MANAGING INVENTORY IN LARGE SCALE MULTI-ECHELON CAPACITATED FULFILLMENT SYSTEMS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Juan Li

August 2012

MANAGING INVENTORY IN LARGE SCALE MULTI-ECHELON CAPACITATED FULFILLMENT SYSTEMS

Juan Li, Ph.D.

Cornell University 2012

When designing and operating an order fulfillment system for an on-line retailer, many factors must be taken into account. We begin by discussing these factors and by proposing an architecture for a fulfillment system. The main goals of this dissertation are to construct a planning model and an execution model for managing inventory levels for each item at each stocking location in the fulfillment system. The planning model represents the system's operating architecture, the order response time requirements associated with customer orders and other operational constraints. This type of model is used in the sales and operations planning activities that are undertaken by an on-line retailer. It is not primarily intended to make daily procurement and allocation decisions. Rather, the planning model is designed to assist management when it makes warehousing, transportation, budgeting and collaboration decisions. The execution models are designed to make daily operational decisions. They indicate when inventories should be procured, how these inventories should be allocated throughout the warehousing network structure, and the timing of fulfilling customer orders. Real on-line retail systems contain millions of items. We describe an algorithm for making these execution decisions for each item at every location within seconds. Hence, both the planning and execution models we propose are of practical value. We report computational results for certain of these algorithms as well.

BIOGRAPHICAL SKETCH

Juan Li was born in Chengdu, China in 1985 and moved to Zhenjiang after a few years. In May 2007, Juan received her Bachelor of Arts degree from Smith College and continued to pursue a doctoral degree in Operations Research at Cornell University in the same year.

Dr. Li's dissertation was supervised by Dr. John Muckstadt. In her thesis, Dr. Li worked on large scale inventory management strategies. She developed scalable optimization models to determine inventory planning and execution strategies for tens of millions of products that are stocked at multiple warehouses at different locations.

Dr. Li will join Xerox Research Center in Webster upon graduation.

This thesis is dedicated to my family for their love, encouragement and endless support.

ACKNOWLEDGEMENTS

I am truly grateful to my advisor, Professor John Mucksdtadt, for his guidance and patience during my study at Cornell. My dissertation would not be possible without his support. I would also like to thank my committee members, Dr. Sidney Resnick and Dr. Huseyin Topaloglue for their instructions and support. In addition to my committee, Dr. Peter Jackson and Dr. Kathryn Caggiano also provided valuable advice to me during my study at Cornell.

Most importantly, I would like to thank my family. My parents, Anying and Long, supported me with their faith in me and unconditional encouragement along these years. I thank my husband, Hao, for his commitment and companionship. He provided the love, support and motivation that kept me going with all the challenges I have faced.

Lastly, I would like to thank my colleagues and friends here at Cornell. They have made my past five years exciting and memorable.

TABLE OF CONTENTS

LIST of TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

E-Commerce has grown rapidly during the past decade. Nowadays, on-line retailers sell products or services directly to customers through internet or other computer network mechanisms. According to United States Census Bureau, E-Commerce retailers accounted for $193.8 billion in sales within the United States in 2010 [12]. The importance of these enterprises in the US economy will continue to increase in scale and sales in the future. Compared with their brick and mortar rivals, on-line retailers can potentially offer more products to customers. Currently, large on-line retailers offer tens of millions of different products to customers anywhere in the United States. In addition, by providing the products and services over the internet, on-line retailers have the opportunity to serve customers living in different geographic regions without having a physical presence in these areas.

To satisfy customer demand throughout the United States, in a timely and cost effective manner, on-line retailers have constructed warehouses in different regions of the country. Deciding what inventories in what quantities to stock in which warehouses and how to fulfill each customer's orders is critical to the operational and financial success of an on-line retailer. For example, in their annual report in 2011, Amazon.com pointed out that "If we do not adequately predict customer demand or otherwise optimize and operate our fulfillment centers successfully, it could result in excess or insufficient inventory or fulfillment capacity, result in increased costs, impairment charges, or both, or harm our business in other ways [1]." An order fulfillment system must be designed so that there is enough warehouse capacity to stock all the inventories to meet

an uncertain and fluctuating demand over time.

Besides warehouse location and inventory considerations, Amazon.com and others are concerned with shipping costs. One of the major costs of operating the fulfillment system is the cost to ship to the customers. In their analysis, Amazon.com has shown that many orders are fulfilled from a warehouse or an external source that is not the most economically desirable one. This increases transportation costs significantly. These costs include the so-called "last mile shipping costs." These costs are much higher when measured on a per pound basis than shipping in full truck loads. The design of the fulfillment system will substantially affect transportation costs.

These factors, and several others, are considered in the models that we develop in this thesis. We note that these models are based on our extensive analysis of one major on-line retailer.

There are two primary goals in this thesis. First, based on an on-line retailers fulfillment system's operating characteristics and multi-echelon architecture, we present a model that can be used to set inventory levels at each location. This model is intended to be used as part of the retailer's sales and operations planning activity. It is not primarily intended to be used to make daily purchasing and allocation decisions, although the stock levels that result from solving this model guide such decisions. Planning models, such as the one we are proposing, assist in planning warehousing and transportation operations, budgeting for inventories and evaluating the consequences of collaborative activities with suppliers. Collaboration is essential for establishing expectations for quantities of each item that will be purchased, the timing of procurement orders placed for items, the supplier to on-line retailer lead times and purchasing

costs as well as other financial and operational issues. The output of the collaborative process yields inputs to the planning model. These models are normally executed on an as-needed basis. The time horizon for these planning models is normally several months to over a year in length.

In our discussion of the planning models, we also show how to set inventory levels efficiently. In the type of on-line retail system that we have analyzed in recent years there are more than a million items that would need to be managed using this algorithm. Computing inventory levels in a timely manner is a challenge given the scale and complexity of the system. Determining optimal stock level is not possible for reasons we will discuss. Consequently, we will focus on describing an approximation approach for setting stock levels. This is a major contribution of this thesis.

Second, we develop a set of models that are the basis for making daily procurement and allocation decisions. The time horizon considered in these models is typically on the order of 10 to 14 days in length. In the planning model we were largely concerned with setting inventory levels given the architecture and constraints found in the fulfillment system. In our execution models we additionally focus on fulfilling individual customer orders. These orders may consist of many different items. The quantity ordered for a single item may be greater than one unit, too. Orders also have different due dates associated with them. Thus all procurement and allocation decisions represented in these execution models reflect the detailed content of existing customer orders. The development of the models is the second major contribution of this thesis.

The organization of this thesis is as follows. After discussing many of the system's operating characteristics in Chapter 2, we describe the fulfillment sys-

tem's architecture. In Chapter 3, we review literature relevant to the problem we are studying. We then present an exact planning model for setting stock levels for each item at each warehouse in Chapter 4. This model is formulated as a dynamic program that cannot be solved directly. Thus, as mentioned, we construct an approach for computing these stock levels that is scalable and that is applicable to the planning activity in the real application that motivated this research. We also discuss our numerical experiments using the proposed algorithm in Chapter 4. In Chapter 5, we discuss the execution models and tractable methods for making order fulfillment decisions for real world applications.

CHAPTER 2

BACKGROUND

We begin this chapter by describing some important attributes we observed in an on-line retailer's fulfillment system.

Once a customer places an electronic order, a fulfillment process is put into motion. This fulfillment process has two attributes. First, a plan for how the item or items will be sourced for shipment to the customer must be developed. Second, the timing of the order's fulfillment must be made consistent with the customer's desires. The customer may request and possibly pay extra for immediate delivery of the order. But it is also possible for the customer to delay shipment because the shipping cost will be waived or reduced.

The timing is of particular importance since customers sometimes order more than one product and sometimes for multiple units of the same product. To understand the customer order patterns, we analyzed data provided to us by an on-line retailer. We first performed a Pareto analysis of the number of items in each customer order and we then studied the number of units of an item that are in an order. The results are shown in Figure 2.1 and Figure 2.2, respectively. The graph in Figure 2.1 shows that more than 65% of orders contain only one or two items. More than 97.5% of the total orders contain fewer than 10 items in an order. The graph in Figure 2.2 shows that even though customers occasionally order more than one unit of an item, more than 97.5% of the orders are for one unit of the item. These observations helped guide the construction of the algorithm we designed to solve our execution model.

Although there are millions of items available for purchase from an on-line

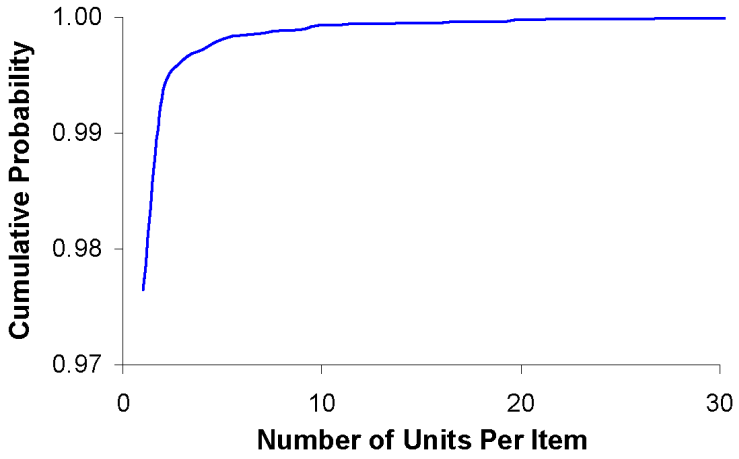

Figure 2.1: Pareto Analysis of Items in An Order

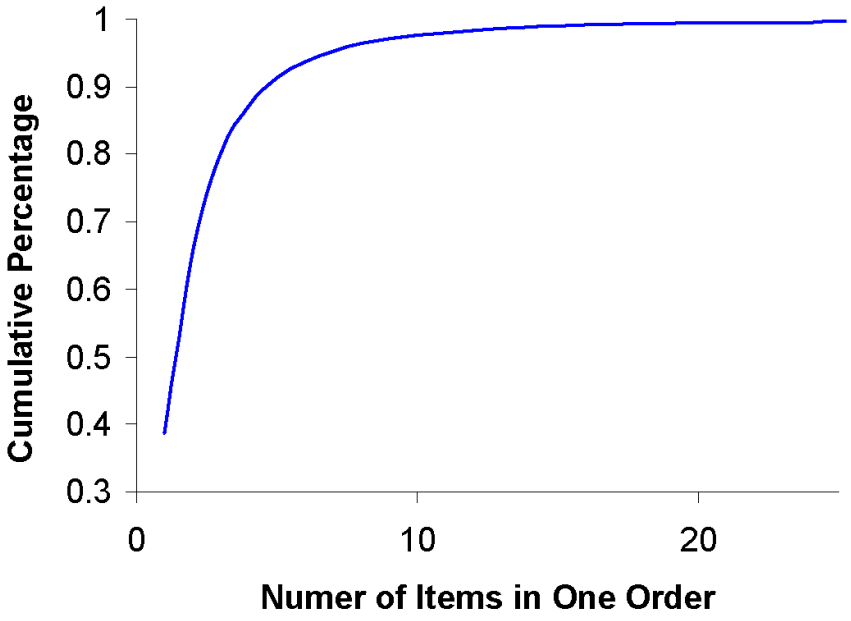

Figure 2.2: Pareto Analysis of Units for Each Item Ordered

retailer, most items have very low demand rates. Figure 2.3 contains a Pareto-like graph for unit sales for the on-line retailer we studied. Most items have annual demands of four or fewer units. These items may be 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 text books are examples of such items. These items are also normally stocked in a single location. As indicated in the graph in Figure 2.3, 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. As indicated in Figure 2.3, 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 thesis. 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. Newly introduced items are referred to as *frontlist* items. Tablets, such as Kindles, are an example of such an item. The demand processes are not stationary for many if not most of these frontlist items. Items that have been marketed in the past are called *backlist* items and often have more stationary demand processes. The graph in Figure 2.4 illustrates how annual daily demand varies through out a year for many backlist items.

As we mentioned, to deal with the large volume of items that are ordered and the widespread geographical locations in which customers live, the fulfill-

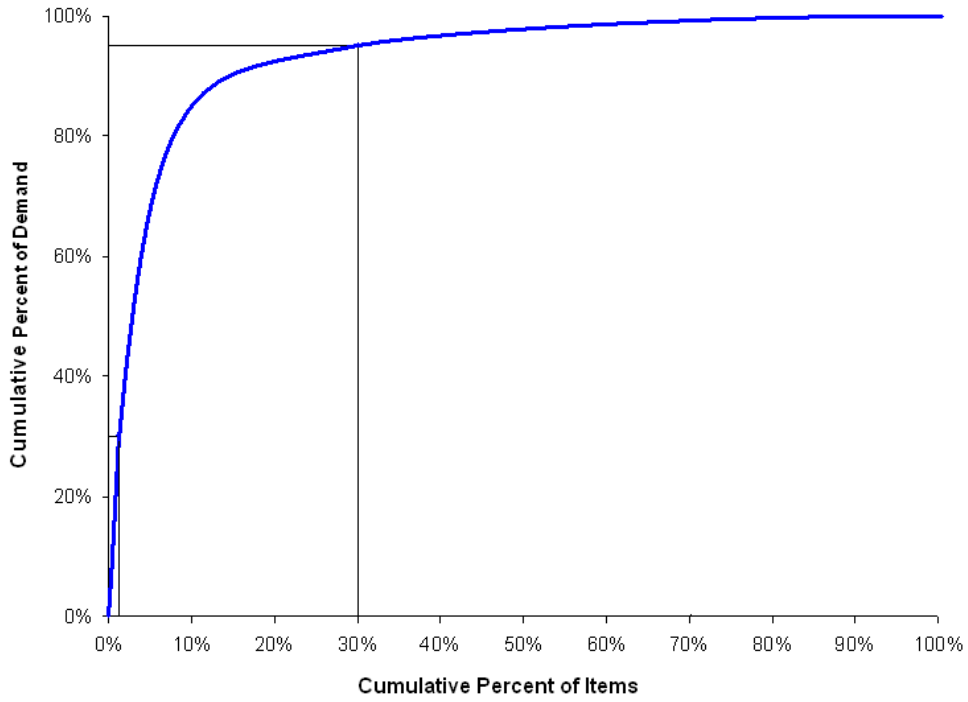

Figure 2.3: Pareto Analysis of an Online Retailer

ment system operated by an on-line retailer contains many warehouses. There are alternative ways to operate a multiple warehouse system.

One way is to have each warehouse operate independently of all the other warehouses. All but the lowest demand rate items are stocked in every warehouse. Demand forecasts for each item would be generated for the geographical region supplied by a regional warehouse. Each regional warehouse would place replenishment orders with the external suppliers. This design has two main shortcomings. First, forecast errors are higher on a regional basis than on a national basis for all items. Second, to take advantage of supplier discounts, regional warehouses often order and stock more inventory than is desirable on a national level. Hence, cycle stocks would become too large. Furthermore, since

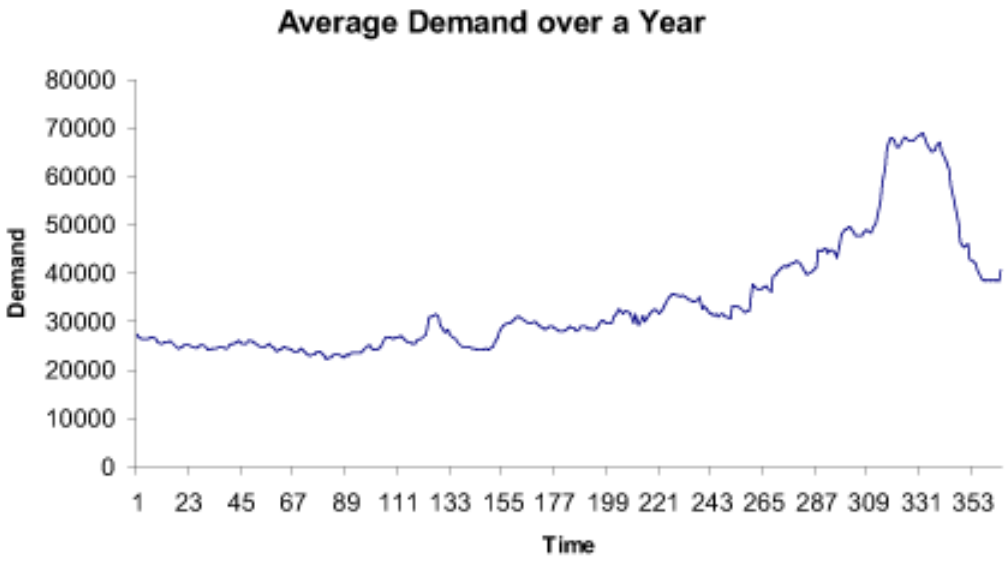

Figure 2.4: Average Daily Demand Over a Year

forecasts of demand are inaccurate, excess stock would accumulate to compensate large forecast errors. The large on-line retailer we studied managed its inventories in this manner at the beginning of our engagement and did accumulate a large amount of excess inventory.

To minimize the risk of over-stocking, the multi-echelon fulfillment system could be operated in a different way. Again, suppose there are many warehouses in the fulfillment system. Each item that was previously managed in each of these regional warehouses will now be managed through a single warehouse, which we call 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 workload and recognizing facility capacities also affects the number of items managed by each warehouse. Once

a primary warehouse is selected for an item, that warehouse becomes responsible for procuring inventory from an external supplier. The supplier then ships the amount procured to the designated primary warehouse. The primary warehouse then distributes inventory to the other warehouses, which we will 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 many hundreds of thousands of items. This structure is similar to the one studied by Eppen and Schrage (1981), although the inventory control policies differ significantly. Figure 2.5 is an example of such a PWS with 5 warehouses.

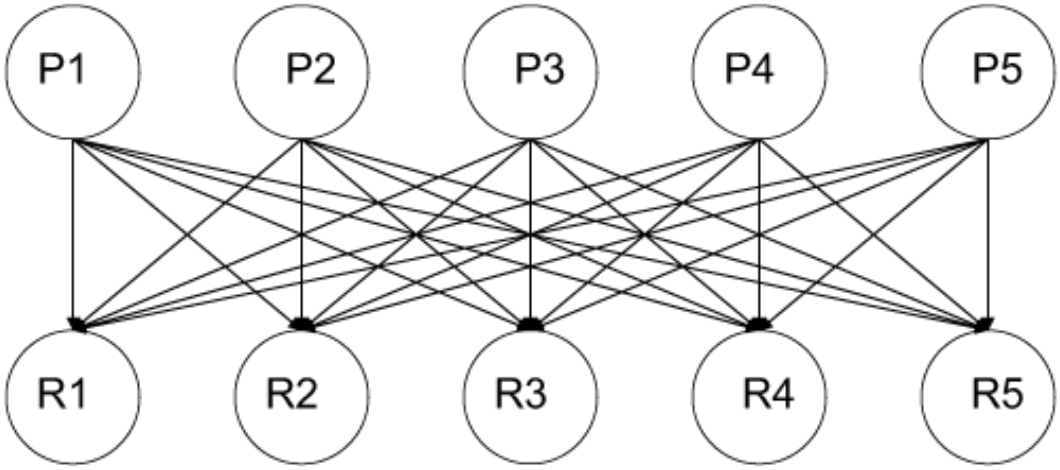

Figure 2.5: PWS with 5 Warehouses

Each primary warehouse is conceptually thought of as two entities, although there is only one physical entity. One entity performs the procurement 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. In the discussion of our models, we will call this virtual

warehouse the co-located warehouse. Since this co-located warehouse is physically located in the same facility as the primary warehouse, shipping to it is assumed to occur instantaneously.

Recall that when customers place orders, they also request a delivery response time. Within the PWS, we categorize demands as either short or long response lead time demands. Orders are satisfied from the regional warehouse closest to the customer's delivery address to minimize last mile transportation costs, which, as we observed earlier, are a substantial component of the retailer's operating cost. Suppose a customer places an order that will be fulfilled from a particular regional 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. As mentioned, the primary warehouse must carry inventory so that it can replenish each regional warehouse's inventory. We note that for the on-line retailer that we have examined, short response lead time demand usually accounted for between 13% – 20% of the total demand for an item. The percentage changes by item and sometimes by the time of the year.

The primary warehouse is also responsible for satisfying long response lead time demand occurring anywhere in the system. Consequently, the primary warehouse must also stock inventory to meet these long-response lead time demands. When a long-response lead time demand occurs, the primary warehouse sends inventory to the regional warehouse that is responsible for satisfy-

ing the customer's order. When inventory is shipped to a regional warehouse from the primary warehouse to satisfy long response lead time demand, the inventory is not placed in the regional warehouse's bins but rather is cross-docked and sent directly to the customer.

There is, in practice, a constraint on the amount of inventory that can be shipped each day to a regional warehouse from the primary warehouse. This constraint exists due to truck capacity, equipment limitations, and the availability of labor. This capacity changes over time as demand patterns change. Since transportation is a major cost, the on-line retailer wants to keep these costs relatively low. Hence, shipments to a regional warehouse from a primary warehouse are normally made in full truck loads so that long haul transportation costs are minimized. As a result, the on-line retailer wants to limit the volume shipped to a regional warehouse per day. The on-line retailer was concerned about the effect the proposed delayed allocation system architecture would have on the goal of limiting the amount shipped daily and the effect on meeting customer delivery requirements in a timely manner. We address these concerns directly.

Based on the attributes we have discussed, we first develop a planning model that can be used to establish target inventory levels for each of the high demand rate items consistent with the fulfillment system's architecture and operating characteristics. Recall that we focus only on items that will have multiple procurements made over the course of a few months and that have relatively high demand rates. Other models can be employed for low demand rate items or the very high demand rate items.

CHAPTER 3

LITERATURE REVIEW

There are four streams of literature that are related to our paper's content. The first pertains to the effect of pooling inventory. Some examples are Eppen and Schrage (1981), Jackson (1988) and Jackson and Muckstadt (1989). The second stream of literature pertains to the discussion of the so called balance assumption. This stream of literatures starts from the seminal paper of Clark and Scarf (1960) and later is extended by Federgruen and Zipkin (1984), Kunnumkal and Topaloglu (2008 and 2011). The third topic focuses on solving capacitated inventory system problems using the shortfall process, such as Roundy and Muckstadt (2000), Muckstadt, Murray and Rappold (2001), Glasserman (1997) and Glasserman and Tayur (1994). Last but not least, the long-response lead time items can be viewed as having advance demand information. There are a sequence of papers that focus on this area, including Hariharan and Zipkin (1995), Gallego and Özer (2003), Özer (2003) and Wang and Toktay (2008). We will discuss these streams 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, Murray and Rappold (2001). They discussed the "No B/C strategy" in two scenarios. Our approach is similar in the sense that the long-lead time demand can be viewed as the 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 item, which is stocked at all regional warehouses across the country. 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 need to have inventory on hand to satisfy the demand.

We assume the primary warehouse places orders from an external supplier under a fixed schedule. Eppen and Schrage (1981) 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. No stock is held at the central warehouse in their model. In our system, this is not the case. Jackson (1988) built both an exact cost model and a computationally tractable approximation cost model to a 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 (1989) extends this idea to a two-echelon, two-period allocation problem.

Our multi-echelon inventory model is based on Clark and Scarf (1960)'s echelon inventory position concept. They found that the difficulty of solving a distribution system type of problem is 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 when the inventories are allocated. 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 (1984) obtain a lower bound on a value function by relaxing the imbalance constraints. Kunnumkal and Topaloglu (2008, 2011) associated Lagrangian multipliers with the balance constraints. By introducing the Lagrangian multipliers, the resulting relaxed problem can be easily solved. However, the balance constraints may be violated. Kunnumkal and Topaloglu (2011) discussed several approximation methods that can be used to select a good set of multiplier values. They also show that the 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.

In this paper the shipping capacity from the primary warehouse to a non-co-located regional warehouse is limited. Many papers have studied capacitated systems. Glasserman and Tayur (1994) discussed conditions under which a capacitated system is stable. They found that a multi-echelon inventory system that operates under a base-stock inventory policy is stable if the mean demand per period is smaller than the capacity at every regional warehouse. Muckstadt, Murray and Rappold (2001) used a shortfall process approach to develop a cost model. In this thesis, when demand is stationary, we extend the idea to the case when the system does not need to satisfy the demand immediately. This approach is usually implemented by finding the transition matrix of the shortfall process and then finding the stationary distribution. This approach is inappropriate when the mean demand is large. Glasserman (1997) and Roundy and Muckstadt (2000) found that the shortfall process can be approximated as a mass exponential function. We will use this idea to approximate the probability distribution of the shortfall process.

As mentioned earlier, the system provides two levels of service, which are the short response lead time demand and long-response lead time demand. As mentioned, 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 (1995) 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 (2003) and Özer (2003) 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 (2008) extend this conclusion to the flexible delivery case. In this paper, we use an alternative approach to analyze the problem by employing the shortfall process to model flexible delivery.

CHAPTER 4

PLANNING MODEL

We will develop our planning model in this chapter. As mentioned, we first construct an exact model, which is formulated as a dynamic program. This model cannot be solved for realistic problems due to the size of the state space. Thus we next construct an approximation model and solution method. Finally we present some computational results.

We begin our model development by stating our assumptions concerning the fulfillment system's operation and by introducing some nomenclature. Additional nomenclature will be 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. Let a period be 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.

The frequency of placing procurement orders depends on several factors. One factor is cost. Costs include fixed and variable procurement costs. The fixed costs arise when placing an order, shipping it and receiving it at the primary warehouse, and putting away inventory at both the primary warehouse and the external supplier. Minimum buy quantity requirements and quantity discounts also influence the frequency of placing orders for an item. While it may be desirable from an inventory holding cost perspective to order frequently, the desire to limit and smooth workloads at the primary warehouse and the external suppliers greatly influences the frequency at which orders are placed and

received. Thus, policies employed for managing inventories must be designed to reflect these costs and constraints.

To manage both costs and workloads, we assume for planning purposes that 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 weekly, monthly and sometimes quarterly. 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 is in itself an interesting problem. We will briefly outline an approach that can be used to create such a schedule.

First, establish the frequency of ordering each item. Employing a power-of-two policy makes the timing of the procurement decisions easier to manage. The solution indicates whether an item should be procured weekly, every two weeks, every month, etc. The frequency of placing orders for an item can change over time as demand rates change.

Second, given the frequency decisions that have been made, employ a bin packing heuristic to determine the timing of the placement of the orders. Roughly speaking, first consider the items that will be ordered every week. Make those procurement decisions so that an equal amount of warehouse workload will arise on each day. That is, if receiving is done seven days a week, then ensure that the workload is spread out appropriately over the week. Next, con-

sider items that are procured every other week. Place orders for them so that total workload (workload for the items ordered weekly and every other week) is smoothed over the two week planning horizon. Continue on in this manner until all items are scheduled. We note that inbound freight is managed through third party logistics providers so that planned receiving schedules can be adhered to.

If the resultant schedule requires more capacity than is available, then the frequency of placing orders must be decreased for some items. A marginal analysis approach could be used to determine which items should be purchased less frequently. We note that capacity does change over a year to reflect the change in the demand processes. Overall, the objective of the approach we have outlined is to keep expected inventory holding costs low while smoothing workloads and adhering to workload capacity constraints. Remember that there are perhaps hundreds of thousands of items being managed at a primary warehouse. The procedure we have outlined will provide a smoothed and cost-based schedule. We assume in the following sections that the cycle lengths are known and the timing of procurement actions have been established for each item.

Inventory is allocated daily 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 in our model that link decisions in one primary warehouse system to another such system. Thus our model will focus on a system consisting of one primary warehouse that manages inventories for

several hundred thousand items. The execution models presented in Chapter 5 do link the allocation decisions, however.

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 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. Also, let $D_{it}^{\alpha} = \sum_{n} D_{it}^{n,\alpha}$, $D_{it}^{\beta} = \sum_{n} D_{it}^{n,\beta}$ and $D_{it} = D_{it}^{\alpha} + D_{it}^{\beta}$.

We assume the inbound shipping capacity from the primary warehouse to a regional warehouse is limited in each period. We let C_{it} represent this inbound capacity for regional warehouse i in period t. We assume that $C_{it} > \mathbf{E}[D_{it}]$ for all i and t.

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 primary warehouse. L_i^β depends on i.

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 $L_i^\alpha > 0$ or $l_i > 0$, the fulfillment system has more flexibility as to when to satisfy a demand. For example, suppose $L_i^\alpha > 0$ and there is no inventory on hand at the regional warehouse that is designated to satisfy a short response lead time demand. Further suppose that a stock replenishment for the item will arrive at the regional warehouse in time to ship the order to the customer within L_i^α periods. Then no backorders will occur. When $l_i > 0$, there is greater flexibility in the timing of shipments to the regional warehouse. This may mitigate the effect of shipping constraints. We first focus on the immediate response time case. In Section (4.2.3), we show how to extend our results to the more flexible delivery case.

Recall that shipments from the primary warehouse to a regional warehouse contain inventories of two types, inventory that is to be cross-docked and sent

directly to customers and inventory that will replenish the receiving regional warehouse's stock. Without loss generality, when $l_i = 0$, we assume when making allocation decisions that shipping capacity will be first allocated to meet long-response lead time demand requirements. After long-response lead time requirements are satisfied, any remaining shipping capacity may be used to ship replenishment stocks. However, when $l_i > 0$, then ensuring that inventories are available to meet short response lead time demand becomes the priority, assuming, of course, that stock will ultimately be available to meet long response lead time demand by its due date.

Our final assumption pertains to the sequence in which events occur in each period. We assume these events occur as follows for each item. First, we observe the echelon inventory positions at all locations for each item. Second, when appropriate, we receive a replenishment order at the primary warehouse corresponding to an order placed a procurement lead time ago. Third, when appropriate, we place a replenishment order from the primary warehouse on an external supplier. Fourth, we observe the demands at all regional warehouses. Fifth, based on availability, the inventory position 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 the primary warehouse. These stocks can be used to satisfy the current period's short-response lead time demand. 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 proportional to the number of outstanding backorders only at the regional warehouses at each period's end.

4.1 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 periods 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 periods 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 $t \in P_n$. Procurement orders can be placed only every τ_n periods for item n, which is the cycle length for item n.

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

$$x_{0t}^n + q_{0t}^n - \sum_{k=t}^{t+L_0^n} \sum_{i=1}^{N} d_{ik}^n, \tag{4.1}$$

where $q_{0t}^n > 0$ only if $t \in P_n$, 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 period t. 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.

4.1.1 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

$$\begin{aligned} & x_{it}^n + q_{it}^n - \sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} - d_{it}^{n,\alpha} - d_{it}^{n,\beta} \\ = \; & y_{it}^n - \sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} - d_{it}^{n,\alpha} - d_{it}^{n,\beta}, \end{aligned} \tag{4.2}$$

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

$$\mathbf{E}[h_i^n(y_{it}^n - \sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} - d_{it}^n)^+ + b^n(\sum_{k=t+1}^{t+L_i} D_{ik}^{n,\alpha} + d_{it}^n - y_{it}^n)^+], \tag{4.3}$$

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.

4.1.2 Holding Costs At the Primary Warehouse

Consider the immediate response case in which $L_i^\alpha = 0$ 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 for more than the shipping lead time from the primary warehouse. By assumption there are no customer demands satisfied directly from the primary warehouse inventory. Hence only holding costs are incurred there. Recall that item n may be ordered from its supplier only every τ^n periods.

Suppose $t' \in P_n$. The echelon inventory position at time $t' + L_0^n$ is equal to

$$x_{0,t'+L_0^n}^n = y_{0,t'}^n - \sum_{i=1}^{N} \sum_{k=t'}^{t'+L_0-1} D_{i,k}^n. \tag{4.4}$$

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

$$x^n_{0,t'+L^n_0} - \left\{ \sum_{i=1}^{N} [x^n_{i,t'+L^n_0} + \sum_{k=t'+L^n_0}^{t} q^n_{i,k}] \right\}$$

$$= \quad y^n_{0,t'} - \textstyle\sum_{k=t'}^{t'+L^n_0-1} D^n_k - \left\{ \sum_{i=1}^{N} [x^n_{i,t'+L^n_0} + \sum_{k=t'+L^n_0}^{t} q^n_{i,k}] \right\}. \tag{4.5}$$

Consequently, the expected holding cost incurred at the primary warehouse at the end of period t, $t \in \left\{ t' + L^n_0, \cdots, t' + L^n_0 + \tau^n - 1 \right\}$, is

$$\mathbf{E}\left[h^n_0 \left\{ (y^n_{0t'} - \sum_{k=t'}^{t'+L^n_0-1} D^n_k) - \sum_{i=1}^{N} [x^n_{i,t'+L^n_0} + \sum_{k=t'+L^n_0}^{t} q^n_{i,k}] \right\} \right]$$

$$= \qquad \mathbf{E}\left[h^n_0 \left\{ (y^n_{0t'} - \sum_{k=t'}^{t} D^n_k) - \sum_{i=1}^{N} y^n_{it} \right\} \right]. \tag{4.6}$$

4.1.3 The Objective Function

The expected cost functions given in (4.3) and (4.6) provide the basis for making procurement and allocation decisions. However, we do not use them directly in our decision model. Rather, we will define another set of functions which is their equivalent. These functions are of the type introduced by Clark and Scarf (1960) in their seminal paper and later used, for example, by Kunnumkal and Topaloglu (2008 and 2011).

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

$$G^n_{it}(y^n_{it}) = -h^n_0 y^n_{it} + \mathbf{E}[h^n_{i,t+L_i}(y^n_{it} - \sum_{k=t+1}^{t+L_i} D^{n,\alpha}_{ik} - d^n_{it})^+ + \mathbf{E}[b^n(\sum_{k=t+1}^{t+L_i} D^{n,\alpha}_{ik} + d^n_{it} - y^n_{it})^+]. \tag{4.7}$$

At the beginning of time period t, we do not know the values of d_{it}^n, the demand which will arise in that period for each regional warehouse i for each item n. But, we do know these values when making the allocation decision later in that period. Thus $G_{it}^n(y_{it})$ reflects the knowledge we have when making the allocation decision in period t.

For regional warehouse N, we define

$$G_{Nt}^n(y_{Nt}^n) = -h_0^n y_{Nt}^n + [b^n(d_{Nt}^n - y_{Nt}^n)^+], \tag{4.8}$$

which again reflects the fact that we know d_{Nt}^n when making the allocation of q_{Nt}^n units.

Let us next turn to the primary warehouse. Let

$$G_{0t}^n(y_{0t}^n) = -h_0^n y_{0t'}^n, \qquad t \in \{t' + L_0^n, \cdots, t' + L_0^n + \tau^n - 1\}, \tag{4.9}$$

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

Observe that $G_{0t}^n(y_{0t}^n) + \sum_{i=1}^{N} G_{it}^n(y_{it}^n)$ yields, except for a constant, the same period t expected costs for item n as would result from using expressions (4.3) and (4.6) to compute these costs. Thus the total expected period t cost function used in our model is

$$\sum_n \{G_{0t}^n(y_{0t}^n) + \sum_{i=1}^{N} G_{it}^n(y_{it}^n)\}. \tag{4.10}$$

Expressing our cost function in this manner permits us to solve the problem more efficiently, as we will observe subsequently and as was observed by Clark and Scarf (1960), Kunnumkal and Topaloglu (2008 and 2011) and by others.

4.1.4 A Dynamic Programming Formulation of the Fulfillment 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 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_{it}^n values at that time.

Four types of constraints exist when making decisions in each period. First, there is a constraint that limits the amount of inventory shipped from the primary warehouse to a particular regional warehouse i. Recall that this capacity is denoted by C_{it}. Second, 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 $q_{it}^n \geq 0$. This is the balance constraint. Third, procurement of each item n can take place only in period $t \in P_n$. Fourth, we cannot ship more than is on hand at the primary warehouse for any item in any period.

Combining the results obtained in Section 4.1.3 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_{it}^n. Then

$$V_t(\bar{x}_t) = \mathbf{E}_{D_t}\{\min \sum_n \{G^n_{0t}(y^n_{0t}) + \sum_{i=1}^{N} G^n_{it}(y^n_{it})\} + V_{t+1}(\bar{y}_t - \bar{D}_t) : \quad (4.11)$$

$$\text{s.t.} \quad \sum_n (y^n_{it} - x^n_{it}) \le C_{it,} \qquad \forall i, \quad (4.12)$$

$$\sum_i^N y^n_{it} \le x^n_{0t}, \qquad \forall n, \quad (4.13)$$

$$y^n_{it} \ge x^n_{it}, \qquad \forall n, i, \quad (4.14)$$

$$y^n_{0t} \ge x^n_{0,t}, \qquad t \in P_n, \forall n, \quad (4.15)$$

$$y^n_{0t} = x^n_{0,t}, \qquad t \notin P_n, \forall n\}. \quad (4.16)$$

While this formulation depends on our assumption that $L^n_0 \le \tau^n$, it can be easily modified. Computational results reported later are not depend on this assumption.

The size of the state space corresponding to this dynamic programming formulation 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.

4.2 An Approximation Approach

The existence of shipping capacity constraints makes the problem presented in the previous section much more difficult to solve. We address this difficulty by obtaining the required stock levels using a two-step process. In the first step, we temporarily relax these shipping capacity constraints. By making this assumption, there is no longer any constraint linking allocation decisions made for each item. Thus, we can compute purchasing and allocation policies among

the items separately. We show how to construct a solution to a single item problem in Section 4.2.1. In Section 4.2.2, we show how to include the shipping capacity constraints in our model. Specifically, we show how to increment the stock levels found in the first step to account for the limited shipping capacity. We also show that in cases of practical significance that the presence of these constraints does not result in additional inventory requirements.

4.2.1 An Approximation Approach for Managing a Single Item

We begin by making some observations and assumptions that are the basis for our approach for setting stock levels for each item.

Demand from cycle-to-cycle may be non-stationary. However, 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 ($q^n_{0t} > 0, t \in P_n$). This is virtually always the case in practice for the type of items we are considering. By assuming so, we are able to formulate the problem as a sequence of independent problems, one for each cycle. Hence a 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 backorder costs are high relative to the holding costs (over a 100 to 1 for most items), inventory levels are high enough to ensure that backorders occur only infrequently. There are two implications of this observation. First, we assume that when a procurement order arrives that it is possible to make allocations so that all regional warehouses will be able to 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, we assume that q_{it}^n could be negative or 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 normally account 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.

Second, a cycle is almost always a week or longer, demand rates for items managed using the PWS are high and relatively stable from cycle-to-cycle, 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.

An Experiment

To test the appropriateness of the balance assumption, we conducted an experiment. In this experiment we assumed that $N = 5$, which is the number of warehouses that the company we studied had in operation at the time we examined it. As we have discussed, the procurement cycles vary in length from a week to a few months. To test our assumption under the worst case scenario, we assumed the length of a cycle is 7 days. It is a worst case since the possibil-

ity of running out of stock, and thereby increasing the possibility of imbalance, occurs more frequently for shorter cycle lengths.

For a typical product, the annual holding cost rate is about 20% of the product cost. Hence, the annual cost to hold a product that costs \$50 for one year is \$10, which implies the approximate per day holding cost is h = \$0.0274. Assume the holding cost at the primary warehouse is 10% of the regional warehouse holding cost, so in this case h_0 = \$0.00274. Additionally, we assume the backorder cost is 25 times greater than the holding cost. This is a conservative estimate of the backorder cost. Since this cost would be higher in practice, the inventory levels computed in this case would be lower bounds on the actual levels. This would also increase the possibility of observing an imbalance situation. In our experiment the per period backorder cost is b = \$0.6849. In practice, the backorder cost would likely be between \$1 and \$5.

We let the shipping lead time from the primary warehouse to the regional warehouses 1 through 5 be $2, 3, 4, 2$ and 0 days, respectively. Again, these reflect the times we observed in the system we studied. Remember that regional warehouse 5 is the co-located regional warehouse and does not hold any inventory. The primary warehouse procurement lead time from the supplier is set at 2 days in length. These data are captured in Table 1.

Table 1: Parameter Values for Five-warehouse, Single Item Example

	RW 1	RW 2	RW 3	RW 4	RW 5	PW 0
Lead Time (Days)	2	3	4	2	0	2
Holding Cost (\$/Unit/Day)	0.0274					0.00274
Backorder cost(\$/Unit/Day)	0.6849					

We considered four items in our experiment. The ranges in the demand rates

and variances are also representative of those we found in our study of an on-line retailer. The demand rates and variance-to-mean (VTM) ratios considered in the experiment are presented in Table 2. One would expect that an item with a low daily demand rate and high VTM ratio would be more likely to experience imbalances in inventory than would other items. However, we will show that our conjecture is valid even when the demand rate is low. Demand is assumed to be negative-binomially distributed. We also assume that short response

lead time demand accounts for 20% of the total demand for each item.

Table 2: Parameters for the Demand Distribution

	RW 1	RW 2	RW 3	RW 4	RW 5	RW 1	RW 2	RW 3	RW 4	RW 5
Mean \ VTM	1.01	1.01	1.01	1.01	1.01	3	3	3	3	3
Item 1	0.1	0.15	0.2	0.25	0.3	0.1	0.15	0.2	0.25	0.3
Item 2	1	2	3	3	4	1	2	3	3	4
Item 3	5	8	9	9	11	5	8	9	9	11
Item 4	20	25	25	31	28	20	25	25	31	28

We found approximately optimal order-up-to values for each item assuming the imbalance constraints were relaxed and that shipping capacity is always available. Our simulation experiment consisted of operating the system for 2000 cycles, that is, for 14000 days for each item and each VTM ratio for each of 10 replications of the experiment.

The results of the experiment are displayed in Table 3.

Table 3: Experimental Results:
Imbalance Periods and Frequency

VTM = 1.01		VTM = 3	
Count	Frequency	Count	Frequency
22.1	0.158%	1.4	0.010%
18.5	0.132%	28.3	0.202%
1.9	0.014%	5	0.036%
0.2	0.001%	1.4	0.010%

The values presented in the table are the average values obtained from the 10 replications of the experiment for each item and variance combination. There are two numbers in each cell. The first is the average number of imbalance periods for the 10 replications and the second is the average percentage of the number of periods in which imbalance occurs. For example, 22.1 and 0.158% in Table 3 indicate that on average 22.1 periods are in an imbalance condition out of the $14,000$ periods, that is, imbalance occurs in about 0.158% of the periods.

In summary, our experiment shows that for the range of means and variances of demands that we tested, the imbalance situation occurs in less than 0.2% of the periods. When expected demands are for more than two units per period, imbalance almost never occurs. Hence we will assume that the balance constraints can be relaxed without affecting the quality of the solution.

We are not implying that the balance constraints can be ignored in all situations. Kunnumkal and Topaloglu (2008 and 2011), for example, show that imbalance can occur and can affect stock level requirements. In their model, the central warehouse orders every period and all demand is short response lead time demand. These are major reasons why an imbalance of inventory can occur in their setting.

We also note that lower demand rate items are normally ordered less frequently than once a week. Thus, in practice, the system is even less likely to experience imbalances than indicated in Table 3.

An Approximation Model for a Single Item

Recall that we assume that decisions made in one cycle do not affect those made in other cycles. Consequently, we now analyze a single cycle for some item. We drop the item identifier from our notation in this section since we focus on a single item. We assume $l_i = 0$, too.

Suppose the item is ordered at time 0 and its echelon inventory position for the primary warehouse, or system, is raised to y_0 units. The amount ordered arrives in period L_0 at which time the system echelon inventory is $x_{L_0} = y_0 - \hat{d}_{0,L_0-1}$ units, where $\hat{d}_{0,L_0-1}$ is the total demand from period 0 through period $L_0 - 1$. The random variable $\tilde{D}_{0,L_0-1}$ measures the demand over this time interval. Thus $P[x_{L_0} = w] = P[\hat{D}_{0,L_0-1} = y_0 - w]$.

Suppose the echelon inventory position at the beginning of period L_0 is x_{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, $q_{i,L_0} \geq x_{i,L_0} - D_{i,L_0}$ 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

$$h_i \mathbf{E}[y_{it} - d_{it} - \hat{D}^{\alpha}_{i,[t+1,t+L_i]}]^+ + b\mathbf{E}[d_{it} + \hat{D}^{\alpha}_{i,[t+1,t+L_i]} - y_{it}]^+, \tag{4.17}$$

where $\hat{D}^{\alpha}_{i,[a,b]}$ is the random variable for the total short response lead time demand at regional warehouse i over the interval of periods a through b and $t \in \{L_0, \cdots, L_0 + \tau - 1\}$. 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 at the primary warehouse is $\mathbf{E}[h_0[y_0 - \sum_i y_{it} - D_{0,[0,t]}]]$ at the end of period t since y_{it} can be negative.

Let

$$G_{0t}(y_{0t}) = h_0 y_0, \tag{4.18}$$

and

$$G_{it}(y_{it}) = -h_0 y_{it} + h_i \mathbf{E}[y_{it} - d_{it} - D^{\alpha}_{i,[t+1,t+L_i]}]^+ + b\mathbf{E}[d_{it} + D^{\alpha}_{i,[t+1,t+L_i]} - y_{it}]^+. \tag{4.19}$$

Then $G_{0t}(y_{0t}) + \sum_{i=1}^{N} G_{it}(y_{it})$ 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 dynamic programming recursion that can be used to find the values of y_{it} and ultimately y_0. Recall that we have relaxed both the shipping and the balance constraints. The recursion we use has a single dimensional state space and is defined as follows.

$$
\begin{aligned}
V_t(x_{0t}) &= \mathbf{E}_{D_t}[\min\{G_0(y_{0t}) + \sum_{i=1}^{N} G_{it}(y_{it}) + V_{t+1}(x_{0t} - D_t)\} : \sum_i y_{it} \le x_{0t}] \\
&= G_0(y_{0t}) + E[\min \sum_{i=1}^{N} G_{it}(y_{it}) + V_{t+1}(y_{0t} - D_t) : \qquad (4.20) \\
&\quad \sum_i y_{it} \le x_{0t}]. \qquad (4.21)
\end{aligned}
$$

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 $\sum V_{L_0}(x_{0,L_0}) \cdot P[D_{[0,L_0-1]} = y_0 - x_{0,L_0}]$, which we call $F(y_0)$. It is easy to show that $F(\cdot)$ is a convex function of y_0 and hence it is easy to determine the optimal value of y_0 using a line search.

Suppose the D_{it} are independent and identically distributed random variables over the cycle for a given regional warehouse and are independent among the warehouses. Then the optimal stationary value of y_{it} can be computed as follows. Define

$$
\tilde{G}_i(\tilde{y}_i) = -h_0\tilde{y}_i + \mathbf{E}[h_i\left\{\tilde{y}_i - D^{\alpha}_{i,[L_0+1,L_0+L_i]}\right\}^+] + \mathbf{E}[b\{D^{\alpha}_{i,[L_0+1,L_0+L_i]} - \tilde{y}_i\}^+]. \qquad (4.22)
$$

This is a newsvendor expression, the minimum of which occurs when $\tilde{y}_i^*$ is the smallest value of $\tilde{y}_i$ for which

$$
P[D_{i,[L_0+1,L_0+L_i]} \le \tilde{y}_i] \ge \frac{h_0 + b}{h_i + b}. \qquad (4.23)
$$

Then $y_{it}^* = \tilde{y}_i^* + d_{it}$. Of course, this desired level can be achieved only if $\sum_i y_{it}^* \le x_{0t}$.

Note, it is easy to determine a lower bound $\bar{y}_0$ on y_0^* in this case. Let $\bar{h} = \max h_i$, and $D_{[0,t]}$ be the total system demand for periods 0 through t. Let

$$
H_t(y_0) = \sum_{k=L_0}^{L_0+\tau-1} \{\bar{h}\mathbf{E}[(y_0 - D_{[0,k]})^+] + b\mathbf{E}[(D_{[0,k]} - y_0)^+]\},
$$

and

$$\Delta H_t(y_0) = H(y_0 + 1) - H(y_0) \tag{4.24}$$

$$= \sum_{k=L_0}^{L_0+\tau-1} (-b + P[D_{[0,k]} \leq y_0] \cdot (\bar{h} + b)) \tag{4.25}$$

$$= -\tau b + (\bar{h} + b) \cdot \sum_{k=L_0}^{L_0+\tau-1} P[D_{[0,k]} \leq y_0]. \tag{4.26}$$

To find $\bar{y}_0$, determine the smallest value of y_0 such that

$$\sum_{k=L_0}^{L_0+\tau-1} P[D_{[0,k]} \leq y_0] \geq \frac{\tau b}{b + h}.$$

This results in a lower bound for y_0^*.

4.2.2 Setting Stock levels When Shipping Capacity Is Limited

Previously we assumed that shipping capacity was infinite to each regional warehouse in each period and that inventories were always balanced among the regional warehouses for every item. We now will see how to include the shipping capacity constraints into our model. Specifically, our goal in this section is to present a method that can be used to determine how much incremental inventory is needed for each item at a regional warehouse to maintain the desired level of service when shipping constraints are active.

There are several ways to determine the desired incremental inventory for each item. One way is to employ a Lagrangian relaxation method similar to the one introduced by Kunnumkal and Topaloglu (2008 and 2011). We found that this approach, in general, requires much more computational effort to find these stock levels than the one we now describe.

To simplify our analysis, assume that both the short and long response lead time demand processes are stationary and demands are independent from period to period for an item at each location and across locations. Furthermore, assume that the shipping capacity at each warehouse is constant throughout the planning horizon, that is, $C_{it} = C_i$ for all t. We also assume the system is stable, that is,

$$\mathbf{E}\left[\sum_n (D_i^{n,\alpha} + D_i^{n,\beta})\right] < C_i, \text{ for all } i, \tag{4.27}$$

where $D_i^{n,\alpha}$ and $D_i^{n,\beta}$ are random variables for the number of short and long response lead time demands that arise daily, respectively, for item n at location i.

In the environment we studied, recall that short response lead time demand normally constituted about $13\% - 20\%$ of total demand. Thus it is reasonable to assume that short response lead time demand does not exceed the available shipping capacity, that is,

$$P\left\{\sum_n D_i^{n,\alpha} < C_i\right\} = 1, \tag{4.28}$$

for all i and all periods.

We also assume that when planning the use of shipping capacity, priority is given to replenishing the regional warehouse's inventory needed to satisfy short response lead time demand. Thus the effective capacity available in a period for shipping long response lead time demand is a random variable, $\tilde{C}_i = C_i - \sum_n D_i^{n,\alpha}$. Note that we ship $\sum_n D_i^{n,\alpha}$ units because we are assuming that a stationary order-up-to policy is employed. By making this assumption, we can

plan the inventory requirements for each item for the short and long response lead time demands separately.

Initially determine the inventory level required to satisfy short response lead time demand using the approach described in the previous section. The extent to which these levels need to be augmented due to the limited amount of shipping capacity depends on two factors. The first is the amount of shipping capacity that exists at a regional warehouse. The second factor is the value of L_i^{α} and l_i, which indicate the amount of time warehouse i has to respond to short response lead time arising there and the length of the grace period, respectively. We initially focus on the immediate response case, that is, where $L_i^{\alpha} = l_i = 0$ for all items.

Augmenting Regional Warehouse Stock Levels

In this section we determine the aggregate amount of inventory required to augment the stock levels by employing the shortfall process. For a complete discussion of the shortfall process, see Glasserman (1997) or Roundy and Muckstadt(2000). Note that in this section, we focus on finding the incremental inventory levels needed to maintain a desired fill rate for long response lead time items only. We have also developed a cost-based approach for estimating the amount of incremental inventory that is required, but this approach is not presented here.

The shortfall process behaves as follows. Let S_{it} denote the "shortfall process random variable" corresponding to regional warehouse i following the allocation decision made to satisfy the long response lead time demand at the primary

warehouse in period t, and let $\hat{y}_{it}^n$ be the inventory position only considering the long response lead time demand for item n for regional warehouse i following that allocation decision. Also, define T_i to be the aggregate target inventory position for regional warehouse i, that is, the sum of the individual item target inventory levels. Then

$$S_{it} = T_i - \sum_n \hat{y}_{it}^n \tag{4.29}$$

and

$$S_{it} = [S_{i,t-1} + \sum_n D_{it}^n - C_{it}]^+. \tag{4.30}$$

In this recursive definition, D_{it}^n is a random variable since we are evaluating S_{it} at the end of period t. Note, importantly, that S_{it} does not depend on T_i in (4.30).

For ease of exposition, we drop the subscript i. Given our assumptions, we could calculate the steady state distribution of S_t by first determining the transition matrix for the shortfall process, $P_{kj} = P\{S_t = j | S_{t-1} = k\}$, and then by calculating the stationary distribution of the shortfall process in the usual manner. Unfortunately, this approach is impractical due to the scale of the problem. Hence, we use the ideas put forth in Roundy and Muckstadt(2000) to determine a continuous approximation of the probability distribution corresponding to the shortfall process.

Roundy and Muckstadt(2000) show that for a fixed capacity c, the following equality holds.

$$P\{S_t > s\} = P\{D_t > s + c\} + \mathbf{E}_{D_t}\left[1(d \leq s + c) \cdot P[S_{t-1} > s + c - d | D_t = d]\right], \tag{4.31}$$

where $1(\cdot)$ is the indicator operator. As they discussed, the steady state distribution function for S can be approximated by a mass exponential distribution. That is,

$$\bar{F}_S(s) = \begin{cases} \bar{P}_0 e^{-\gamma s}, & s \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{4.32}$$

where $\bar{F}_S(s)$ is the complementary cumulative distribution function of the random variable S, $\bar{P}_0 = 1 - P[S = 0]$, and γ satisfies $e^{\gamma c} = \mathbf{E}[e^{\gamma D_t}]$ for a fixed capacity c.

Given that D_t is the cumulative demand per period over all items, it is reasonable to approximate its distribution with a normal distribution. In this case, Glasserman (1997) shows that for a given value of c, $\gamma = 2(c - E[D_t])/\sigma^2$, where σ^2 is the variance of the random variable D_t. Then we see that

$$\bar{F}_S(s) = \begin{cases} \bar{P}_0 e^{-[2(c-E[D_t])/\sigma^2]s}, & s \geq 0, \\ 0, & \text{otherwise}. \end{cases} \tag{4.33}$$

To use this approach, we let $c = \mathbf{E}[\tilde{C}_i]$ in the following section. This will underestimate the tail probabilities of the true shortfall distribution and therefore, in doing so, we will find a lower bound on the true best order-up-to level.

Determining T^*

We now show how to determine the optimal aggregate incremental inventory level, which we denote by T_i^* for regional warehouse i. We first consider the case where $l_i = 0$, that is $L_i^\beta = L_i$, and then consider the case where l_i is positive.

When $l_i = 0$, the number of backoders of long response lead time demand at the end of period t that correspond to demands that occurred in period $t-1$ or earlier is given by

$$[S_{t-1} - T - c]^+. \tag{4.34}$$

Thus, the number of the long response lead time demands that occurred in period t that are backordered at the end of period t is

$$[S_t - T]^+ - [S_{t-1} - T - c]^+. \tag{4.35}$$

Suppose $\eta(T)$ represents the expected number of long response lead time demand that arrived in period t that is backordered at the end of that period . Then

$$\eta(T) = \mathbf{E}_S[[S_t - T]^+ - [S_{t-1} - T - c]^+]. \tag{4.36}$$

In Section 4.2.1 we showed how to obtain an order-up-to value for each item. Corresponding to this stock level for item n is a fill rate, which we denote by f^n. Then the average regional warehouse fill rate is

$$\bar{f} = \sum_n \left\{ \frac{\mathbf{E}[D^n]}{\sum_n \mathbf{E}[D^n]} \right\} f^n. \tag{4.37}$$

To find T^*, we seek the value of T that satisfies

$$\eta(T) = \mathbf{E}(D)(1 - \bar{f}). \tag{4.38}$$

T^* can be found using a line search.

That is, we seek to achieve the same customer service level that was planned in the absence of capacity constraints. Suppose the per period demand random variables D_t are normally distributed. By applying Glasserman's approximation (4.29)when $l_i = 0$ and $c = \mathbf{E}[C]$,

$$\eta(T) = \frac{\bar{P}_0}{\gamma} e^{-\gamma T}(1 - e^{-\gamma c}), \tag{4.39}$$

where $\bar{P}_0$ and γ are determined as indicated in the previous section.

Again, suppose we have a target average fill rate $\bar{f}$ for a regional warehouse. As before, T can be found using a line search or approximated using the following expression, assuming the cumulative long response lead time demand over all items is approximated by a normal distribution. In this case,

$$T \approx \left[-\frac{1}{\gamma} ln \frac{(1 - \bar{f})\gamma \mathbf{E}(D)}{\bar{P}_0(1 - e^{-\gamma c})} \right]^+ \tag{4.40}$$

as discussed in Roundy and Muckstadt(2000).

We note that T^* can be determined quickly using this approach. Given that there are hundreds of thousands of items managed at each warehouse, computational efficiency is important. Remember the value of T^* is a lower bound on the optimal aggregate incremental amount of inventory required to achieve a desired fill rate when the shipping constraint is active.

Determining T^*

We now show how to determine the optimal aggregate incremental inventory level, which we denote by T_i^* for regional warehouse i. We first consider the

case where $l_i = 0$, that is $L_i^{\beta} = L_i$, and then consider the case where l_i is positive.

When $l_i = 0$, the number of backoders of long response lead time demand at the end of period t that correspond to demands that occurred in period $t - 1$ or earlier is given by

$$[S_{t-1} - T - c]^+. \tag{4.41}$$

Thus, the number of the long response lead time demands that occurred in period t that are backordered at the end of period t is

$$[S_t - T]^+ - [S_{t-1} - T - c]^+. \tag{4.42}$$

Suppose $\eta(T)$ represents the expected number of long response lead time demand that arrived in period t that is backordered at the end of that period . Then

$$\eta(T) = \mathbf{E}_S[[S_t - T]^+ - [S_{t-1} - T - c]^+]. \tag{4.43}$$

In Section 4.2.1 we showed how to obtain an order-up-to value for each item. Corresponding to this stock level for item n is a fill rate, which we denote by f^n. Then the average regional warehouse fill rate is

$$\bar{f} = \sum_n \left\{ \frac{\mathbf{E}[D^n]}{\sum_n \mathbf{E}[D^n]} \right\} f^n. \tag{4.44}$$

To find T^*, we seek the value of T that satisfies

$$\eta(T) = \mathbf{E}(D)(1 - \bar{f}). \tag{4.45}$$

T^* can be found using a line search.

That is, we seek to achieve the same customer service level that was planned in the absence of capacity constraints. Suppose the random variables D_k are normally distributed. In this case, Glasserman's approximation (4.29) may be applied. When $l_i = 0$ and $c = \mathbf{E}[C]$,

$$\eta(T) = \frac{\bar{P}_0}{\gamma} e^{-\gamma T}(1 - e^{-\gamma c}), \tag{4.46}$$

where $\bar{P}_0$ and γ are determined as indicated in the previous section.

Again, suppose we have a target average fill rate $\bar{f}$ for a regional warehouse. As before, T can be found using a line search or approximated using the following expression, assuming the cumulative long response lead time demand over all items is approximated by a normal distribution. In this case,

$$T \approx \left[-\frac{1}{\gamma} ln \frac{(1 - \bar{f})\gamma \mathbf{E}(D)}{\bar{P}_0(1 - e^{-\gamma c})} \right]^+ \tag{4.47}$$

as discussed in Roundy and Muckstadt (2000).

We note that T^* can be determined quickly using this approach. Given that there are hundreds of thousands of items managed at each warehouse, computational efficiency is important. Remember the value of T^* is a lower bound on the optimal aggregate incremental amount of inventory required to achieve a desired fill rate when the shipping constraint is active.

Allocating the T^* Among Items

We now discuss how to disaggregate the T^* units needed to compensate for the shipping constraint. That is, we provide a method to find the values of y^n, the amount of incremental stock needed for item n at a regional warehouse such that $\sum y^n = T^*$.

There are many ways to determine the values of y^n. We do so by minimizing the sum of the total current period's holding plus backorder costs plus the sum of future expected holding costs resulting from the choice of y^n. We include the term for the expected future holding costs so that items whose demand processes have high coefficients of variation will not be stocked as heavily as those having lower coefficients of variation. Stated differently, we want to add this capacity-protecting inventory in items whose demands are most predictable and likely to be needed in the near future.

As discussed in Chan (1999), the function

$$Q^n(x) = \sum_{k=2}^{\infty} \mathbf{E}[x - D^n_{[1,k]}] \tag{4.48}$$

measures the expected number of future unit inventory periods that will result from stocking x units of item n in the current period, where $D^n_{[1,k]}$ represents the random variable for the cumulative demand for item n for periods 1 through k.

The optimization problem we propose solving to find the values of y^n is

$$\min \quad \left[\sum_n \{G^n(y^n + y^{n*}) + h^n Q^n(y^n + y^{n*})\} \right] \tag{4.49}$$

$$\text{s.t.} \quad \sum_n y^n = T^* \tag{4.50}$$

$$y^n \geq 0, \tag{4.51}$$

where y^{n*} is the optimal order-up-to value determined using the method discussed in Section 4.2.1. We assume that current period demand d_{it} is replaced by the expected demand $\mathbf{E}[D_{it}]$ when setting the value of y^{n*}. This problem can be solved easily using a marginal analysis method.

4.2.3 Flexible Delivery

We have assumed that both long response lead time demand and short response lead time demand had to be sent to a regional warehouse on the day a customer's order was received. In practice, when a customer places an order online, he is usually told to expect to receive the item within a few days range. This provides the fulfillment system a few days of flexibility in fulfilling the demand without incurring any backorder costs. Intuitively, by extending the fulfillment date of an order, the system can utilize its capacity more effectively. In this section we study the impact of flexible delivery on the required inventory levels when shipping capacity between the primary warehouse and a regional warehouse is limited.

Recall that regional warehouses carry inventories for two reasons. One is to satisfy short response lead time demand. The other is to satisfy the long response lead time demands that cannot be fulfilled by the primary warehouse

due to limited shipping capacity. When $l_i > 0$ the system can utilize future shipping capacity to respond to today's orders for long response lead time demand without penalty. Hence, the regional warehouses would carry no more inventory than in the immediate response case.

Let us compute the number of expected per period backorders when $l_i > 0$. Observe that the number of units of long response lead time demand that arrived in period $t - l_i$ that are backordered at the end of period t is

$$[S_t - \sum_{k=t-l_i+1}^{t} D_k^\beta - T]^+ - [S_{t-1} - \sum_{k=t-l_i}^{t-1} D_k^\beta - T - c]^+, \tag{4.52}$$

where D_k^β is the long response lead time demand arising on day k and $c = \mathbf{E}[C_t]$. Recall that S_t is the shortfall process defined in Equation 4.29. Assume that the total demand follows a normal distribution with mean and variance equal to μ and σ^2, respectively. Let $k \leq 1$ denote the percentage of the long response lead time demand. Assume that long response lead time demand follows a normal distribution with mean and variance of $k\mu$ and $k\sigma^2$, respectively.

Redefine $\eta(T)$ to be the expected number of units of long response lead time demand that arrived in period $t - l_i$ that remain backordered at the end of period t, where the expectation is taken with respect to the stationary shortfall random variable and the random variable $\sum_{k=t-l_i+1}^{t} D_k^\beta$.

When $l_i = 1$,

$$\eta(T) = \mathbf{E}\left[[S_t - D_t^\beta - T]^+ - [S_{t-1} - D_{t-1}^\beta - T - c]^+\right]. \tag{4.53}$$

Let $S_t^1 = S_t - D_t^\beta$, then

$$\eta(T) = \mathbf{E}\left[[S_t^1 - T]^+ - [S_{t-1}^1 - T - c]^+\right] \tag{4.54}$$

$$= \mathbf{E}\left[[S^1 - T]^+ - [S^1 - T - c]^+\right], \tag{4.55}$$

where S^1 is the stationary distribution of S_t^1.

We show how to approximate the distribution function for S^1. Specifically we prove the following.

Theorem 4.2.1. *For $m > 0$,*

$$P(S_t^1 > m) \approx \bar{P}_0 e^{\gamma(1-k)c} e^{-\gamma(m+c)} \Phi(\frac{m + k\mu}{\sqrt{k}\sigma}) + \bar{P}_0 \Phi(\frac{-2kc - m + k\mu}{\sqrt{k}\sigma}). \tag{4.56}$$

Proof. Recall that $S_t^1 = S_t - D_t^\beta$ when $l_i = 1$, where S_t is the shortfall process. The shortfall process can be approximated by a mass exponential function $f_S(s) = \bar{P}_0 e^{-\gamma s}$, where as before $\gamma = \frac{2(c-\mu)}{\sigma^2}$. The distribution functions of short and long response lead time demand, D_t^α and D_t^β, are $\mathcal{N}\left(k\mu, k\sigma^2\right)$ and $\mathcal{N}\left((1-k)\mu, (1-k)\sigma^2\right)$.

Observe that for $m \geq 0$,

$$\begin{aligned} P(S_t^1 > m) =& P\left\{[[S_{t-1} + D_t - c]^+ - D_t^\beta]^+ > m\right\} \\ =& P\left\{(S_{t-1} + D_t^\alpha > m + c) \cap (S_{t-1} + D_t^\alpha + D_t^\beta \geq c)\right\} \\ &+ P\left\{(-D_t^\beta > m) \cap (S_{t-1} + D_t^\alpha + D_t^\beta < c)\right\}. \end{aligned}$$

We will compute the two probabilities separately. Let $S_t' = S_{t-1} + D_t^\alpha$, which

is the convolution of $S_{t-1}$and D_t^{α}. Hence, the stationary distribution of S' is

$$f_{S_t'}(z) = \int_{-\infty}^{\infty} f_{D^{\alpha}}(\tau) f_S(z-\tau) d\tau \tag{4.57}$$

$$= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi(1-k)\sigma^2}} e^{-\frac{(\tau-(1-k)\mu)^2}{2(1-k)\sigma^2}} \gamma \bar{P}_0 e^{-\gamma(z-\tau)} d\tau \tag{4.58}$$

$$= \bar{P}_0 \gamma e^{\gamma(1-k)\mu + \frac{1}{2}\gamma^2(1-k)\sigma^2} e^{-\gamma z} \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi(1-k)\sigma^2}} e^{-\frac{(\tau-(1-k)\mu+\gamma(1-k)\sigma^2)^2}{2(1-k)\sigma^2}} d\tau \tag{4.59}$$

$$= \bar{P}_0 \gamma e^{\gamma(1-k)\mu + \frac{1}{2}\gamma^2(1-k)\sigma^2} e^{-\gamma z} \Phi\left(\frac{z-(1-k)\mu-(1-k)\gamma\sigma^2}{\sqrt{1-k}\sigma}\right) \tag{4.60}$$

$$\approx \bar{P}_0 \gamma e^{\gamma(1-k)\mu + \frac{1}{2}\gamma^2(1-k)\sigma^2} e^{-\gamma z} \tag{4.61}$$

$$= \bar{P}_0 \gamma e^{\gamma(1-k)c} e^{-\gamma z}, \tag{4.62}$$

when

$$\frac{z-(1-k)\mu-(1-k)\gamma\sigma^2}{\sqrt{1-k}\sigma} \geq 3. \tag{4.63}$$

In the following calculation, condition (4.63) holds when $z \geq c$ for VTM ranges between 1.01 and 5 when the cumulative demand is more than 100 units, which is clearly the case for the environment we have studied. Hence,

$$P(S_t^1 > m) = P\left\{(S'_{t-1} > m+c) \cap (S'_{t-1} + D_t^{\beta} \geq c)\right\} + P\left\{(-D_t^{\beta} > m) \cap (S'_{t-1} + D_t^{\beta} < c)\right\}. \tag{4.64}$$

For $m \geq 0$,

$$P\left\{(S'_t > m+c) \cap (S'_{t-1} + D_t^{\beta} \geq c)\right\} \tag{4.65}$$

$$= \int_{m+c}^{\infty} \int_{c-y}^{\infty} f_{D^{\beta}}(x) f_{S'}(y) dx dy \tag{4.66}$$

$$= \int_{m+c}^{\infty} (1 - F_{D^{\beta}}(c-y)) f_{S'}(y) dy \tag{4.67}$$

$$= \int_{m+c}^{\infty} f_{S'}(y) dy + \int_{m+c}^{\infty} F_{D^{\beta}}(c-y)) d\bar{P}_0 e^{\gamma(1-k)c} e^{-\gamma y}. \tag{4.68}$$

We find that

$$\int_{m+c}^{\infty} f_{S'}(y) dy = \bar{P}_0 e^{\gamma(1-k)c} e^{-\gamma(m+c)}, \tag{4.69}$$

and by employing integration by parts,

$$\int_{m+c}^{\infty} F_{D^\beta}(c-y)d\bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma y} \tag{4.70}$$

$$=F_{D^\beta}(c-y)\cdot \bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma y}|_{m+c}^{\infty} \tag{4.71}$$

$$-\int_{m+c}^{\infty} \bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma y} f_{D^\beta}(c-y)\cdot(-1)dy \tag{4.72}$$

$$=-\Phi(\frac{-m-k\mu}{\sqrt{k}\sigma})\cdot \bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma(m+c)}+\int_{m+c}^{\infty} \bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma y} f_{D^\beta}(c-y)dy. \tag{4.73}$$

$$\tag{4.74}$$

$$\int_{m+c}^{\infty} \bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma y} f_{D^\beta}(c-y)dy \tag{4.75}$$

$$=\bar{P}_0 e^{\gamma(1-k)c}\int_{m+c}^{\infty}\frac{1}{\sqrt{2\pi k\sigma^2}}e^{-\frac{(c-y-k\mu)^2}{2k\sigma^2}}e^{-\gamma y}dy \tag{4.76}$$

$$=\bar{P}_0 e^{\gamma(1-k)c}e^{-\frac{2c(c-\mu)(1-k)}{\sigma^2}}\int_{m+c}^{\infty}\frac{1}{\sqrt{2\pi k\sigma^2}}e^{-\frac{(y-(k\mu-(2k-1)c))^2}{2k\sigma^2}}dy \tag{4.77}$$

$$=\bar{P}_0 e^{\gamma(1-k)c}e^{-\frac{2c(c-\mu)(1-k)}{\sigma^2}}\Phi(\frac{-2kc-m+k\mu}{\sqrt{k}\sigma}) \tag{4.78}$$

$$=\bar{P}_0\Phi(\frac{-2kc-m+k\mu}{\sqrt{k}\sigma}). \tag{4.79}$$

Combining Equations (4.68), (4.69), (4.73) and (4.79), we find that

$$P\left\{(S_t'>m+c)\cap(S_{t-1}'+D_t^\beta\geq c)\right\}$$

$$\approx \bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma(m+c)}\Phi(\frac{m+k\mu}{\sqrt{k}\sigma})+\bar{P}_0\Phi(\frac{-2kc-m+k\mu}{\sqrt{k}\sigma}) \tag{4.80}$$

$$=\bar{P}_0 e^{\gamma(1-k)c}e^{-\gamma(m+c)}\Phi(\frac{m+k\mu}{\sqrt{k}\sigma})+\bar{P}_0\Phi(\frac{-2kc-m+k\mu}{\sqrt{k}\sigma}). \tag{4.81}$$

In addition, we have

$$P\left\{(-D_t^\beta>m)\cap(S_{t-1}+D_t<c)\right\} = \int_{-\infty}^{-m}\int_{0}^{c-x} f_{D^\beta}(x)f_S(y)dydx$$

$$\approx 0,$$

since $P(D_t^\beta = 0) \approx 0$ for any practical case. Hence

$$P(S_t^1 > m) = \bar{P}_0 e^{\gamma(1-k)c} e^{-\gamma(m+c)} \Phi(\frac{m + k\mu}{\sqrt{k}\sigma}) + \bar{P}_0 \Phi(\frac{-2kc - m + k\mu}{\sqrt{k}\sigma}).$$

□

When $\frac{\mu}{\sigma}$ is large, which is the case in the system under study, where k is between 0 and 1,

$$P(S_t^1 > m) \approx \bar{P}_0 e^{-\gamma m - \gamma c k}, m \geq 0. \tag{4.82}$$

This implies that S_t^1 also can be approximated by using a mass-exponential function. Hence, the expected number of per period backorders can be estimated using (4.55) and (4.83).

Suppose capacity utilization rate is over 99%. When $l_i = 1$ the incremental inventory requirement reduces significantly as the percentage of long response lead time increases. As the utilization rate is lowered, no incremental inventory is needed.

Note that when $\frac{\mu}{\sigma}$ is large, which is the case in the system under study, where k is between 0 and 1.

$$P(S_t^1 > m) \approx \bar{P}_0 e^{-\gamma m - \gamma c k}, m \geq 0. \tag{4.83}$$

This implies that S_t^1 also can be approximated by using a mass-exponential function. Hence, the expected number of per period backorders can be estimated using (4.55) and (4.83).

When $l_i = k, (k \geq 2)$, define $S_t^k = [S_t^{k-1} - D_{t-k+1}^{\beta}]^+$. Then,

$$\eta(T) = \mathbf{E}\left[[S_t^k - T]^+ - [S_{t-1}^k - T - c]^+\right] \tag{4.84}$$

$$= \mathbf{E}\left[[S^k - T]^+ - [S^k - T - c]^+\right], \tag{4.85}$$

where S^k is the stationary distribution of S_t^k. Note that S_t^k and D_a^{β}, $(a \in [t-k+1, t-k+2, \cdots, t-1])$ are independent random variables.

Using 4.55 and 4.85 we can determine $\eta(T)$ for $k \geq 1$. By employing expression similar to 4.53 through 4.83 we can efficiently determine an approximation of the aggregate requirement of incremental stocks needed when shipping capacity is active in the flexible delivery case.

The result found in this section suggests that when the capacity is very limited, having higher percentage of long response lead time demand with 1 day of grace period , will reduce the incremental stocks requirements significantly. Figure 4.1 shows an example when the aggregate daily demand is a normal distribution with mean and variance equal 1000 and 5000, respectively. The capacity is set at 1000 units for day.

4.2.4 Other Experiments

In Section 4.2.1 we described an experiment in which we tested the appropriateness of our assumption that the balance constraints could be ignored. We saw that this assumption is appropriate as long as shipping capacity is adequate. Suppose shipping capacity is limited. Would imbalance occur more frequently?

Another important question to address pertains to the importance of including the shipping constraint in the model when $l_i > 0$. This question, as we noted,

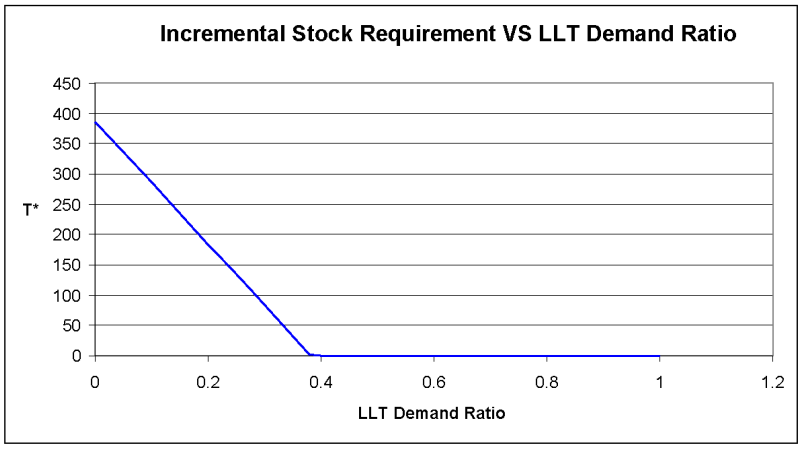

Figure 4.1: Incremental Stock Requirement VS LLT Demand Ratio

was of particular importance to the on-line retailer. We describe experiments designed to address these two questions.

Testing the Balance Assumption When Shipping Capacity is Limited

We first test the effect of ignoring the balance constraints in the optimization model when shipping constraints exist. To do this, we use the demand, cost, lead time and other data that were used in the experiment described in Section 4.2.1 .

We estimate the impact of shipping capacity on the number of imbalance incidents that occur in the experiment. The experiment was conducted as follows. For each of the four items, we assumed that there was a limit on the amount of that item that could be shipped in any period. For each item, we ran the experiments for four sets of capacity levels, labeled Case A through Case D. In Case A, the capacities exceed the expected per period demand by one unit in most cases. The exception is for item 1 for which the expected per-period demand

is less than one unit. Capacities are increased in each successive case. The capacities in Case D are equal to twice the per period expected demand for each item, except for item 1. For item 1, the capacity in Case D is 2 units, which is more than twice the expected per period demand. The capacity levels are given in Table 4.

Table 4: Capacity Levels

	Case A					Case B				
	RW 1	RW 2	RW 3	RW 4	RW 5	RW 1	RW 2	RW 3	RW 4	RW 5
Item 1	1	1	1	1	10000	N/A	N/A	N/A	N/A	10000
Item 2	2	3	4	4	10000	3	4	5	5	10000
Item 3	6	9	10	10	10000	7	10	11	11	10000
Item 4	21	26	26	32	10000	22	27	27	33	10000
	Case C					**Case D**				
	RW 1	RW 2	RW 3	RW 4	RW 5	RW 1	RW 2	RW 3	RW 4	RW 5
Item 1	N/A	N/A	N/A	N/A	10000	2	2	2	2	10000
Item 2	4	5	6	6	10000	5	6	7	7	10000
Item 3	8	12	13	13	10000	10	16	18	18	10000
Item 4	24	30	30	38	10000	40	50	50	62	10000

We assume $l_i = 0$ in this set of experiments as well. Given these capacity levels, we determined the order-up-to levels for each item. We then simulated 2000 cycles for each item (14000) days for each of 10 replications of the experiment. The results of the experiments are given in Table 5. As before, we report two numbers for each combination of items, capacities and VTM ratios. The first number represents the average number of periods in which imbalance occurred over the 10 replications of the experiment. The second number is the average percentage of the periods in which imbalance occurs. We again observe that imbalance rarely occurs. The maximum percentage is about 0.2% of the periods. Note that the percentage does not necessarily improve as capacity increases.

This is because the optimized inventory levels are higher when the capacity is lower.

Table 6: Flexible Delivery Impact

	$l_i = 0$				$l_i = 1$			
Capacity (Utilization)	1010 99.01%	1050 95.24%	1100 90.91%	1200 83.33%	1010 99.01%	1050 95.24%	1100 90.91%	1200 83.33%
(VTM) 1.01	0	0	0	0	0	0	0	0
2	51	0	0	0	0	0	0	0
3	147	0	0	0	0	0	0	0
5	386	0	0	0	0	0	0	0
10	1088	7	0	0	281	0	0	0

Measuring the Effect of l_i on Inventory Requirements

We now focus on determining the effect of the magnitude of l_i on incremental inventory requirements. To do so, we constructed a test in which the aggregate daily expected demand for a regional warehouse was scaled to 1000 units. We then determined the value of T^* using the methods described earlier in this section for various combinations of the VTM and the available capacity per period for long response lead time items.

We determined the value of T^* for four per-period capacity levels: 1010, 1050, 1100 and 1200 units. These levels correspond to utilization rates of 99.01%, 95.24%, 90.92% and 83.33%, respectively. Although, in practice, the planned utilization rate does not normally exceed 90%, we wanted to see how the incremental inventory levels would increase as a consequence of very high utilization rates.

Obviously the VTM ratio will also impact the amount of required incremental inventory. We considered five values for VTM: $1.01, 2, 3, 5$ and 10. Data that

we examined from an on-line retailer indicated that the VTM of the distribution of forecast errors for most items ranged from 1.1 to 3. We considered larger values in the experiment to estimate the consequences of increased uncertainty on the inventory requirements.

In all cases we assumed the short response lead time demand accounted for 20% of the total demand on average. Finally, we considered three values for l_i, $l_i = 0$, $l_i = 1$, and $l_i = 2$ periods. In the practical environment we examined, $l_i \geq 2$.

The resultant values for T^* are given in Table 6 for all combinations of the aforementioned factors. Keep in mind that these values are lower bounds. These results indicate that for practical problem environments, when the capacity utilization rate is around 80%, no incremental inventory is required. However, it is clear, and not surprising, that when capacity is just above the expected demand and there is substantial uncertainty concerning the aggregate demand process, then inventory levels will increase significantly when $l_i = 0$. The data in Table 6 illustrate this point when the capacity utilization rate is 99.01%. This observation is an important one for planners to comprehend.

Table 6: Flexible Delivery Impact

	l_i= 0					l_i= 1			
	Capacity (Utilization)	1010 99.01%	1050 95.24%	1100 90.91%	1200 83.33%	1010 99.01%	1050 95.24%	1100 90.91%	1200 83.33%
(VTM)	1.01	0	0	0	0	0	0	0	0
	2	51	0	0	0	0	0	0	0
	3	147	0	0	0	0	0	0	0
	5	386	0	0	0	0	0	0	0
	10	1088	7	0	0	281	0	0	0

Recall that the values of T^* that we calculated are lower bounds. Thus, we

constructed an experiment to test the impact of the shipping capacity using these values in cases of practical interest. While keeping costs and shipping lead times the same as before, assume each regional warehouse has an aggregate daily demand rate of 1000 units. In addition, we assume that the procurement lead time is 30 periods in length and the cycle length is 20 periods. We consider the worst case scenario when the system operates a single item with infinite supply at the primary warehouse. The goal is to test if our conjecture that no incremental inventory is needed in practical situations. Specifically, we let the capacity utilization be 83.33%. The VTM values were 1.01 and 3. In these cases $T^* = 0$. We simulated the demands incurred in 2000 periods and allocated the the inventories based on the method introduced in Section 4.2.1.

The results of these experiments are given in Table 7 and Table 8. The data in Table 7 show the average number of backorders during the $2,000$ periods and those in Table 8 show the average variance in the amount allocated to the regional warehouses each day over the simulation horizon. The reported values are the average of the result from 50 independent replications. The percentage is evaluated against the fixed delivery case when $l_i = 0$. The expected aggregate demand at each regional warehouse during the 100 cycles with 20 periods in each cycle, is approximately 2 million units.

Table 7 suggests two interesting points. First, when the utilization rate is 83.3%, the total number of backorders in each regional warehouse is at most 1665 units, which is about 0.08% of the total expected demand. Second, the number of backorders is slightly larger when $l_i = 0$. When $l_i > 0$, the shipping capacity is sufficient, long response lead time demands are always satisfied on time using the inventory at the primary warehouse. The backorders result from

the variation in short response lead time demand. These results support our conclusion that when capacity utilization rates are 83.3% or lower, inventory levels can be determined without considering the capacity constraints directly in the optimization process. Table 8 shows the advantage of having flexible delivery. The variance of daily allocation is about 17%−20% lower when flexible delivery is allowed.

Table 7: Total Backorders in 200 Cycles

		RW 1		RW 2		RW 3		RW 4	
VTM = 1.01	$I_0 = 0$	364.00	(100.0%)	806.00	(100.0%)	484.00	(100.0%)	382.00	(100.0%)
	$I_0 = 1$	364.00	(100.0%)	806.00	(100.0%)	484.00	(100.0%)	382.00	(100.0%)
	$I_0 = 2$	364.00	(100.0%)	806.00	(100.0%)	484.00	(100.0%)	382.00	(100.0%)
VTM = 3	$I_0 = 0$	877.00	(100.0%)	767.00	(100.0%)	859.00	(100.0%)	891.00	(100.0%)
	$I_0 = 1$	877.00	(100.0%)	767.00	(100.0%)	859.00	(100.0%)	891.00	(100.0%)
	$I_0 = 2$	877.00	(100.0%)	767.00	(100.0%)	859.00	(100.0%)	891.00	(100.0%)

Table 8: Variance of Daily Allocation

		RW 1		RW 2		RW 3		RW 4	
VTM = 1.01	$I_0 = 0$	1000.87	(100.0%)	1052.97	(100.0%)	1007.22	(100.0%)	1021.54	(100.0%)
	$I_0 = 1$	809.18	(80.8%)	863.42	(82.0%)	830.93	(82.5%)	828.46	(81.1%)
	$I_0 = 2$	845.85	(84.5%)	895.84	(85.1%)	857.83	(85.2%)	859.98	(84.2%)
VTM = 3	$I_0 = 0$	3031.07	(100.0%)	3002.23	(100.0%)	3045.89	(100.0%)	2887.46	(100.0%)
	$I_0 = 1$	2462.24	(81.2%)	2432.74	(81.0%)	2445.57	(80.3%)	2348.32	(81.3%)
	$I_0 = 2$	2521.22	(83.2%)	2493.71	(83.1%)	2529.47	(83.0%)	2390.10	(82.8%)

We performed another set of experiments by reducing the capacity to 1050, which corresponds to a utilization rate of 95.2%. Table 6 suggests that no incremental inventory is needed in this case. The results are shown in Table 9 and Table 10.

Table 9: Total Backorders in 200 Cycles

		RW 1		RW 2		RW 3		RW 4	
VTM = 1.01	$I_0 = 0$	436.00	(100.0%)	424.00	(100.0%)	759.00	(100.0%)	540.00	(100.0%)
	$I_0 = 1$	254.00	(58.3%)	343.00	(80.9%)	599.00	(78.9%)	443.00	(82.0%)
	$I_0 = 2$	254.00	(58.3%)	343.00	(80.9%)	599.00	(78.9%)	443.00	(82.0%)
VTM = 3	$I_0 = 0$	3608.00	(100.0%)	1728.00	(100.0%)	1942.00	(100.0%)	4905.00	(100.0%)
	$I_0 = 1$	785.00	(21.8%)	705.00	(40.8%)	810.00	(41.7%)	1002.00	(20.4%)
	$I_0 = 2$	785.00	(21.8%)	705.00	(40.8%)	810.00	(41.7%)	1002.00	(20.4%)

Table 10: Variance of Daily Allocation

		RW 1		RW 2		RW 3		RW 4	
VTM = 1.01	$I_0 = 0$	855.05	(100.0%)	930.98	(100.0%)	891.35	(100.0%)	951.42	(100.0%)
	$I_0 = 1$	719.99	(84.2%)	776.08	(83.4%)	769.52	(86.3%)	787.18	(82.7%)
	$I_0 = 2$	797.12	(93.2%)	835.65	(89.8%)	852.33	(95.6%)	861.73	(90.6%)
VTM = 3	$I_0 = 0$	2086.53	(100.0%)	1916.51	(100.0%)	2102.01	(100.0%)	1909.95	(100.0%)
	$I_0 = 1$	1846.29	(88.5%)	1681.81	(87.8%)	1832.09	(87.2%)	1671.42	(87.5%)
	$I_0 = 2$	2521.09	(120.8%)	2138.35	(111.6%)	2415.81	(114.9%)	2317.54	(121.3%)

The results suggest that the fill rates obtained when operating the system without augmenting the target inventory levels are still very high. Note that the maximum number of total backorders at a regional warehouse is 5929 units when VTM = 3 under the immediate response case. This is merely 0.30% of the total expected demand in the 2 million units of expected demand.

In this experiment, the shipping capacity utilization rate is high. Hence, under the immediate response case, there are more backorders due to the insufficiency of the shipping capacity. When the grace period is 1 or 2 days, the number of backorders is the same levels as observed. Hence the capacity is sufficient.

In addition, the data displayed in Table 10 suggests that when VTM = 3, the variance of daily allocation when $l_i = 2$ is higher than in the case when $l_i = 0$.

This is because the shipping capacity is limiting the variation of daily allocation in the immediate response time case.

The above experiment confirms our conjecture that under the delayed allocation system, when the short response lead time items only account for 20% of the total demand, no incremental inventory is needed under the presence of shipping capacity constraints. Thus, the question posed by the on-line retailer concerning the impact of using the delayed allocation system to manage inventories is answered. When the grace period is two or more days, that is, $l_i \geq 2$, then it is possible to smooth the flow of inventory throughout the system. This observation is made as a result of our simulation experiments in which the rules for satisfying customer orders that are discussed in detail in Chapter 5 are used.

4.3 Final Comments

We have described a modeling approach that can be employed to plan inventory, space and shipping capacity requirements for an online retailer's multi-location fulfilment system. We showed that balance and shipping constraints can be safely ignored when planning stock levels for the type of system we studied. Thus, the desired order-up-to levels can be determined one cycle at a time for each item.

We initially stated that our goal is to create a scalable computational procedure. For a primary warehouse system consisting of 5 regional warehouses, which as we noted corresponds to one we studied, we determined order-up-to levels for approximately $250,000$ items for a 15-month planning horizon. To calculate these levels required approximately 9.8 minutes on a PC with an Intel®

Xeon® Processor E5520 (2.26GHz). Thus the approach we have presented is one that planners can use in practical environments.

CHAPTER 5

EXECUTION MODEL

We have emphasized that the model discussed in Chapter 4 is designed to be used when planning fulfillment system operations. That is, it is a planning model rather than an execution model. Different models are needed to make daily procurement and allocation decisions. For example, our analysis shows that oftentimes it is not desirable to ship long response lead time demand to a regional warehouse at the end of a cycle. Rather it is sometimes better to wait when the grace period is positive and when demand during a cycle is higher than expected. By shipping the long response lead time demand before it must be shipped, a situation may arise in which it is impossible to replenish stocks needed to satisfy short response lead time demand for an item at other regional warehouses. Thus, it may be better to wait until the next replenishment order is received at a primary warehouse before some long response lead time demands are shipped. Of course, shipments must be made to ensure the customer delivery due dates are not violated.

Another example pertains to the fixed intervals of time between placing procurement orders. Obviously, there are times when it may be necessary to order stock sooner than planned. There are also times when no order needs to be placed in a planned period since demand was much lower than anticipated. An execution based model would consider such tradeoffs.

The planning model focuses on setting inventory levels. Customers, however, place orders which may contain many different item types. Thus allocation and fulfillment actions must recognize that orders must be satisfied as they arrive.

For these and many other reasons we consider another modeling approach to address day-to-day decision making in this chapter. There are three decisions to make in this execution model. The first decision pertains to the procurement decision for each item in a cycle. The procurement decision is made in anticipation of fulfilling both short and long response lead time demands that would occur before the next procurement arrives. The second decision pertains to the period-to-period allocation decisions from the primary warehouses to the regional warehouses. This allocation includes the inventory that will be cross-docked and used to satisfy the long response lead time demand orders and the inventories that are used to replenish the regional warehouse. The last decision is to fulfill customer orders at the regional warehouse level.

Note that once the procurement decisions are made, the allocation decisions are constrained by the available inventories. So are the order fulfillment decisions. Hence, we constructed three models that focus on one decision at a time. We are not saying that we are ignoring the interactions among the three decisions. Rather, in combination, these models address the sequence of decisions that must be made as time progresses and orders are ultimately fulfilled.

5.1 A Procurement Model

In the planning model, we assumed that the primary warehouse only places orders according to a fixed schedule. In reality, the primary warehouse does not strictly adhere to the planned schedule. As we have mentioned, if too much demand occurs for an item since the arrival of the last procurement, the primary warehouse will place an order to reduce the probability of incurring a stockout.

Similarly, if fewer customer orders are received, the primary warehouse can place the 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 are close to the fixed cycle-length discussed in the planning model. If there is a large discrepancy between the planned cycle length and the actual cycle length, it means the planning model is no longer accurate. The planning model would normally 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 each item in the current time period and the quantity of the procurement. The procurement decision is made based on the current inventory position and the target inventory level computed in the planning model.

5.1.1 Procurement Model Formulation

The procurement and the allocation and order fulfillment decisions become decoupled due to the lengthy procurement lead time. The procurement lead times are several weeks or longer. 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 inventory position computed from the planning model as a guidance of the execution strategy. From the target inventory position, we are able to compute the planned in-cycle service level. The goal of the procurement model is maintain the ser-

vice level during the execution stage.

Let y_j^* denote the target inventory position for item j at the beginning of each cycle. The planned in-cycle service level of item j at primary warehouse $m = m(j)$ can be computed as

$$p_j^* = P[D_{[0,L_m+\tau-1]} \leq y_j^*], \tag{5.1}$$

which is the probability that the total system demand over the time horizon of length $L_m + \tau - 1$ days does not exceed y_j^*.

Suppose t is some time in a cycle. Let y_t^j represent the system stock at time t. At time t, we calculate

$$p_t^j = P[D_{[t,t+L_m]} \leq y_t^j], \tag{5.2}$$

which is the probability that the expected demand within a lead time exceeding the current system inventory position.

If $p_t^j \leq p_j^*$, then the primary warehouse will place an order up to raise the echelon inventory position of item j to the target y_j^*.

5.2 Inventory Allocation

We now focus on the period-to-period inventory allocation decisions. There are three important attributes we address in this model. First, as mentioned, the fulfillment system consists of several warehouses located across country. Each warehouse serves as a primary warehouse for some item, and also serves as regional warehouse for the other items. In addition, the inventory to satisfy long

response lead time demand is stocked at the primary warehouses. When an order contains items stocked at different primary warehouses, actions from several warehouses must be coordinated to allocate inventories to satisfy a long response lead time order. Therefore, we cannot consider the primary warehouses individually in this case.

Second, the inventory planning and the procurement decisions are determined based on demand for items. As mentioned, a customer order may be for more than one item type and for more than one unit of an item type. When we make allocation decisions, we have to take the known order information into account to reduce the "last mile shipping cost".

Third, towards the end of the inventory cycle, we need to allocate inventory to satisfy long response lead time orders cautiously when the inventory for an item is running low. It may be desirable to not satisfy long response lead time demand before its due date so that several short response lead time orders that may occur can be satisfied.

The major contribution of this section is to present a model that can be used to allocate inventories from a primary warehouse to the regional warehouses when flexible delivery is allowed. In addition, this model is designed to incorporate the order information when allocating inventories.

5.2.1 Assumptions and Nomenclature

We start by introducing the assumptions underlying the the allocation model. One of the major constraints when making allocation decisions is the shipping

capacity constraint between each pair of the warehouses. The allocation includes three types of inventories, and we have to assign priorities to each type. First, the 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. Otherwise, the order will be backordered if the regional warehouse fails to do so. Second, the capacity is used to replenish the regional warehouse stock. Remember, the regional warehouse 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 long response lead time orders in advance. The problem is to coordinate the timing of allocating inventories to regional warehouse to fulfill the orders both known and unknown.

The objective in this stage is to minimize the inventory holding cost, backorder cost and the delivery cost to satisfy each order in a short planning horizon consisting of a few days. The delivery cost is roughly proportional to the number of packages used to satisfy one order. To simplify the problem, we charge an additional cost to orders that are not satisfied fully in one shipment.

Note that we only determine what quantity of each item to allocate in the current period. For the purpose of smoothing daily allocation, we will allocate inventories to fulfill long response lead time items only after orders for the items are known. Hence, the planning horizon now equals the length of the grace period, which is l periods in length.

We use I_{nt}^{j} to denote the echelon net inventory of item j at location n. Let L be the shipping time required to send any item from the primary warehouse to

a non-co-located regional warehouse $n = 1, \cdots, N-1$. As before, we assume that the shipping time from the primary warehouse to the co-located regional warehouse N is instantaneous.

For every order i, we use a_i^j to denote the number of units of item j that are requested in order i. Let $\kappa(i)$ denote the time period when customer order i is placed. We also record a time $\tau(i)$ which is the time period by which order i must be sent from regional warehouse $n(i)$ to a customer, where $n(i)$ is the regional warehouse that will fulfill order i. In the actual operation, orders with $\tau(i) < \kappa(i) + L$ are the short response lead time demand. Orders where $\tau(i) \geq \kappa(i) + L$ are the long response lead time orders. In particular, we assume that short response lead time demand must be shipped to the customer from the regional warehouse on the day the order is received, that is $\kappa(i) = \tau(i)$. At the non-co-located regional warehouse n, the short response lead time orders are satisfied from inventories on hand there. We use $d_{nt}^{j\alpha} = \sum_{i:n(i)=n,\tau(i)=\kappa(i)=t} a_i^j$ to denote the observed short response lead time demands of item j at regional warehouse n in time period t. $D_{nt}^{j\alpha}$, a random variable, i denotes the future short response lead time demand in period t.

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 in that order that has not been shipped previously cannot be sent to the customer on time using stock located at the primary warehouse. Let $d_{n\tau(i)}^{j\beta}$ denote the outstanding long response lead time demand that is due at time $\tau(i)$. 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 $\tau(i) - L < t$.

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 proportional to the number of units of unfilled demand that are not sent out from its regional warehouse by time $\tau(i)$. We have assumed in the previous chapter that backorders are satisfied from the regional warehouse stock in the subsequent period. In the actual operation of a fulfillment system, when an item is backlogged, the primary warehouse inventories will be used to satisfy those backorders directly if it is available. As a consequence, the fulfillment of backorders does not take up the shipping constraints from the primary warehouse to regional warehouses. We will address this assumption again when we introduce the constraints.

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 out today from the primary warehouse, that is, $\tau(i) = t + L_{n(i)}$, when $n(i) \in \{1, 2, \cdots, N-1\}$, $L_{n(i)} = L$ and when $n(i) = N$, $L_{n(i)} = 0$. The decisions are denoted by u_i^j, where u_i^j denotes the number of units of item j shipped to satisfy order i. Clearly, $u_i^j \leq a_i^j$, since we do not allocate more inventory than is requested by a customer. When there is not enough stock at the primary warehouse to fulfill an order completely, the inventory on hand at regional warehouse $n(i)$ will be used to fill the order. Let w_i^j denote the units of inventory of regional warehouse $n(i)$ on-hand stock used to fill order i, where $n(i) \neq N$. Since the co-located regional warehouse does not hold any inventory, $w_i^j = 0$ when $n(i) = N$. If the orders that are due at time $t + L$ are expected to be completely filled at time t, then $u_i^j + w_i^j = a_i^j$ for all items j for which $a_i^j > 0$. Let the binary variable $\tilde{z}_i = 1$ if order i is fulfilled completely.

The second decision pertains to the replenishment stock, which is denoted by y_{nt}^{j}. It represents the number of units of item j that are allocated from the primary warehouse to replenish the stock at regional warehouse n in time period t. The goal is to raise the inventory level at regional warehouse 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 longer response lead orders for which $\tau(i) > t+L_{n(i)}$. 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

$$x_{it} = \begin{cases} 1, & \text{if inventory and capacity are allocated in period } t \text{ to fulfill order } i \text{ completely,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Inventories are not allocated from the primary warehouse to a regional warehouse in anticipation of order fulfillment. Rather, these allocations are made only if an order is 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 time period $\tau(i)$. We have

$$\sum_{k=1}^{\tau(i)-L-1} x_{ik} + \tilde{z}_i + z_i = \tilde{x}_i + \tilde{z}_i + z_i = 1, \tag{5.4}$$

where $\tilde{x}_i = \sum_{k=1}^{\tau(i)-L-1} x_{ik}$, which indicates whether or not order i is planned to be fulfilled before $\tau(i)$. When order i is not expected to be satisfied, a penalty cost Q_i is charged, which means the expected incremental cost is due to the partial

fulfillment of the order. Note that we say planned not to be fulfilled. It still may be fulfilled even though $z_i = 1$. This can occur because short response lead time demand at regional warehouse $n(i)$ is less than anticipated during the lead time of length L.

Remember, $d_{nt}^{j\beta}$ is the outstanding long response lead time demand for item j due in period t. Hence, $d_{nt}^{j\beta} = \sum_{i:\tilde{x}_i=0,\tau(i)=t} (a_i^j - u_i^j)$. At regional warehouse N, recall that the allocation decisions are made following 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 $d_{Nt}^j = \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} (a_i^j - u_i^j)$. Note that d_{Nt}^j includes the short response lead time demand as well as the remaining long response lead time demand at regional warehouse N.

There are two goals we shall keep in mind in the allocation stage. We would like to minimize backorders while maximizing the number of orders filled completely. In the next section, we will construct a dynamic program that could be used to determine the optimal allocations for all three decision problems.

5.2.2 Single Primary Warehouse System

Recall that each warehouse serves as primary warehouse for some items as well as a regional warehouse for the other items. As a result, the allocation model is complicated since allocation decisions of whether 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 the ease of notation, 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 with N regional warehouses. We generalize our models to the N primary warehouse system case subsequently.

Dynamic Program

In this section, we formulate a dynamic program that can be used on a daily basis to make the three allocation decisions. When making these allocations, three types of costs are considered. At the primary warehouse, a holding cost h_0^j is charged at the end of each period proportional to on-hand inventories. Regional warehouse n charges a holding cost h_i^j 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. As noted, when a long response lead time order i is not fully fulfilled by time $\tau(i) - L$ from the primary warehouse, an incremental cost Q_i is charged to order i. This penalty cost is also charged to the short response lead time orders unfilled at the co-located warehouse.

The One Period Cost Model To formulate the problem as a dynamic program, we first need to construct a one-period cost model that calculates the holding costs, backorder costs and penalty shipping costs at all warehouses. The dynamic program's objective can be expressed as the one-period cost plus future expected costs.

Let I_{0t}^j represent the net inventory of item j at the primary warehouse at the end of period t and let q_{0t}^j denote the replenishment orders placed on the outside vendor in period t. The replenishment lead time from the vendor to the primary warehouse is L_j'. At the end of period t the net inventory of item j at the

primary warehouse is 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

$$I_{0t}^j = I_{0t-1}^j + q_{0t-L_j'}^j - \sum_{i:\tau(i)=t+L_{n(i)},\tilde{x}_i=0} u_{it}^j - \sum_{n=1}^{N-1} y_{nt}^j - \sum_{i:\tau(i)>t+L_{n(i)}} x_{it} a_i^j. \tag{5.5}$$

In addition to the holding cost, when we send out the inventories to fill long response lead time demand, we should plan to minimize the number of packages used to fulfill one order. Therefore, when there is not enough inventory to fill all orders from the primary warehouse, we shall 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 is a planned fulfillment rather than an actual fulfillment of the order. If during the shipping lead time an unexpectedly large amount of short lead time demand occurs for item j, then some portion of the w_i^j units may be used to satisfy these orders.

Let G_{0t} be the one period cost at the primary warehouse in time period t. Holding costs are charged proportional to the on-hand inventory of each item. In addition, for order i such that $\tau(i) = t + L$, the incremental shipping cost is charged when $z_i = 1$. Hence, G_{0t} can be expressed as

$$G_{0t} = \sum_j h_0^j(I_{0t}^j) + \sum_{i:\tau(i)=t+L} Q_i \cdot z_i \tag{5.6}$$

$$= \sum_j h_0(I_{0t-1}^j + q_{0t-L_0'}^j - \sum_{i:\tau(i)=t+L_{n(i)},\tilde{x}_i} u_i^j - \sum_{n=1}^{N-1} y_{nt}^j - \sum_{i:\tau(i)>t+L_{n(i)}} x_{it} a_i^j) + \sum_{i:\tau(i)=t+L_{n(i)}} Q_i \cdot z_i, \tag{5.7}$$

where

$$\tilde{z}_i \leq 1 - \frac{\sum_j (a_i^j - u_i^j - w_i^j)}{\sum_j a_i^j}, \tag{5.8}$$

$$\tilde{x}_i + \tilde{z}_i + z_i = 1, \tag{5.9}$$

and

$$\tilde{x}_i = \sum_{k=1}^{\tau(i)-L_{n(i)}-1} x_{it}. \tag{5.10}$$

x_{it}, z_i, and $\tilde{z}_i$ are binary variables.

Next let us analyze the one-period cost at regional warehouse $n \in \{1, 2, \cdots, N-1\}$. The inventory level at the end of period t is the 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 long response lead time demands that are not satisfied from the stock at the primary warehouse $d_{nt}^{j\beta}$ and the demands that are satisfied from the regional warehouse n stock $d_{nt}^{j\alpha}$.

Thus,

$$I_{nt}^j = I_{nt-1}^j + y_{nt-L}^j - d_{nt}^{j\alpha} - d_{nt}^{j\beta} \tag{5.11}$$

$$= I_{nt-1}^j + y_{nt-L}^j - d_{nt}^{j\alpha} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a_i^j - u_i^j) \tag{5.12}$$

$$= I_{nt-1}^j + y_{nt-L}^j + \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} u_i^j - d_{nt}^{j\alpha} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} a_i^j. \tag{5.13}$$

Let $G_{nt}(d_{nt}^{\alpha}, I_{n,t-1}^j, y_{n,t-L}^j, u_i^j, a_i^j)$ represent the one period cost incurred at regional warehouse n. Then

$$G_{nt}(d_{nt}^{\alpha}, I_{n,t-1}^{j}, y_{n,t-L}^{j}, u_i^j, a_i^j) = \sum_j \left[h_n^j (I_{nt}^j)^+ + b^j (-I_{nt}^j)^+ \right], \tag{5.14}$$

$$= \sum_j \left[h_n^j (I_{nt-1}^j + y_{nt-L}^j + \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} u_i^j - d_{nt}^{j\alpha} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} a_i^j)^+ \right. \tag{5.15}$$

$$\left. + b^j (-I_{nt-1}^j - y_{nt-L}^j - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} u_i^j + d_{nt}^{j\alpha} + \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} a_i^j)^+ \right]. \tag{5.16}$$

For regional warehouse N all allocations are made after customer orders are received, since there is no replenishment stock. Thus

$$I_{Nt}^j = I_{Nt-1}^j + \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} u_i^j + y_{Nt}^j - \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} a_i^j. \tag{5.17}$$

The backorder cost at the co-located regional warehouse N is

$$G_{Nt}(I_{N,t-1}^j, u_i^j, a_i^j) = \sum_j \left[b^j (-I_{Nt-1}^j - y_{Nt}^j - \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} u_i^j + \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} a_i^j)^+ \right].$$

Therefore, the one-period costs charged across all warehouses at the end of period t are

$$G_{0t}(d_{nt}^{j\alpha}, I_{n,t-1}^j, u_i^j, a_i^j) + \sum_{n=1}^{N-1} G_{nt}(d_{nt}^{j\alpha}, I_{n,t-1}^j, y_{n,t-L}^j, u_i^j, a_i^j) + G_{Nt}(I_{N,t-1}^j, y_{Nt,}^j, u_i^j, a_i^j). \tag{5.18}$$

The objective is to determine the allocation strategy that minimizes the cost over the time horizon consisting of periods t through $t + L + l$, where l is the length of the grace period. Let $\mathbf{I}_t$ denote the vector of the net inventory levels

at the end of period t at all locations for all items. The corresponding dynamic programming recursion for the allocation problem is

$$V_t(\mathbf{I}_t) = \mathbb{E}\min[\, G_{0t}(d_{nt}^{j\alpha}, I_{n,t-1}^j, u_i^j, a_i^j) + \sum_{n=1}^{N-1} G_{nt}(d_{nt}^{j\alpha}, I_{n,t-1}^j, y_{n,t-L}^j, u_i^j, a_i^j) + G_{Nt}(I_{N,t-1}^j, y_{Nt}^j u_i^j, a_i^j) + V_{t+1}(\mathbf{I}_{t+1}) \,].$$

The solutions must also satisfy the following constraints.

- The primary warehouse cannot allocate to regional warehouses more than the amount it has on hand

$$\sum_n y_{nt}^j + \sum_{i:\tau(i)=t+L_{n(i)},\tilde{x}_i=0} u_i^j + \sum_{i:\tau(i)>t+L_{n(i)}} x_t^i a_j^i \le I_{0,t-1}^j + q_{0,t-L_j'}^j. \tag{5.19}$$

- The shipping capacity from the primary warehouse to regional warehouse $n \neq N$ is limited by C_{nt} from period-to-period. Hence

$$\sum_j (y_{nt}^j + \sum_{i:n(i)=n,\tau(i)=t+L,\tilde{x}_i=0} u_i^j + \sum_{i:n(i)=n,\tilde{x}_i=0,\tau(i)>t+L} x_{it} a_i^j) \le C_{nt}. \tag{5.20}$$

- An order is either fulfilled fully in advance or not.

$$\sum_{k=1}^{\tau(i)-L_{n(i)}-1} x_{ik} + \tilde{z}_i + z_i = 1. \tag{5.21}$$

- Order i is anticipated to be fulfilled completely in period $\tau(i) - L_{n(i)}$

$$\tilde{z}_i \le 1 - \frac{\sum_j (a_{i,}^j - u_i^j - w_i^j)}{\sum_j a_{i,}^j}. \tag{5.22}$$

- A regional warehouse can only allocate on-hand inventory to orders that are expected to be filled at the due date, which is L periods in the future, $n \neq N$.

$$\sum_{i:n(i)=n,\tau(i)=t+L,\tilde{x}_i=0} w_i^j \le I_{nt}^j + q_{n,t-L}^j + \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_{it}=0} u_i^j - d_{nt}^{j\alpha} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_{it}=0} a_i^j. \tag{5.23}$$

To summarize the above, the dynamic programming recursion is

$$V_t(\mathbf{I}_t) = \min\,[\; G_{0t}(d_{nt}^{j\alpha}, I_{n,t-1}^{j}, u_i^j, a_i^j) + \sum_{n=1}^{N-1} G_{nt}(d_{nt}^{j\alpha}, I_{n,t-1}^{j}, y_{n,t-L}^{j}, u_i^j, a_i^j) + G_{Nt}(I_{N,t-1}^{j}, u_i^j, a_i^j) + V_{t+1}(\mathbf{I}_{t+1})\;]$$

$$\text{s.t.} \sum_n y_{nt}^j + \sum_{i:\tau(i)=t+L_{n(i)},\tilde{x}_{it}=0} u_i^j + \sum_{i:\tau(i)>t+L_{n(i)}} x_t^i a_j^i \le I_{0,t-1}^j + q_{0,t-L_j'}^j, \tag{5.24}$$

$$\sum_j (y_{nt}^j + \sum_{i:n(i)=n,\tau(i)=t+L,\tilde{x}_i=0} u_i^j + \sum_{i:n(i)=n,\tilde{x}_i=0,\tau(i)>t+L} x_{it} a_i^j) \le C_n, \tag{5.25}$$

$$\sum_{k=1}^{\tau(i)-L_{n(i)}-1} x_{ik} + \tilde{z}_i + z_i = 1, \tag{5.26}$$

$$\tilde{z}_i \le 1 - \frac{\sum_j (a_{i,}^j - u_i^j - w_i^j)}{\sum_j a_{i,}^j}, \tag{5.27}$$

$$\sum_{i:n(i)=n,\tau(i)=t+L,\tilde{x}_i=0} w_i^j \le I_{nt}^j \tag{5.28}$$

$$I_{nt}^j = I_{nt-1}^j + y_{n,t-L}^j + \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_{it}=0} u_i^j - d_{nt}^{j\alpha} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_{it}=0} a_i^j, n \ne N, \tag{5.29}$$

$$I_{Nt}^j = I_{Nt-1}^j + \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} u_i^j + y_{Nt}^j - \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} a_i^j, \tag{5.30}$$

$$I_{0t}^j = I_{0t-1}^j + q_{0t-L_0'}^j - \sum_{i:\tau(i)=t+L_{n(i)},\tilde{x}_i} u_{it}^j - \sum_{n=1}^{N-1} y_{nt}^j - \sum_{i:\tau(i)>t+L_{n(i)}} x_{it} a_i^j. \tag{5.31}$$

Note that this dynamic program formulation cannot be solved directly due to the size of the state space. Hence, we will show how to construct a sequence of approximate models that provide the desired allocation and fulfillment quantities.

An Approach for Making Allocation and Order Fulfillment Decisions

In this section, we present an approach for obtaining solutions to the above dynamic program. Recall that there are three distinct decisions that must be made each day. We have created sub models to address them in sequence. In the first sub-model, we determine which orders we are expecting to fill after L periods. In the second sub-model, we determine how to allocate inventory of each item to replenish regional warehouse stocks. In the last sub-model we determine which orders to fulfill completely prior to their due dates. We now formulate three separate problems, one corresponding to each of the three types of decisions.

Sub-model 0: Satisfy Unfilled Backorders At Co-located Regional Warehouse
Before we start allocating inventory to fill orders that are required to be satisfied today or later, we first allocate on-hand inventory at the primary warehouse to satisfy as much of the the backlogged demand as possible at each regional warehouse. Let y^j_{n0} denote the amount of item j allocated to regional warehouse n to satisfy backorders in time period 1 and recall $q^j_{0,t-L'_j}$ denote the known replenishment order for item j that is scheduled to arrive at primary warehouse in period t.

The available on-hand inventory of item j at the primary warehouse is I^j_{00} at the beginning of the time horizon. After fulfilling existing backorders, the effective inventory level there for item j at the beginning of the planning horizon is $I^{j\prime}_{00} = (I^j_{00} - \sum_{n=1}^{N} y^j_{n0} + q^j_{0,t-L'_j})$. The inventory level of item j at regional warehouse n increases by y^j_{n0}. Hence, $I^{j\prime}_{n0} = I^j_{n0} + y^j_{n0}$. Note that $\sum_n y^j_{n0} \leq I^j_{00} + q^j_{0,t-L'_j}$.

Sub-model 1: Fulfill Orders that Must be Sent Out Today In this sub-model we determine how inventories at the primary warehouse and regional warehouses should be allocated to satisfy orders that must be filled at the regional warehouse a lead time $L_{n(i)}$ in the future. Thus, we are only interested in allocating inventory to satisfy demands that are due on day $L_{n(i)} + 1$.

To begin, we first identify those outstanding orders for which $\tau(i) = t + L_{n(i)}$ that can be completely satisfied from the primary warehouse stock. For each item j, determine if $I_{00}^{j\prime} \geq \sum_{i:\tau(i)=t+L_{n(i)},\tilde{x}_i=0} a_i^j$, where $L_{n(i)} = L$ when $n(i) \neq N$ and $L_N = 0$ when $n(i) = N$. Let J be the set of items j for which the above inequality holds. Next, let I be the set of unsatisfied orders $i : \tau(i) = t + L_{n(i)}$ for which $a_i^j > 0$ only for items $j \in J$. Allocate stocks to these orders and decrement stocks accordingly. $\forall i \in I$, set $\tilde{x}_i = 1$. Then $j \notin J$ implies there is not enough stock on hand at the primary warehouse to satisfy all orders for which $a_i^j > 0$.

Recall that u_i^j denotes the number of units of item j sent to $n(i)$ from the primary warehouse and allocated to order i . The objective in this second part of sub-model 1 is to allocate inventories so that we maximize the number of orders satisfied completely. Define O to be the set of long response lead time orders that are due at time $\tau(i) = 1 + L$ at regional warehouse $n \neq N$, and the short and long response lead time orders that are due at the co-located regional warehouse N for which $\tau(i) = 1$ and that remain unfilled.

The regional warehouse allocation made when solving this first sub-model is a preliminary one, which may be revised subsequently. This revision may occur because short response lead time demand occurring between the current day, day 1, and day $L + 1$ may make the planned allocation impossible to carry out. Recall that w_i^j units of item j stock on hand at regional warehouse $n(i)$

were planned to be used to fill order i. Suppose further that some of these units were needed to satisfy several short response lead time orders completely. Thus, rather than using them to satisfy the single long response lead time order i, they were used to satisfy several orders fully.

Recall that $z_i = 1$ if order i is not filled completely. Thus, the objective in this sub-model is to minimize the sum of the penalty costs charged based on the number of orders not in set I, that is, our goal is to minimize the proportion of unfilled orders, that is $\min \sum_{i \in O} Q_i \cdot \left(\frac{\sum_j (a_i^j - u_i^j - w_i^j)}{\sum_j a_i^j} \right)$.

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. Second, z_i is 1 if a_i^j units of item j are allocated to order i. The final 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.

Hence, the formulation of the model is

$$\min \sum_{i \in O} Q_i \cdot \left(\frac{\sum_j (a_i^j - u_i^j - w_i^j)}{\sum_j a_i^j} \right) \tag{5.32}$$

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

$$\sum_{i \in O} u_i^j \leq I'^j_{00}, \qquad \forall j, \tag{5.33}$$

$$\sum_{i \in O : n(i) = n} w_i^j \leq I_{n0}^j + \sum_{k=2}^{1+L} S_{nk}^j - d_{n1}^{j\alpha} - \sum_{k=2}^{L+1} E[D_{nk}^{j\alpha}], \forall j \tag{5.34}$$

$$\text{for which the right hand side is positive}, n(i) \neq N, \tag{5.35}$$

where S_{nt}^j is the number of units of item j received at regional warehouse n in period t corresponding to shipments made from the primary warehouse in

period $t - L$, when $n \neq N$.

Sub-Model 2: Allocate Replenishment Stock to Regional Warehouses After making allocations that minimize the number of unsatisfied orders due on day 1, we next allocate stocks to replenish regional warehouse inventories. Recall that regional warehouses carry inventories to fulfill short response lead time orders. We call the inventory position for each item at the regional warehouse that minimizes the echelon inventory holding and backorder costs the target inventory level. If the inventory position at regional warehouse n is below its target inventory level entering period 1 the goal is to raise it to the target level. These levels are determined in the planning model. Thus, the goal of the second sub model is to determine the amount of replenishment stocks to allocate of each item type to each regional warehouse.

The supply at the primary warehouse that is available for allocation is

$$I_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j. \tag{5.36}$$

The decision variable value that we next determine is y_{n1}^j, which is the amount of item type j 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 $I_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j - \sum_n y_{n1}^j$.

The net inventory of item j at regional warehouse $n \neq N$ at the end of period $L+1$ is $I_{n0}^{j\prime} + \sum_{k=2}^{L} S_{nk}^j + y_{n1}^j - d_{n1}^{j\alpha} - \sum_{k=2}^{L+1} D_{nk}^{j\alpha} - \sum_{k=1}^{L+1} \sum_{i:\tau(i)=k,\tilde{x}_i=0} (a_i^j - u_i^j)$. Note that we are looking at the expected cost incurred in period $L+1$ instead of period 1 since the the consequence of the decision y_{n1}^j is realized in period $L+1$. In this expression, the only random variables are the short response lead time demand random

variables, $D_{nk}^{j\alpha}$. To simplify the notation, let $\bar{I}_{n0}^{j} = I_{n0}^{j\prime} + \sum_{k=2}^{L} S_{nk}^{j} - d_{n1}^{j\alpha} - \sum_{k=1}^{L+1} \sum_{i:\tau(i)=k,\tilde{x}_i=0} (a_i^j - u_i^j))$. Hence, the net inventory level of item j at regional warehouse n in time period $L+1$ can be rewritten as $\bar{I}_{n0}^{j} = \bar{I}_{n0}^{j} + y_{n1}^{j} - \sum_{k=2}^{L+1} D_{nk}^{j\alpha}$.

Finally, the net inventory at the end of period 1 at regional warehouse N is $I_{N0}^{j\prime} + y_{N1}^{j} + \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} u_i^j - \sum_{i:\tau(i)=1,n(i)=N} a_i^j$. Similarly, we can simplify the notation by $\bar{I}_{N0}^{j} = I_{N0}^{j\prime} + \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} u_i^j - \sum_{i:\tau(i)=1,n(i)=N,\tilde{x}_i=0} a_i^j$.

The one-period cost across all warehouses is therefore

$$\sum_{j} h_0^j (I_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j - \sum_{n} y_{n1}^{j}) \tag{5.37}$$

$$+ \sum_{n=1}^{N-1} \sum_{j} \mathbf{E}[h(\bar{I}_{n0}^{j} + y_{n1}^{j} - \sum_{k=2}^{L+1} D_{nk}^{j\alpha})^{+}] \tag{5.38}$$

$$+ \sum_{n=1}^{N-1} \sum_{j} \mathbf{E}[b^j(-\bar{I}_{n0}^{j} - y_{n1}^{j} + \sum_{k=2}^{L+1} D_{nk}^{j\alpha})^{+}] \tag{5.39}$$

$$+ \sum_{j} b^j(-\bar{I}_{N0}^{j} - y_{N1}^{j})^{+}. \tag{5.40}$$

We propose to find the value of y_{n1}^{j} by minimizing one period holding and backorder costs at all locations. The replenishment decisions are constrained by the remaining capacity and available inventory at the primary warehouse. These two constraints can be expressed as

$$\sum_{n=1}^{N} y_{n1}^{j} \leq I_{00}^{\prime j} - \sum_{i:\tau(i)=L_{n(i)}+1,,\tilde{x}_i=0} u_i^j, \tag{5.41}$$

$$\sum_{j} y_{n1}^{j} \leq C_{nt} - \sum_{i:n(i)=n,\tau(i)=1+L,\tilde{x}_i=0} u_i^j. \tag{5.42}$$

In addition to these constraints, we do not wish to allocate too much inventory in advance to regional warehouses and subsequently to have the inventory in one regional warehouse when it is needed elsewhere. This is the imbalance situation we discussed in the previous chapter. Imbalance is more likely to occur at the end of a cycle when the primary warehouse inventory is low. It may be beneficial to hold inventory back at the primary warehouse at this stage, and to allocate inventories to the remaining known long response lead time demand. Let us consider an example. Suppose the primary warehouse only has one unit of item j left on-hand prior to making the allocation decision. Suppose the next procurement order will arrive at the primary warehouse two days later. Suppose only one regional warehouse n requires the one unit to achieve its target inventory level. Furthermore, suppose there is one unit of outstanding long response lead time demand of item j at regional warehouse $k \neq n$ that must be sent from the primary warehouse to regional warehouse k in the next period. In this case, we would use this one unit of item j to satisfy the long response lead time demand at regional warehouse k rather than use it to replenish the stock at regional warehouse n.

To prevent too much inventory being allocated to a regional warehouse and to reduce the desire to redistribute the inventories between regional warehouses, we introduce the following constraint. Let χ_t^j denote the maximum number of units of item j that we permit to be available to replenish regional warehouse stocks in period t. Let Γ denote the period before the next procurement shipment of item j arrives at the primary warehouse. χ_t^j is also the total net supply of item j that is available for allocation through period Γ, where the net supply of item j is defined to be the current on-hand inventory at the primary warehouse minus the sum of outstanding known and expected long response

lead time demand of item j sent out from the primary warehouse by time period Γ.

When $l + L_{n(i)} \leq \Gamma + L_{n(i)}$,

$$\chi_1^j = \bar{I}_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j - \sum_{i:L_{n(i)}+2\leq\tau(i)\leq L_{n(i)}+l,\tilde{x}_i=0} a_i^j - \mathbb{E}\sum_{k=l+L}^{\Gamma+L} D_{0k}^{j\beta}, \tag{5.43}$$

where $D_{0k}^{j\beta}$ is the random variable of long response lead time demand for item j due on day k at all regional warehouses.

When $l + L_{n(i)} > \Gamma + L_{n(i)}$, all long response lead time orders are known. Thus

$$\chi_1^j = I_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j - \sum_{i:L_{n(i)}+2\leq\tau(i)\leq\Gamma+L,\tilde{x}_i=0} a_i^j. \tag{5.44}$$

The constraint we have is

$$\sum_n y_{n1}^j \leq \chi_1^j.$$

To summarize, the model is

$$\min \sum_j h_0^j (I_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j - \sum_n y_{n1}^j) \tag{5.45}$$

$$+ \sum_{n=1}^{N-1} \sum_j \mathbf{E}[h_n^j (\bar{I}_{n0}^j + y_{n1}^j - \sum_{k=2}^{L+1} D_{nk}^{j\alpha})^+] \tag{5.46}$$

$$+ \sum_{n=1}^{N-1} \sum_j \mathbf{E}[b^j (-\bar{I}_{n0}^j - y_{n1}^j + \sum_{k=2}^{L+1} D_{nk}^{j\alpha})^+] \tag{5.47}$$

$$+ \sum_j b^j (-\bar{I}_{N0}^{\prime j} - y_{N1}^j)^+. \tag{5.48}$$

$$\text{s.t.} \sum_{n=1}^{N} y_{n1}^j \leq I_{00}^{j\prime} - \sum_{i:\tau(i)=L_{n(i)}+1,\tilde{x}_i=0} u_i^j, \tag{5.49}$$

$$\sum_j y_{n1}^j \leq C_n - \sum_{i:n(i)=n,\tau(i)=L_{n(i)}+1,\tilde{x}_i=0} u_i^j, \tag{5.50}$$

$$\sum_n y_{n1}^j \leq \chi_1^j. \tag{5.51}$$

This myopic model does not necessarily provide an optimal solution since we do not account for the impact the current decision has on future ones. Thus, we assume that the decision in one period does not affect those made in subsequent periods.

Sub-model 3: Allocate Inventory to Fill Orders Due in the Future In this model, we determine whether we should fulfill some orders in advance with the remaining inventories and shipping capacities. Let us first recompute the available inventories $\bar{I}_{02}^j = I_{00}^{j\prime} - \sum_{i:\tau(i)=1+L_{n(i)},\tilde{x}_i=0} u_i^j - \sum_n y_{n1}^j - \gamma_1^j$, where γ_1^j is the amount of inventory of item j that the primary warehouse is expected to require to raise the inventories at the regional warehouses to their target levels until the next period in which the primary warehouse is replenished by the vendor.

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 $L_{n(i)}+t$. We constructed a model to determine which of the orders in O_t should be shipped to the regional warehouse in their entirety in period 1. Let us start with $t = 2$. The objective is to maximize the number of these orders that can be fulfilled completely using the remaining available inventory and shipping capacity. Let $y_{nt}^{j\prime} = \sum_{i:\tau(i)=t+L_{n(i)},n(i)=n} x_{it}a_i^j$. Then our third sub-model is :

$$\max \sum_{i:\tau(i)=t+L_{n(i)}} x_{it} \tag{5.52}$$

$$\text{s.t.} \sum_n y_{nt}^{j\prime} \leq \bar{I}_{0t}^j, \tag{5.53}$$

$$\sum_j y_{nt}^{j\prime} \leq \bar{C}_{nt}, \tag{5.54}$$

$$y_{nt}^{j\prime} = \sum_{i:\tau(i)=t+L_{n(i)}} x_{it}a_{i,}^j, \tag{5.55}$$

$$\tag{5.56}$$

where $\bar{C}_{nt}$ is the remaining shipping capacity from the primary warehouse to regional warehouse n.

Let $\bar{I}_{0,t+1}^j = \bar{I}_{0,t}^j - \sum_n y_{nt}^{j\prime}$ and $\bar{C}_n = \bar{C}_n - \sum_j y_{nt}^{j\prime}$. Solve the above problem (5.52)-(5.55) with $t = t + 1$. Continue in this manner until capacity or inventories run out.

5.2.3 Complete System with N Warehouses

We are now ready to extend our results from the one primary warehouse system to the case where each warehouse serves as a primary warehouse for some

items and serves as a regional warehouse for the other items. We mentioned that a customer may order multiple items in the same order. The fulfillment system wants to the send all items to the customers in the same shipment. If the items in the same order are managed by different warehouses, then the allocation decisions made at one primary warehouse will affect the decision made at another primary warehouse. In this section, we develop an allocation model that coordinates the allocation decisions among all primary warehouses.

Let $m(j)$ denote the primary warehouse of item j and $n(i)$ denote the regional warehouse designated to fill order i. When $n(i) = m(j)$, the regional warehouse is a co-located regional warehouse for item j in order i.

Orders can be categorized into three types in this system. We will illustrate the allocation policies for each type of orders.

1. Let θ_1 denote the set of orders i, when $m(j) = n(i)$ for all items j in order i. In this case, regional warehouse $n(i)$ serves as a co-located regional warehouse for all items requested in order i. We will first try to fill these orders completely before $\tau(i)$ from primary warehouse $m(j)$. We discuss this pre-filling idea more fully in Section 5.3. If not filled previously, we need to determine the amount of inventory to allocate to this order at time $\tau(i)$ for each item j. Note that the policy is the same for both long response and short response lead time orders. These orders can be satisfied anytime through its due date without incurring backorder costs.
2. Let θ_2 denote the set of orders i for which $m(j) \neq n(i)$ for all items j. The regional warehouse is a non-co-located regional warehouse for items j requested within order i. Recall that $\kappa(i)$ is the period in which order i is placed. For a long response lead time orders i, where $\tau(i) \geq \kappa(i) + L$, the pri-

mary warehouses will first attempt to fill this order completely before the end of the the grace period. If it is not possible to fill this order completely either due to insufficient inventory or shipping capacity before $\tau(i) - L$, we will *allow* a partial fulfillment of this order in time period $\tau(i) - L$ with a penalty cost. Inventories at the regional warehouse $n(i)$ may also be used to fill this order at time $\tau(i)$ to avoid a backorder cost. Similar to the one primary warehouse model, we use w_j^i to denote the number of units of item j that is planned to be allocated to order i. For short response lead time order i, where $\tau(i) = \kappa(i)$, regional warehouse $n(i)$ will use its inventory to fill this order.

3. Let θ_3 denote the set of orders defined as follows. For $i \in \theta_3$. $\exists j, m(j) = n(i)$ and $\exists j, m(j) \neq n(i)$. In this case, the warehouse $n(i)$ is a co-located regional warehouse for some items, and it also serves as non-co-located regional warehouse for the other items. For short response lead time orders, the regional warehouse $n(i)$ will use its stock to satisfy the demand for item j that is requested in order i when $m(j) \neq n(i)$. When $m(j) = n(i)$, then the primary warehouse will allocate its inventory to satisfy the order. For long response lead time orders, the decisions are made during the grace period, which is from $\kappa(i)$ through $\tau(i) - L$. Note that we will decrement the inventories at all primary warehouses once the fulfillment decision is made.

Single Period Model Similar to the single primary warehouse case, there are three allocation decisions to make at each primary warehouse in time period t. For each order, we use L_i to denote the lead time. If the order $i \in \theta_2 \cup \theta_3$, let $L_i = L$. Otherwise, $L_i = 0$. The allocation of units of item type j to order i

that is sent from the primary warehouse is denoted by u_i^j, where $\tau(i) = t - L_i$. Following these allocations, inventories and capacities are allocated to replenish regional warehouse stocks from the appropriate primary warehouse. If the primary warehouse is denoted by m and the regional warehouse by n, let y_{mnt}^j be the number of units of item j allocated to n from m. This is replenishment stock when $m \neq n$. When $n = m$, then it represents the amount used to satisfy backorders existing at the beginning of period t at regional warehouse n. In addition, if $n = m$, and $i \in \theta_3$, then y_{mmt}^j also includes fulfillment of short response lead time demand. Finally, the last decision is the amount of inventory allocated to fulfill orders completely that are not yet due.

In reality, backlogged orders at any regional warehouse are broken up into suborders by the primary warehouse locations and will be satisfied directly using the inventory there. In this dynamic program, we assume that all backorders are recorded and fulfilled from regional warehouses for accounting purposes. We use I_{mt}^{jP} and I_{nt}^{jR} to denote the ending inventory levels of item j in period t at primary warehouse m and regional warehouse n, respectively.

We start by defining the one period cost function at a primary warehouse. The ending net inventory in period t for item j at primary warehouse $m = m(j)$ can be recursively defined as

$$I_{m,t}^{jP} = I_{m,t-1}^{jP} + q_{m,t-L'_j}^{j} - \sum_{i:\tau(i)=t+L_i,\tilde{x}_i=0} u_i^j - \sum_{n=1}^{N} y_{mnt}^j - \sum_{i:\tau(i)>t+L_i} x_{it} a_i^j. \tag{5.57}$$

Hence, the one period cost expression for item j at the end of period t at primary warehouse $m = m(j)$ is

$$G^{jP}_{m(j)t}(I^{jP}_{mt-1}, q^{j}_{mt-L_j}, a^{j}_{i}, u^{j}_{i}, y^{j}_{mnt}) \tag{5.58}$$

$$= \sum_{j} h^{j}_{m}(I^{jP}_{mt}) + \sum_{i:\tau(i)=t+L} Q_i \cdot z_i, \tag{5.59}$$

$$= \sum_{j} h^{P}_{m}(I^{jP}_{m,t-1} + q^{j}_{mt-L'_j} - \sum_{i:\tau(i)=t+L_i,\tilde{x}_i=0} u^{j}_{i} - \sum_{n=1}^{N} y^{j}_{mnt} - \sum_{i:\tau(i)>t+L_i} x_{it} a^{j}_{i}) \tag{5.60}$$

$$+ \sum_{i:\tau(i)=t+L_i} Q_i \cdot z_i, \tag{5.61}$$

where $z_i = 1$ if order i is not expected to be filled by time $\tau(i)$.

At the non-co-located regional warehouse n, where $m(j) \neq n$, the net inventory at the end of period t for item j is

$$I^{jR}_{n,t} = I^{jR}_{n,t-1} + y^{j}_{m(j)n,t-L} - d^{j\alpha}_{nt} - d^{j\beta}_{nt} \tag{5.62}$$

$$= I^{jR}_{n,t-1} + y^{j}_{m(j),n,t-L} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a^{j}_{i} - u^{j}_{i}) - d^{j\alpha}_{nt}. \tag{5.63}$$

Let $G^{jR}_{nt}(d^{\alpha}_{nt}, I^{jR}_{n,t-1}, y^{j}_{m(j)n,t-L}, u^{j}_{i}, a^{j}_{i})$ represents the one period cost incurred at a regional warehouse n, $m(j) \neq n$. Then

$$G^{jR}_{nt}(d^{\alpha}_{nt}, I^{jR}_{n,t-1}, y^{j}_{m(j)n,t-L}, u^{j}_{i}, a^{j}_{i}) = h^{j}_{n}(I^{jR}_{nt})^{+} + b^{j}(-I^{jR}_{nt})^{+}, m(j) \neq n. \tag{5.64}$$

At the co-located regional warehouse, where $m(j) = n$, the ending net inventory in period t for item j is

$$I^{jR}_{n,t} = I^{jR}_{n,t-1} + y^{j}_{mnt} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a^{j}_{i} - u^{j}_{i}). \tag{5.65}$$

It is important to recall that y^{j}_{mnt} measures only the amount of item j used to satisfy backorders when $n(i) = n$. However, when $n(i) = n$ and $i \in \theta_3$, then y^{j}_{mnt} also includes the allocation made to satisfy short response lead time demand at warehouse n.

The backorder cost at the co-located regional warehouse n, where $m(j) = n$, is

$$G_{nt}^{jR}(I_{n,t-1}^{jR}, u_i^j, a_i^j, y_{mnt}^j) = b^j(-I_{nt-1}^{jR} + y_{mnt}^j - \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_{it}=0} u_i^j + \sum_{i:n(i)=N,\tau(i)=t,\tilde{x}_i=0} a_i^j)^+, \quad (5.66)$$

Therefore, the one-period costs charged across all warehouses at the end of period t are

$$\sum_j G_{m(j)t}^{jP}(I_{m,t-1}^{jP}, q_{m,t-L'_j}, a_i^j, u_i^j, y_{mnt}^j) + \sum_{n=1}^{N} \sum_j G_{nt}^{jR}(d_{nt}^{j\alpha}, I_{n,t-1}^{jR}, y_{m(j)n,t-L_n}^j, u_i^j, a_i^j). \quad (5.67)$$

Constraints The constraints are similar to the ones in the single primary warehouse case. Except that the lead time of one order is a function of i, rather than $n(i)$.

The supply at the primary warehouse is always non-negative:

$$\sum_{i:\tau(i)=t+L_i,\tilde{x}_i=0} u_i^j + \sum_{n=1}^{N} y_{mnt}^j + \sum_{i:\tau(i)>t+L_i} x_{it} a_i^j \le I_{m,t-1}^{jP} + q_{m,t-L'_m}^j. \quad (5.68)$$

Shipping capacity is limited from the primary warehouse to a regional warehouse:

$$\sum_j \left\{ \sum_{i:n(i)=n,\tau(i)=t+L_i,tildex_i=0} u_i^j + y_{mnt}^j + \sum_{i:n(i)=n,\tau(i)>t+L_i} x_{it} a_i^j \right\} \le C_{mn}. \quad (5.69)$$

Indicate whether an order is filled completely or not:

$$\sum_{k=1}^{\tau(i)-L_i-1} x_{ik} + \tilde{z}_i + z_i = 1. \quad (5.70)$$

Recall that $\tilde{z}(i) = 1$ when order i is filled in period $\tau(i) - L_i$.

$$\tilde{z}_i \le 1 - \frac{\sum_j (a_i^j - u_i^j - w_j^i)}{\sum_j a_i^j}. \quad (5.71)$$

A regional warehouse can only use the on-hand inventory to satisfy an order.

$$\sum_{i:n(i)=n,\tau(i)=t+L,\tilde{x}_i=0} w_i^j \le I_{nt}^{jR}. \tag{5.72}$$

The recursive definition of the ending inventory levels at the co-located warehouse when $n = m(j)$ is:

$$I_{n,t}^{jR} = I_{n,t-1}^{jR} + y_{mnt}^j - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a_i^j - u_i^j). \tag{5.73}$$

The recursive definition of the ending inventory levels at regional warehouse when $n \neq m(j)$ is:

$$I_{n,t}^{jR} = I_{n,t-1}^{jR} + y_{m(j)n,t-L}^j - d_{nt}^{j\alpha} - d_{nt}^{j\beta} = I_{n,t-1}^{jR} + y_{m(j),n,t-L}^j - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a_i^j - u_i^j) - d_{nt}^{j\alpha}. \tag{5.74}$$

The recursive definition of the ending inventory levels at primary warehouse $m = m(j)$ is:

$$I_{m,t}^{jP} = I_{m,t-1}^{jP} + q_{m,t-L_j'}^j - \sum_{i:\tau(i)=t+L_i} u_i^j - \sum_{n=1}^{N} y_{mnt}^j - \sum_{i:\tau(i)>t+L_i} x_{it} a_i^j. \tag{5.75}$$

Dynamic Program Formulation In this section we summarize the above into a dynamic program recursion. Let $\mathbf{I}_t$ denote the vector of ending inventory levels for all item j at all warehouses.

$$V_t(\mathbf{I}_t) = \mathbb{E}\min\Big[\sum_j \sum_j G^{jP}_{m(j)t}(I^{jP}_{m,t-1}, q_{m,t-L'_j}, a^j_i, u^j_i, y^j_{mnt}) \tag{5.76}$$

$$+ \sum_{n=1}^{N}\sum_j G^{jR}_{nt}(d^{j\alpha}_{nt}, I^{jR}_{n,t-1}, y^j_{m(j)n,t-L_n}, u^j_i, a^j_i) + V_{t+1}(\mathbf{I}_{t+1})\Big]$$

$$\text{s.t.} \sum_{n=1}^{N} y^j_{mnt} + \sum_{i:\tau(i)=t+L_i,\tilde{x}_{it}=0} u^j_i + \sum_{i:\tau(i)>t+L_i} x^i_t a^i_j \le I^{jp}_{m,t-1} + q^j_{m,t-L'_m}, \tag{5.77}$$

$$\sum_{j:m(j)=m}(y^j_{mnt} + \sum_{i:n(i)=n,\tau(i)=t+L_i,\tilde{x}_i=0} u^j_i + \sum_{i:i=n,\tilde{x}_i=0,\tau(i)>t+L} x_{it}a^j_i) \le C_{mn}, \tag{5.78}$$

$$\sum_{k=1}^{\tau(i)-L_i-1} x_{ik} + \tilde{z}_i + z_i = 1, \tag{5.79}$$

$$\tilde{z}_i \le 1 - \frac{\sum_j (a^j_{i,} - u^j_i - w^j_i)}{\sum_j a^j_{i,}} = 0, \tag{5.80}$$

$$\sum_{i:n(i)=n,\tau(i)=t+L,\tilde{x}_i=0} w^j_i \le I^{jR}_{nt}, \tag{5.81}$$

$$I^{jP}_{m,t} = I^{jP}_{m,t-1} + q^j_{m,t-L'_j} - \sum_{i:\tau(i)=t+L_i} u^j_i - \sum_{n=1}^{N} y^j_{mnt} - \sum_{i:\tau(i)>t+L_i} x_{it}a^j_i. \tag{5.82}$$

$$I^{jR}_{n,t} = I^{jR}_{n,t-1} + y^j_{m(j),n,t-L} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a^j_i - u^j_i) - d^{j\alpha}_{nt}, \forall n \ne m(j), \tag{5.83}$$

$$I^{jR}_{n,t} = I^{jR}_{n,t-1} + y^j_{mn,t} - \sum_{i:n(i)=n,\tau(i)=t,\tilde{x}_i=0} (a^j_i - u^j_i), \forall n = m(j), \qquad y^j_{mnt} \text{are integers.} \tag{5.84}$$

Note that this dynamic program is very similar to one used in the the single primary warehouse case, except that the shipping lead time L_i is evaluated based on orders rather than on the regional warehouse. Hence, the approximation models we introduced in Section 5.2.2 also apply here.

5.2.4 Conclusion

In this section, we formulated a dynamic program to describe the allocation operation. It is impossible to solve this dynamic program using the traditional recursive method. Consequently, we present three sub-models that can be used to make daily allocation decisions for all items. When making the daily allocation decisions, we have also made tentative order fulfillment decisions based on the current demand and inventory information. This tentative order fulfillment decision could be changed when the allocated inventory arrives at the regional warehouses. In the next section, we will introduce a model that focuses on fulfilling customer orders with available inventory at each regional warehouse.

5.3 Order Fulfillment

A final decision must be made pertaining to the fulfillment of a customer's order by its due date. On this day, each regional warehouse uses its own inventory to satisfy customer orders. In the allocation model, we have created a tentative order fulfillment plan. More short response lead time demand information is unveiled during the transportation lead time. Given the newly obtained information on demand and inventory at the time the allocation arrives at a regional warehouse, the order fulfillment plan may need to be adjusted.

In this section, we introduce a model designed to make daily order fulfillment decisions at a regional warehouse. The orders that are waiting to be filled include backlogged orders, both short and long response lead time orders that must be sent out today, and long response lead time orders that are not yet due.

Recall that these orders may contain multiple units of multiple items. Hence, we need to develop an algorithm to determine how to assign the available inventory to each order.

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 of them will be satisfied in advance. When making these order fulfillment decisions, we also want to minimize the number of packages used to satisfy each order. This is done to reduce the expensive "last mile shipping cost". We will address these concerns in our model.

The model and heuristics we develop are intended for use in large scale systems when there may be a few hundred thousand orders due every day.

5.3.1 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.

At this point, the available inventory to fulfill customer orders is a result of allocation decisions that are made in previous periods. Each warehouse can only use its own on-hand inventory to fulfill customer orders. Hence, we focus on one regional warehouse, say regional warehouse n, at a time. Recall that this regional warehouse serves as primary warehouse for some items, which

are denoted as set R_n. Assume we are in period 1. Let us use I_{j1} to denote the on-hand inventory of item j in period 1 and use q_{jt} to denote the allocation of item j at the regional warehouse that is scheduled to arrive in period t, where $t = \{1, 2, 3, \cdots, l\}$, where l is the grace period. The long response lead time orders contain items that are not managed by this primary warehouse. The available inventory includes the on-hand inventory stocked to meet short response lead time demand and the daily allocations that arrived in the current and future time periods.

We first construct a list of outstanding orders that need to be satisfied by each period in the planning horizon. The order information is then updated when a customer order is received, or 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. Recall that a_i^j denotes the number of units of item j requested in order i.

Not all outstanding orders may be fulfilled due to the limited inventories. We evaluate the order fulfillment decisions sequentially. Let O_0 denote the orders i that are backordered, and O_1 denote the orders that are due in the current period, that is $\tau(i) = 1$. The priority is first given to the backlogged orders. In formulating the allocation model, we assumed that all backorders would be sent directly to the customer from the primary warehouse via express delivery. Hence, the backlogged orders only include items which belong to set R_n. Then we will use the remaining available inventory to satisfy orders that must be sent out from regional warehouse n 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. The last step is to determine which of

the orders that are not yet due can be fulfilled in advance with the remaining inventory after setting aside safety stocks needed to ensure satisfaction of short response lead time demand over the regional warehouse n's lead time. We examine orders based on the due date subsequently. Note that we do not fill an order in advance if we cannot fulfill it fully.

Let x_{it} denote whether order i is fulfilled completely in the current time period t.

$$x_{it} = \begin{cases} 1 & \text{if order } j \text{ is fulfilled completely in period } t, \\ 0 & \text{otherwise.} \end{cases} \tag{5.85}$$

If an order that is due today is not fulfilled completely, we use y_i^j to denote the amount of item j that is shipped to satisfy customer order i.

5.3.2 Two Step Order Fulfillment Models

We now introduce the order fulfillment models. We have discussed that we will analyze the order types separately, since they have different priorities. For orders that are due on the same day, we develop two models to make order fulfillment decisions. The first one is developed to fill orders completely. The second one is developed to determine how to allocate inventory to partially satisfy an order.

Model 1: Complete Order Fulfillment

Given the set of orders O, let the set of the supply of inventories that is available to fulfill orders in set O be $S = \{S_j\}$. Note that S_j, the amount of item j that is available for allocation, may be smaller than the actual on-hand inventory since

the regional warehouse needs to maintain some level of safety stock to meet future short response lead time orders that will be placed in the next periods. We first allocate the inventories to satisfy as many orders in O as possible. We formulate the model as a integer program

$$\max \sum_{i \in O} x_i \tag{5.86}$$

$$\text{s.t.} \sum_{i} a_{ij} x_i \leq S_j, \tag{5.87}$$

$$\text{s.t.} x_i \in \{0, 1\}. \tag{5.88}$$

This integer program may be slow to solve or may need a large memory to run when there are a few hundred thousand orders to process. Therefore we present an algorithm that reduces the number of variables in the model significantly.

First, find the set J where for any *item* $j \in J$ there is sufficient supply to satisfy all demand in set O, that is, $\sum_i a_i^j \leq I_j$. Based on set J, define set $O' \subset O$ to be the set orders where the orders in O' contain items only from set J, that is $a_i^j > 0$ only when $j \in J$. Every order $i \in O'$ can be fulfilled completely. Hence, let $x_i = 1$ for $i \in O'$. The available inventory to fulfill the remaining orders is $S'_j = S_j - \sum_{i \in O'} a_{ij}$. Let O_F be the set of orders that are fulfilled completely. Then the set of remaining orders to be filled is $O'' = O \setminus O_F$.

Second, we will focus on the single item and single unit orders. This is the optimal strategy, since our target in this model is to satisfy as many orders as possible. Therefore, in this step, determine the order $i \in O''$, $a_i^j > 0$ for only one item j and $a_i^j = 1$ for that item. We can choose arbitrarily which orders to satisfy until either all orders are satisfied or the item runs out of inventory. Let $x_i = 1$ if order i is satisfied and add order i to set O_F. The remaining supply is

$S''_j = S'_j - \sum_{i \in O''} a_{ij} x_i$.

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

$$\max \sum_{i \in O \setminus O_F} x_i \tag{5.89}$$

$$\text{s.t.} \sum_i a_{ij} x_i \leq S''_j, \tag{5.90}$$

$$\text{s.t.} x_i \in \{0, 1\}. \tag{5.91}$$

After solving this problem, let $S'''_j = S''_j - \sum_{i \in O \setminus O_F} a_{ij} x_i$. We add order i to set O_F if $x_i = 1$.

The state space of this problem is greatly reduced by the above two steps for two reasons. First of all, most items have enough inventory to satisfy demand throughout a cycle except possibly at the end of a cycle. Therefore, most orders are satisfied in step 1. Second, recall from the Chapter 2, we stated that over 30% of total orders only request a single item and among these orders, 97% are single unit orders. Therefore, for practical problems, the method we have introduced for fulfilling orders is computationally tractable.

After executing the complete order fulfillment model, we have some remaining inventory $\bar{S}_j = S'''$, and a list of orders $O_P = O \setminus O_F$. There is not enough inventory to fill any order completely with the remaining inventories. We will use the remaining supply to fill these orders partially.

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 cost low. Recall that the "last-mile shipping cost" is charged based on the volume or weight of the package. This shipping cost is a concave function of volume or weights. One method to keep a low shipping cost is to maximize the sum of volumes satisfied among all orders based on percentage. The total amount of inventories used to fill the remaining orders cannot exceed the supply. Our model is

$$\max \sum \left\{ \frac{\sum_j v_j y_{ij}}{\sum_j v_j a_i^j} \right\} = \sum_i \sum_j \frac{v_j}{C_i} y_{ij} \tag{5.92}$$

$$\text{s.t.} \sum_{i \in O_P} y_{ij} \leq \bar{S}_j, \tag{5.93}$$

$$0 \leq y_{ij} \leq a_i^j, \tag{5.94}$$

$$y_{ij} \text{ is an integer}, \tag{5.95}$$

where v_j is the volume of item j and $C_i = \sum_j v_j a_i^j$ is the total volume of order i.

In the above integer program, $\sum_j v_j y_{ij}$ is the volume or weight associated with partially shipping order i. Due to the concavity of the cost function, the greater the fraction of the volume or weight shipped produces a lower per unit transportation cost and hence is desirable.

Note, we can relax the constraint that y_{ij} is an integer due to the structure of the problem. The problem becomes linear program, for which there is an optimal integer solution. This solution can be obtained using a greedy algorithm

and executed one item at a time. Hence, the execution of optimization process can be carried out in a parallel manner.

5.3.3 Order Fulfillment Execution

In the previous two sections, we introduced two order fulfillment models. In this section, we will state an algorithm that uses the two models to find the order fulfillment decisions.

As we have mentioned, orders that are due at different dates are ranked with different priorities. Therefore, we will 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 analyze the orders that are due today which are in set O_1. The remaining unfilled items in the orders are added to the backlogged list to the primary warehouses that manage the items. Next, we execute model 1 for orders in sets $O_2, \cdots, O_l$ subsequently. Model 2 is not executed for these orders, since we do not fill orders partially before they are due so as to avoid an unnecessary incremental "last-mile delivery shipping costs".

5.4 Final Remarks

In this chapter, we focus on developing day-to-day inventory execution strategies to operate the fulfillment system. The three execution decisions- procurement, allocation and order fulfillment decisions, are decoupled since they are made by different people at different locations. In this chapter, we presented three separate models that are designed for making these decisions. Along with

the models, we also presented scalable algorithms that are capable of obtaining close to optimal solutions efficiently.

CHAPTER 6

CONCLUSIONS

In this thesis, we analyzed a multi-echelon capacitated fulfillment system that offers short response and long response lead time delivery services of tens of millions of products to customers. We constructed a planning model and developed a scalable approximation algorithm that is computational tractable. This model is designed for planning warehouse, inventory and staffing requirements. Our major contribution is that the algorithm we developed extends the current literature by setting target inventory level using short fall methods with advance demand information. Furthermore, the approximation algorithm is very efficient. Our numerical experiments show the algorithm performs well under the various demand process.

Next, using the inventory target levels obtained in the planning model, we constructed an order fulfillment process. Three models are developed to make inventory procurement, allocation and order fulfillment decisions. To our knowledge, very few people have ever constructed models for executing an order fulfillment process. Most research focused only on the planning inventory requirements. For a fulfillment system that is the size of the one we are studying, making execution decisions is not trivial. A good strategy, such as the one we developed, must carefully decide where and when to send the inventory.

In this thesis, we have built models that can be basis for further studies. For example, we plan to test the performance of the system when the ratio between short and long response lead time demands increases.

In our execution model, we made the assumption that the shipping lead time from the primary warehouse to a non-co-located regional warehouse is the same among locations. While this assumption is reasonable, we would also like to relax this assumption. We will also examine how effective our fulfillment approach is when demand has higher variance to mean ratios.

BIBLIOGRAPHY

[1] Amazon.com. 2011 annual report. April 2012. http://phx.corporate-ir.net/phoenix.zhtml?c=97664p=irol-reportsannual.

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

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

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

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

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

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

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

[9] D. Murray J. Muckstadt 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.

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

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

[12] W.Davie K. Allen and D. Weidenhamer. Quarterly retail e-commerce sales 1st quarter 2012. May 2012.

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

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

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

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

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

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