub-Model 3 holds in the N primary warehouse case.

After making all the allocations of stocks to satisfy backorders, fulfill orders due today, and replenish regional warehouse stock levels, our final priority is to determine, which, if any, long response lead time orders to fulfill in advance given the remaining inventories and shipping capacities. These allocations are made using the approach presented in our earlier discussion of Sub-Model 4 combined with our discussion concerning the modification we must make to Sub-Model 2 in the N primary warehouse case.

Note that our fulfillment strategy for $i \in \theta_3$ may not be an optimal one. We may prefer to use the allocated inventory of an item j, $m(j) = n(i)$, that has been set aside to fill short response lead time orders incurred during the shipping lead time. For example, suppose 3 units of an item are needed to completely satisfy order i. Then, by policy, these 3 units are set aside, or allocated, to order i. When the other items j, with $m(j) \neq n(i)$, arrive at $n(i)$, these 3 units will be withdrawn from stock and, along with those items arriving from the other warehouses, will be shipped to the customer. During the L-period lead time, these 3 units will still be physically on the shelf of warehouse $n(i)$. However, they could be needed to satisfy 3 single unit, single item short lead time orders completely that arise these during these L periods. If our objective is to maximize the number of orders satisfied on time, then by allocating the 3 units to order i would result in more orders being unsatisfied on their due date than necessary. While this case is possible, recall that only 2% of the orders in the environment we examined contained items for which $a_i^j > 1$. Hence the approach we have taken is a reasonable one to follow.

Thus we have demonstrated how the four sub-models introduced in Section 6.2.2 can be modified to create the N-primary warehouse allocation and order fulfillment plan.

7. Order Fulfillment

Each day a decision must be made pertaining to the actual fulfillment of outstanding customer orders at each regional warehouse. In the allocation discussion in Section 6, we created a tentative order fulfillment plan. Since more short response lead time demand information is revealed during the transportation lead time, a final order fulfillment plan needs to be created. We now show how to make these final daily order fulfillment decisions at each regional warehouse.

The objective of the daily order fulfillment process is first to satisfy the backlogged orders and then to satisfy orders that must be sent out that day. Of course, there may not be enough inventory on hand to achieve this objective. Then, based on the remaining on-hand inventory and long response lead time demand information, we will decide which orders will be satisfied in advance.

7.1. Assumptions And Nomenclature

We begin our order fulfillment model development by stating our assumptions concerning the operation of the fulfillment system and introducing some nomenclature.

The inventory available to fulfill customer orders each day is a result of allocation decisions made in previous periods. Since each warehouse can only use its own on-hand inventory to fulfill customer orders, we focus on one regional warehouse at a time. Consequently, we drop the subscript n that designates the regional warehouses. Recall that a regional warehouse serves as a primary

warehouse for some items. Let M denote this set of items. As before, assume we are in period 1. Let S_t^j denote the allocation of item j that is scheduled to arrive in period t at the regional warehouse, $t = 1, \cdots, L$. These allocations include inventory to meet both short and long response lead time demands.

We first construct a list of outstanding orders that need to be satisfied by the end of each period in the planning horizon. The order information is then updated when a customer order is received or is filled either completely or partially. For each order i, we record the period $\tau(i)$ by which the order must be sent from the local warehouse to the customer.

We make the order fulfillment decisions in the following manner. Let O_0 denote the orders i that are backordered at the beginning of period 1, and O_1 denote the orders that are due in the current period, that is, $\tau(i) = 1$. Consistent with the allocation strategy, fulfillment priority is given first to the backlogged orders. Remember, in formulating the allocation model, we stated that all backorders would be sent directly to the customer from the primary warehouse via express delivery. Hence, the backlogged orders only include items belonging to set M. Then we will use the remaining available inventory to satisfy orders that must be sent from the regional warehouse today. Among these orders, we first determine which of them can be fulfilled completely. Then we use the available inventory to satisfy the remaining orders partially. In the last step, we determine which of the orders that are not yet due can be fulfilled in advance after setting aside stocks needed to ensure satisfaction of short response lead time demand over the regional warehouse's replenishment lead time. We examine orders based on their due date sequentially. Remember, we do not fill an order in advance if we cannot fulfill it fully.

Let x_i denote whether or not order i is fulfilled completely in the current time period.

$$x_i = \begin{cases} 1, & \text{if order } i \text{ is fulfilled completely in the current period,} \\ 0, & \text{otherwise.} \end{cases} \tag{49}$$

7.2. Two Step Order Fulfillment Models

We next develop two models to make order fulfillment decisions. The first one is employed to determine which orders to fill completely and the second one is developed to determine how to allocate inventory to partially satisfy an order. We state the models corresponding to a generic set of open orders, O. Then we show how the models are executed to make the actual fulfillment decisions.

7.2.1. Model 1: Complete Order Fulfillment Given a set of open orders O, let the inventory levels available to fulfill orders in set O be denoted by $\tilde{I}_j$, the available stock of item j following the receipt of the S_1^j units. To satisfy as many orders in O as possible, solve

$$\max \sum_{i \in O} x_i \tag{50}$$

$$\text{s.t.} \sum_{i} a_i^j x_i \leq \tilde{I}_j, \qquad \forall j \text{ s.t. } a_i^j > 0 \text{ for some } i, \tag{51}$$

$$x_i \in \{0,1\}. \tag{52}$$

Rather than solving this very large scale integer program directly, we employ the approach used in Sub-Model 2. There will likely be hundreds of thousands of variables in this formulation along with potentially many tens of thousands of constraints.

As we did in Sub-Model 2, we first find the set J. Item $j \in J$ if there is sufficient supply to satisfy all demand in set O, that is, $\sum_i a_i^j \leq \tilde{I}_j$. Based on set J, define set $O_F \subset O$ to be the set of orders that contain items only in set J, that is, $i \in O_F$ only if $a_i^j > 0$ and $j \in J$. Since every order $i \in O_F$ can be fulfilled completely, $x_i = 1$ for $i \in O_F$. The available inventory to fulfill the remaining orders is $\tilde{I}'_j = \tilde{I}_j - \sum_{i \in O_F} a_i^j$. Then the set of remaining orders to be filled is $O' = O \setminus O_F$.

Next, as in our approach to allocating inventories in Sub-Model 2, we focus on the single item and single unit orders, since our goal is to satisfy as many orders completely as possible. Therefore, in this step, determine the orders $i \in O'$ for which $a_i^j > 0$ for only one item j and $a_i^j = 1$ for that item. We satisfy orders depending on the value of Q_i, beginning with the largest value of Q_i and continuing until either all such orders are satisfied or there is no remaining inventory. Let $x_i = 1$ if order i is satisfied and add order i to set O_F. The remaining supply is $\tilde{I}''_j = \tilde{I}_j - \sum_{i \in O_F} a_{ij} x_i$.

The rest of the order fulfillment decisions are made by solving the following integer program

$$\max \sum_{i \in O \setminus O_F} x_i Q_i \tag{53}$$

$$\text{s.t.} \sum_{i} a_i^j x_i \leq \tilde{I}''_j, \qquad \forall j \text{ s.t. } a_i^j > 0 \text{ for some } i, \tag{54}$$

$$x_i \in \{0,1\}, \qquad \forall i \in O \setminus O_F, \tag{55}$$

where Q_i, as before, represents the incremental cost of shipping order i one unit at a time from the appropriate primary warehouses.

This problem will have far fewer variables than does problem $(55) - (57)$ for two reasons. First, most items have enough inventory to satisfy demand throughout a cycle except possibly at the end of a cycle. This is the case since the probability of running out stock during a cycle is usually less than 0.03 in the environment we studied. Therefore, most orders are satisfied in step 1. Second, recall that only a single item is requested in about 40% of the orders and about 98% are for a single unit of an item. Nonetheless, there may be several thousand orders that remain in this model. Hence, we determine orders $i \in O/O_F$ to fill completely using a greedy heuristic Algorithm 5.

Algorithm 5

Step 0: Sort the orders $i \in O/O_F$ according to the ratio $\frac{Q_i}{\sum_j a_i^j}$, largest to smallest. Renumber the orders, beginning with the order with the largest ratio, which is designated as order 1. Set $i = 1$.

Step 1: If $a_i^j \leq \tilde{I}_j'' \ \forall j$, such that $a_i^j > 0$, set $x_i = 1$. Otherwise, set $x_i = 0$ and go to step 2.

If $x_i = 1$, set $\tilde{I}_j'' \leftarrow \tilde{I}_j'' - a_i^j \ \ \forall j$ such that $a_i^j > 0$ and set $O_F \leftarrow O_F \cup \{i\}$.

Go to step 2

Step 2: If all orders have been evaluated, stop; otherwise, increment i and return to step 1.

This heuristic provides a computationally tractable method for determining which of the remaining orders should be filled completely. Of course, the resulting allocation need not be optimal.

After executing the heuristic, we may have some remaining inventory and a list of unfilled orders $O_P = O \setminus O_F$. There is not enough inventory to fill any order $i \in O_P$ completely with the remaining inventories. We will use the remaining supply to partially fill a portion of these remaining orders.

7.2.2. Model 2: Partial Order Fulfillment We now construct an integer program to determine how to assign the remaining inventory to the unfilled orders O_P. The goal is to send out as many requested items as possible while keeping the shipping costs low. Recall that the last-mile shipping cost is charged based on the volume or weight of the shipment. The greater the fraction of the volume or weight shipped produces a lower per unit transportation cost and hence is desirable. We choose to maximize the sum of the percentage of volumes satisfied among all orders as a surrogate for minimizing the incremental shipping costs.

Let y_{ij} represent the number of units of item j allocated to order i. There are two types of constraints in our model. One limits the amount of an item that can be allocated to the orders and the other limits the amount allocated to a particular order. Then our model is

$$\max \sum_i \left\{ \frac{\sum_j v_j y_{ij}}{\sum_j v_j a_i^j} \right\} = \sum_i \sum_j \frac{v_j}{V_i} y_{ij} \tag{56}$$

$$\text{s.t.} \sum_{i \in O_P} y_{ij} \leq \bar{I}_j, \tag{57}$$

$$0 \leq y_{ij} \leq a_i^j, \tag{58}$$

$$y_{ij} \text{ is an integer}, \tag{59}$$

where v_j is the volume(weight) of item j, $V_i = \sum_j v_j a_i^j$ is the total volume(weight) of order i and $\bar{I}_j$ is the remaining stock of item j.

Note, we can relax the constraint that y_{ij} is an integer due to the structure of the problem. An optimal integer solution can be obtained using a greedy algorithm and executed one item at a

time. Hence, the execution of the optimization process can be carried out efficiently in a parallel manner.

7.3. Order Fulfillment Execution

In the previous two sections, we introduced two order fulfillment models. We now summarize how the two models are executed when making the order fulfillment decisions.

As we have mentioned, orders that are due at different dates are ranked with different priorities. Therefore, we start with the backlogged orders O_0 and execute both models to satisfy these orders. The remaining unfilled items continue to be backlogged. Then we execute the two models for the orders that are due today, that is, those that are in set O_1. The remaining unfilled items in these orders are added to the backlogged lists for the primary warehouses that manage the items. Next, we execute Model 1 for orders in sets $O_2, \cdots, O_R$ sequentially, where $R = \max(R_j^{\max})$. Model 2 is not executed for orders in $O_2, O_3, \cdots, O_R$ since we do not fill orders partially before they are due. When executing Model 1 for the future orders, we need to ensure that inventories are not used to the detriment of satisfying future short response lead time orders at each warehouse. We employ the method described in the section on Sub-Model 4 to determine how much inventory to hold in reserve.

8. Final Remarks

A few comments are in order pertaining to the operational environment in which our approach is intended to be applied.

We have stated that it is essential to employ an information system, a transportation system and operating practices that ensure the effective coordination of picking and transportation activities across warehouses. If such systems and practices are not in place, the models and the approach to fulfilling customer orders discussed in this and our companion paper would not be possible to implement in a cost effective manner.

The company with which we worked had systems in place to manage all logistics activities pertaining to each warehouse. Before employing the PWS, each regional warehouse operated essentially as an independent entity. That is, picking was done in waves at each warehouse. All items needed to fulfill a collection of orders were picked in each wave. The picked items were then sorted by customer order and then usually placed in one box, thereby lowering the per pound freight costs. These waves were determined separately for each warehouse. Unfortunately, only about 50% of the orders could be filled from stocks in the regional warehouses closest to the customer.

In the PWS system, the timing of picking had to be coordinated so that picking to fulfill a given long response lead time order would occur at the appropriate time in each affected regional

warehouse. Hence, significant changes to the information system and to the picking and packing practices were needed. We will briefly describe some of the key changes that were required.

Items sent to other regional warehouses were placed in totes in the revised system. Each tote had an ID tag. The contents of each tote were known, since each item placed in a tote was scanned. Thus the contents of each tote received at another regional warehouse were known. Specifically, the items in the tote were known as well as the customer order to which they were allocated. Replenishment stocks sent to a regional warehouse were also placed in totes; however, these totes were a different color than the totes containing items that were to be to cross-docked. Totes that contained inventory that would be cross-docked were loaded in sequence on trucks so that they would be removed first. The sequence was determined so that items needed to fulfill an order would be removed from trucks and joined with items picked within warehouse $n(i)$ at approximately the same time. In essence, the types of movement of the items from bins in primary warehouses to the regional warehouse from which the customer orders were to be fulfilled were very similar to the way they were carried out initially. Conceptually, the transportation on trucks of items from a primary warehouse to a regional warehouse was an extension of their movement within a warehouse. It just took longer to go from picking to order consolidation. By coordinating the unloading of trucks made cross-docking possible.

The primary warehouse for an item was selected largely on the location of the supplier. In total, the cost of inbound freight plus transshipped material and last mile shipping costs were less than the transportation costs were initially. Furthermore, the predictability of the lead times improved dramatically due to improved collaboration with suppliers and logistics providers. This resulted in lower safety stock requirements. Since suppliers ship to only one warehouse for each item, cycle stocks were reduced substantially. In total, costs were lower, inventory levels were lower, the number of on-time shipments increased and orders were delivered in fewer packages.

References

Chan, E. (1999). *Markov Chain Models for Multi-echelon Supply Chains.* School of Operations Research and Industrial Engineering, Cornell University 14850. Dissertation.

Clark, A. and Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management Science*, 6:475–490.

Eppen, G. and Schrage, L. (1981). Centralized ordering policies in a multi-warehouse system with lead times and random demand. *Management Science*, 30:69–84.

Federgruen, A. and Zipkin, P. (1984). Approximations of dynamic, multiocation production and inventory programs. *Management Science*, 30(1):69–84.

Gallego, G. and O.Ozer (2003). Optimal replenishment policies for multiechelon inventory problems under advance demand information. *Manufacturing and Service Operations Management*, 5(2):157–175.

Glasserman, P. (1997). Bounds and asymptotics for planning critical safety stocks. *Operations Research*, 45(2):244–257.

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

Harihar, R. and Zipkin, P. (1995). Customer-order information, leadtimes, and inventories. *Management Science*, 41:1599–1607.

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

Jackson, P. and Muckstadt, J. (1989). Risk pooling in a two-period, two-echelon inventory stocking and allocation problem. *Naval Research Logistics*, 31(1):1–26.

Kunnumkal, S. and Topaloglu, H. (2008). A duality-based relaxation and decomposition approach for inventory distribution systems. *Naval Research Logistics Quarterly*, 55(7):612–631.

Kunnumkal, S. and Topaloglu, H. (2011). Linear programming based decomposition methods for inventory distribution systems. *European Journal of Operational Research*, 211(2):282–297.

Li, J. and Muckstadt, J. (2013a). *Fulfilling Orders in a Multi-Echelon Capacitated On-line Retail System: PART ONE, Planning Inventory Levels.* School of Operations Research and Information Engineering, Cornell University 14850. Technical Report, No. 1483.

Li, J. and Muckstadt, J. (2013b). *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. Technical Report, No. 1482.

Muckstadt, J., Murray, D., and Rappold, J. (2001). *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.

Muckstadt, J. and Sapra, A. (2009). *Principles of Inventory Management: When You Are Down to Four, Order More.* Springer, 1 edition.

Ozer, O. (2003). Replenishment strategies for distribution systems under advance demand information. *Management Science*, 49(3):255–272.

Roundy, R. and Muckstadt, J. (2000). Heuristic computation of periodic-review base stock inventory policies. *Management Science*, 46(1):104–109.

Wang, T. and Toktay, B. (2008). Inventory management with advance demand information and flexible delivery. *Management Science*, 54(4):716–732.