Inventory Planning for Low Demand Items in Online Retailing

by

Pallav Chhaochhria

S.M. Manufacturing Systems & Technologies
Nanyang Technological University, 2002

B.E. Electrical & Computer Engineering
National University of Singapore, 2001

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of
Master of Science in Operations Research
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

Author ....
Sloan School of Management
May 17, 2007

Certified by..
Stephen C. Graves
Abraham J. Siegel Professor of Management Science & Engineering Systems
Thesis Supervisor

Accepted by ....
Cynthia Barnhart
Professor, Civil and Environmental Engineering
Co-director, Operations Research Center

Inventory Planning for Low Demand Items in Online Retailing

by

Pallav Chhaochhria

Submitted to the Sloan School of Management on May 17, 2007, in partial fulfillment of the requirements for the degree of Master of Science in Operations Research

Abstract

A large online retailer strategically stocks inventory for SKUs with low demand. The motivations are to provide a wide range of selections and faster customer fulfillment service. We assume the online retailer has the technological capability to manage and control the inventory globally: all warehouses act as one to serve the global demand simultaneously. The online retailer will utilize its entire inventory, regardless of location, to serve demand.

We study inventory allocation and order fulfillment policies among warehouses for low-demand SKUs at an online retailer. Thus, given the global demand and an order fulfillment policy, there are tradeoffs involving inventory holding costs, transportation costs, and backordering costs in determining the optimal system inventory level and allocation of inventory to warehouses. For the case of Poisson demand and constant replenishment lead time, we develop methods to approximate the key system performance metrics like transshipment, backorders and average system inventory for one-for-one replenishment policies when warehouses hold exactly one unit of inventory. We run computational experiments to test the accuracy of the approximation. We develop extensions for cases when more than one unit of inventory is held at a warehouse. We then use these results to develop guidelines for inventory stocking and order fulfillment policies for online retailers.

We also compare warehouse allocation policies for conditions when an order arrives but the preferred warehouse does not have stock although there is stock at more than one other location in the system. We develop intuition about the performance of these policies and run simulations to verify our hypotheses about these policies.

Thesis Supervisor: Stephen C. Graves
Title: Abraham J. Siegel Professor of Management Science & Engineering Systems

To the memory of my father ('44 – '94)

Acknowledgments

This dissertation could not have been successfully completed without the help and support of a number of people. I take this opportunity to thank my supervisor Professor Graves for providing me the opportunity to work with him on such an innovative project. I find it hard to imagine anyone being a more sincere and better research supervisor than what he has been. I also wish to express my gratitude to Dr Russell Allgor and Dr Ping Xu for their interest and effort. Their guidance has helped this research to be of practical value. Finally, I thank the Singapore-MIT Alliance (SMA) for their generous financial support of this research.

I take this opportunity to thank my parents and loved ones for teaching me values that are precious, irrespective of time and place. Finally, all praise and thanks is due to God for all that He has gifted us, although, He can never be praised or thanked enough.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

1.1 INTRODUCTION AND MOTIVATION

[1]A large online retailer strategically stocks inventory for items with low demand for several reasons. One reason for an e-tailer to keep low demand items in its catalog is to provide customers with a wide range of product choices. A second motivation to hold inventory is to fill customer orders faster as orders are filled from stock rather than through a drop-shipper. The third incentive is to gain a competitive advantage from other online retailers. Suppose that an e-tailer only drop-ships the low-demand SKUs, then its drop-shipper who serves many online retailers, may choose to satisfy a competitor's demand. These reasons become significantly more important in the online retail context as the customer has very low switchover cost from one e-tailer to another.

[1] This work builds on the research done by Ping Xu for her doctoral thesis [Xu05]. Some of the material in this section follows closely that in her thesis.

Efficient inventory planning and order fulfillment for low-demand items is important in the retailing setting. Often over 90% of a retailer's catalog comprises slow moving items with demand in the range of 0.2 – 0.8 units per week. Therefore, the inventory planning for low-demand items is very critical for the ultimate success of an online retailer.

For many of these low demand items, the e-tailer may only stock a few inventory units across all warehouses and use centralized order fulfillment to provide faster response time to the customer. Thus, we assume that if a warehouse is out of stock, its demand can be satisfied by on-hand units in other warehouses. We also assume that when all warehouses are out of stock, a customer demand is met by the warehouse that would first receive an on-hand unit regardless of its location. Such order fulfillment policies are consistent with the practice of online retailers, due to their emphasis on fast customer response times.

Inventory planning for low-demand items is challenging primarily because of these reasons – the discrete effect in deciding whether to stock 2 or 4 units in the system makes a significant difference in costs given the large number of such low demand items; current inventory models often assume all variables are continuous. We illustrate this with an example.

Suppose that we have two demand regions in the system, and one has 30% of the total demand and another has 70%. The total demand during the replenishment lead-time is a Poisson random variable with rate d. We want to stock enough inventory in the system so that the fill rate (probability of serving a customer immediately by on-hand inventory) is at least 90%. We can plan inventory according to two ways: global planning (plan for the entire system) or regional planning (plan for the two regions separately).

For global planning, in order for the probability that an order is filled immediately to be at least 90% , we set the system inventory of as follows:

$$\arg\min_{k} \{\Pr[\text{System Demand during lead time} \geq k] \leq 0.1\}$$

Similarly, for regional planning, we find the on-hand inventory required for the probability that an order is filled immediately to be at least 90% by:

$$\sum_{\text{Regions } i=1}^{2} \arg\min_{ki} \{\Pr[\text{Regional Demand during lead time} \geq k_i] \leq 0.1\}$$

Two examples for different values of d are given in the table below:

System Demand during Lead Time	Global Planning	Regional Planning
0.5	2	4
10	15	17

TABLE 1: EXAMPLE OF THE DISCRETE EFFECT

In these examples, when system demand is low, the regional planning case holds twice the inventory that the global planning case, although there is some compensation for the regional planning due to lower delivery costs.

We study these trade-offs between inventory holding, penalty for backorders, and transportation costs for an online retailer. We assume that the e-tailer has several warehouses in the system. We also assume that it has the technological capability to manage and control the inventory globally: all warehouses act as one to serve the global demand. Specifically, the e-tailer will utilize its entire inventory, regardless of location, to serve demand. We develop methods to calculate key performance metrics and determine the optimal inventory policy for such a system given the global demand rate and distribution.

1.2 LITERATURE REVIEW

This work builds on the research done by Ping Xu as part of her doctoral dissertation. Her thesis, [Xu05], provides a detailed literature review of this problem.

Although we did not come across any inventory planning research in the online retail environment, a related body of research studies lateral transshipments, often in a context of spare parts inventory distribution systems. The goal is to develop operational rules for joint order and transshipment policies. Many papers have considered a single-item, multi-location, periodic review inventory system with lateral transshipments. Notable papers in this category are: Gross [Gro63], Krishnan and Rao [KR65], Das

[Das75], Karmarkar and Patel [KP77], Robinson [Rob90], Tagaras and Cohen [TC92], and Archibald [AST97]. However a common problem in most of these is that lead times for replenishment are assumed to be zero, and hence have limited applicability for non-zero lead times.

Other papers consider continuous review inventory systems with lateral transshipments. Lee [Lee87] considered a two-echelon model in the context of spare parts inventory. The first echelon had a repair center while the second echelon comprised of service centers that received customer requests for parts. Lateral transshipments were allowed between the service centers when one of them ran out of stock and had a customer request to fill. Lee developed approximations for the number of backorders and lateral transshipments for the case when all service centers faced identical demand processes. Axsater [Axs90] extended Lee's model to include non-identical service centers. He used a different modelling approach and developed a more accurate approximation to predict the system performance parameters. Alfredsson and Verrijdt [AV99] built on Axsater's work and allow for more order fulfillment options like direct delivery from the upper echelon repair center and the manufacturing plant. Dada [Dad92] worked on a two-echelon model with prioritized lateral transshipments among the lower echelon service centers, emergency shipments from the upper echelon repair center and if no inventory was available in the system, then the demand was lost. His model assumed the replenishment lead times from the repair center to the service centers to be exponentially distributed. Hence, he developed an exact markov process model and a fast approximation to estimate system performance parameters. Grahovac and Chakravarty [GC01] allowed for lateral transshipments not only when a warehouse is out of stock, but also pro-actively based on on-hand and in-transit inventory information.

Most papers in the spare parts inventory literature assume one-for-one or (S-1, S) replenishment policies; we also use this policy. Lateral transshipment lead-times in the above literature are assumed to be instantaneous but with additional cost. We assume that if a warehouse is out of stock, on-hand units in other warehouses can fill its demand, which is equivalent to an instantaneous transshipment from other warehouses with an additional transportation cost. This instantaneous transshipment assumption is realistic for retailers with good IT infrastructure. However, unlike these papers, we also assume that even if all warehouses are out of stock, a lateral transshipment is allowed if another warehouse would have an on-hand unit earlier. Such order fulfillment policies are consistent with online retailers due to the emphasis on fast customer response times. This is the main difference between our model and those in the literature.

Recent work by Xu [Xu05], focused on the effect of inventory allocation on outbound transportation costs for an online retailer. Her model is built on the same assumptions as our model. Given that the e-tailer stocks a certain number of units of inventory in the system, Xu studied how best to allocate inventory to warehouses by considering outbound transportation costs from the warehouses to customers. She develops a 2-state Markov Chain based model that encompassed all the possible states of a 2-unit 2-warehouse system, with the transition probabilities being functions of the demand rate and demand distribution among regions. Then, the transshipment proportions from one warehouse to another could be derived as functions of the steady state probabilities of the Markov Chain and the transition probabilities. Her approach produced exact solutions for the 2-unit 2-warehouse case, but was not tractable for the general N-unit N-warehouse case.

In chapter 2, we describe the model, its assumptions and the 2-unit 2-location case solved by Xu [Xu05]. We then develop a method to calculate transshipment in an N-unit N-location (one unit in each location) system for two special cases of the demand distribution over the regions. In chapter 3 we use these exact results to develop an interpolation-based approximation method for transshipment for the case of a general demand distribution over the regions. Then we describe another method to approximate the transshipment in a k-unit N-location system (k>N), where each warehouse holds at least one unit of inventory. This is described in chapter 3. In chapter 4, we build models to compare different inventory policies that can help determine the optimal policy for an e-tailer. We also compare different order fulfillment policies and provide recommendations.

Chapter 2

In this chapter, we describe the online retailer's model that we work with in this dissertation. We first explain the assumptions in the model, properties arising out of these assumptions, and then introduce some notation that will be used throughout the rest of the dissertation. We then examine the simplest non-trivial model for the online retailer, the 2-unit 2-location model that was developed by Xu [Xu05].

2.1 MODEL

We want to find methods to estimate key performance metrics like transportation costs, backorders, average system inventory, and determine the optimal inventory policy for an online retailer. We start with the following assumptions for the model.

A-1 The system demand process is Poisson with rate λ.

A-2 The demand process is split into N independent processes, 1 to N. With probability α_1, a demand arrival is from region 1; with probability α_2, a demand arrival is from region 2; and so on. The α_i are non-negative and sum to 1.

A-3 There are N warehouses, one for each region. The replenishment lead-time for each warehouse is the same constant L.

A-4 The inventory policy is one-for-one replenishment at each warehouse: at each demand epoch, we assign the demand to a warehouse, as specified in A6, A7 and A8; this assignment triggers a replenishment for the selected warehouse.

A-5 Demand is backlogged when there is no on-hand inventory in the system.

In the context of online retailing, the e-tailer can utilize any warehouse or fulfillment center to serve the customer demand. Specifically, a demand is always served by an on-hand inventory unit in the system if there is any; if there are no on-hand inventory units in the system, the demand is served by and triggers replenishment at the warehouse that has the next arriving unassigned replenishment. We then have the following assumptions on how the system operates for all stocking scenarios.

A-6 When a customer arrives and its closest warehouse has on-hand inventory, then its closest warehouse serves the demand and triggers a replenishment. (The closest warehouse is the warehouse in the same region as that for the customer.)

A-7 When a customer arrives and its closest warehouse does not have inventory on-hand, the system will assign the demand to another warehouse if there is on-hand inventory elsewhere in the system. A warehouse with on-hand unit is chosen according to an order fulfillment policy, *P,* to serve demand; this assignment triggers a replenishment for the chosen warehouse.

A-8 If a customer arrives and the system has no on-hand units, then the policy is to assign the demand to the warehouse with the next arriving unassigned replenishment. This assignment triggers a replenishment for the chosen warehouse.

Note that assumption A-8 is possible because we assume deterministic supply lead-times, so we know exactly when all future replenishments arrive. Also, assumption A-7 and A-8 are analogous to an emergency transshipment.

By our model assumptions, every demand is matched with the next available unit and a replenishment is triggered at each demand epoch. We see that the system inventory level, fill-rate and the customer waiting times are the same as in an aggregate model where all inventory in the system is stored in one warehouse. Thus, the system-level inventory holding costs, ordering costs, and backorder costs are independent of how the inventory is allocated among the warehouses. On the other hand, outbound transportation costs depend on the location from which demands are served. Therefore, we will need to examine how inventory allocations to the warehouses influence the outbound transportation costs. We now introduce some notation to be used throughout the dissertation.

Notation

N : Number of regions in the system; each region has a warehouse.

λ : System demand rate from a Poisson process

L : Replenishment lead-time for each warehouse; L is a constant, and same for all warehouses in the system

D_L : Random variable for the system demand over the lead time. $E[D_L] = \lambda L$

SI : Random variable for the on-hand system inventory, $SI = (N - D_L)^+$

We define the following probabilistic events:

$F_{i,j}$: Event that an order from region i is filled immediately from on-hand stock at warehouse j

$B_{i,j}$: Event that an order from region i is backordered and filled subsequently from a replenishment to warehouse j

$A_{i,j}$: Event that an order from region i is filled from warehouse j. Hence, $A_{i,j} = F_{i,j} \cup B_{i,j}$

The system performance metrics such as the fill rate and average inventory can be calculated for general cases under the assumption that lead-time demand D_L is Poisson. These metrics directly affect the backorder costs and the inventory holding costs.

$$\text{System fill rate} = 1 - \Pr[D_L \geq N]$$

$$\text{System inventory } SI = (N - D_L)^+$$

$$\text{Average system inventory} = \sum_{i=1}^{N} i \times \Pr[SI = i]$$

However, the outbound transportation cost is affected significantly by the number of transshipments in the system. (We will use transshipment to denote when a customer demand in a region is served by a warehouse from outside the region) The proportion of transshipments in the system in turn depends on the demand distribution across the regions and the specific order fulfillment policy in the system. In this dissertation, we describe methods to estimate the probability of an order being transshipped when its preferred warehouse does not have stock on-hand. The next section provides a model to exactly calculate the transshipments between regions in a 2-location scenario where each warehouse holds one unit of stock.

2.2 2-UNIT 2-LOCATION MODEL

Xu [Xu05] developed a 2-state Markov Chain to model all the possible states of a 2-unit 2-warehouse system, with the transition probabilities being functions of the demand rate and demand distribution among regions. Most of the content of this section have been taken from her dissertation. This section explains this initial approach, and why it is not tractable for cases of more than 2 units and 2 locations.

Suppose the e-tailer decides to stock two units of inventory in two warehouses in the system, A and B, where warehouse A is the preferred server for region 1 and warehouse B is the preferred server for region 2. We intend to find the proportion of shipments from warehouse A to region 2 and from warehouse B to region 1. As a result of our assumptions, we see that the system inventory position at each warehouse is always one, where inventory position is on-hand and on-order inventory minus backorders. Hence, given our one-for-one replenishment policy, there is always exactly one unit of inventory associated with each warehouse that has not yet been assigned to any demand. This unassigned unit can be either on-hand or on-order. This gives us three scenarios that can occur for the relative positions of the two unassigned units in A and B:

(i) both units are on-hand

(ii) unit for A is on hand or will arrive before the unit for B

(iii) unit for B is on hand or will arrive before the unit for A

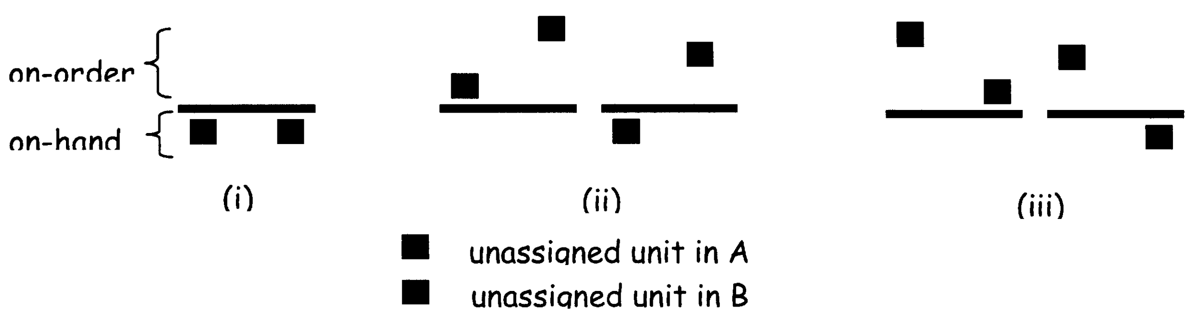

FIG 1: POSITIONS OF UNASSIGNED UNITS IN 2-UNIT 2-LOCATION SYSTEM

We can then visualize the order fulfillment process as a "race" between the two unassigned replenishment units.

Now consider a system with demand rate λ, that is split into independent processes with probabilities r_1 and r_2, for the two regions being served by warehouses A and B respectively. We define a Markov Chain with 2 states: State A indicates that the most recent order was assigned to warehouse A, while state B indicates that the most recent order was assigned to warehouse B. Let the probability of zero orders during the lead time be indicated by q (which is equal to $e^{-\lambda L}$). The Markov Chain is as illustrated in the figure below.

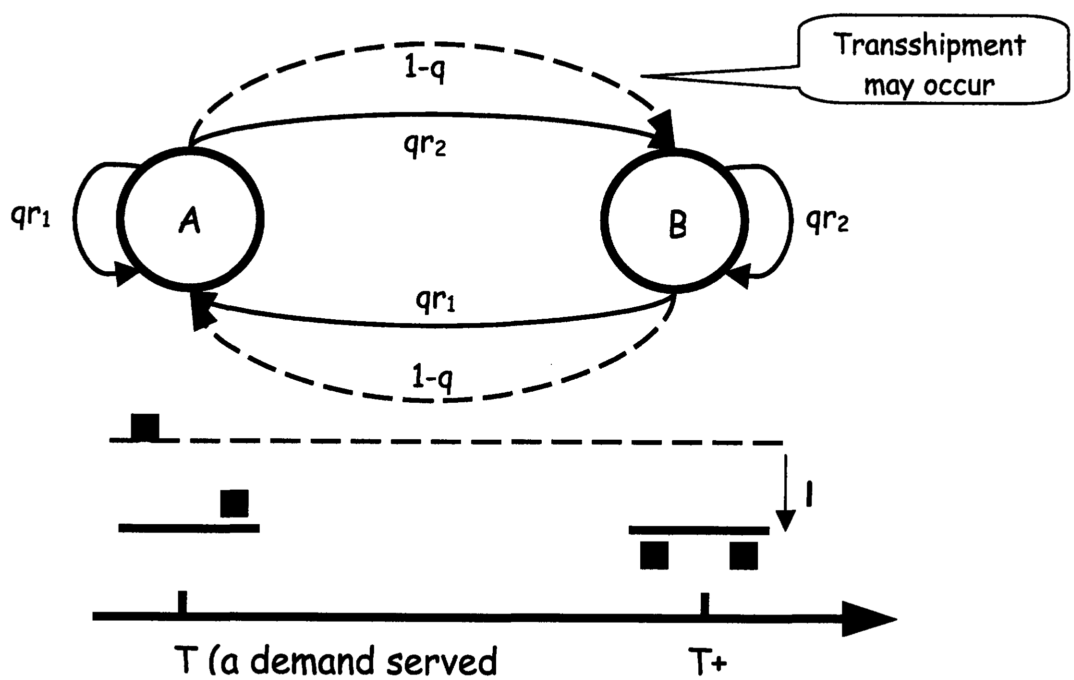

FIG 2: MARKOV CHAIN FOR 2-UNIT 2-LOCATION SYSTEM

We describe the transition out of state A only, since the same logic applies for those out of state B. Suppose the k^{th} demand epoch occurs at time t_k and is assigned to warehouse A. Then, we start at state A at t_k. The k^{th} demand also triggers a replenishment at t_k for A. This replenishment unit would not arrive to A until $t_k + L$. The dotted-line transitions represent the next demand arriving before $t_k + L$, $t_{k+1} < t_k + L$. The solid-line transitions represents the next demand arriving after $t_k + L$, $t_{k+1} > t_k + L$. The state of the system at t_k is of case (iii) in Figure 1 with the unit in A being L time units away; the unassigned unit for B must be either on-hand or on-order within L time units of delivery. If $t_{k+1} < t_k + L$ (with probability $1-q$), by our policy, the $(k+1)^{st}$ demand would be assigned to B and the system transitions to state B. If $t_{k+1} > t_k + L$ (with probability q), then at time $t_k + L$ we know that both units are on-hand, i.e, case (i) in Figure 1. Then, with probability r_1, the system transitions to state A and with probability r_2, the system transitions to state B since we assign the demand at time $t_{k+1} > t_k + L$ to its closest warehouse.

This is an aperiodic, single recurrent class markov chain. Hence, there exist steady state probabilities for being in states A and B. Let these steady state probabilities of the markov chain be indicated by p_A and p_B. Then, the probability of a transshipment from warehouse i to region j is given by P_{ij}:

$$P_{A2} = p_B(1-q)r_2 \qquad\qquad P_{B1} = p_A(1-q)r_1$$

The above formula can be easily verified. Note that P_{A2}, the probability of a transshipment from warehouse A to region 2, occurs if and only if the following events take place– the most recent order at t_k was assigned to warehouse B (which has probability, p_B), and the next order came in before t_k +L from region 2 (which has probability, $(1-q)r_2$). Since, warehouse A's unassigned unit is ahead of warehouse B's unassigned unit when the demand from region 2 occurs, warehouse A ships its unit to region 2.

This approach produces exact solutions for transshipment proportions for the 2-unit 2-location case. A similar approach in [Xu05] also allows for analysis when the replenishment lead times for the 2 warehouses are different.

Note that in the 2-unit 2-location case, whenever a demand is assigned to a warehouse, the other warehouse will then have the next arriving unassigned replenishment unit. However, when we try to extend this model to a larger system of say 3-units and 3-locations, we notice that the transition probabilities need to take into account the time of the last order filled at each of the warehouses in order to track the earliest arriving unassigned replenishment unit. Thus the state now changes with time, hence requiring the system to have as many states as there are points in the real line. This makes the markov chain approach intractable for the general N-unit N-warehouse case.

This problem motivates us to explore other methods to calculate or estimate the proportions of orders being shipped from warehouses that are outside the region of the customer demand. In chapter 3, we develop some methods to calculate or estimate this quantity.

Chapter 3

In this chapter, we first develop methods to estimate the probability of transshipment of an order for the general case of N-units and N-locations where each warehouse holds exactly one unit of inventory. We then extend this methodology to estimate the probability of transshipment of an order for the case of k-units and N-locations (k>N) where each warehouse may hold one or more units of inventory. This will provide the tools necessary to analyze a range of inventory configurations for the online retailer, enabling us to determine the optimal inventory policy for it.

3.1 N-UNIT N-LOCATION MODEL

We solve the N-unit N-location transshipment estimation problem by decomposing it into a few different cases that can be solved individually using specific probabilistic approaches. In sections 3.1.1, we calculate the probability of transshipment of an order for the case of balanced demand, when all regions have the same demand rate. Next, in section 3.1.2, we consider the case of an extreme demand distribution, when all warehouses hold one unit of inventory but all the demand is concentrated in one region only. This is not a realistic scenario but it allows us to exactly calculate the transshipment probability from each region to the region that has all the demand. Then, in section 3.1.3, we use a monotonicity argument to justify using interpolation between the probabilities of transshipment developed earlier as an estimate for the transshipment probability for other cases of demand distribution among the regions. We compare the estimates obtained by this interpolation method with simulation results for a wide variety of system fill-rates, configurations and demand distributions, and show that these estimates perform extremely well.

3.1.1 BALANCED DEMAND CASE

We first analyze the case in which each of the N-warehouses faces a demand rate of λ/N from its local region. We consider the order fulfillment policy as stated in A-6 and A-7, with the feature that if a customer arrives and its closest warehouse does not have inventory on-hand, but one or more of the other warehouses do have inventory on hand, then a warehouse with an on-hand unit is chosen randomly (with equal probability) to fill the order. We call this policy P_1. This case allows for a neat analytical solution making use of certain properties of the Poisson process.

Result 1

The probability that an order from region i is filled from warehouse j immediately from stock under the above condition is given by:

$$\Pr\left[F_{i,j}\right]=\sum_{k=1}^{N}\Pr\left[F_{i,j}\mid SI=k\right]\times\Pr\left[SI=k\right]$$

$$where \qquad \Pr\left[SI=k\right]=\Pr\left[D_L=N-k\right]=\frac{e^{-\lambda L}\left(\lambda L\right)^{N-k}}{N-k!}$$

$$\Pr\left[F_{i,j}\mid SI=k\right]=\frac{\begin{pmatrix}k-1\\N-1\end{pmatrix}}{\begin{pmatrix}k\\N\end{pmatrix}}=\frac{k}{N} \quad for\ i=j$$

$$\Pr\left[F_{i,j}\mid SI=k\right]=\left(1-\frac{k}{N}\right)\frac{1}{N-1} \quad for\ i\neq j$$

The probability that an order from region i is backordered and filled from a replenishment to warehouse j under the above condition is given by:

$$\Pr\left[B_{i,j}\right]=\Pr\left[B_{i,j}\mid SI=0\right]\times\Pr\left[SI=0\right]$$

$$where \qquad \Pr\left[SI=0\right]=\Pr\left[D_L\geq N\right]$$

$$and \qquad \Pr\left[B_{i,j}\mid SI=0\right]=\frac{1}{N}$$

Proof:

The results follow from application of the Total Probability Theorem, and properties of the Poisson process. The key insight used here is that for a given level of system inventory, each inventory state is equally likely. This result depends on the assumptions that the demand process is Poisson with equal rates for each region and that we use the allocation policy P_1. For example, if N=3 and SI=2, then the inventory states (1,1,0), (1,0,1) and (0,1,1) are equally likely to occur. Hence, conditioned on SI=2, we can argue that the probability that a demand from region 1 is served from warehouse 1 is

$\Pr\left[\boldsymbol{F}_{1,1} \mid \boldsymbol{SI}=2\right]=2/3$, as this will happen for inventory states (1,1,0) and (1,0,1). Similarly, conditioned on SI=2, a demand from region 1 is served by warehouse 2 or 3 only if the inventory state is (0,1,1), where each warehouse has an equal probability; thus, we have $\Pr\left[\boldsymbol{F}_{1,2} \mid \boldsymbol{SI}=2\right]=\Pr\left[\boldsymbol{F}_{1,3} \mid \boldsymbol{SI}=2\right]=1/6$.

.

A customer order is back-ordered if and only if none of the warehouses in the system has inventory, i.e, SI=0. Again, since demand is Poisson with equal rates for each region, when the system inventory is 0, each warehouse is equally likely to have the next arriving unassigned replenishment unit. Thus:

$$\Pr\left[B_{i,j} \mid SI=0\right]=\frac{1}{N}$$

QED

With the above results, we can find the probability a demand in region i is served by a transshipment from warehouse j: Pr[Fi,j]+Pr[Bi,j] for $i \neq j$.

3.1.2 EXTREME DEMAND CASE

We now suppose that all demand originates from one region, e.g., $\alpha_1=1$, while $\alpha_j=0$ for j=2,..N

We now analyze the case where one of the N-warehouses faces a demand rate of λ

from its local region, while the other warehouses do not face any demand but still carry a single unit of inventory. The order fulfillment policy for this analysis is P_1, as described in the previous section. Although this is not a realistic scenario, we can analyze it exactly using a renewal theory based approach.

Consider the demand arrival process with $\alpha_1=1$. We define a renewal as occurring whenever an order is filled by warehouse 1 either immediately from stock or as a backorder. We define the inter-renewal interval (M_t) as the number of demands that occur between renewal epochs. Then the counting process that looks at the number of orders served by warehouse 1, is a renewal process, and M_t are IID RVs for renewals occurring at t.

Result 2

The probability that an order from region 1 is filled from warehouse 1 immediately from stock under the above condition is given by:

$$\Pr\left[F_{1,1}\right]=\Pr\left[F_{1,1}\mid A_{1,1}\right]\times\Pr\left[A_{1,1}\right]$$

where

$$\Pr\left[F_{1,1}\mid A_{1,1}\right]=\Pr\left[D_L<N\right]$$

$$\Pr\left[A_{1,1}\right]=\frac{1}{1+E[M]}$$

$$E[M]=\sum_{k=0}^{N-1}k\times\Pr\left[D_L=k\right]+(N-1)\Pr\left[D_L\geq N\right]$$

The probability that an order from region 1 is backordered and filled from a replenishment to warehouse 1 under the above condition is given by:

$$\Pr\left[B_{1,1}\right] = \Pr\left[B_{1,1} \mid A_{1,1}\right] \times \Pr\left[A_{1,1}\right]$$
$$where$$
$$\Pr\left[B_{1,1} \mid A_{1,1}\right] = 1 - \Pr\left[F_{1,1} \mid A_{1,1}\right] = 1 - \Pr\left[D_L < N\right]$$
$$\Pr\left[A_{1,1}\right] = \frac{1}{1 + E\left[M\right]}$$
$$E\left[M\right] = \sum_{k=0}^{N-1} k \times \Pr\left[D_L = k\right] + (N-1)\Pr\left[D_L \geq N\right]$$

Proof:

The results follow from application of the Total Probability Theorem and Renewal-Reward Theory [Gal96].

Recall that $A_{1,1} = F_{1,1} \cup B_{1,1}$; then by applying the Total Probability Theorem, we get:

$$\Pr[F_{1,1}]=\Pr[F_{1,1} \mid A_{1,1}] \times \Pr[A_{1,1}]$$

Next, to see how $\Pr[F_{1,1} \mid A_{1,1}] = \Pr[D_L < N]$, consider the following argument:

Without loss of generality, suppose a demand occurs at time t and is assigned to warehouse 1. Then this event triggers replenishment for warehouse 1, which will arrive at time t + L. We consider how this replenishment will be used. There are two cases to consider:

1. Let D(t, t + L] denote the demand over the interval (t, t+L], and suppose D(t, t+ L] < N. Then at time t + L, the system on-hand inventory is non-negative and the

item that arrives at time t + L to warehouse 1 enters the on-hand inventory. By the assumed assignment rule, this item will be used to serve the first demand after time t + L.

2. Suppose $D(t, t+L] \geq N$. Then the on-order item to warehouse 1 will be assigned to the N^{th} order that occurs within the time interval (t, t + L]; when this item arrives at time t+L, it will be immediately used to serve the earlier demand.

Case 1 corresponds to using the item to serve a demand from stock, whereas case 2 corresponds to using the item to fill a backorder. Furthermore, this is a general characterization of how we allocate a replenishment to warehouse 1, as each replenishment to warehouse 1 is triggered by a prior demand assignment that occurred a lead time earlier. Thus, we see that $Pr[F_{1,1} \mid A_{1,1}] = Pr[D_L < N]$ holds.

Finally, we need to show the probability that a random demand is assigned to warehouse 1 is given by:

$Pr[A_{1,1}] = \frac{1}{1+E(M)}$ where the average inter-renewal interval is [1+E(M)]

From the renewal process as defined in the beginning of this section, the inter-renewal interval is the number of orders that come in between consecutive renewals. This is the number of orders that are assigned to warehouses other than warehouse 1, just after an order was assigned to warehouse (triggering a renewal), plus the final order in this interval that is assigned to warehouse 1, which triggers the next renewal. Let M be the number of orders assigned to warehouses other than warehouse 1, between renewals. Then the mean inter-renewal interval is given by [1+E(M)]. The value of M varies between 0 and N-1, with the probability depending on the number of demands during the replenishment lead-time. There are two cases to consider:

1. Suppose an order is assigned to warehouse 1 at time t and let D(t, t + L] denote the demand over the interval (t, t+L], and suppose D(t, t+ L] < N. Then M equals D(t, t + L], as the next demand after time t + L is assigned to warehouse 1, which is the next renewal point.

2. Suppose an order is assigned to warehouse 1 at time t and suppose D(t, t+ L] ≥ N. Then M equals N -1 as we will assign the first N – 1 demands to the other warehouses and will assign the N^{th} demand to warehouse 1, which is the next renewal point.

Thus, we get:

$$E[M]=\sum_{k=0}^{N-1}k\times\Pr[D_L=k]+(N-1)\Pr[D_L\geq N]$$

Let us define a reward function, R(k) = 1 if order k is assigned to warehouse 1. Then, the reward accumulated during every inter-renewal interval is exactly one, since a renewal occurs immediately after an order is assigned to warehouse 1. Thus, E[R(n)]=1.

By applying the Key Renewal Theorem, we get:

$$\lim_{n \to \infty} E[R(t)] = \frac{E[R(n)]}{\overline{X}}$$

where $\overline{X}$ is the average inter-renewal interval, $\overline{X} = 1 + E(M)$
and E[R(n)] is the average reward accumulated during X

However, E[R(t)] is the expected rate of reward accumulation, which in this model, is the probability of an order being assigned to warehouse 1 to be fulfilled either immediately from stock or from a replenishment unit when it arrives, i.e., P[$A_{1,1}$]. Hence, the result follows.

The only difference in the proof for Pr[$B_{1,1}$] is that:

$$\Pr\left[B_{1,1} \mid A_{1,1}\right] = 1 - \Pr\left[F_{1,1} \mid A_{1,1}\right] = 1 - \Pr\left[D_L < N\right]$$

This follows immediately by recalling that $F_{1,1}$ and $B_{1,1}$ are mutually exclusive events, and $A_{1,1} = F_{1,1} \cup B_{1,1}$.

QED

For the other warehouses, we can similarly show that,

$$\begin{aligned}
&\textit{for } j \neq 1: \\
&\Pr\left[F_{1,j}\right] = \frac{\Pr\left[D_L < N\right] - \Pr\left[F_{1,1}\right]}{N-1} \\
&\Pr\left[B_{1,j}\right] = \frac{\left(1 - \Pr\left[D_L < N\right]\right) - \Pr\left[B_{1,1}\right]}{N-1}
\end{aligned}$$

As explanation, we note that the fill rate from the non-local warehouses to serve demand in region 1 equals the system fill rate, net of the fill rate from warehouse 1; by symmetry, we then divide the total fill rate associated with the non-local warehouses equally across these N-1 warehouses. Similarly we find the probability that a non-local warehouse serves a backordered demand from region 1.

With the above results, we can find the probability a demand in region i is served by a transshipment from warehouse j; $Pr[Fi,j] + Pr[Bi,j]$ for $i \neq j$.

3.1.3 INTERPOLATION METHOD

For other demand distribution across the regions, exact solutions could not be found either by similar methods or using the method Xu [Xu05]. However, we expect Pr[Fi,i] to be monotonically decreasing in α_i for a given system demand rate, λ, since we are effectively increasing the demand rate for the region while maintaining the same inventory level of one unit at the regional warehouse. This holds true provided that demand in all other regions remains proportional, and the order fulfillment policy in place is P_1. Thus, we propose to approximate the Pr[Fi,i] for other cases of demand using some form of monotonic interpolation. Using the known results for α_i=1 and α_i=1/N, we considered a linear interpolation and exponential interpolation approximation for Pr[Fi,i] using the values of Pr[Fi,i] for the balanced and extreme demand distribution cases. Let y_k = Pr[Fi,i] for $\alpha_i = k$, $0<k\leq1$.

The linear interpolation for y_k is given by:

$$y_k = y_1 + \frac{(y_1 - y_{1/N}) \times (k-1)}{(1-1/N)}.$$

The exponential interpolation for y_k is given by:

$$y_k = a \times \exp(b \times k) \quad \text{where } b = \frac{1}{1-1/N}\log\left(\frac{y_1}{y_{1/N}}\right),\ a = y_1 \times \exp(\text{-}b)$$

We compared these approximations with Monte-Carlo simulation results for Pr[Fi,i] for a wide range of scenarios. A sample case for a 4-unit 4-warehouse scenario with λ=1 and L=3 is shown below. We set $\alpha_i = (1 - \alpha_1)/(N - 1)$ for i=2, 3, ... N.

TABLE 2: COMPARISON OF SIMULATION AND INTERPOLATION RESULTS

	α1									
Pr[$F_{1,1}$]	1	0.8	0.6	0.5	0.4	0.3	0.25	0.2	0.1	0.05
Simulation	0.1958	0.2202	0.2501	0.2719	0.291	0.3161	0.3282	0.3386	0.37	0.3761
Linear approx	0.1945	0.2306	0.2667	0.2847	0.3028	0.3208	0.3298	0.3389	0.3569	0.3659
Exponential approx	0.1945	0.2239	0.2578	0.2766	0.2968	0.3184	0.3298	0.3417	0.3666	0.3797

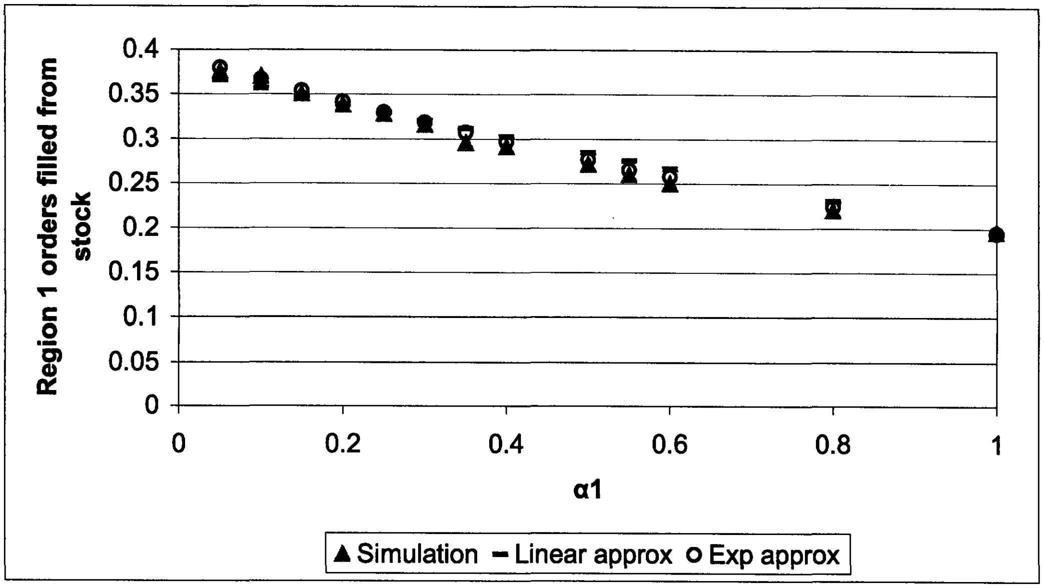

FIG 3: COMPARISON OF SIMULATION AND INTERPOLATION RESULTS

We observe empirically that the exponential approximation performs slightly better than the linear approximation. In general, the approximations are both reasonably good, within 5% of the simulation results. Furthermore, the error seems systematic with the approximation overestimating Pr[Fi,i] when $\alpha_i \in [1/N, 1]$ and underestimating Pr[Fi,i] when $\alpha_i \in [0, 1/N]$. This allows us to estimate the proportion of system-wide transshipment better since some regions have lower than average demand rates while other have greater than average demand rate, thus balancing the underestimation with the overestimation.

However, this approximation method does not account for the effects due to the demand distribution across the warehouses. For instance, consider a 3-unit 3-location scenario. If α_1=0.33, α_2=0.67, and α_3=0, then Pr[F1,1 | α_1=0.33, α_2=0.67, α_3=0] is clearly less than $Pr[F_{1,1} \mid \alpha_1=0.33, \alpha_2=0.33, \alpha_3=0.33]$. To see this, consider an order arriving from region 2 when warehouse 2 does not have stock but Warehouses 1 and 3 have stock. Under the order fulfillment policy P_1, warehouse 1 may be chosen to fill the order with probability ½. This event occurs more often when $\alpha_2=0.67$ compared to $\alpha_2=0.33$, thus lowering warehouse 1's probability of filling region 1's orders from stock, $Pr[F_{1,1}]$. Clearly, the better allocation policy is to fill the order from warehouse 3 stock which does not have any demand in its region. We use this concept in developing and comparing different order fulfillment policies in Chapter 4.

We expect these approximation methods perform best when local demand faced at the other warehouses is equal, i.e., $\alpha_i = (1 - \alpha_1)/(N - 1)$ for i=2, 3, ... N. If demand at other warehouses is not equal, then under the order fulfillment policy P_1, the approximation underestimates $Pr[F_{i,i}]$. This effect can be illustrated using the same example as in the previous paragraph.

We define another performance metric, Service Failure, as the probability that an order is not filled immediately by its local warehouse. We say that a Service Failure occurs when a demand is backordered or when a demand is filled immediately by some warehouse other than its local warehouse. Thus,

$$\text{Service Failure for region } i,\ SF_i = 1 - Pr[F_{i,i}]$$

$$\text{Service Failure for system, SF} \cong 1 - \sum_{i=1}^{N} \alpha_{i^i} \times \Pr[F_{i,i}]$$

Note that the Service Failure for the system is just the demand-weighted service failure in the regions. We can estimate the Service Failure for the system quite accurately using the above formula despite the errors in estimating Pr[Fi,i]. This is due to the cancellation of the systematic errors in the approximation of Pr[Fi,i] as some $\alpha_i \in [1/N, 1]$ while other $\alpha_i \in [0, 1/N]$.

We approximate the probability of a backorder filled by its local warehouse, Pr[Bi,i], as being almost equal for each warehouse in the system, then:

$$\Pr[B_{i,\,i}] \approx \frac{1}{N} \Pr[D_L \geq N]$$

Thus, we can estimate the probability of transshipment for each region and for the system as:

$$\boldsymbol{TS}_i = \boldsymbol{SF}_i - \Pr[\boldsymbol{B}_{i,i}]$$

$$\boldsymbol{TS} = \sum_{i=1}^{N} \alpha_i \times \boldsymbol{TS}_i$$

We compared these estimates of Service Failure and Transshipment for the system with Monte-Carlo simulation results under a wide range of conditions. We considered systems with 3 to 5 warehouses, each warehouse holding one unit of inventory. The fill rates for each system was varied ranging from 70% to 99%, by changing the system demand rate λ for a constant value of L=3. Finally, the demand across regions was changed from almost balanced to extremely unbalanced by varying $\alpha_k = 1/N + m\delta$ for $m=k-[(N+1)/2]$ where $k=1,2,\ldots,N$ and $\delta>0$. We set δ based on the scenario we want to model. For instance, to model a scenario of unbalanced demand, we set δ such that α_1 is close to 0. Similarly, to model a scenario of almost balanced demand, we set δ such that α_1 is close to 1/N. Finally, for the middle case, we set δ to a value between the two previous values. To give an example, one such scenario (Table 3, row 13) was for a 4-unit 4-warehouse system that had a system demand rate of $\lambda=0.58$ items per week and lead time of 3 weeks (giving a fill rate of 90%) and with demand across regions being $\alpha_4=0.4$, $\alpha_3=0.3$, $\alpha_2=0.2$, $\alpha_1=0.1$ (demand distribution spread is high) for $\delta=0.1$.

Each simulation scenario was run 100 times for 500,000 orders to reduce variability inherent in the simulation. The mean and standard deviation of the estimates of the transshipment and service failure rates in simulation runs for each scenario were recorded, and confidence intervals derived for these parameters. For example, consider the scenario from (Table 3, row 13); From the simulation, we computed an estimate of 0.4463 for service failure with a standard deviation of 0.0009, thus giving a 95% confidence interval that the actual service failure is in [0.4445, 0.4481]. These confidence intervals were found to be quite tight, with interval length being around 0.3%, thus giving a good estimate of the actual service failure value for comparison with the approximation. A sample table (Table 3) is shown on the next page.

The summarized results of the computational tests for the service failure and transshipment values using the linear approximation are shown in tables 5 – 10. The parameter values for the demand rate, corresponding fill-rate and demand distribution for each of the test scenarios is listed in Table 4. Tables 5 and 8 show the service failure and transshipment approximation values for each of the scenarios. Tables 6 and 9 show the mean service failure and transshipment values obtained from simulation for each of the scenarios. Finally, Tables 7 and 10 show the relative error between the approximation and simulation values as a percentage [(Approximation value – Simulation mean value)/Simulation mean value %]. These tables show that the relative error between the approximation and the simulation result systematically increases with the number of warehouses in the system, with increasing fill-rates, and as demand gets increasingly unbalanced between regions. However, these errors are small, with absolute error below 2%, and relative error below 5% for the most part. Thus, we conclude that there is a generally good fit between the approximation and the simulation. This gives us greater confidence in using this method as a tool in estimating transshipment proportions and using it to determine optimal inventory policies.

TABLE 3: SIMULATION AND APPROXIMATION RESULTS FOR SERVICE FAILURE
(4-UNIT 4-WAREHOUSE SYSTEM WITH VARYING FILL RATES AND DEMAND DISTRIBUTIONS)

Policies / Parameters		Equiprobable Random Selection of WH		
Fill Rate	**Demand Rate and Distribution across Regions**	**Mean**	**Std Dev**	**Approx Method**
99%	Λ=0.28 (0.28,0.26,0.24,0.22)	0.2103	0.0006	0.2102
96%	Λ=0.42 (0.28,0.26,0.24,0.22)	0.3129	0.0007	0.3127
90%	Λ=0.58 (0.28,0.26,0.24,0.22)	0.4246	0.0009	0.4247
80%	Λ=0.76 (0.28,0.26,0.24,0.22)	0.5396	0.0008	0.5399
70%	Λ=0.28 (0.28,0.26,0.24,0.22)	0.6297	0.0009	0.6301
99%	Λ=0.28 (0.34,0.28,0.22,0.16)	0.2175	0.0006	0.2155
96%	Λ=0.42 (0.34,0.28,0.22,0.16)	0.3205	0.0008	0.3181
90%	Λ=0.58 (0.34,0.28,0.22,0.16)	0.4317	0.0009	0.4295
80%	Λ=0.76 (0.34,0.28,0.22,0.16)	0.5456	0.0009	0.5439
70%	Λ=0.42 (0.34,0.28,0.22,0.16)	0.6346	0.0008	0.6334
99%	Λ=0.28 (0.4,0.3,0.2,0.1)	0.2320	0.0006	0.2261
96%	Λ=0.42 (0.4,0.3,0.2,0.1)	0.3361	0.0007	0.3289
90%	Λ=0.58 (0.4,0.3,0.2,0.1)	0.4463	0.0009	0.4393
80%	Λ=0.76 (0.4,0.3,0.2,0.1)	0.5576	0.0009	0.5512
70%	Λ=0.42 (0.4,0.3,0.2,0.1)	0.6444	0.0009	0.6399

TABLE 4: PARAMETER VALUES FOR COMPUTATIONAL TESTS

Fill Rate	Demand Distribution Spread	3U-3L	4U-4L	5U-5L
99%	Low	λ=0.14 (0.367,0.333,0.3)	λ=0.28 (0.28,0.26,0.24,0.22)	λ=0.43 (0.24,0.22,0.2,0.18,0.16)
96%	Low	λ=0.25 (0.367,0.333,0.3)	λ=0.42 (0.28,0.26,0.24,0.22)	λ=0.62 (0.24,0.22,0.2,0.18,0.16)
90%	Low	λ=0.37 (0.367,0.333,0.3)	λ=0.58 (0.28,0.26,0.24,0.22)	λ=0.81 (0.24,0.22,0.2,0.18,0.16)
80%	Low	λ=0.51 (0.367,0.333,0.3)	λ=0.76 (0.28,0.26,0.24,0.22)	λ=1.02 (0.24,0.22,0.2,0.18,0.16)
70%	Low	λ=0.64 (0.367,0.333,0.3)	λ=0.28 (0.28,0.26,0.24,0.22)	λ=1.21 (0.24,0.22,0.2,0.18,0.16)
99%	Medium	λ=0.14 (0.4,0.333,0.267)	λ=0.28 (0.34,0.28,0.22,0.16)	λ=0.43 (0.3,0.25,0.2,0.15,0.1)
96%	Medium	λ=0.25 (0.4,0.333,0.267)	λ=0.42 (0.34,0.28,0.22,0.16)	λ=0.62 (0.3,0.25,0.2,0.15,0.1)
90%	Medium	λ=0.37 (0.4,0.333,0.267)	λ=0.58 (0.34,0.28,0.22,0.16)	λ=0.81 (0.3,0.25,0.2,0.15,0.1)
80%	Medium	λ=0.51 (0.4,0.333,0.267)	λ=0.76 (0.34,0.28,0.22,0.16)	λ=1.03 (0.3,0.25,0.2,0.15,0.1)
70%	Medium	λ=0.64 (0.4,0.333,0.267)	λ=0.42 (0.34,0.28,0.22,0.16)	λ=1.21 (0.3,0.25,0.2,0.15,0.1)
99%	High	λ=0.14 (0.467,0.333,0.2)	Λ=0.28 (0.40,0.30,0.20,0.10)	λ=0.28 (0.36,0.28,0.2,0.12,0.04)
96%	High	λ=0.25 (0.467,0.333,0.2)	Λ=0.42 (0.40,0.30,0.20,0.10)	λ=0.43 (0.36,0.28,0.2,0.12,0.04)
90%	High	λ=0.37 (0.467,0.333,0.2)	Λ=0.58 (0.40,0.30,0.20,0.10)	λ=0.81 (0.36,0.28,0.2,0.12,0.04)
80%	High	λ=0.51 (0.467,0.333,0.2)	Λ=0.76 (0.40,0.30,0.20,0.10)	λ=1.02 (0.36,0.28,0.2,0.12,0.04)
70%	High	λ=0.64 (0.467,0.333,0.2)	Λ=0.42 (0.40,0.30,0.20,0.10)	λ=1.21 (0.36,0.28,0.2,0.12,0.04)

TABLE 5: SERVICE FAILURE APPROXIMATION

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**
99%	0.14	0.21	0.26	0.14	0.22	0.27	0.15	0.23	0.28
96%	0.25	0.31	0.37	0.25	0.32	0.38	0.26	0.33	0.39
90%	0.36	0.42	0.48	0.36	0.43	0.48	0.37	0.44	0.50
80%	0.48	0.54	0.59	0.48	0.54	0.59	0.49	0.55	0.60
70%	0.58	0.63	0.67	0.58	0.63	0.67	0.58	0.64	0.68

TABLE 6: SERVICE FAILURE SIMULATION MEAN

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**
99%	0.14	0.21	0.26	0.14	0.22	0.27	0.15	0.23	0.30
96%	0.25	0.31	0.37	0.25	0.32	0.38	0.26	0.34	0.41
90%	0.36	0.42	0.48	0.36	0.43	0.49	0.37	0.45	0.51
80%	0.48	0.54	0.59	0.48	0.55	0.60	0.49	0.56	0.62
70%	0.58	0.63	0.67	0.58	0.63	0.68	0.58	0.64	0.69

TABLE 7: SERVICE FAILURE APPROXIMATION ERROR PERCENTAGE
(RELATIVE TO SIMULATION RESULTS)

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**
99%	0.14	-0.08	-0.36	-0.06	-0.95	-2.12	-0.70	-2.63	-5.18
96%	0.14	-0.06	-0.21	0.02	-0.75	-1.63	-0.56	-2.18	-4.04
90%	0.19	0.01	-0.16	0.08	-0.51	-1.16	-0.38	-1.58	-2.94
80%	0.16	0.04	-0.07	0.12	-0.30	-0.73	-0.27	-1.17	-1.93
70%	0.14	0.06	-0.01	0.10	-0.20	-0.47	-0.15	-0.71	-1.30

Statistics taken over 100 simulation runs of 500k orders

TABLE 8: TRANSSHIPMENT APPROXIMATION

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**
99%	0.14	0.21	0.26	0.14	0.21	0.27	0.15	0.22	0.28
96%	0.23	0.30	0.36	0.24	0.31	0.37	0.24	0.32	0.38
90%	0.33	0.40	0.46	0.33	0.40	0.46	0.33	0.41	0.48
80%	0.41	0.49	0.55	0.41	0.49	0.55	0.42	0.50	0.56
70%	0.48	0.56	0.61	0.48	0.56	0.61	0.48	0.57	0.62

TABLE 9: TRANSSHIPMENT SIMULATION MEAN

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**
99%	0.14	0.21	0.26	0.14	0.21	0.27	0.15	0.23	0.29
96%	0.23	0.30	0.36	0.24	0.31	0.38	0.25	0.33	0.40
90%	0.33	0.40	0.46	0.33	0.41	0.47	0.34	0.42	0.49
80%	0.41	0.49	0.55	0.41	0.50	0.56	0.42	0.51	0.57
70%	0.48	0.56	0.61	0.48	0.56	0.62	0.48	0.57	0.63

TABLE 10: TRANSSHIPMENT APPROXIMATION ERROR PERCENTAGE
(RELATIVE TO SIMULATION RESULTS)

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**	**3U-3L**	**4U-4L**	**5U-5L**
99%	0.02	-0.12	-0.37	-0.16	-0.95	-2.10	-0.70	-2.59	-5.13
96%	-0.02	-0.14	-0.25	-0.09	-0.76	-1.61	-0.51	-2.09	-3.94
90%	0.00	-0.08	-0.20	-0.05	-0.52	-1.15	-0.30	-1.47	-2.82
80%	-0.01	-0.06	-0.11	0.01	-0.33	-0.71	-0.16	-1.06	-1.79
70%	0.00	-0.03	-0.05	-0.01	-0.20	-0.46	-0.07	-0.60	-1.18

Statistics taken over 100 simulation runs of 500k orders

3.2 MULTIPLE UNITS AT ONE WAREHOUSE

Until now we have only considered the case where one unit of inventory is held at a warehouse. However, for cases where demand between regions is extremely unbalanced or when service level measures such as fill rate are high, holding more than one unit in a warehouse may be required. We can represent a warehouse that holds say m units (m>1) as m warehouses that hold one unit each, and in addition, use a specific order fulfillment policy, say P_2. This order fulfillment policy has a priority list whereby these m warehouses have equal preference to fulfill orders from each other in the event that any of them receives an order but does not have inventory on-hand.

For example, consider a system with 2 warehouses where warehouse 1 receives 70% of customer demand and warehouse 2 receives 30% of customer demand. Let us assume that the system demand requires 3 units of inventory to be held in order to satisfy a fill rate target set by the e-tailer with warehouse 1 stocking 2 units of inventory and warehouse 2 stocking 1 unit of inventory. Then we can analyze this system as a 3-unit 3-warehouse system where warehouse 1 is represented as two warehouses, say warehouses 1A and 1B, each receiving 35% of the customer demand, and holding 1 unit of on-hand inventory. Warehouse 2 remains with 1 unit of on-hand inventory and receives 30% of the customer demand. The priority list for order fulfillment is shown in the table below (1 being highest priority):

Warehouse Filling Order / Warehouse Receiving Order	1A	1B	2
1A	1	2	3
1B	2	1	3
2	2	2	1

TABLE 11: PRIORITY LIST FOR ORDER FULFILLMENT FOR 3-UNIT 3-WAREHOUSE SYSTEM

The first row of this table says that when an order comes from region 1A, then the first preference is to assign the order to warehouse 1A. If warehouse 1A does not have stock, then the second preference is to assign the order to warehouse 1B. If warehouse 1B also does not have stock, only then does the system assign the order to warehouse 2. If none of the warehouses in the system have stock, then the earliest arriving unassigned replenishment unit in the system is assigned to this order.

This system is equivalent to the 3-unit 2-warehouse system as described above. Thus, we can extend this method and describe any k-unit N-warehouse system (k>n) as an equivalent k-unit k-warehouse system, where the warehouses may have different demand rates and a priority list for order fulfillment that gives strict preference to warehouses that were derived from the same parent warehouse in the original system.

The present methodology (described in Section 3.1) only allows us to estimate the proportion of local demand that is filled immediately from stock at the local or preferred warehouse from a k-unit k-warehouse system when the order fulfillment policy in effect is P_1. However, since some of the warehouses are physically the same, and the order fulfillment policy in effect is P_2, we need to develop a new methodology to estimate the proportion of orders that are transshipped.

The new method entails two steps:

1. Use the present methodology (described in Section 3.1) to estimate the proportion of local demand that is filled immediately from stock at the local warehouse from a k-unit k-warehouse system when the order fulfillment policy in effect is P_1.

2. Apply a correction factor to these estimates to account for fact that some of the warehouses are physically the same, and the order fulfillment policy in effect is P_2,

We now describe this new method in more detail. Given only the demand rates seen by each warehouse and the replenishment lead time, we can calculate the system fill rate, FR. We can use the method from section 3.1 to estimate the proportion of local demand that is filled immediately from stock at the local warehouse, $P(F_{ii})$, under the assumption that the order fulfillment policy in effect is P_1. The transshipment into each region depends on the proportion of the regional demand that is not filled by its local warehouse, multiplied by a correction factor for the regional demand that is actually assigned to the dummy warehouse(s) at the same physical warehouse. In the example that we are using, this correction ratio is the ratio of region 1A's demand filled by warehouse 2 to region 1A's demand filled by warehouses 1B and 2. Thus, we estimate the percentage transshipment into each region and for the system as:

$$TS_i \approx (1 - P(F_{ii}) - P(B_{ii})) \times [\textit{Correction Ratio}]$$

$$\text{Correction Ratio} = \text{Ratio of local region's demand filled by foreign}$$
$$\text{warehouse to local region's demand filled by dummy and foreign warehouse}$$

$$TS = \sum_{i=1}^{k} \alpha_i \times \mathrm{TS}_i$$

The Service Failure in each region depends on the proportion of the regional demand that is not filled by its local warehouse immediately, with a correction term for the regional demand filled immediately by the dummy warehouse(s) at the same physical warehouse. Since the proportion of demand filled immediately from stock in any region is equal to the system fill-rate, this correction term depends on (FR - Pr[Fi,i]). In the context of the current example, this correction ratio is the ratio of region 1A's demand that is filled immediately by warehouse 1B to region 1A's demand that is filled immediately by warehouses 1B and 2. The service failure for each region and for the system can be estimated as:

$$SF_i \approx [\{1 - P(F_{ii})\} - \{FR - P(F_{ii})\} \times \{1 - \textit{Correction Ratio}\}]$$

$$\text{Correction Ratio} = \text{Ratio of local region's demand filled immediately by}$$
$$\text{foreign warehouse to local region's demand filled immediately by}$$
$$\text{dummy and foreign warehouse}$$

$$SF = \sum_{i=1}^{k} \alpha_i \times SF_i$$

We note that the above approximations require us to calculate the proportion of local demand that is filled by the foreign warehouses under the order fulfillment policy, P_2. This is not easy to calculate in the current scenario. However, If we assume that the replenishment lead time follows an exponential distribution with the mean equal to the actual constant replenishment lead time, then the k-unit k-warehouse system with the priority shipment policy, P_2, can be modeled as a markov process. This markov process model enables us to approximate the proportion of local demand that is filled by foreign warehouses as a function of the steady state probabilities and flow rates. The markov process model for the current 3-unit 2-warehouse system is shown in Figure 4.

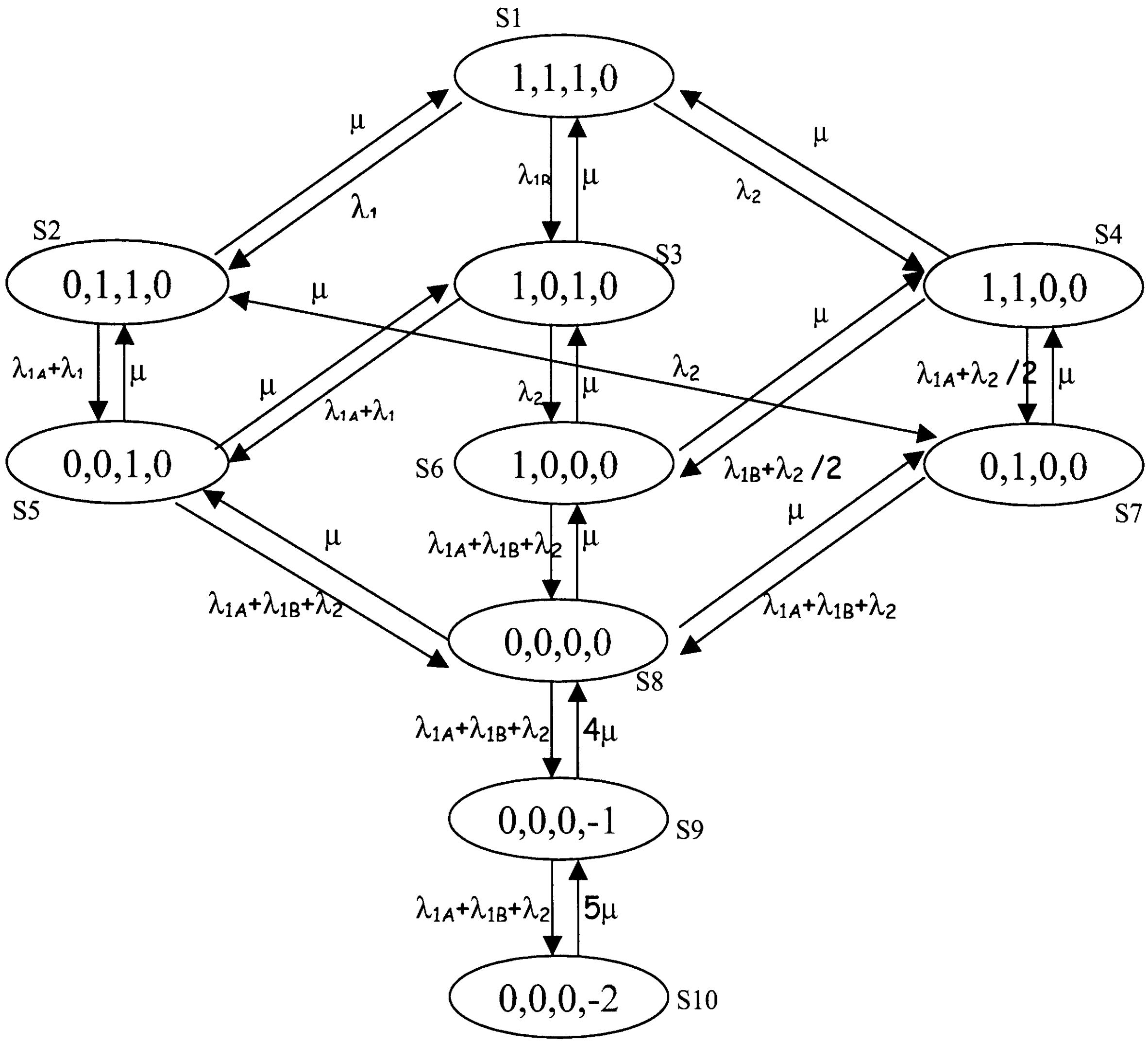

FIG 4: MARKOV PROCESS MODEL FOR 3-UNIT 3-LOCATION MODEL WITH PRIORITIZED TRANSSHIPMENTS (POLICY P_2)

Each state of this system is represented as (x_1, x_2, x_3, x_4) where x_i indicates the level of on-hand inventory at warehouse i, for i=1, 2, 3. The value of x_4 indicates the number of backorders in the system. The flow rates between the states of this markov process model incorporate the priority shipment rules for the order fulfillment policy, P_2. For example, the flow rate from state (1,1,0,0) to (0,1,0,0) is $\lambda_{1A}+\lambda_2/2$ since this transition occurs only if an order comes from region 1A or if the order comes from region 2 and warehouse 1A is chosen to fill the order (which occurs with probability ½ since warehouse 1B has equal priority in filling warehouse 2's order).

We note that the online retailer model allows for any number of backorders, not just two, as shown in the markov process model above. This truncation is done for computational purposes, and can be justified since the probability of having greater than 2 backorders is negligible in the above case. For other cases, we can extend the markov process model up to a stage where the probability of exceeding that many backorders is negligible, less than 0.05.

We first calculate the steady state probabilities, P(Si), of this system using the flow balance and probability normalization equations. We then calculate the ratio of orders from region 1A filled by warehouse 2 immediately to orders from region 1A filled by warehouses 1B and 2 immediately, as:

$$P[F_{1A,2}] = P(S5)/[P(S2) + P(S5) + P(S7)]$$

The explanation of this factor is as follows: The rate of orders from region 1A filled by warehouse 2 is $[\lambda_{1A} \times P(S5)]$ since S5 represents the state where only warehouse 2 has stock, and if any order from region 1A comes, then it must be assigned to warehouse 2. Similarly, the rate of orders from region 1A filled by warehouse 1B is $[\lambda_{1A}$ x

{P(S2)+P(S7)}] since S2 and S7 represents the states where warehouse 1B has stock while warehouse 1A does not have stock, and if any order from region 1A comes, then it must be assigned to warehouse 1B.

$$\text{Correction Ratio for Service Failure} = P[F_{1A,2}]$$

Similarly, the proportion of region 1A's orders that are filled by region j as backorders is given by:

$$P[B_{1A,j}] \cong P(D_L \geq 3)/3$$

Then the proportion of orders from region 1A filled by warehouse 2 (either immediately or as a backorder) is:

$$\text{Correction Ratio for transshipment} =$$

$$\{P[F_{1A,2}] + P[B_{1A,2}]\} / \{P[F_{1A,2}] + P[B_{1A,2}] + P[F_{1A,1B}] + P[B_{1A,1B}]\}$$

We use this approach for the above example, and then compare its results with simulation results obtained for the 3-unit 2-warehouse system with 70% demand from region 1 and 30% from region 2. The following results were obtained for system fill rate ranging from 99.9% to 0.6%, by setting L=3 and varying the system demand rate, λ. The table 12 shows the simulation and approximation results for this case. More computational tests were conducted for 3-unit 2-warehouse systems and 4-unit 2-warehouse systems (shown in tables 13 – 15). The approximation yields very good results, usually within 1% of the value obtained by simulation. Moreover, the approximation is very good for fill rates for practical systems (>90%). However, simulation results also show that this estimate worsens as the number of units in the system, k, increases.

Fill-Rate (%)	Transshipment (Simulation) %	Transshipment (Approx) %	Service Failure (Simulation) %	Service Failure (Approx) %
99.9	2.070	2.122	2.120	2.178
99.1	5.770	5.919	6.280	6.425
90	17.150	17.756	22.900	23.417
80	23.180	24.100	34.780	35.353
70	28.020	28.672	45.340	45.727
55	33.090	33.538	58.520	58.794
42	36.320	36.897	69.310	69.571
25	40.080	40.308	82.920	82.885
6	42.890	42.816	95.780	96.074
0.6	43.400	43.297	99.570	99.612

TABLE 12: COMPARISON OF SIMULATION AND APPROXIMATION RESULTS FOR 3-UNIT 2-LOCATION SYSTEM (70:30 DEMAND DISTRIBUTION)

Fill-Rate (%)	Transshipment (Simulation) %	Transshipment (Approx) %	Service Failure (Simulation) %	Service Failure (Approx) %
99.9	1.420	1.493	1.470	1.550
99.1	4.550	4.897	5.120	5.430
90	15.380	16.311	21.810	22.300
80	21.380	22.234	33.930	34.156
70	25.300	26.624	43.880	44.686
55	30.360	31.151	57.670	57.905
42	33.650	34.240	68.860	68.846
25	36.390	37.324	82.410	82.413
6	39.250	39.550	95.740	95.836
0.6	39.840	39.949	99.610	99.577

TABLE 13: COMPARISON OF SIMULATION AND APPROXIMATION RESULTS For 3-Unit 2-Location System (80:20 Demand Distribution)

Fill-Rate (%)	Transshipment (Simulation) %	Transshipment (Approx) %	Service Failure (Simulation) %	Service Failure (Approx) %
99.9	3.150	3.072	3.190	3.123
99.1	7.550	7.510	7.980	8.004
90	19.560	20.032	24.860	25.433
80	25.830	26.412	36.240	37.083
70	30.910	31.212	46.920	47.344
55	36.060	36.271	59.720	60.112
42	39.560	39.797	70.370	70.605
25	43.420	43.409	83.320	83.515
6	46.380	46.101	95.950	96.138
0.6	46.710	46.626	99.680	99.628

TABLE 14: COMPARISON OF SIMULATION AND APPROXIMATION RESULTS
For 3-Unit 2-Location System (60:40 Demand Distribution)

Fill-Rate (%)	Transshipment (Simulation) %	Transshipment (Approx) %	Service Failure (Simulation) %	Service Failure (Approx) %
99.9	2.650	2.886	2.730	2.948
99.1	5.620	6.310	6.240	6.877
90	14.590	15.956	20.930	22.171
80	19.650	21.158	32.050	33.567
70	23.520	25.039	42.610	43.831
55	27.870	29.189	56.090	57.064
42	30.880	32.014	67.160	68.059
25	34.530	34.926	81.340	81.911
6	37.160	37.057	95.510	95.687
0.6	37.790	37.468	99.630	99.588

TABLE 15: COMPARISON OF SIMULATION AND APPROXIMATION RESULTS
For 4-Unit 2-Location System (75:25 Demand Distribution)

Chapter 4

In this chapter, we use the performance metrics developed in chapters 2 and 3 to create guidelines for optimal inventory stocking for online retailers. Given regional demand rates and the replenishment lead-time, we develop methods to determine how much inventory should be held in the system and how should this be distributed among the warehouses. Finally, we study various order fulfillment policies and find the best order fulfillment policy to service demand under specific conditions.

4.1 INVENTORY PLANNING

The objective here is to develop an optimal inventory stocking policy for each item. We consider a scenario where the e-tailer has N warehouses. Essentially we want to find under what conditions should the e-tailer hold k units of inventory and in which of its N warehouses.

There are three primary dimensions when considering the inventory holding policy for this e-tailer: the inventory holding cost, the transportation cost of orders to the customer and the loss of customer goodwill if the customer order has to be backordered (when there is no on-hand inventory in the system to fulfill the order). We call this loss of customer goodwill as backorder cost. Furthermore, the transportation cost can be divided into the local transportation cost (cost of delivery from local warehouse to the customer) and the additional transshipment cost when the local warehouse does not have on-hand inventory and the order has to be shipped by a warehouse in some other region to the customer.

Low inventory holding cost requires the e-tailer to hold less system-wide inventory but that increases both the transshipment cost and the backorder cost due to the centralized order fulfillment policy of the system. The best inventory holding policy develops the optimal trade-off between these costs.

When the system demand rate, λ, and the replenishment lead-time, L, is known, we can calculate the average inventory held in the system and the percentage of system orders that are backordered. Let the backorder cost per order be denoted by B, and the holding cost per unit per period be denoted by H. Then,

$$E[BackorderCost/period] = \lambda * B * (1 - FR) = \lambda * B * (1 - \sum_{i=0}^{N-1} P(D_L = i))$$

$$E[InvCost/period] = H * AvgInv = H * \sum_{i=0}^{N} i \times P(D_L = N - i)$$

Estimating the transportation cost for an item is more complicated as it is affected by the inventory holding configuration of that item in the system. For example, if the e-tailer has a single warehouse and holds just one unit of inventory in the system, then the local transportation cost is high since the single warehouse serves all demand; however, for the same reason there is no transshipment cost. However, if the e-tailer holds N units of inventory in N warehouses, then the local transportation cost goes down but transshipment cost needs to be considered. For inventory holding configurations between 1 and N, the e-tailer will need to revise the fulfillment regions for each warehouse that holds inventory of that item, and then estimate local transportation and system-wide transshipment costs. We perform inventory planning under such a general model for transportation costs in Section 4.1.1. Since we are studying low demand items, we only consider k-unit k-warehouse inventory configurations at first ($k \leq N$), although this can be easily extended to evaluate costs for m-unit k-warehouse ($m > k$) inventory configurations.

On the other hand, if the e-tailer already has established warehouses and logistics processes for transshipment, then the inventory planning problem reduces to determining how many units of an item to stock in each warehouse given the holding, backorder and transportation costs (for local and cross-region shipments). If the e-tailer can estimate the demand split between the regions in the system for this item, we can estimate the percentage of system orders that are transshipped by using the approximation developed in Chapter 3. This approach is explained in Section 4.1.2.

4.1.1 GENERAL TRANSPORTATION COST MODEL

Transportation cost depends on a number of factors including distance traveled or transportation zones, and fixed costs. However, for the sake of simplicity, we develop a general method to estimate the transportation cost for a k-unit k-warehouse system. Let the system-wide area be A. Then assuming a uniform demand split over the k regions, each warehouse needs to cover an area of size approximately A/k. The average distance traveled to deliver a customer order in the local region is then approximated as $\sqrt{(A/k)}$. Thus, if the local transportation cost of delivery is R/unit distance, then the average local delivery cost is $R\sqrt{(A/k)}$. However, we assume that the average transshipped distance does not change significantly when the number of warehouses is increased or decreased, and also involves significant fixed costs. Hence, in this model, we keep the additional transshipment cost as a constant, say T per order being transshipped.

$$TransportationCost\,/\,period \approx \lambda * \left[R * \sqrt{\frac{A}{k}} + T * P(Transshipment) \right]$$

In order to decide how many units of inventory to hold in the system and in which of the N warehouses, we need to evaluate the costs for all possible inventory holding configurations among the N warehouses. Thus, for N=2, we would have to evaluate the costs for the (1,0), (0,1), (1,1) inventory configurations. This may seem as a large computational problem with 2^{N-1} evaluations. However, e-tailers generally don't have more than 10 warehouses and so this brute force method has a relatively small number of numerical calculations.

Thus, the total cost of a k-unit k-warehouse system can be estimated as the sum of the Inventory Holding cost, Backorder cost and the Transportation cost. This can be optimized numerically for different values of k, and the best inventory holding configuration determined.

Some examples of developing optimal inventory holding policies using the procedure described above are shown below. For the given parameter values, we determined the total cost of the system for various values of k, and we selected the value of k that minimized the total cost.

TABLE 16: PARAMETER VALUES/RANGE AND OPTIMAL STOCKING POLICY FOR INVENTORY PLANNING

Parameter	**Value/Range**
System demand rate, λ	1
Demand during Lead Time,	1
System Area, A	1
Inventory Holding Cost per unit per period, H	1
Backorder Cost per order, B	1 to 3
Local Transportation Cost per order per unit distance, R	1 to 3
Additional Fixed Transshipment Cost per order, T	1 to 3
Number of warehouses holding inventory, k	1 to 5

R	B	T	Optimal k	Min Total Cost
1	1	1	1	2
1	1	2	1	2
1	1	3	1	2
1	2	1	1	2.6321
1	2	2	1	2.6321
1	2	3	1	2.6321
1	3	1	2	2.9195
1	3	2	2	3.2356
1	3	3	1	3.2642
2	1	1	1	3
2	1	2	1	3
2	1	3	1	3
2	2	1	2	3.3624
2	2	2	1	3.6321
2	2	3	1	3.6321
2	3	1	2	3.6266
2	3	2	2	3.9427
2	3	3	2	4.2588
3	1	1	2	3.8053
3	1	2	1	4
3	1	3	1	4
3	2	1	2	4.0695
3	2	2	2	4.3856
3	2	3	1	4.6321
3	3	1	3	4.2951
3	3	2	3	4.5939
3	3	3	3	4.8927

We notice that k=1 is optimal when B is low and R is low. This agrees with intuition, as there is no incentive to hold more inventory or distribute inventory if the backorder and local transportation costs are low. Similarly, k=2 is optimal when B and/or R is high, and T is medium or low. This can be explained by the incentive to hold either more or distributed inventory for the optimal policy. Again, k=3 is optimal when R and B are high, and T is relatively lower. This can be attributed to the incentive to hold larger quantities of regional inventory.

4.1.2 INVENTORY PLANNING FOR ESTABLISHED E-TAILER

Another situation that often arises in practice is when the e-tailer has established warehouses and centralized order fulfillment procedures, and wants to decide how many units of inventory to hold and in which warehouses, for a particular item. The decision then requires the analysis of the trade-off between the inventory holding, backorder and transportation costs under various inventory configurations. We use the methods developed in the chapters 2 and 3 to calculate the costs, and illustrate this with an example.

Consider a two-warehouse system where 70% of the demand comes from the first region and 30% from the second. The system demand rate is given by λ=0.25 orders/ week, and the replenishment lead-time, L=2 weeks. The costs of transportation within a local region and to the foreign region are known, as are the holding and backorder costs. We consider seven inventory configurations for the system – (1, 0), (0, 1), (2, 0), (0, 2), (1, 1), (2, 1), and (3, 1) where (a, b) indicates the amount of inventory held in each warehouse. Table 13 illustrates the trade-offs in this example.

LT	2
Λ	0.25
A1	0.7
A2	0.3

Trpt Cost	To R1	To R2
From R1	3	4
From R2	4	3

H	0.1
B	1

Scenario #	Inventory Config	AvgInv /period	Holding Cost /period	Fill Rate	Backorder Cost /period	Proportion R2 Demand shipped from R1	Proportion R1 Demand shipped from R2	TrptCost /period	TotalCost /period	TotalCost /unit
1	(1,0)	0.607	0.06065	0.6065	0.0983	1	0	0.825	0.9840	3.9360
2	(2,0)	1.516	0.15163	0.9098	0.0225	1	0	0.825	0.9991	3.9967
3	(0,1)	0.607	0.06065	0.6065	0.0983	0	1	0.925	1.0840	4.3360
4	(0,2)	1.516	0.15163	0.9098	0.0225	0	1	0.925	1.0991	4.3967
5	(1,1)	1.516	0.15163	0.9098	0.0225	0.1625	0.2309	0.8026	0.9767	3.9071
6	(2,1)	2.502	0.25019	0.9856	0.0036	0.1529	0.0384	0.7015	0.9553	3.8212
7	(3,1)	3.5	0.35002	0.9982	0.0004	0.1384	0.0048	0.7008	1.0512	4.2051

TABLE 17: PARAMETER VALUES AND STOCKING POLICY COSTS FOR VARIOUS INVENTORY PLANNING SCENARIOS

Table 13 shows that when the local transportation costs is $3 and transshipment cost to the other region is $4, when holding cost is $0.1 per week and a backorder costs $1, then the optimal inventory configuration is (2,1), i.e, holding 2 units of inventory in the first warehouse and 1 unit of inventory in the second warehouse. Such a cost structure is typical for media items. This example illustrates the usefulness of such a methodology for inventory planners at online retailers.

In fact, we can use this methodology to develop an efficient frontier of inventory policies – inventory stocking policies that minimize total cost per unit sold over a range of system demand rate, λ, for a given replenishment lead-time, L. This will enable online retailers to better understand the trade-offs of the various policies and determine which policy is optimal for them. For example, continuing with the same example as earlier, we vary λ from 0.1/week to 0.65/week for L=2 weeks. We compare the total cost per unit for the previously studied inventory policies over this range of λ and graph it. Then the lower envelope of the total cost per unit over all inventory policies is the efficient frontier of inventory policies. This is illustrated in Table 14 and Figure 5.

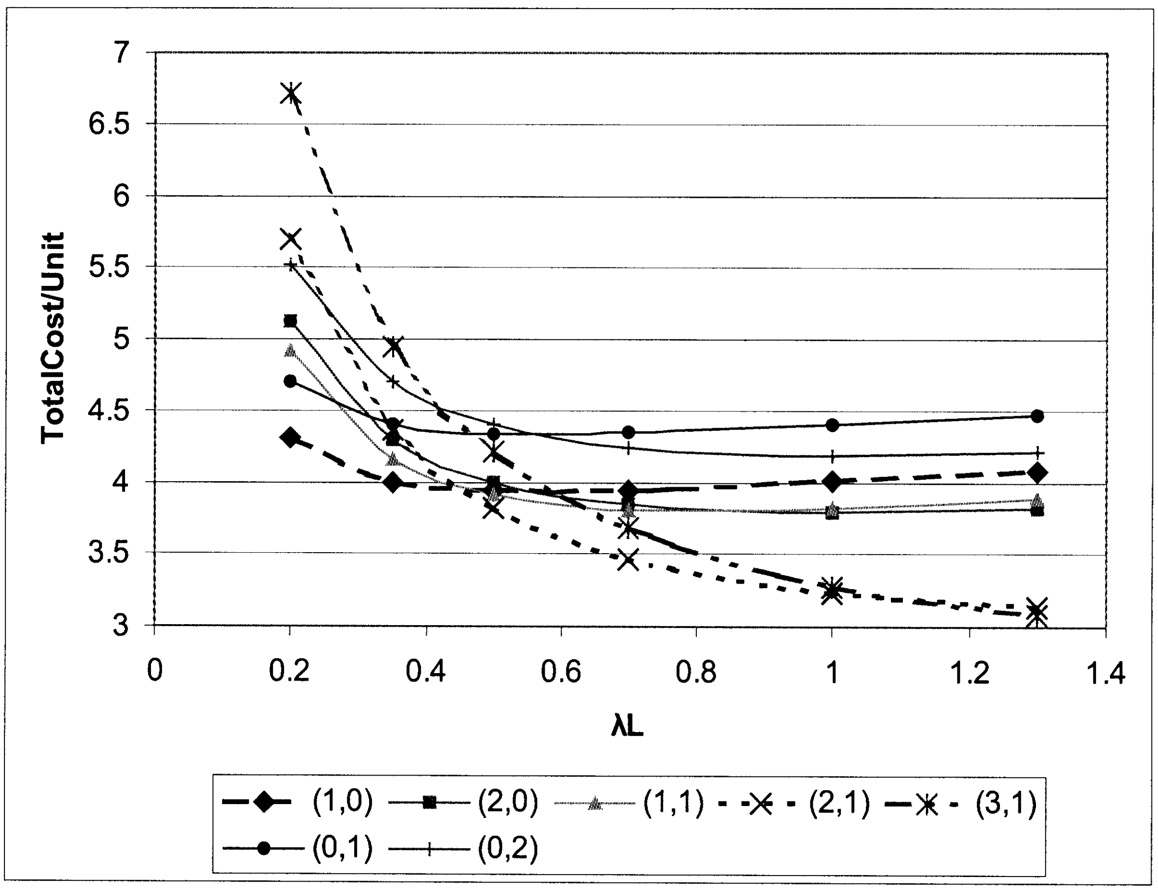

FIG 5: TOTAL COST PER UNIT FOR VARIOUS INVENTORY STOCKING POLICIES

Policy / λL	(1,0)	(0,1)	(2,0)	(0,2)	(1,1)	(2,1)	(3,1)
0.2	4.300	4.700	5.119	5.519	4.919	5.697	6.720
0.35	3.998	4.398	4.295	4.695	4.156	4.362	4.947
0.5	3.936	4.336	3.997	4.397	3.907	3.821	4.205
0.7	3.945	4.345	3.839	4.239	3.804	3.465	3.684
1	4.006	4.406	3.785	4.185	3.812	3.227	3.268
1.3	4.069	4.469	3.812	4.212	3.884	3.140	3.076

TABLE 18: TOTAL COST PER UNIT FOR VARIOUS INVENTORY STOCKING POLICIES

The efficient frontier of inventory stocking policies is the policy (1,0) for $\lambda < 0.45$, then policy (2,1) for $0.45 < \lambda < 1.10$, and then policy (3,1) for $\lambda > 1.10$. Note that policies like (0,1) and (0,2) have total costs per unit that are always higher than those of policies (1,0) and (2,0) respectively. Thus, these policies are completely dominated and should not be used.

For an online retailer who faces an uncertain demand rate for an item (within some bounds), these graphs enable the e-tailer to make an informed decision by showing how the various policies perform over the range of interest. For example, suppose that an e-tailer estimates the system demand rate for an item to be in the range [0.3, 0.6]. Then the policy (1,1), which is fairly robust in this range, might be preferred to the policies (1,0) and (2,1) even though it is not on the efficient frontier.

We also compare the performance of a centralized order fulfillment policy with that of a policy that does not allow transshipments across regions. Consider a two-warehouse system where 70% of the demand comes from the first region and 30% from the second. The system demand rate is given by λ=0.25 orders/ week, and the replenishment lead-time, L=2 weeks. The cost of transportation within a local region is known, as are the holding and backorder costs. We consider seven inventory configurations for the system – (1, 1), (2, 1), (3, 1), (2, 2) and (3, 2) where (a, b) indicates the amount of inventory held in each warehouse. Table 15 illustrates the trade-offs in this example.

LT	2
λ	0.25
a1	0.7
a2	0.3

Trpt Cost	To R1	To R2
From R1	3	
From R2		3

H	0.1
B	1

Scenario #	Inventory Config	AvgInv /period R1	AvgInv /period R2	Total Holding Cost /period	Fill Rate R1	Fill Rate R2	Total Backorder Cost/period	Transportation Cost / Period R1	Transportation Cost / Period R2	Total Cost /period	Total Cost /unit
1	(1,1)	0.705	0.861	0.157	0.705	0.861	0.062	0.525	0.225	0.969	3.875
2	(2,1)	1.656	0.861	0.252	0.951	0.861	0.019	0.525	0.225	1.021	4.083
3	(3,1)	2.651	0.861	0.351	0.994	0.861	0.011	0.525	0.225	1.113	4.450
4	(2,2)	1.656	1.119	0.277	0.951	0.990	0.009	0.525	0.225	1.037	4.147
5	(3,2)	2.651	1.119	0.377	0.994	0.990	0.002	0.525	0.225	1.129	4.515

TABLE 19: TOTAL COST PER UNIT FOR VARIOUS INVENTORY STOCKING POLICIES WITHOUT TRANSSHIPMENT

This policy not only has poorer service in terms of longer waiting times for customers and lower fill rates, but also has higher inventory holding costs when compared to a policy that allows transshipments (Table 13). Only the transportation cost is lower. This suggests that when transshipment cost is relatively much higher than holding and backorder costs, then this policy will perform better than the policy that allows transshipments.

4.2 ORDER FULFILLMENT POLICIES

We have only considered the order fulfillment policy, P_1, in the previous cases due to its computational tractability. However, we find that other order fulfillment policies can perform better than P_1 in terms of reducing the transshipments in the system, without affecting the average system inventory and backorders.

In this section, we consider two other types of order fulfillment policies that have the same rules as stated in A-6 and A-7, but now with a different fulfillment policy when a customer arrives and its closest warehouse does not have inventory on-hand. When one or more of the other warehouses do have inventory on hand, then a warehouse with an on-hand unit is chosen by a certain rule to fill the order.

In the first type of policy, say P_3, the warehouse is randomly chosen but with higher probabilities for warehouses facing lower local demand rates. In the second type of policy, say P_4, the warehouse is chosen from a priority list that orders the warehouses according to their local demand rates, with lower local demand having a higher priority.

We perform a computational study of the proportion of orders transshipped in the system under the order fulfillment policies, P_1, P_3 and P_4. We vary λ while keeping L=3 constant, in order to study the performance under different fill-rates. We also compare performance for different system sizes (3-units 3-locations to 5-units 5-locations), and also under different demand distributions among the regions ranging from almost balanced to extremely unbalanced. We compare the mean value and coefficient of variation of transshipment obtained for each scenario over 100 Monte-Carlo simulation runs of 500,000 orders each, for the order fulfillment policies, P_1, P_3 and P_4. Results from simulation indicate that the strict priority policy P_4 has the best system-wide performance in terms of service failure and transshipments. However, there was not a significant difference in system performance between these policies. A sample set of simulation results for transshipments in the 3-unit 3-location case is presented below.

TABLE 20: TRANSSHIPMENT PERCENTAGE UNDER VARIOUS SYSTEM CONFIGURATIONS AND POLICIES

Fill Rate	**Policies / Parameters** **Demand Rate and Distribution across Regions**	**P1** Mean	**P1** CV	**P3** Mean	**P3** CV	**P4** Mean	**P4** CV
96%	Λ=0.25 (0.65,0.3,0.05)	0.2906	0.0021	0.2815	0.0019	0.2787	0.0020
42%	λ=1 (0.65,0.3,0.05)	0.6012	0.0010	0.5956	0.0011	0.5940	0.0010
96%	λ=0.25 (0.367,0.333,0.3)	0.2344	0.0027	0.2344	0.0028	0.2331	0.0028
42%	Λ=1 (0.367,0.333,0.3)	0.5839	0.0012	0.5838	0.0012	0.5832	0.0014
96%	λ=0.25 (0.3433,0.3333,0.3233)	0.2336	0.0028	0.2337	0.0027	0.2334	0.0026

Statistics taken over 100 simulation runs of 500k orders

Chapter 5

Conclusion

In this dissertation, we present an inventory-planning problem for low demand items motivated by the customer fulfillment process in online retailing. This problem underscores the variety of issues that are particularly important in managing an efficient supply chain in online retailing. In particular, we show how analytical tools can assist in this complex decision making process – strategic, tactical or operational.

In chapter 2, we explain the assumptions and key performance metrics of the online retailer model with centralized order fulfillment. We also study the special case of when the system comprises only two warehouses with each warehouse holding exactly one unit of inventory of an item, and explain how Xu [Xu05] modeled this as a two-state markov chain. Unfortunately, this approach could not be extended to larger systems of three or more warehouses as the state space explodes. In chapter 3, we present methods to calculate the transshipment performance metric across regions under special cases of demand distribution. We then develop an approximation for this metric in the case of general demand distribution across the regions for N-units N-warehouses. Finally, we extend this method to estimate the transshipment in a system of k-units and N-warehouses where each warehouse holds one or more units of inventory. Comparing the performance of this approximation against Monte-Carlo simulation results indicate that these are good estimates. In chapter 4, we use these performance metrics to develop guidelines for optimal inventory stocking for online retailers. Given regional demand rates and the replenishment lead-time, we developed methods to determine how much inventory should be held in the system and how should this be distributed among the warehouses. Finally, we studied various order fulfillment policies and found the best order fulfillment policy to service demand under specific conditions.

So far we have examined inventory planning for low demand items individually. However, customers often place orders for more than one item. Then the overall cost of shipping is lower for the online retailer if shipping can be consolidated for the various items in the order into as few shipments as possible. This will require investigation into the correlations between such low demand items being ordered together, and then the development of an aggregate inventory-planning model that takes into account multi-item orders.

Another key issue of interest is whether the demand for low-demand items, as seen by the online retailer, actually follows a Poisson process. This will involve testing actual order data from an online retailer. Another issue to examine is how well the methods developed in this paper perform when the demand distribution is not Poisson.

Bibliography

[AST97] T.W. Archibald, S.A.E. Sassen, and L.C. Thomas. An optimal policy for a two depot inventory problem with stock transfer. Management Science, 43(2):173–183, 1997.

[AV99] P. Alfredsson and J. Verrijdt. Modeling emergency supply flexibility in a two-echelon inventory system. Management Science, 45(10):1416–1431, 1999.

[Axs90] S. Axsater. Modeling emergency lateral transshipments in inventory systems. Management Science, 36(11):1329–1338, 1990.

[Axs03] S. Axsater. A new decision rule for lateral transshipments in inventory systems. Management Science, 49(9):1168–1179, 2003.

[Dad92] M. Dada. A two-echelon iventory system with priority shipments. Management Science, 38(8):1140–1153, 1992.

[Das75] C. Das. Supply and redistribution rules for two-location inventory systems: One period analysis. Management Science, 21:765–776, 1975

[Gal96] R.G. Gallager. Discrete Stochastic Processes. pp. 57-81, Kluwer Academic Publishers, 1996

[GC01] J. Grahovac and A. Chakravarty. Sharing and Lateral Transshipment of Inventory in a Supply Chain with Expensive Low-Demand Items. Management Science, 47(4):579–594, 2001.

[Gro63] D. Gross. Centralized inventory control in multilocation supply systems. In *Multistage Inventory Model and Techniques*. Stanford University Press, 1963.

[KP77] U.S. Karmarkar and N.R. Patel. The one-period n-location distribution problem. Navel Research Logistics Quarterly, 24:559–575, 1977.

[KR65] K.S. Krishnan and V.R.K. Rao. Inventory control in n warehouses. Journal of Industrial Engineering, 16:212–215, 1965

[Lee87] H. Lee. A multi-echelon inventory model for repairable items with emergency lateral transshipments. Management Science, 33(10):1302–1316, 1987.

[Rob90] L.W. Robinson. Optimal and approximate policies in multiperiod, multilocation inventory models with transshipments. Operations Research, 38(2):278–295, 1990.

[TC92] G. Tagaras and M. Cohen. Pooling in two-location inventory systems with non-negligible replenishment lead times. Management Science, 38(8):1067–1083, 1992.

[Xu05] P. Xu, "Order Fulfillment in Online Retailing: What goes where," *Doctoral Dissertation at MIT*, pp. 89–125, 2005.