Order Fulfillment in Online Retailing: What Goes Where

by

Ping Josephine Xu

M.S. Operations Research
Massachusetts Institute of Technology, 2003

B.S. Industrial Engineering and Management Sciences
Northwestern University, 1999

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2005

Author .....
Sloan School of Management
August 11, 2005

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

Accepted by .....
James Orlin
Edward Pennell Brooks Professor of Operations Research
Co-director, Operations Research Center

Order Fulfillment in Online Retailing:
What Goes Where

by

Ping Josephine Xu

Submitted to the Sloan School of Management
on August 11, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

We present three problems motivated by order fulfillment in online retailing.

First, we focus on one warehouse or fulfillment center. To optimize the storage space and labor, an e-tailer splits the warehouse into two regions with different storage densities. One is for picking customer orders and the other to hold a reserve stock that replenishes the picking area. Consequently, the warehouse is a two-stage serial system. We investigate an inventory system where demand is stochastic by minimizing the total expected inventory-related costs subject to a space constraint. We develop an approximate model for a periodic review, nested ordering policy. Furthermore, we extend the formulation to account for shipping delays and advance order information. We report on tests of the model with data from a major e-tailer.

Second, we focus on the entire network of warehouses and customers. When a customer order occurs, the e-tailer assigns the order to one or more of its warehouses and/or drop-shippers, so as to minimize procurement and transportation costs, based on the available current information. However, this assignment is necessarily myopic as it cannot account for any subsequent customer orders or future inventory replenishments. We examine the benefits from periodically re-evaluating these real-time assignments. We construct near-optimal heuristics for the re-assignment for a large set of customer orders by minimizing the total number of shipments. Finally, we present saving opportunities by testing the heuristics on order data from a major e-tailer.

Third, we focus on the inventory allocation among warehouses for low-demand SKUs. A large e-tailer 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 e-tailer has the technological capability to manage and control the inventory globally: all warehouses act as one to serve the global demand simultaneously. The e-tailer will utilize its entire inventory, regardless of location, to serve demand. Given we stock certain units of system inventory, we allocate inventory to warehouses by minimizing outbound transportation costs. We analyze a few simple cases and present a methodology for more general problems.

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

To the memory of my father ('48 – '01) and joan connor ('33 – '05)

Acknowledgments

First and foremost, I would like to thank my advisor Professor Steve Graves for the opportunity to work with him. I truly appreciate his ability and willingness to discuss research at a big-picture level as well as at the very nitty-gritty level. I appreciate his patience and wisdom at key points in the process. I would also like to thank Dr. Russell Allgor for his great input. The many valuable meetings with him allowed this research to be of practical value. To my committee members Professor Cindy Barnhart and Jeremie Gallien, I am grateful for their interest, effort, and helpful comments. I also thank Professor Jim Orlin for his ideas and comments. Thanks to Russell and his team for providing the real-world data, and him and Dan Stratila for providing the test data benchmark solutions in Chapter 3. Finally, I thank the Singapore-MIT Alliance (SMA) for their generous financial support of this research.

I will miss my time at MIT and Cambridge. MIT is a great place, filled with so much intelletual energy and research dynamics. Cambridge is just awesome, what can I say. I thank my friends at the Operations Research Center and Sloan Operations Management group. Not only are they the best people to turn to for research, academic, or Latex questions, they've shown me what it meant to be the best and the brightest, in work and life. I thank the ORC administrators for making it such a great place. Thanks to my two roommates for the past three years. I doubt I could ever find better roommates to match with you two. Finally, I thank my friends in the Bishop Allen Drive Coop and their satellites. Thanks for all the delicious dinners, fun times, and support.

To my mother, I thank her for always thinking about me and for letting me leave her so that I may spread my own wings. I thank Ben for always being there for me at good and bad times. At last, I thank my father for instilling in me the love of math and value of creativity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the decade since the Dotcom boom, many online retailers or e-tailers have come of age. The existence and growth of these companies pose new challenges to efficient supply chain management. While their market segmentations, operational scales, and supply chain structures may differ, they share some common characteristics.

Large scale Unlimited by physical space in the store front, e-tailers often pride themselves in having a universal selection of products and in providing a very customer-friendly shopping experience. As a result, online retailers and/or their drop-shippers have very large scale operations with hundreds of thousands of Stock-Keeping Units (SKUs) in stock. The sheer size of the operations and catalogs poses a challenge to sound decision making.

Logistics as a matter of trust Brynjolfsson and Smith [BS00] conclude that trust is one of the major criteria that customers use to evaluate online retailers. Moreover, empirical research by Keeney [Kee99], Torkzadeh and Dhillon [TD02] shows that customers consider the timely delivery of products to be a significant component of trust. Evidently, the reliability and efficiency of the supply chain is crucial in online retailing.

High visibility With strong information technology capabilities, these companies can collect virtually any data from the time customers start to browse on the website to the time customers receive their orders. With this ability to collect an overwhelming amount of data, making sound data-driven decisions requires sophisticated tools.

This availability of information also raises questions about how e-tailers should share real-time information with their customers or suppliers. For instance, e-tailers may want to display real-time inventory availability on their websites. However, this information may decrease demand at the time of an inventory shortage.

Assemble-to-order system Unlike bricks-and-mortar retailing, a major element of the online retailing operations is an assemble-to-order system. Some customers order

multiple items. Online retailers profit from bundling items by sending one shipment to customers. The components of the assembly are the items in a customer order, and the final product is an individual customer order. Compared with other assemble-to-order systems, such as Dell, the assembly process is much simpler. The total number of possible final products, however, explodes because the number of items offered is large and customers can and will order any combination of items. The challenges in general assemble-to-order systems still exist here: coordinating all items in an order to be packed around the same time and allocating items among orders. A cluster of literature on assemble-to-order flourished in the recent few years. Song and Zipkin offer a comprehensive survey of general assemble-to-order systems in 2003 [SZ03].

Delay in demand fulfillment Unlike in physical stores, in online retailing there is typically a time delay between when a demand occurs and when inventory is consumed or deployed to meet a customer order. At the time a demand occurs, the online retailer and the customer reach an agreement on all aspects of the transaction. The completion of the transaction, or the time at which the customer receives the products, however, may take place days after the initial transaction. Consequently, e-tailers can exploit this order-to-delivery time window to reduce their costs. That is, by delaying the decisions on inventory allocation, shipping methods, etc., e-tailers can utilize more resources and information to make better decisions. Furthermore, e-tailers can employ pricing schemes to entice certain customers to give an even longer order-to-delivery time window.

This delay in demand fulfillment is in some ways analogous to having advance order information, which has recently received attention in the research literature. For instance, Hariharan and Zipkin [HZ95] show that advance customer orders improve system performance the same way that replenishment leadtimes degrade it. Chen [Che01] develops an optimal pricing-replenishment policy for different classes of customers with different shipping preferences. Gallego and Ozer [GO01, GO03] show an optimal inventory stocking policy for stochastic systems with advance demand information.

Retailer-directed demand allocation In contrast to large retail chains, online retailers have only one storefront, namely their web portal. In bricks-and-mortar retailing, demand occurs at specific stores; a customer picks a store to visit and expects his/her demand to be served at that store. In online retailing, customers cannot control how their demand will be served. Rather, the e-tailer will decide which warehouse or drop-shipper serves what demand. As a consequence, the e-tailer can utilize all of its warehouses or fulfillment centers to serve the customer demand. This centralized demand allocation poses new challenges and opportunities to minimize operating costs. For instance, when should e-tailers use drop-shipper or in-house fulfillment? Which

warehouse should fulfill which demand? What subset of SKUs should be stocked in each warehouse?

Netessine and Rudi [NR04b] analyze a game theoretical model involving a retailer and a wholesaler, where either the retailer carries inventory or the wholesaler carries inventory and the retailer drop-ships. They find both models to be system sub-optimal. They extend the models to include multiple retailers [NR04a], and examine a dual strategy of the retailer carrying inventory as well as drop-shipping. They find that the two options have the potential to be Pareto optimal: retailer drop-ships only or adopts the dual strategy.

In summary, in this increasingly competitive online marketplace with few barriers to entry, the success or dominance of an e-tailer will depend on building an efficient customer fulfillment process. This challenge can be seen as an opportunity to rethink the current assumptions and to extend the current models in the literature. Even though hundreds of billions of dollars of goods are sold on the Internet, there has been relatively little research focusing on the issues particular to e-tailing supply chain management. In addition to the ones mentioned above, some others include an overview of models in e-business by Swaminathan Tayur [ST03] and a survey of research papers and case studies by Johnson and Whang [JW02]. In this thesis, we present three problems motivated by the online retailing fulfillment process. While these problems only cover a few of the important issues in fulfillment, we believe they provide a glimpse of the variety of problems in online retailing. In particular, we show how analytical tools can assist in this complex decision making process, tactical or operational.

In Chapter 2, we focus on a single warehouse or fulfillment center. After a customer orders online, the e-tailer assigns the order virtually to one of its order fulfillment centers. An order fulfillment center is a large warehouse that might store several hundreds of thousands of SKUs in a floor space of several hundred thousand square feet. The key objectives of such a warehouse are to achieve a high utilization of its storage space and, at the same time, be able to fill orders quickly and reliably with the least amount of effort.

To optimize storage space and labor, an e-tailer splits the warehouse into two storage regions with different storage densities. One region is for picking customer orders and the other holds reserve stock. The picking area is laid out to facilitate efficient picking by a person; this limits the height and depth of the storage racks, as well as the quantity of each SKU stored in the picking area. As a consequence, the storage density in the picking area is relatively low. On the other hand, the reserve or deep-storage area has high storage density; the purpose of the reserve area is to store larger quantities with the most efficient use of space. Replenishment from outside suppliers will typically come in pallet loads and be first stored in reserve storage. The inventory in reserve storage is then used to replenish the picking area, usually in smaller quantities like cases or cartons. Consequently, the

inventory in the warehouse flows in a serial, two-stage fashion. We investigate a multi-item inventory problem for a two-stage serial system where demand is stochastic and the objective is to minimize the expected long-run average cost under space constraints. We derive an approximate formulation for the serial two-stage model, and we generate a solution procedure for computing multi-item periodic-review ordering policy under space constraints. Moreover, we model to account for advance order information and shipping delays. Finally, we report tests of the model on real data and the resulting managerial insights.

In Chapter 3, we focus on the entire network of warehouses and customers. When a customer places an order on an e-tailer's website, the e-tailer, in real time, searches for available fulfillment options from its order fulfillment centers or drop-shippers. The e-tailer assigns the order to one or more warehouses virtually, mainly based on the transportation cost of shipping the order from the warehouse(s) to the customer location and on the current warehouse inventory availability. Depending on the inventory availability and customer preferences, the e-tailer then quotes a promise-to-ship date to the customer. The promise-to-ship date is the date by which the e-tailer promises to ship the order from the warehouse(s). After the e-tailer assigns the order, the order enters the picking queue at the warehouse. The order might wait six to eighteen hours before the items in the order are picked and assembled into a shipment that is then given to a third party carrier to deliver the package(s) to the customer location.

We show with examples that the real-time decision is necessarily myopic because the e-tailer does not anticipate any future customer orders or inventory replenishment. We can reduce the total transportation cost of shipping orders from warehouses by re-evaluating the real-time assignment decisions, subject to the constraint that there is no violation of the promise-to-sip date commitment for any customer order.

We formulate the re-evaluation problem as a network design problem, and we show that the problem is NP-hard in complexity. By designing simple but effective heuristics, we are able to improve greatly upon the real-time assignments. Our solution can generate large savings to the e-tailer without any revamping of the current systems. Finally, we perform worst-case analysis of the heuristics.

In Chapter 4, we focus on the inventory allocation among warehouses for low-demand SKUs. A large e-tailer strategically stocks inventory for SKUs with low demand. One motivation is to provide a wide range of selections; indeed such SKUs actually constitute a significant portion of the total SKUs. The second incentive, of course, is to provide faster customer fulfillment service. For many of these SKUs, the e-tailer may only stock a handful of inventory units across all warehouses.

Here we focus on the effect of outbound transportation costs on the inventory allocation. We assume that an 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 simultaneously. Specifically, the e-tailer will utilize its entire inventory, regardless of location, to serve demand. Given that we stock certain units of inventory in the system, we allocate inventory to warehouses by minimizing outbound transportation costs from the warehouses to customers.

We propose an inventory planning process for low-demand SKUs. For some simple cases, we find the inventory stocking policy that minimizes the outbound transportation costs. We also present a methodology that is the first step in analyzing the general problems.

Chapter 2

Inventory System in a Serial Warehouse

2.1 Introduction and Motivation

For e-tailers, which operate with no physical stores, the efficient utilization of inventory, storage space, and labor is paramount to achieving high levels of customer service and company profits. After a customer orders online, an e-tailer assigns the order virtually to one of its order fulfillment centers. An order fulfillment center is a large warehouse that might store several hundred thousands of SKUs in a floor space of several hundred thousand square feet. The key objectives of such a warehouse are to achieve a high utilization of its storage space and, at the same time, be able to fill orders quickly and reliably with the least amount of effort. These objectives are often conflicting as efficient space utilization entails high-density storage, whereas efficient order picking requires ready access to the full portfolio of SKUs, which results in low-density storage. Furthermore, to provide reliable service at the minimum cost, an e-tailer needs to hold the right amount of inventory.

To optimize the storage space and labor, an e-tailer splits the warehouse into two storage regions with different storage densities. One region is for picking customer orders and the other to hold reserve stock. The picking area is laid out so as to facilitate efficient picking by a person; this limits the height and depth of the storage racks, as well as the quantity of each SKU stored in the picking area. As a consequence, the storage density in the picking area is relatively low. On the other hand, the reserve or deep-storage area has high storage density; the purpose of the reserve area is to store larger quantities with the most efficient use of space. In the reserve area, most items are stored in pallet loads and moved by fork-lift trucks. The pallets are stored in high-rise storage racks, with a depth of one or two pallets.

Replenishment from outside suppliers will typically come in pallet loads and be first stored in reserve storage. The inventory in reserve storage is then used to replenish the

picking area, usually in smaller quantities like cases or cartons. Consequently, the inventory in the warehouse flows in a serial, two-stage fashion, as illustrated in Figure 2-1. To determine the right amount of inventory entails the optimization of this inventory system. In this research we investigate this multi-item inventory problem for a two-stage serial sys-

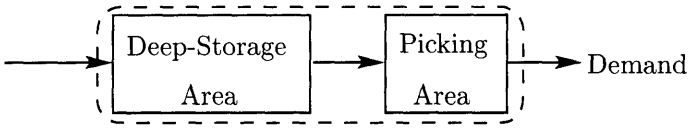

Figure 2-1: A Serial, Two-Stage Warehouse

tem with stochastic demand and the objective is to minimize the expected long-run average cost under space constraints.

Our objective is to support both tactical and strategic decisions in the order fulfillment center. At the tactical level, we intend for the model to guide decisions on the amount of inventory and its deployment between the picking and reserve storage areas, as well as on the replenishment frequency of the picking area. At the strategic level, in designing a warehouse, one needs to decide how much space to allocate to reserve storage versus for picking. We intend to explore how the operational effectiveness of the warehouse depends on the amount of available storage space and its division between picking and reserve storage.

Clark and Scarf [CS60] consider an unconstrained serial inventory system and observe that for the system with set-up costs at each stage, the optimal ordering policy, if one exists, must be extremely complex. This observation remains true today and has driven subsequent research to focus on heuristic policies. There is an extensive literature on the evaluation and analysis of heuristic policies in various multi-echelon, stochastic inventory systems. Here we contribute to this literature by providing an analysis and solution procedure for a periodic review, heuristic ordering policy in a multi-item two-stage problem, from which we can generate insights about the intrinsic trade-offs in a constrained warehouse operation.

The choice of a periodic review ordering policy for this inventory system is motivated by several practical considerations. An e-tailer stocks an extremely wide range of products, and as a consequence, needs to coordinate the replenishment of these products from the various outside suppliers. For instance, the number of products an e-tailer orders from a single supplier is often very large. Employing a periodic review policy reduces the fixed replenishment costs by combining order replenishment for different products from the same supplier. Furthermore, it facilitates the coordination of transportation, and other logistical considerations, since it results in a regular repetitive schedule for replenishments [Rao03].

2.1.1 Literature Review

The field of stochastic multi-echelon inventory systems started in the 1960's with two papers from Clark and Scarf [CS60, CS62]. In the first paper [CS60], they characterize the optimal policy for a single-item, discrete-time, multi-echelon system by solving a finite-horizon dynamic program. The optimal policy is a function of the total on-hand and on-order inventory. The major assumptions of the model are: demand at each echelon is backlogged and the ordering cost is linear, except a fixed ordering cost is permitted at the most upstream stage. In the second paper [CS62], they examine a two-stage example in which the set-up cost appear at both stages. They show that the optimal policy, if exists, may be quiet complex.

A considerable body of research has evolved in the field since the publications by Clark and Scarf. The subsequent research focuses on characterizing optimal policies for more generalized models as well as generating bounds and heuristic policies or approximate models. We can view the literature by three principal inventory control policies: (s, S), (Q, R), and (R, T). The first two policies have been studied extensively, while we employ a less studied (R, T) policy. We discuss the three categories below.

The (s, S) policy, or basestock policy, is optimal in the Clark-Scarf model, which is a discrete-time, periodic review model. At every period, if the inventory position is less than or equal or s, then we order up to S. This model has been studied extensively. Federgruen [Fed93] provides a comprehensive review of this literature from serial to assembly systems. In particular, Federgruen and Zipkin [FZ84c] streamline the Clark-Scarf model as well as extend it to the infinite horizon case. Schmidt and Nahmias [SN85] consider the simplest assembly system of two components, and they characterize the optimal policy assuming all ordering costs are linear. With the same assumption on ordering costs and an additional assumption of the initial inventory levels satisfying certain simple conditions, Rosling [Ros89] shows that an assembly system can be transformed into an equivalent serial system. The basestock policy, therefore, is also optimal for all nodes in the assembly system. Chen and Zheng [CZ94b] show an alternative proof of the above known optimality results for serial and assembly systems.

Most recently, Chen [Che00] derives the optimal policies for serial and assembly system with batch ordering. A basestock policy (modified to accommodate the base order quantities), he suggests, is still optimal in the serial model with every stage ordering in batches. He also shows the transformation from assembly to serial systems with batch ordering. All of the discrete-time models mentioned so far assume that demand in each period is *i.i.d.*. To extend the Clark-Scarf model to nonstationary demand or, specifically, Markov-modulated demand, Chen and Song [CS01] show that the optimal policy is an echelon basestock policy with state-dependent order-up-to levels. Taking a step further, Muharremoglu and Tsitsiklis [MT03] extend the results of Chen and Song to finite horizon and infinite horizon

discounted cost problems with stochastic leadtimes.

Some examples of heuristic policies include Eppen and Schrage [ES81], Federgruen and Zipkin [FZ84b, FZ84a], Jackson [Jac88]. The area of constructing effective bounds has also been very active recently. To name a few, we have Gallego and Zipkin [GZ99], Zipkin [Zip00], and Dong and Lee [DL03]. Shang and Song [SS03] generate newsvendor-type lower and upper bounds on the optimal echelon stocking policies of a serial infinite horizon problem. They devise a simple heuristics that is within 1.5% of the optimal.

The (Q, R) policy is a continuous review policy. Inventory positions are monitored continuously and we order Q units whenever the inventory position is below R. The (Q, R) policy is also well studied and optimal for the continuous time case. Axsater [Axs93a] reviews the literature on continuous review policies for multi-echelon, stochastic systems. First, there is the literature on one-for-one replenishment policies or $(S-1, S)$ policy in a one-warehouse-n-retailer setting. All locations have $(S-1, S)$ policies. When an item at a retailer or local site fails, it is sent to the warehouse to repair. At the same time, a local inventory unit replaces the failed item and the local site orders a replenishment from the warehouse. The METRIC approximation by Sherbrooke [She68] assumes that replenishment leadtimes for the local sites are independent. It then led to modeling the number of outstanding orders at a local site as a Poisson random variable, which is completely characterized by its mean. Therefore, the METRIC approximation is a single parameter approximation. Graves [Gra85] determines the mean and the variance of the number of outstanding orders, and, therefore characterizes a two-parameter approximation. Muckstadt [Muc73] extends the METRIC model to indentured parts (MODMETRIC). Sherbrooke [She86] extends Grave's approximation to the multi-indenture case. Lee [Lee87] considers the same framework with the added dimension of lateral transshipment: the retailers are grouped such that retailers among the same group can transship among each other. Later, Svoronos and Zipkin [SZ91] allow stochastic leadtimes but preserve the order sequence. Recently, Wang, Cohen and Zheng [WCZ00] relax the *i.i.d.* assumption on depot replenishment leadtimes and allow the depot replenishment leadtimes to depend on the local site.

A large cluster of literature also exists on general (Q, R) or batch-ordering policies. Some examples of heuristic policies include De Bodt and Graves [DG85], Deuermeyer and Schwarz [DS81], Svoronos and Zipkin [SZ88], Axsater [Axs90b], and Chen and Zheng [CZ94a]. Closest to our model in spirit is the De Bodt & Graves [DG85] paper. They develop a similar two-stage serial model for a continuous review (Q, R) policy. They provide approximate performance measures under a nested policy assumption: whenever a stage receives a shipment, a batch must be immediately sent down to its downstream stage. They do not make an assumption about the form of the demand distribution. We, however, consider a periodic review (R, T) policy. There has been some progress to establish near-optimal heuristic

policies with a guaranteed, worst-case performance. Chen [Che99] characterizes a continuous review heuristic policy for a two-stage inventory system. The long-run average cost is guaranteed to be within 6% of optimality, where demand is Poisson, leadtime at stage 2 is zero, and both stages incur a fixed order cost.

There are some exact results for continuous review models. Axsater [Axs93b] provides exact cost results for a two-echelon system with one central warehouse and multiple identical retailers. He assume that leadtimes are constant and the retailers face independent Poisson demand. Later in 2000, he generalizes the solution to Compound Poisson demand and nonidentical retailers [Axs00]. Chen and Zheng [CZ97] provide exact results where the central warehouse uses echelon stock reorder point policies. Cheung and Hausman [CH00] show the exact results for the central warehouse where the retailers are nonidentical.

As we have shown, because of their optimality, both (s, S) and (Q, R) policies have been extensively studies. In comparison, the (R, T) policy does not receive as much attention. Hadley and Whitin [HW63] and Naddor [Nad82] have studied the (R, T) policy in their books. Graves [Gra96] analyzes a multi-echelon system with general system topology. He assumes that at each location there is a schedule of preset replenishment times, and argues that such scheduled shipments are common in practice to utilize the transportation resources efficiently. Most recently, Rao [Rao03] analyzed the properties of the single-stage (R, T) model, as a counterpart of Roundy [Rou86] and Zheng [Zhe92] for a deterministic periodic review model and stochastic (Q, R) model, but with certain demand function restrictions. In the extension, he develops a two-stage serial system which is similar to our model but has different assumptions on the interaction between echelons.

In section two, we review the single-stage periodic review model and its most recent results, and present the two-stage serial model. In section three, we show numerical results. In section four, we test the solution procedure on real data sets. In section five, we introduce an extension that accounts for shipping delays and advance demand information.

2.2 Model Formulation and Solution Approach

We first present the key assumptions in the single-stage and two-stage models, while additional assumptions apply only to the two-stage model will be introduced later.

A-1 The demand process is stationary for the relevant time horizon.

A-2 Each stage has a constant known nonzero lead time.

A-3 When on-hand inventory at stage 1 is depleted, demand at stage 1 is backlogged and a penalty cost per backorder is charged.

A-4 Backorder costs are high. As a result, demand backorder quantities are small. We will provide more details on this assumption later in the section.

A-5 Each echelon follows a periodic-review (R, T) policy, where R is the order-up-to level and T is the review period.

We assume stationary demand in A-1, whereas in the e-tailing setting, there are usually two distinct demand patterns for the off-peak and peak season. Within each season, it is reasonable to assume stationary demand. We can treat the two seasons as two separate models. We can relax A-2 to allow stochastic lead times, and will comment on this extension later in the paper. Our assumption on the cost penalty is more applicable when the fixed cost component of backorder is much larger than the time variable component, as is the case in e-tailing. We note that the formulation under our backorder cost assumption may be less convenient for theoretical analysis, but it is easier in computation.

To facilitate the discussion, we list the following standard definitions:

$I(t)$	on-hand inventory or inventory physically in the warehouse at time t,
$B(t)$	amount of unfulfilled customer demand at t,
$IL(t)$	inventory level or net inventory at t,
	equivalent to $I(t) - B(t)$,
$O(t)$	amount of on-order or inventory in transit to warehouse at t,
$IP(t)$	inventory position at t,
	equivalent to $IL(t) + O(t)$.

Following the literature convention, we denote stage 1 as the downstream stage that serves external demand, and stage 2 as the upstream stage that replenishes stage 1 and is replenished by outside suppliers. In the e-tailing setting, stage 1 is the picking area and stage 2 is the deep storage area. We state the control policy in terms of *echelon stock*, which is the total inventory in the current stage and all its downstream stages.

We first review the single-item single-stage periodic-review (R, T) model [HW63, p. 237-245]. Extending the single-stage model, we then present the two-stage model, an unconstrained single-item serial model.

2.2.1 Single-Stage Model Review

We denote

$C(\cdot)$	expected total cost per unit time,
l	replenishment lead time,
d	expected demand per unit time,
a	fixed order, or replenishment, cost,

h	holding cost per item per unit time,
b	backorder cost per item,
$f(x\|l)$	probability density function of demand over a time interval of length l nominally corresponding to the lead time demand.

We assume that discrete units of inventory can be approximated by continuous quantities. We follow the inventory literature [e.g., HW63, p. 237-245], and approximate the expected total cost per unit time as:

$$C(R,T) \cong \frac{a}{T} + h\left(R - d(l + \frac{T}{2})\right) + \frac{b}{T}\int_R^\infty (x-R)f(x|l+T)dx \tag{2.1}$$

We compare Equation (2.1) with the exact model for Poisson demand in Appendix A.1, and we observe that the main approximation is the underestimation of the holding-cost term. Figure (2-2) is an inventory diagram for the single-stage model. The x-axis represents time,

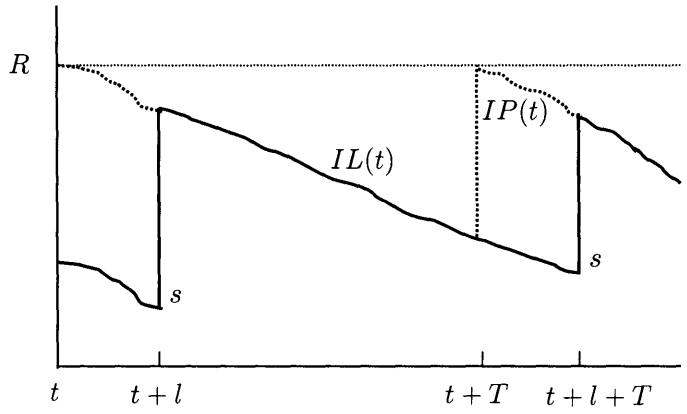

Figure 2-2: Single-Stage Inventory Level and Position

and the y-axis represents inventory amount. The solid line indicates inventory level, and the dotted line indicates inventory position when it differs from the inventory level. The time between $[t+l,\ t+l+T]$ is a typical replenishment cycle, assuming we receive a replenishment at $t + l$. The holding-cost term in Equation (2.1) is equivalent to $\frac{h}{T}\int_l^{l+T} E[IL(t)]\ \mathrm{d}t$, whereas in the exact model it is $\frac{h}{T}\int_l^{l+T} E[I(t)]\ \mathrm{d}t$. That is, we approximate on-hand inventory with net inventory. This error is small, however, when backorder is small and infrequent as assumed in 4, thus on-hand inventory is nearly the same as net inventory. The backorder term is slightly overestimated, since in Equation (2.1) we assume that we start each replenishment cycle with zero backorders.

For a given value of T, $C(R,T)$ in Equation (2.1) is convex in R. We can obtain the

optimal value of R for a given value of T:

$$\int_R^\infty f(x|l+T)\,\mathrm{d}x = \frac{hT}{b}, \qquad T \le \frac{b}{h} \tag{2.2}$$
$$R = 0, \qquad T > \frac{b}{h}$$

Given a value of R, $C(R,T)$ is not convex in T.

We search over values of T in the range $(0, \frac{b}{h})$, and use Equation (2.2) to find the best choice of R for a given value of T, and Equation (2.1) to determine the minimal value of expected total cost.

2.2.2 A Two-Stage Serial Model

Now we consider our approximate two-stage serial (R,T) model based on §2.2.1. We use the subscript 2 to indicate echelon stock in IL, I, IP, whereas we use the subscript 2 to indicate stage 2 inventory on-order in O. We define $IL_2(t)$, the echelon inventory level at time t for stage 2, by:

$$IL_2(t) = I_2(t) - B(t),$$

where $I_2(t)$ is the echelon inventory at stage 2, which is the sum of on-hand inventory at stage 2, on-hand inventory at stage 1, and inventory in transit from stage 1 to stage 2. Similarly, we define the echelon inventory position for stage 2 as

$$IP_2(t) = IL_2(t) + O_2(t),$$

where $O_2(t)$ is the inventory amount in transit from outside suppliers to the warehouse.

Here we first present the remaining assumptions of the two-stage model.

A-6 To coordinate the replenishment of both stages, we impose a constraint on the review periods of both echelon, $T_2 = nT_1$, where n is a positive integer. Furthermore, the ordering policies are time-phased so that stage 1 places a replenishment order when stage 2 receives its replenishment.

This assumption results in a periodic-review version of a nested policy [e.g., Lov72, WCW73]: whenever a stage reorders, its downstream stages also reorder. Whereas this assumption on the ordering policy simplifies analysis, it is not unreasonable in our periodic-review context where one would desire to coordinate the less-frequent replenishment of the deep-storage area with that of the picking area.

We demonstrate the policy behavior through an example in Figure (2-3) for $n = 3$. In this figure, echelon 2 orders at time 0 and T_2, and receives its replenishments l_2 leadtime later, at time l_2 and $l_2 + T_2$. Echelon 1 orders at time l_2, $l_2 + T_1$, and $l_2 + 2T_1$, and receives

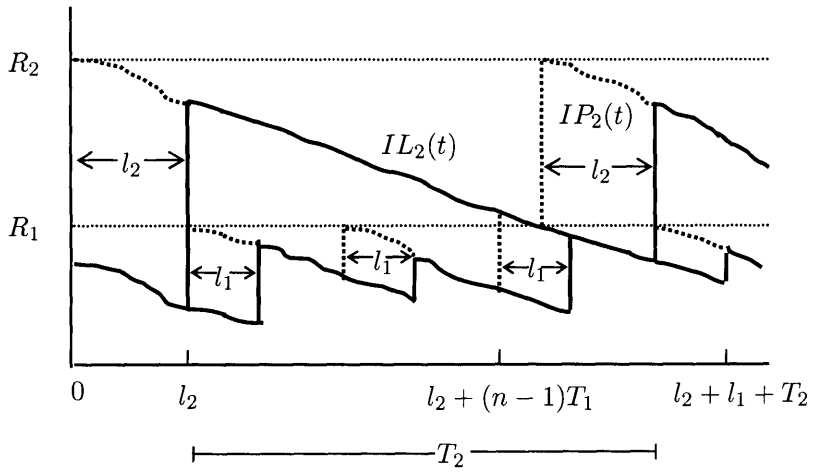

Figure 2-3: A Two-Stage Inventory Diagram for $n = 3$

its replenishment l_1 leadtime later, at time $l_2 + l_1$, $l_2 + l_1 + T_1$, and $l_2 + l_1 + 2T_1$. Since $n = 3$, there are three inventory replenishment (reviews) of echelon 1 for each replenishment of echelon-2 inventory. We define an order cycle of length T_2 as the time between consecutive echelon-2 inventory replenishments, such as $[l_2, l_2 + T_2]$.

A-7 At the last replenishment of echelon 1 in each cycle, such as at $t = l_2 + (n-1)T_1$, stage 1 orders *all* of the remaining on-hand inventory from stage 2. That is, as in Figure (2-3), the inventory position of stage 1 merges with the inventory level of echelon 2 at the last replenishment in a cycle, $IP_1(l_2 + (n-1)T_1) = IL_2(l_2 + (n-1)T_1)$.

We call the last echelon-1 replenishment in a cycle an *exhaustive replenishment*, and the other $(n-1)$ echelon-1 replenishment *normal replenishment*. We have $(n-1)$ normal replenishments for every exhaustive replenishment in this ordering policy. There are two reasons for this assumption. **i)** Holding inventory in reserve at the last replenishment of the picking area in an order cycle has very limited value. One might as well put everything into the picking area, given that a replenishment of the reserve area will arrive soon. Furthermore, we expect that the "extra stock" that is moved from reserve into the picking area, if any, should be quite small on average. If the extra stock is large, then we could decrease the value of R_2. **ii)** This assumption simplifies the analysis of the model, as we avoid having to keep track of the remnant inventory left in reserve at the end of each order cycle. With this assumption, inventory in reserve is depleted when an order arrives from the outside supplier.

To ensure the $(n-1)$ normal replenishment of stage 1 are well-behaved, or, they order up to R_1 at the start of the cycle, we make two additional assumptions. We state both assumptions for an order cycle in which stage 2 orders at time $t = 0$, as depicted in Figure (2-3).

A-8 $IL_2(l_2 + (n-2)T_1) > R_1$.

This assumption ensures that echelon 2 has sufficient inventory to raise the stage 1 inventory position to R_1 for every normal replenishment. Since $IL_2(\tau)$ is nonincreasing in an order cycle for $l_2 \leq \tau \leq l_2 + T_2$, this assumption also implies that echelon-2 inventory level is greater than R_1 for the first $(n-1)$ replenishment in a cycle, $IL_2(l_2 + mT_1) > R_1$, $\forall\ 0 \leq m \leq n-2$. However, we make no assumption on the inventory level at the exhaustive replenishment, such as on $IL_2(l_2 + (n-1)T_1)$.

A-9 $IP_1^-(l_2) = IL_2^-(l_2) \leq R_1$.

We denote $IP^-(t), IL^-(t)$ as the inventory positive or level before an event occurs at time t. Due to A-7, the echelon inventory level in stage 2 is equivalent to the inventory position in stage 1 just before the shipment from the outside supplier arrives at time l_2. We assume that this inventory level is less than or equal to R_1, so that the inventory position at stage 1 can be returned to its order-up-to point R_1 at time l_2, but not more than R_1.

We denote $D(t, t+\tau)$ as the total demand from time t to $t+\tau$. If echelon 2 orders up to R_2 at time t, then for A-8 to be valid, we must have that demand during $l_2 + (n-2)T_1$ is no more than $R_2 - R_1$:

$$D(t, t + l_2 + (n-2)T_1) \leq R_2 - R_1. \tag{2.3}$$

For A-9 to be valid, we must have that demand during $l_2 + T_2$ time period is greater than $R_2 - R_1$:

$$D(t, t + l_2 + T_2) \geq R_2 - R_1. \tag{2.4}$$

We expect that the accuracy of our cost expressions will depend on the probability that the above two equations hold true. We argue here that in the e-tailing setting, these probabilities should be quiet high. The order-up-to point for the reserve area, R_2, needs to cover demand over an interval of length $T_2 + l_1 + l_2$; the order-up-to point for the picking area, R_1, needs to cover demand over the interval of length $T_1 + l_1$. Thus, we expect their difference $R_2 - R_1$ to be roughly the expected demand over an interval of length $(n-1)T_1 + l_2$, and Equation (2.3) and (2.4) are likely to be true as long as the leadtime l_2 is not too large relative to the review periods T_1 and T_2.

To develop the cost expressions, we derive the cost elements separately. The expected fixed order cost per unit time is:

$$\frac{a_1}{T_1} + \frac{a_2}{T_2}. \tag{2.5}$$

Since by assumptions, we order after every review period.

To derive the holding cost element, we examine echelon 1 and 2 separately. We approximate the echelon 2 holding cost as in the single-stage model, $\frac{h_2}{T_2}\int_{l_2}^{l_2+T_2} E[IL_2(t)]\,\mathrm{d}t$, that

is

$$h_2\left(R_2 - d\left(l_2 + T_2/2\right)\right). \tag{2.6}$$

The holding cost for echelon 1 needs more discussion. For the $(n-1)$ normal replenishment cycles, we can approximate the holding cost for each cycle just as in the single-stage model. Here, we use assumptions A-8 and A-9 to ensure that the inventory position for stage 1 is exactly the order-up-to point, R_1, at the start of each of the $(n-1)$ normal replenishment cycles. The expected inventory level for an exhaustive replenishment cycle, however, requires a slightly different development. In Figure (2-3), the time during $[l_2 + l_1 + (n-1)T_1, l_2 + l_1 + T_2]$ is an exhaustive replenishment cycle for stage 1. The inventory level at the start of the cycle is $R_2 - D\left(0, l_2 + l_1 + (n-1)T_1\right)$. The inventory level at the end of the cycle is $R_2 - D\left(0, l_2 + l_1 + T_2\right)$. The average net inventory in the cycle is, therefore,

$$\frac{1}{T_1}\int_{l_1+l_2+(n-1)T_1}^{l_1+l_2+T_2} E[IL_1(t)]\,\mathrm{d}t = R_2 - d\left(l_1 + l_2 + T_2 - \frac{T_1}{2}\right),$$

which we will use as an approximation for the on-hand inventory at stage 1 in an exhaustive replenishment cycle. We can then write the holding cost at stage 1 as

$$h_1\left(\frac{n-1}{n}\big(R_1 - d(l_1 + T_1/2)\big) + \frac{1}{n}\big(R_2 - d(l_1 + l_2 + T_2 - T_1/2)\big)\right) \tag{2.7}$$

Similarly, we derive the backorder costs for normal and exhaustive replenishment separately. The expected number of backorders during each normal replenishment cycle is $\int_{R_1}^{\infty}(x - R_1)f(x|T_1 + l_1)dx$. The expected number of backorders during an exhaustive replenishment cycle is $\int_{R_2}^{\infty}(x - R_2)f(x|T_2 + l_1 + l_2)dx$. We express the expected backorder cost per unit time as

$$\frac{b}{T_1}\left(\frac{n-1}{n}\int_{R_1}^{\infty}(x - R_1)f(x|T_1 + l_1)dx + \frac{1}{n}\int_{R_2}^{\infty}(x - R_2)f(x|T_2 + l_1 + l_2)dx\right) \tag{2.8}$$

Summing up Equations (2.5) to (2.8), we have the expected average total cost $C(R_1, R_2, T_1, T_2, n)$. Substituting the constraint nT_1 for T_2, we have the cost function $C(R_1, R_2, n, T_1)$. We write the optimization problem $\mathcal{P}$ as:

$$\begin{aligned} \min \quad & C(R_1, R_2, T_1, n) \\ & n, R_1, R_2 \in \mathbb{Z}^+ \\ & T_1 \geq 0 \end{aligned}$$

where

$$
\begin{aligned}
C &= \frac{a_1}{T_1} + \frac{a_2}{T_2} + h_2\left(R_2 - d\left(l_2 + \frac{T_2}{2}\right)\right) \\
&\quad + h_1\left(\frac{n-1}{n}\left(R_1 - d\left(l_1 + \frac{T_1}{2}\right)\right) + \frac{1}{n}\left(R_2 - d\left(l_1 + l_2 + T_2 - \frac{T_1}{2}\right)\right)\right) \\
&\quad + \frac{b}{T_1}\left(\frac{n-1}{n}\int_{R_1}^{\infty}(x - R_1)f(x|T_1 + l_1)dx + \frac{1}{n}\int_{R_2}^{\infty}(x - R_2)f(x|T_2 + l_1 + l_2)dx\right)
\end{aligned}
$$

For given values of (T_1, n), the cost function $C(R_1, R_2, n, T_1)$ is a convex function in R_1, R_2. We can find solutions of R_1, R_2 according to the following equations:

$$\int_{R_1}^{\infty} f(x|T_1 + l_1)dx = \frac{h_1 T_1}{b}, \tag{2.9}$$

$$\int_{R_2}^{\infty} f(x|T_2 + l_1 + l_2)dx = \frac{h_1 T_1}{b} + \frac{h_2 T_2}{b}. \tag{2.10}$$

Equations (2.9) and (2.10) are a result of setting $\frac{\partial C}{\partial R_1}, \frac{\partial C}{\partial R_2}$ to be zero. For Equation (2.9) and (2.10) to have unique minimums of R_1, R_2 given T_1, n, we need to have $\frac{\partial^2 C}{\partial R_1^2} = \frac{b}{T_1}\frac{n-1}{n}f(R_1|T_1 + l_1) > 0$ and $\frac{\partial^2 C}{\partial R_2^2} = \frac{b}{T_2}f(R_2|T_2 + l_1 + l_2) > 0$. As in the single-stage model, for demand distributions that have $f(x|t) > 0,\ \forall x > 0,\ t > 0$, Equations (2.9) and (2.10) have unique solutions. However, the cost function $C(R_1, R_2, n, T_1)$ is not convex in T_1 or n.

We can search over given values of T_1 and n. The value of n is a positive integer. Note that for large value of T_1 or n in Equation (2.10), we set $R_2 = 0$. Therefore, we search over the range of values of (T_1, n) such that $(h_1 + nh_2)T_1 < b$. If the value of T_1 is restricted to be a multiple of some minimal review period (e.g., a day), it is simple just to tabulate over the values of T_1 and n. For problems with a large range of (T_1, n), we consider using simple gradient methods like Newton's method or Steepest Descent method where the step size can be determined by Amijo's rule. We can use the starting value of $T_1^D = \frac{EOQ}{d} = \sqrt{\frac{2a_1}{h_1 d}}$ from the single-stage deterministic problem. We can use the starting value of $n^D \approx \sqrt{\frac{a_2}{a_1}\frac{h_1}{h_2}}$ from the deterministic demand two-stage problem. The starting values of R_1 and R_2 can be determined accordingly given T_1^D and n^D.

For $n = 1$, we can solve the problem as a single-stage problem whose cost parameters are $h = h_1 + h_2, a = a_1 + a_2$, and $l = l_1 + l_2$. However, the cost of the $n = 1$ two-stage problem is not equivalent to such a single-stage problem due to a minor accounting difference in holding cost. Specifically, the holding cost term in the $n = 1$ two-stage problem is $h_2 d l_1$ more than that of the single-stage problem. In the single-stage problem, a replenishment

cycle starts (i.e. inventory arrives at the warehouse) at $l_2 + l_1$ if we order up to R at time 0. Recall the two-stage problem, at l_2, inventory arrives at stage 2, but then it takes an additional l_1 time periods to arrives to stage 1. That is, during $(l_2, l_2 + l_1)$, inventory is in transit from stage 2 to stage 1. Therefore, we charge holding cost h_2 during $(l_2, l_2 + l_1)$ in the two-stage problem, whereas no holding cost was charged in the single-stage problem in the same time interval. Because of this minor accounting difference, we charge an additional of $h_2 d l_1$ in the $n = 1$ two-stage problem.

2.2.3 Multi-Item Two-Stage Model with Space Constraints

In the context of an order fulfillment center in e-tailing, we need to solve the two-stage inventory problem for each SKU. When the warehouse has limited space, a space constraint couples together all of the SKUs. Here we consider two different space constraints: *i*) on the total space in an order fulfillment center, covering both the picking and deep-storage area, and *ii*) on the space in stage 1 only, the picking area. We introduce additional notations:

M	number of SKUs in storage,
γ_{ik}	storage space required by a unit of SKU i, (e.g., cubic in. per item), in stage k. Typically, $\gamma_{i1} > \gamma_{i2}$.
A_{ij}	average inventory per unit time of SKU i in echelon j,
S_j	available space in echelon j.

Space Constraint on Echelon 2

In e-tailing, stage 1 and stage 2 are in the same warehouse, and, therefore, share the total space in the warehouse. Imposing a constraint on the total space seems natural. However, there may be flexibility in deciding how much space to devote to picking and how much for reserve storage. Denote C_i as the total expected cost per unit time of SKU i, then we formulate the problem as:

$$\begin{aligned} \min \quad & \sum_{i=1}^{M} C_i(R_{i1}, R_{i2}, T_{i1}, n_i) \\ \text{s.t.} \quad & \sum_{i=1}^{M} \gamma_{i1} A_{i1} + \gamma_{i2}(A_{i2} - A_{i1}) \le S_2 \\ & n_i, R_{i1}, R_{i2} \in \mathbb{Z}^+, \qquad \forall i \\ & T_{i1} \ge 0, \qquad \forall i, \end{aligned} \tag{2.11}$$

where for each SKU i, we determine the average inventory in stage 1 A_1 from Equation (2.7). Equation (2.7) is equivalent to $h_1 A_1$. We find the average inventory in echelon 2 A_2 in

Equation (2.6), which gives $h_2 A_2$. We use the average inventory in the space constraint as proxy for the actual space requirements, which will depend upon warehouse-specific utilization factors.

We solve the problem by solving the dual problem. Denote θ to be the Lagrangian Multiplier. Given θ, the Lagrangian function is:

$$\begin{aligned} L(\bar{R}_1, \bar{R}_2, \bar{n}, \bar{T}_1, \theta) &= \sum_{i=1}^{M} C_i(R_{i1}, R_{i2}, n_i, T_{i1}) \\ &\quad + \theta \left(\sum_{i=1}^{M} \gamma_{i1} A_{i1} + \gamma_{i2}(A_{i2} - A_{i1}) \right) - \theta S_2 \\ &= \sum_{i=1}^{M} \tilde{C}_i(R_{i1}, R_{i2}, n_i, T_{i1}, \theta) - \theta S_2, \end{aligned} \tag{2.12}$$

where $\bar{R}_1, \bar{R}_2, \bar{n}, \bar{T}_1$ are vectors whose ith component is for SKU i, and the cost function $\tilde{C}_i$ has the same cost structure as C_i but with modified holding costs. Specifically, we set the holding costs in $\tilde{C}_i$, denoted as $\tilde{h}_{ij}$, as:

$$\begin{aligned} \tilde{h}_{i1} &\leftarrow h_{i1} + \theta(\gamma_{i1} - \gamma_{i2}) \\ \tilde{h}_{i2} &\leftarrow h_{i2} + \theta\gamma_{i2}. \end{aligned}$$

The dual function q can be written as:

$$q(\theta) = \min_{\substack{T_{i1} \geq 0, \\ R_{i1}, R_{i2}, n_i \in \mathbb{Z}^+}} \sum_{i=1}^{M} \tilde{C}_i(R_{i1}, R_{i2}, n_i, T_{i1}, \theta) - \theta S_2. \tag{2.13}$$

For a given value of θ, we solve Equation (2.13) as M separable problems, each of which is a single-item problem with modified holding costs. We can then solve the dual problem:

$$\begin{aligned} \max \quad & q(\theta) \\ \text{s.t.} \quad & \theta \geq 0. \end{aligned}$$

Space Constraint on Echelon 1

In e-tailing, the larger the picking area, the more difficult it is to pick items efficiently. For example, a worker picks items from a list of customer orders. The larger the picking area, the longer the route he or she may have to walk to complete the task. Therefore, other things begin equal, labor costs are higher per customer order when the picking area is larger. We impose a space constraint on echelon 1 to ensure efficient picking or efficient utilization of labor. It may be possible to augment reserve space by, say, adding some trailers in the

yard or finding a close storage building. Here we impose a constraint only on echelon 1, and we formulate the problem as:

$$\begin{aligned} \min \quad & \sum_{i=1}^{M} C_i(R_{i1}, R_{i2}, T_{i1}, n_i) \\ \text{s.t.} \quad & \sum_{i=1}^{M} \gamma_{i1} A_{i1} \leq S_1, \\ & R_{i1}, R_{i2}, n_i \in \mathbb{Z}^+, \qquad \forall i \\ & T_{i1} \geq 0, \qquad \forall i, \end{aligned} \tag{2.14}$$

Similar to the procedures in the previous section, we again solve for the dual problem. Given the value of θ, the dual function can be solved by solving M separable single-item minimization problems, where we set the holding costs as $\hat{h}_{ij}$:

$$\begin{aligned} \hat{h}_{i1} &\leftarrow h_{i1} + \theta\gamma_{i1} \\ \hat{h}_{i2} &\leftarrow h_{i2} \end{aligned}$$

2.3 Numerical Study

Our numerical study addresses a few important questions on the single-stage and two-stage single-item model: how computationally efficient are the approximate models, especially in comparison to the exact models? How sub-optimal are the approximate models relative to the exact models? Can we identify any parameters for predicting a priori the relative performance?

We implement the solution of four single-item models for Poisson demand in Matlab: *1)* the approximate single-stage model in § 2.1, *2)* the exact single-stage model in Appendix A.1, *3)* the two-stage model in § 2.2, and *4)* the exact two-stage model derived in Appendix A.2. For each model, we employ the steepest descent method, and determine the step size using Amijo's rule. We randomly selected a set of starting values in addition to using the optimal values from the deterministic models.

Since the single-stage model is the basis for the two-stage model, we first compare it with the exact single-stage model. For the sake of comparison, we perform the comparison using the input data sets from Zheng [Zhe92] and Rao [Rao03]. For all problems, the lead time is normalized to 1. Demand is Poisson with rate $d = 5, 25, 50$. Set-up cost is taken to be $a = 1, 5, 25, 100$, backorder cost to be $b = 5, 10, 25, 100$, and holding cost to be $h = 1, 10, 25$. We note that the units for the backorder cost here differ from that in Zheng [Zhe92] and Rao [Rao03]. Their units is per item per unit time, whereas ours is per item. We solve a total of 144 problems for both the approximate and exact model.

The term $\frac{hT}{b}$ in Equation (2.2) represents the probability of stock-out during a replenishment cycle for a given review period T. Suppose that we set the review period T to the deterministic problem optimal value ($T^D = \sqrt{\frac{2a}{dh}}$), and define $\boldsymbol{p_1} = \frac{hT^D}{b}$. We expect that p_1 is a good estimate of the optimal stock-out probability in the stochastic demand problem, and the accuracy of the approximate model depends on the magnitude of p_1. That is, we expect the approximate model to be less accurate for larger values of p_1.

Table (2.1) presents the exact and approximate solutions for four examples, which all share the same parameter values of a, h, d, and l but have different values of b.

Numerical Examples: $a = 25, h = 10, d = 25, l = 1$

Exact

b	T^{E}	R^{E}	TotalCost$^{\mathrm{E}}$	AvgInven$^{\mathrm{E}}$	SafetyStock$^{\mathrm{E}}$
100	0.46	47	244.2	16.21	10.38
25	0.52	43	201.2	11.65	4.97
10	0.63	40	164.4	7.81	-0.82
5	0	0	125.0	0	0

Approximate

b	T^{A}	R^{A}	TotalCost$^{\mathrm{A}}$	AvgInven$^{\mathrm{A}}$	SafetyStock$^{\mathrm{A}}$	p_1
100	0.47	47	244.20	16.20	10.35	0.04
25	0.53	43	201.27	11.52	4.70	0.18
10	1	38	187.16	4.13	-12	0.45
5	0.5	34	137.49	4.30	-3.50	0.89

Table 2.1: The Single-Stage Exact and Approximate Solutions

We use superscript E and A to denote solutions from the exact and approximate model, respectively. The columns *TotalCost* and *AvgInven* contain the long-run average cost and the average inventory. The column *SafetyStock* is the expected inventory level at the end of a cycle (denoted by s in Figure 2-2). The columns *TotalCost*^{A}, *AvgInven*^{A}, and *SafetyStock*^{A} are computed using the exact model cost function given the resulting (R, T)'s from the approximate model. In the `Exact` table, for $b = 5$, the optimal solution is to carry no inventory, therefore, we have no values for all terms except the total cost term. We list p_1 in the `Approximate` table for each example. We observe that the gap between the exact and approximate solutions varies directly with p_1; for these examples, the approximate model appears quite accurate for $p_1 \leq 0.18$, but is not accurate for $p_1 \geq 0.45$. In Table (2.2), we summarize the comparison of the approximate and exact single-stage models for the 144 problems. We group the test problems according to their p_1 value. For each interval of p_1, we denote **J** as a subset of all problems in this range, and **I** as the subset of J whose optimal solution carries non-zero inventory (i.e., $R > 0$). Column $|J|$ indicates the number of problems that have their p_1 value in the range specified by the p_1 column;

p_1	$\|I\|$	$\|J\|$	$\bar{T}$	$\widehat{R}$	$\bar{\text{TC}}$	$\bar{\text{AvgInven}}$
< 0.04	30	31	0.03	2	0.37%	2.47%
0.04 - 0.11	29	31	0.04	2	0.47%	2.20%
0.12 - 0.29	31	33	0.16	3	5.41%	17.8%
0.30 - 0.80	21	27	0.39	5	5.18%	25.1%
> 0.80	0	22	-	-	-	-

Table 2.2: Summary of the Approximate and Exact Single-Stage Model Comparison

column $|I|$ indicates the number of problems that have their p_1 value in the specified range and that have non-zero inventory in their optimal solution. To evaluate the quality of the approximate model, we report in the table the following measures:

$$\begin{aligned}
\bar{T} &= \frac{1}{|I|}\sum_{\forall i\in I}|T_i^E - T_i^A| \\
\widehat{R} &= \max_{\forall i\in I}|R_i^E - R_i^A| \\
\bar{\text{TC}} &= \frac{1}{|I|}\sum_{\forall i\in I}\frac{|\text{TotalCost}_i^E - \text{TotalCost}_i^A|}{\text{TotalCost}_i^E}\,100 \\
\bar{\text{AvgInven}} &= \frac{1}{|I|}\sum_{\forall i\in I}\frac{|\text{AvgInven}_i^E - \text{AvgInven}_i^A|}{\text{AvgInven}_i^E}\,100
\end{aligned}$$

As predicted, we see from Table 2.2 that the approximate model is quite accurate for small values of p_1, e.g., $p_1 \leq 0.11$. In an e-tailing setting, we expect such small values for p_1, as the fixed order cost a and the holding cost h are quite low relative to the backorder cost b. We will discuss this more in the next section. Both the exact and approximate models run less than a second.

To compare the approximate two-stage single-item model with the exact model (given in Appendix A.2), we solved a total of 36 problems. For these test problems we set the fixed replenishment cost to be $(a_1, a_2) = (1, 4)$, the echelon holding cost to be $(h_1, h_2) = (0.2, 0.8), (0.8, 0.2), (2, 8), (8, 2), (5, 20), (20, 5)$, the echelon lead time to be $(l_1, l_2) = (1, 2), (1, 8)$, the backorder cost to be $b = 10$, and the demand rate to be $d = 5, 25, 50$. In addition to p_1, we define p_2 as $\frac{h_2 T_2^D}{b} + \frac{h_1 T_1^D}{b} = \frac{1}{b}\left(\sqrt{\frac{2a_2h_2}{d}} + \sqrt{\frac{2a_1h_1}{d}}\right)$.

Table 2.3 displays three numerical examples of the exact and approximate two-stage model. All examples share the same value of a, h, l, b but have different values of demand rate d. Again, we use superscript `E` and `A` to label solutions from the exact and approximate model. The $\texttt{AvgInven}_1$ column is the average inventory in echelon 1 and the $\texttt{AvgInven}_2$ column is the average inventory in echelon 2. Also, we list the value of p_2 in the approximate table for each example. We observe that, at least for these three examples, the exact and approximate solutions are close.

Numerical Examples:
$(a_1, a_2) = (1, 4), (h_1, h_2) = (0.8, 0.2), (l_1, l_2) = (1, 2), b = 10$

Exact

d	T_1^{E}	n^{E}	R_1^{E}	R_2^{E}	TotalCost$^{\mathrm{E}}$	AvgInven$_1^{\mathrm{E}}$	AvgInven$_2^{\mathrm{E}}$
5	0.81	4	14	36	12.87	6.48	17.91
25	0.34	4	45	123	32.04	15.13	56.00
50	0.25	4	79	221	48.63	22.03	96.00

Approximate

d	T_1^{A}	n^{A}	R_1^{A}	R_2^{A}	TotalCost$^{\mathrm{A}}$	AvgInven$_1^{\mathrm{A}}$	AvgInven$_2^{\mathrm{A}}$	p_2
5	0.79	4	14	37	12.96	6.69	19.10	0.11
25	0.34	4	45	126	32.26	15.39	59.00	0.05
50	0.24	5	79	235	49.30	22.61	105.00	0.04

Table 2.3: The Two-Stage Exact and Approximate Solutions

Again, we denote **J** as a subset of all problems and **I** as the subset of J whose optimal solution carries inventory. Here each subset has 12 problems. Let

$$\begin{aligned}
\hat{n} &= \max_{\forall i \in I} |n_i^E - n_i^A| \\
\bar{R}_1 &= \frac{1}{|I|} \sum_{\forall i \in I} \frac{|R_{i1}^{\mathrm{E}} - R_{i1}^{\mathrm{A}}|}{R_{i1}^{\mathrm{E}}} 100 \\
\bar{R}_2 &= \frac{1}{|I|} \sum_{\forall i \in I} \frac{|R_{i2}^{\mathrm{E}} - R_{i2}^{\mathrm{A}}|}{R_{i2}^{\mathrm{E}}} 100
\end{aligned}$$

We summarize the results in Table 2.4. The approximate model is computationally more ef-

p_2	$\lvert I\rvert$	$\lvert J\rvert$	$\bar{T}$	$\hat{n}$	$\bar{R}_1$	$\bar{R}_2$	$\bar{\mathrm{TC}}$	$\bar{\mathrm{AvgInven}}_1$	$\bar{\mathrm{AvgInven}}_2$
< 0.12	10	12	0.05	1	0.02	0.04	4.52%	6.88%	10.58%
0.12 - 0.23	4	12	0.25	0	0.07	0.05	11.75%	18.96%	17.23%
0.24 - 0.71	2	12	0.14	0	0	0.10	3.87%	47.95%	44.81%

Table 2.4: Summary of the Approximate and Exact Two-Stage Model Comparison

ficient than the exact model. We observe that the optimal solution does not carry inventory when p_2 is large, and that a better approximation is more likely when p_2 is small.

2.4 Application to Industry Data

We test the two-stage model with space constraints and its solution approach to data from a major global e-tailer. The purpose of the study is to examine how the model applies to this setting, and to obtain managerial insights. Of particular interest are the questions of how to

allocate space between the reserve and picking area, and how the structure of the inventory policies depends upon the space allocation as well as on various problem parameters.

2.4.1 Data

Weekly demand data were collected from a warehouse in a six-week period for about 400,000 SKUs. Rather than computing the inventory policies for this large number of SKUs, we first group the SKUs into aggregate product categories. For each category, we then derive a *"typical" SKU* whose average demand rate and standard deviation are input parameters to the inventory model. We assume that all SKUs in a category have the same cost parameters and replenishment lead times. The assumption is quite realistic because the products are very homogenous within a category, e.g., high-demand books, low-demand DVDs. The optimal inventory policy of the "typical" SKU is then applied to each SKU in the category. We justify this aggregation based on our intent to explore the applicability of the model in this context, and to uncover managerial insights. We divide SKUs into eight product types and three demand volumes to create twenty-four mutually exclusive categories. The product types include books, DVD, music, software, video, and video games. The demand volume is divided into fast, medium, and slow.

For each category, we estimate the demand rate and standard deviation for a "typical" SKU to ensure a good approximation for the resulting total average inventory of the category. We propose to do this with a simple average for both the demand rate and the standard deviation:

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mu_i, \quad \bar{\sigma} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i. \tag{2.15}$$

where i is the SKU index, N is the number of SKUs in the category, μ_i is the demand rate, and σ_i is the standard deviation of demand per unit time for SKU i. We expect this will provide a reasonable estimate of the total average inventory for the category, since a good rough estimate of the total average inventory, assuming a periodic review order-up-to policy for all of the SKUs in the category, is:

$$\sum_{i=1}^{N}\mu_i\frac{T}{2} + \sum_{i=1}^{N}\sigma_i k\sqrt{l+T} = \bar{\mu}\sum_{i=1}^{N}\frac{T}{2} + \bar{\sigma}\sum_{i=1}^{N}k\sqrt{l+T}$$

where T is the review period, l is the lead time, and k is some positive constant. That is, we expect the cycle stock to vary linearly with the average demand rate $\bar{\mu}$, and the safety stock to vary linearly with the average standard deviation $\bar{\sigma}$.

We apply this approximation method to each category as along as the Coefficient of Variation (CV) of SKUs in a category is fairly consistent. To examine that, we plotted a CV histogram for each category, where each SKU's CV is tabulated. Figure 2-4 is an

example. For those categories that have more than one spike or have long tails in the CV

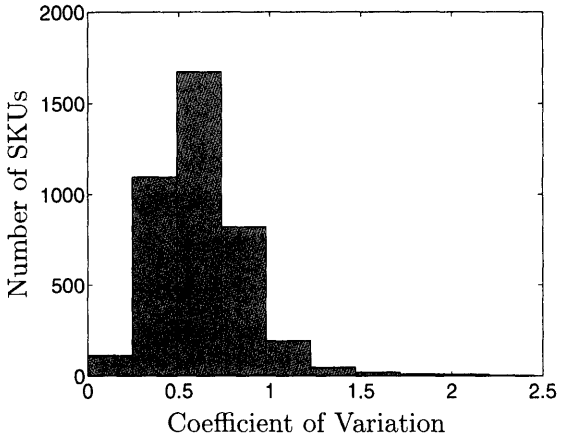

Figure 2-4: Coefficient of Variation Histogram for a Category

histogram, we further divide the category into sub-categories.

2.4.2 Results

We report on the fast and medium demand categories. The optimal inventory policy for the items in the slow-demand categories is to store in only one area typically, either picking or reserve, or not stock at all. Therefore, we have a total of 16 categories as input to our constrained two-stage model.

The cost parameters are obtained from the retailer in consideration of the warehouse's actual operation. The set-up cost or fixed replenishment cost of echelon 2 for an individual SKU is zero or near zero. This is because each supplier replenishes a large number of SKUs in a coordinated way, so that it is quite difficult to discern the ordering cost at an individual item level. However, there are economies of scale associated with the joint replenishment of a set of items by a single supplier, which takes the form of a lower bound on T_2, the review period for echelon 2. The ratio of holding over backorder cost, $\frac{h}{b}$, ranges from 0.07 to 0.56. We normalize the echelon 1 lead-time to be one time unit for all categories; this lead-time represents the time for the picking area to be replenished by the reserve area. The echelon 2 lead time, namely the lead time from external suppliers, ranges from 2.1 to 10.7 time units.

To examine the intrinsic trade-offs in the warehouse, we parameterize the echelon 1 set-up cost a_1 and the Lagrangian Multiplier θ associated with a space constraint. The results are shown in Figure 2-5. Note that the axis are scaled to disguise the real data. For all figures in Figure 2-5, each line corresponds to a value of a_1 as indicated in the legend, and the points on each line are associated with $\theta = [2.4, 0.4, 0.2, 0]$ from left to right. We test the same input data on two problems. One problem has a space constraint on the picking area; the total warehouse space versus the total cost are shown in the top-left figure

in Figure 2-5, and the picking-area space versus the total cost are shown in the top-right figure. Similarly, the other problem has a space constraint on the total warehouse area; the results are shown on the bottom two figures.

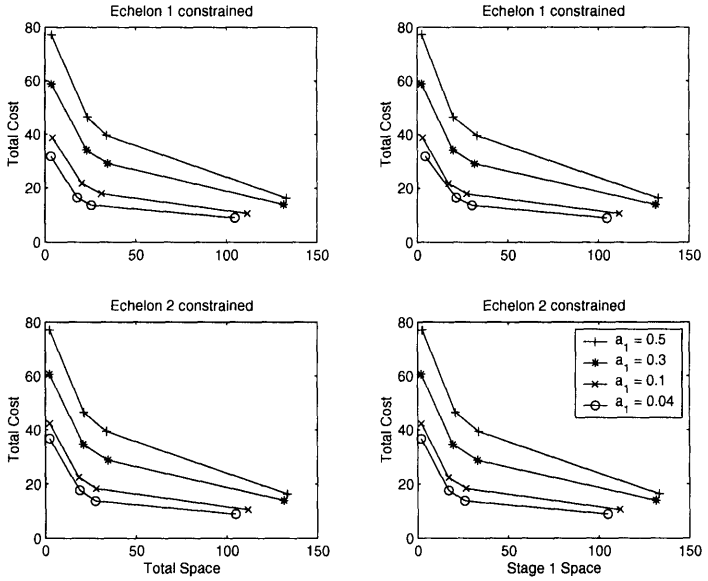

Figure 2-5: Echelon 1 or 2 Constrained Problem

As expected, as the warehouse space tightens, that is, as θ increases, the total cost increases and the total space decreases due to the tighter constraint. In the unconstrained problem results ($\theta = 0$), all categories have single-stage solutions in both problems. This is because the echelon 2 set-up cost a_2 is zero, and we did not impose a constraint on T_2.

Transition to a two-stage solution

In this section, we investigate how the inventory policies for different items change as a space constraint tightens. In a constrained multi-item two-stage warehouse, we wish to know which SKUs are to be stored in both stages and which SKUs are only in stage 1. It is also useful to predict this type of profiling in the event that demand increases or the number of SKUs increases, which results in a more constrained warehouse. We first examine the deterministic demand case for some insight.

We first consider a space constraint on picking area only. According to Schwarz [Sch73], the two-stage cost for constant deterministic demand of a single item is:

$$C(n,T) = \frac{a_2 + na_1}{nT} + (nh_2 + h_1)\frac{dT}{2}. \tag{2.16}$$

The optimal cost is then

$$C^* = \sqrt{\frac{2(a_2 + n^* a_1)(n^* h_2 + h_1)d}{n^*}}, \tag{2.17}$$

where n^* is the optimal n satisfying

$$n^*(n^* - 1) \leq \frac{a_2}{a_1}\frac{h_1}{h_2} \leq n^*(n^* + 1). \tag{2.18}$$

For the multi-item constrained problem, we solve

$$\begin{aligned} \min \quad & \sum_i C_i(n_i, T_i) \\ \text{s.t.} \quad & \sum_i \gamma_i A_i(T_i) \leq S \\ & n_i \in \mathbb{Z}^+,\ T_i \geq 0, \quad \forall i \end{aligned}$$

where C_i is as in Equation (2.16) and $A_i = \dfrac{d_i T_i}{2}$ is the average inventory in picking area for SKU i. Let the total picking area space be $S = \hat{S}$, $\widehat{C}$ be the corresponding optimal cost, $\hat{\theta}$ be the corresponding Lagrangian multiplier. Then,

$$\begin{aligned} \hat{C} &= \min_{\substack{n_i \in \mathbb{Z}^+ \\ T_i \geq 0}} \left\{ \sum_i C_i + \hat{\theta}\left(\sum_i \gamma_i A_i - \hat{S} \right) \right\} \\ &= \sum_i \min_{\substack{n_i \in \mathbb{Z}^+ \\ T_i \geq 0}} C_i + \hat{\theta}\gamma_i A_i - \hat{\theta}\hat{S} \\ &= \sum_i \sqrt{\frac{2(a_{2i} + \hat{n}_i a_{1i})(\hat{n}_i h_{2i} + (h_{1i} + \hat{\theta}\gamma_i))d_i}{\hat{n}_i}} - \hat{\theta}\hat{S} \end{aligned}$$

and $\hat{n}_i$ is:

$$\hat{n}_i(\hat{n}_i - 1) \leq \frac{a_{2i}}{a_{1i}}\frac{h_{1i} + \hat{\theta}\gamma_i}{h_{2i}} \leq \hat{n}_i(\hat{n}_i + 1). \tag{2.19}$$

For a different value of the pick area area $S = \bar{S}$, let $\bar{C}$ be the corresponding cost , $\bar{\theta}$ be the corresponding Lagrangian multiplier, and $\bar{n}_i$ be the optimal n. If the space constraint tightens from $\hat{S}$ to $\bar{S}$, $\bar{S} \leq \hat{S}$, then the corresponding Lagrangian multiplier increases $\bar{\theta} \geq \hat{\theta}$ [Ber99]. As a results, the optimal value of n_i increases $\bar{n}_i \geq \hat{n}_i$, $\forall i$.

As the space constraint tightens, the optimal n_i moves from 1 to 2. We want to find the θ at which a SKU first turns into two stage, i.e, $n_i = 2$, from Equation (2.19). The resulting θ is (we omit the index i here):

$$\tilde{\theta} = \frac{h_1}{\gamma_1}\left(2\frac{a_1 h_2}{a_2 h_1} - 1 \right), \tag{2.20}$$

where the term γ_1 is the space taken by an item in stage 1. The value $\tilde{\theta}$ is the threshold at which the item transitions from a single-stage solution (store only in picking) to a multi-stage solution (store in both the reserve and picking area). The larger the value of $\tilde{\theta}$ is, the tighter the space constraint must be before the SKU is stored in reserve.

Similarly, the threshold value $\tilde{\theta}$ for the problem where both the picking and reserve areas are constrained is,

$$\tilde{\theta} = \frac{h_1}{\gamma_1}\left(2\frac{a_1 h_2}{h_1 a_2} - 1\right)\frac{1}{1 - \frac{\gamma_2}{\gamma_1}\left(1 + 2\frac{a_1}{a_2}\right)}, \tag{2.21}$$

where γ_2 is the space taken by an item in stage 2.

Remark. The deterministic problem shows that regardless of whether the picking-area is constrained or the entire warehouse is constrained, among items that have the same $\frac{a_1}{a_2}$, $\frac{h_1}{h_2}$, and $\frac{\gamma_1}{\gamma_2}$ ratios, SKUs with small holding cost and large volume are more likely to move to a two-stage solution first as the space tightens.

In an attempt to find similar insights in the stochastic models, we compute the constrained models based on the exact two-stage model in Appendix A.2 for the 16 categories. For each value of θ in an increment of 0.1, we find the optimal n for each category. However, often, there is no unique $\tilde{\theta}$. We denote $\bar{\theta}$ as the smallest θ such that $n^* = 2$ for each category. It is most evident that five of the six categories with large echelon 2 lead times have the smallest such $\bar{\theta}$ among all categories. We conclude that SKUs with long echelon-2 lead times tend to move to a two-stage solution first as a space constraint tightens.

2.5 Extension – Allocating Space for WIP

In most of the inventory models in the literature, inventory is immediately depleted when demand occurs. As we mentioned in Chapter 1, this assumption is not entirely accurate in the e-tailing setting. There is usually a random delay between when demand occurs and when the demand is fulfilled. For instance, in order to qualify for free shipping, an e-tailer might impose a service delay of, say, five to ten days to fill the order. Also, for demanded items that are part of a multi-item order, there might be a random delay to fill the order if some of the items are out of stock; the entire order might wait in the warehouse until all items are available. In this sense, the order fulfillment process is more like a make-to-assemble process: products can be any subset of all items, and customer orders are assembled after their replenishment are received.

There are two implications from this. First, we need to account for the space consumed by these items. Second, we have an opportunity to modify our inventory policies to take advantage of this delay, as it is much like advance order information.

We call items in the warehouse that were assigned to customer orders but are waiting

to be shipped as Work In Process (WIP). We propose to model the WIP for an item as a $M/G/\infty$ queue. That is, we assume that demand is from a Poisson process, and the delay follows a general distribution. We also effectively assume that the item has a high service level, so that all demand immediately enters the $M/G/\infty$ queue; in reality, when the item stocks out, subsequent demand is delayed before it could become WIP. More formally, for single SKU, we denote

d	Poisson demand arrival rate,
Y	time from a customer order placement
	to the time until all items in the order are assembled

That is, random variable Y represents the extra delay in demand fulfillment in online retailing. Let $\bar{Y}$ be the expected value of Y. Thus, in steady state, the amount of WIP has a Poisson distribution with mean $d\bar{Y}$.

The actual distribution of Y may depend on the ordering policy, multi-item demand patterns, and assembly priority policy. As an approximation, we can determine Y from historical data. Then, we can incorporate the term into our models to account for the WIP queue in stage 1:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{M} C_i(R_{i1}, R_{i2}, T_{i1}, n_i) \\
\text{s.t.} \quad & \sum_{i=1}^{M} \gamma_{i1} \left(A_{i1} + d_i \bar{Y}_i\right) \leq S_1, \\
& n_i, R_{i1}, R_{i2} \in \mathbb{Z}^+, && \forall i \\
& T_{i1} \geq 0, && \forall i
\end{aligned}$$

where d_i is the arrival rate of SKU i and Y_i is the random relay of SKU i. The first constraint represents the space of the average inventory amount and the space taken by the WIP. This extended model, however, only incorporates the use of space. Future research should also consider the impact of WIP on the inventory model.

In this chapter, we formulate an approximate two-stage serial model for a multi-item inventory system with space constraints. We benchmark the approximate single-stage and two-stage models with the exact models, and show that the approximate models performs well under our assumptions. We also report tests on the model using real-world data. We conclude that allocation of picking and deep-storage area space depend on the product mix, where long-echelon 2 leadtime is an important factor.

Chapter 3

Order–Warehouse Assignments

3.1 Introduction

When a customer places an order on an e-tailer's website, the e-tailer, in real time, searches for available fulfillment options from its order fulfillment centers (warehouses) or drop-shippers. The e-tailer assigns the order to one or more warehouses virtually, mainly based on the transportation cost of shipping the order from the warehouse(s) to the customer location and on the current warehouse inventory availability. Depending on the inventory availability and customer preferences, the e-tailer then quotes a *promise-to-ship date* to the customer. The promise-to-ship date is the date by which the e-tailer promises to ship the order. If the order has multiple items, then the e-tailer may not be able to ship the order from one location. As a result, it may assign the order to multiple warehouses or drop-shippers and the order is split. After the e-tailer assigns the order, each item in the order enters the picking queue at its designated warehouse. The order might wait *six to eighteen hours* before the items in the order are picked and assembled into a shipment. The shipment is then given to a third party carrier to deliver the package(s) to the customer location.

We present Example 3.1.1 to illustrate the real-time assignment decision.

Example 3.1.1. Suppose a customer located at Chicago orders one unit of CD, as indicated in the dash box in Figure 3-1. Seconds later, a customer from Boston orders a unit of the same CD and a book as in the solid box. In real time, the e-tailer searches for its available inventory in all of its warehouses: Warehouse 1 near New York and Warehouse 2 by San Francisco. Both warehouses have one unit of the CD available, and the e-tailer will make the assignment to minimize transportation costs. When O1 arrives, both warehouses can satisfy the order. The e-tailer chooses the cheaper option to ship the CD from New York, so the e-tailer assigns the CD inventory in Warehouse 1 to O1. When O2 arrives, there is only one fulfillment option. Without placing an inventory replenishment order, the e-tailer

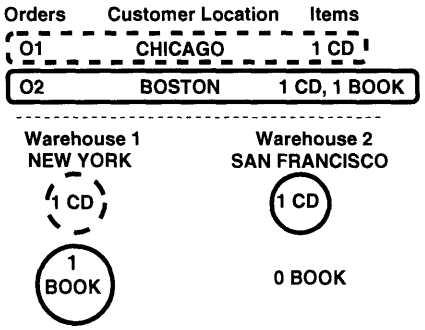

Figure 3-1: Real-Time Assignments, Three Shipments – Example 3.1.1.

fulfills the second order with two shipments: Warehouse 2 can ship the CD and Warehouse 1 can ship the book to the second customer. We have a total of three shipments for the two orders.

In the transportation cost of shipping a package, the fixed cost component is very significant. We display the current Ground Commercial rates within the US continent from UPS in Figure (3-2) [UPS05]. We display both the rates for shipping to Zone 1 from Zone 2 (the closest zone) and from Zone 8 (the farthest zone). We see that in both instances

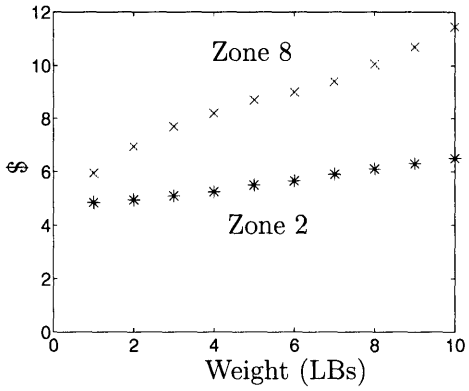

Figure 3-2: UPS Ground Commercial Rates Within the US Continent

the shipping cost consists of a fixed cost of about $5 per shipment, plus a variable cost that is linear in the weight of the package. Furthermore, for small shipments the fixed cost represents the majority of the shipping costs. As a consequence, reducing the number of shipments is a very good proxy for minimizing the transportation costs in the e-tailing setting. For example, consider an order that weighs about eight pounds. It is cheaper to ship a single package of eight pounds from Zone 8 than to ship two four-pound packages from Zone 2. The difference is even more pronounced at smaller weights. For example, shipping a two-pound package and a six-pound package from Zone 2 costs $10.60, while shipping one eight-pound package from Zone 8 costs $10.05. For items that can typically

be fit into the few standard packages, their weight is at most a few pounds, e.g., books, CDs, DVDs. Therefore, the e-tailer minimizes its transportation costs by minimizing the number of shipments.

If we consider the two orders in Example 3.1.1 alone, we can reduce the number of shipments to two, as illustrated in Figure 3-3 by changing the order-warehouse assignments. We assign the first customer order O1 to Warehouse 2 and the second O2 to Warehouse 1. Example 3.1.1 is a bit extreme and the modification to the initial assignments is very

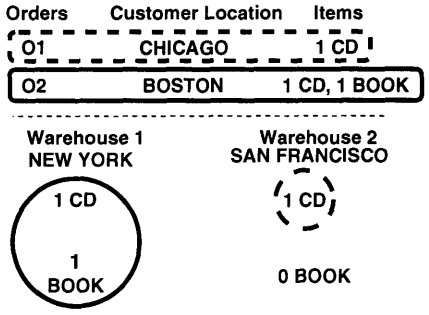

Figure 3-3: Re-Evaluation Reduces No. of Shipments to 2 – Example 3.1.1.

straightforward. To appreciate the difficulty and the subtlety of the problem, we discuss Example 3.1.2 next.

Example 3.1.2. We have four customer orders, labeled as O1, O2, O3, O4 in the sequence of arrival, and three warehouses, labeled as W1, W2, W3. The warehouses carry five SKUs, with the names CD, book, toy, camera, and DVD. The real-time assignment is as indicated in Figure 3-4. The first customer order is for one unit of CD, and the e-tailer assigns it to

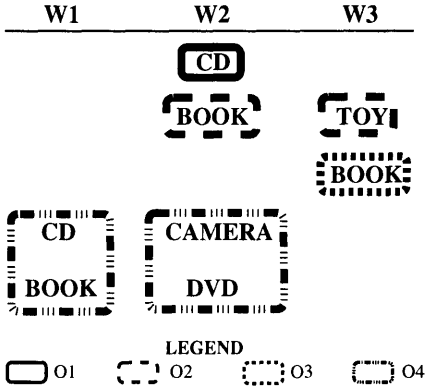

Figure 3-4: Read-Time Assignments, 6 Shipments – Example 3.1.2.

W2, possibly because the first customer is nearest to W2 or W2 is the only warehouse with the CD in stock at the time. The second order O2 consists of the book and the toy. The e-tailer assigns the book to W2 and the toy to W3. Suppose that an inventory replenishment

of books is received at W3 between the arrivals of O2 and O3. The e-tailer then assigns O3 to W3. Finally, there are two shipments for customer O4: the CD and book from W1 and the camera and DVD from W2. Thus, there are six shipments for the four orders, and it may be unclear how we can shuffle the assignments to reduce the number of shipments. In

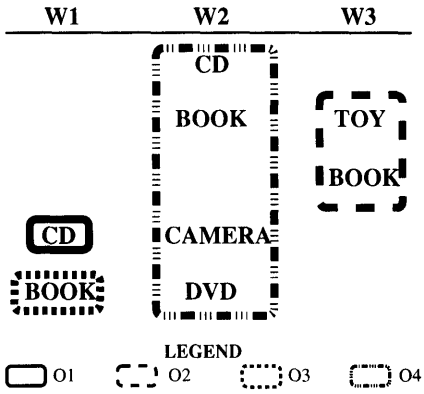

Figure 3-5: Re-Evaluation Reduces Number of Shipments from 6 to 4 – Example 3.1.2.

Figure (3-5), we reduce the number of shipments from six to four, which is clearly the best we can do.

We show with examples that the real-time decision is necessarily myopic because the e-tailer does not anticipate any future customer orders or inventory replenishment. The real-time assignment is myopic in practice because the e-tailer wants to reserve the inventory for the customer, then inform the customer with confidence that inventory is available and that the order can be fulfilled by the promise-to-ship date. The real-time assignment is myopic also because of two main challenges. The large number of customer orders and the need to provide customers with a very quick response in real time make efficient assignment difficult. We conjecture that we can reduce the total transportation cost of shipping orders from warehouses by re-evaluating the real-time assignment decisions, subject to the constraint that the promise-to-ship date commitment for any customer order is guaranteed.

This shuffling of assignments is also practically feasible. Even when all items in an order are available at the warehouse, the order may wait 8 to 16 hours until the order is released to be picked and sent for shipping. If one or more of the items in the order is not available, then the rest of the order is reserved and waits until the missing items arrive. By re-evaluating the real-time decision, the e-tailer can also afford more decision making time. We pose a problem to re-evaluate the real-time decisions. We consider the queue of *not-yet-picked* customers orders at a random time and their real-time warehouse assignments. We re-evaluate these real-time decisions to reduce the shipping cost without violating the promise-to-ship date commitments for these orders. The not-yet-picked orders are the orders that have not yet been released to be picked at each warehouse. We take

inventory availability and the real-time quoted promise-to-ship dates as given. We call this optimization of snapshot order assignments as the re-evaluation problem.

We will show in later sections that this snapshot optimization problem is difficulty theoretically (belong to the NP-hard complexity class) and in practice. For now, exact methods cannot solve realistically dimensioned cases. In the e-tailing setting, the problem size is especially large. For an off-season snapshot at a large e-tailer, there are 1 million orders with 2 to 3 million units waiting to be picked. There are up to 10 warehouses. The total number of SKUs in those orders ranges from 500,000 to 800,000. In the peak season, the number of orders can reach three or five times of the off-season.

We, therefore, develop efficient and easy-to-implement sub-optimal heuristics to solve the re-evaluation problem. Given the real-time assignment decisions, we take the natural path to construct an improvement heuristic that starts with a feasible solution and iteratively finds better solutions. We also derive bounds to determine the sub-optimality of our heuristics.

In the following sections, we discuss the problem formulation and our heuristic solution approach. We also summarize some computational experiments on sets of real data from a global e-tailer.

3.2 Problem Formulation

We present two formulations of the re-evaluation problem, where one is based on the set partitioning problem, and another is a network design formulation. Both formulation shed light on the underlying structure and difficulty of the problem.

3.2.1 Formulation 1

For this set-partitioning based formulation, we first examine the real-time assignment decision for a single order. For now, we assume that we have enough inventory across warehouses in the network to satisfy the order. Without this assumption, we have a set-packing problem, where some of the items in an order may be unassigned. We start with some notation.

k	index for warehouses
I	set of SKUs, where $\|I\| = m$ and i is the index.
N	$= \{1, \ldots, n\}$, a collection of all possible subsets of the order, i.e., C_l, $l \in N$, is the l^{th} subset of the order
A	a m by n matrix such that a_{il} is the number of of SKU i included in subset C_l

d_i	units of SKU i in the order
u	order size, or the number of units in the order, $u = \sum_i d_i$
e_n	a n by 1 vector of 1's
y_{lk}	$= 1$ if subset C_l is shipped out of warehouse k; $=0$ otherwise.
s_{ik}	inventory units of SKU i available at warehouse k

We denote the following formulation of assigning an order to warehouses as $\mathcal{P}$.

$$\begin{aligned} \min \quad & \sum_{\forall\ l,k} y_{lk} \\ \text{s.t.} \quad & \sum_{\forall k} \sum_{\forall\ l} a_{il}\ y_{lk} = d_i, \quad \forall\ i \\ & \sum_{\forall l} a_{il}\ y_{lk} \leq s_{ik}, \qquad \forall\ i,k \\ & y_{lk} \in \{0,1\}, \qquad \forall\ l,\ k \end{aligned}$$

The first constraint guarantees that the required number of each SKU in the order is shipped. The second constraint is a supply constraint: the amount of SKU i shipped from warehouse k cannot exceed the supply of SKU i in warehouse k.

Suppose we substitute index r for (l,k), and restrict each SKU to have at most one unit in the order ($d_i = 1$) and allow supply to be infinite. This problem is a set partitioning problem:

$$\begin{aligned} \min \quad & \sum_{\forall r} y_r \\ \text{s.t.} \quad & Ay = e_n \\ & y_r \in \{0,1\}. \end{aligned}$$

Now we examine all orders in the re-evaluation problem. Again, we assume that inventory in the network can satisfy all orders. E-tailers can fulfill orders either by inventory physically in the warehouse or inventory on order. We modify the previous notations and introduce new notations.

j	index for customer orders, J is the order set
I_j	set of unique SKUs in order j
N_j	$= \{1, ..., n_j\}$, a subset collection of order j with n_j subsets
N	$= \{N_i, \ldots, N_j, \ldots\}$, and the total number of subsets is $n = \sum_j n_j$

For each SKU, we denote J_i as the set of orders with at least one unit of SKU i. We

call the re-evaluation problem as $\mathcal{Q}$:

$$\begin{aligned}
\min \quad & \sum_{\forall j} \sum_{l \in N_j, k} y_{lk}^j \\
\text{s.t.} \quad & \sum_{\forall k} \sum_{l \in N_j} a_{il}^j \, y_{lk}^j = d_i^j, \quad \forall \; j, \; i \in I_j \\
& \sum_{j \in J_i} \sum_{l \in N_j} a_{il}^j \, y_{lk}^j \leq s_{ik}, \quad \forall \; i, k \\
& y_{lk}^j \in \{0,1\}, \qquad\qquad \forall \; j, \; l \in N_j, \; k
\end{aligned}$$

The major difference from problem $\mathcal{P}$ is the second constraint. Here the total amount of SKU i shipping from warehouse k for all orders should be no more than the supply in warehouse k.

Let u_j be the size, or number of units, of order j and m_j be the number of SKUs in order j. In problem $\mathcal{Q}$, the number of binary decision variable is nK, or $K \sum_j 2^{u_j}$ in the worst case. The number of first constraints is $m = \sum_j m_j \leq |I||J|$, the number of second constraints is $|I||K|$. Notice n could be exponential in the input data. In problem $\mathcal{Q}$, the number of binary variables can be exponential and the number of constraints is linear in the input data.

3.2.2 Formulation 2

We can also formulate the re-evaluation problem as a network design problem, specially, a fixed-charge multi-commodity flow problem [AS04]. We redefine the decision variables, but keep the previously defined notations.

x_{jk}^i	units of SKU i shipped from warehouse k to customer j
y_{jk}	binary variable to indicate a shipment from k to j

We also denote set K_i as the set of warehouses carrying SKU i inventory, J_i as the set of customer orders containing nonzero units of SKU i.

We denote the following formulation as $\mathcal{MIP}$.

$$\begin{aligned}
\min \quad & \sum_{j,k} y_{jk} \\
\text{s. t.} \quad & \sum_{k \in K(i)} x_{jk}^i = d_j^i, \qquad \forall \; i \in I, \; j \in J_i \\
& \sum_{j \in J_i} x_{jk}^i \leq s_k^i, \qquad \forall \; i \in I, \; k \in K_i \\
& 0 \leq x_{jk}^i \leq d_j^i \; y_{jk}, \quad \forall \; i \in I, \; j \in J_i, \; k \in K_i \\
& y_{jk} \in \{0,1\}, \qquad \forall \; j, \; k
\end{aligned}$$

Notice that a commodity is a SKU. Variable x is a continuous variable here because for any given choice of y, we can decompose the problem into a transportation problem by SKU, and there exists an optimal integer solution for each transportation problem. The first constraint assures that the demand is met for each SKU in each order. The second constraint assures that the amount of each SKU shipped from each warehouse does not exceed the supply. Problem $\mathcal{MIP}$ has $|J||K|$ binary variables and $|I||J||K|$ continuous variables. It has $|I||K|+|I||J|+|I||J||K|$ number of constraints. The number of constraints and variables is linear in the input data.

3.2.3 Complexity

In this section, we aim to show that our general re-evaluation problem is NP-hard by proving that some simple special cases of the problem are NP-hard. First, we introduce a decision problem that we use in the following proof.

Minimum Edge Coloring

INSTANCE: $G = (V, E)$, positive integer K

QUESTION: Can the edge set E be partitioned into disjoint sets $E_1, ..., E_k$, with $k \leq K$, such that for every subset $1 \leq i \leq k$, no two edges in E_i share a common endpoint or node in G?

The Minimum Edge Coloring problem is a known NP-Complete problem [Hol81]. Figure 3-6 is an example of edge coloring. The graph can be partitioned into two colors: one formed by the solid edges $(1,4),(2,3)$ and another by the dotted edges $(1,2),(3,4)$.

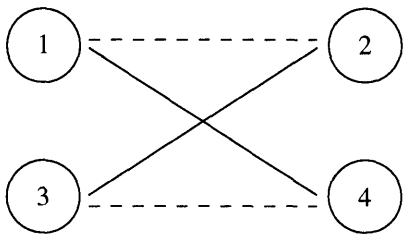

Figure 3-6: Edge Coloring

2-SKU K-Warehouse

We state a special case of the general re-evaluation problem where each customer orders exactly two items and there are K warehouses.

INSTANCE: n distinct items or SKUs, K warehouses and each warehouse has one unit of the n SKUs, m orders and each order has one unit each of 2 distinct SKUs.

QUESTION: Can the m orders be satisfied from the K warehouses with at most m shipments?

Proposition 3.2.1. *The* 2-SKU K-Warehouse *decision problem is NP-complete.*

Proof. We transform Minimum Edge Coloring to 2-SKU K-Warehouse.

Suppose we have $G = (V, E)$, let each item be a node in G and $n = |V|$ and let each order be an edge in G and $m = |E|$.

Clearly, this transformation can be done in polynomial time. It remains to show that E can be partitioned into $k \leq K$ disjoint sets iff the m orders can be satisfied from the K warehouses with at most m shipments.

First, suppose E can be partitioned into $k \leq K$ disjoint sets, $E_1, ..., E_k$. For $1 \leq i \leq k$, no two edges in E_i share a common endpoint in G and E_i contains a matching of the n elements. For the set of orders correspond to edges in E_i, no two orders share a common element. In other words, E_i represents orders that can be satisfied *entirely* from Warehouse i. We then can satisfy the orders from k warehouses and each order can be satisfied from one warehouse.

Conversely, suppose we have m shipments. Then each order can be satisfied from one of the $k \leq K$ warehouses. This corresponds to partitioning the orders into $k \leq K$ sets, where no two orders in the same set share a common item. Therefore, all edges can be partitioned into $k \leq K$ disjoint sets. ■

Corollary 3.2.1. *The problem of minimizing the number of shipments by assigning 2-SKU orders to K warehouses is NP-hard.*

Proof. From Proposition 3.2.1, we see that a special case of assigning 2-SKU orders to K warehouses, the 2-SKU K-warehouse problem, is NP-complete. By restriction, the general 2-SKU K-warehouse problem is also NP-complete. Therefore, the optimization problem is NP-hard. ■

Next we introduce a decision problem that we use in the following proof.

Exact Cover by 3SETS (X3C)

INSTANCE: A finite set X with $|X| = 3q$, a collection C of 3-element subsets of X

QUESTION: Does C contain an exact cover for X, that is, a subset collection $C' \subseteq C$ such that every element of X occurs in exactly one member of C'?

The X3C problem is a known NP-Complete problem [GJ79].

3-SKU 2-Warehouse

We state a special case of the general re-evaluation problem where each customer has exactly three items and there are two warehouses.

INSTANCE: m distinct elements or SKUs, n distinct orders and each order has 3 distinct elements, Warehouse 1 has one of each m elements, and Warehouse 2 has $3n - m$ elements just enough to fulfill the rest of the $n - m/3$ orders

QUESTION: Can the n orders be satisfied from the two warehouses with at most n shipments?

Proposition 3.2.2. *The* 3-SKU 2-Warehouse *decision problem is NP-complete.*

Proof. We show the NP-completeness by transforming X3C to 3-SKU 2-Warehouse.

Let the m elements be the set X and $m = 3q$.

Let the n orders be C the collection of 3-element subsets of X, and each order corresponds to a 3-element subset.

Clearly, this transformation can be done in polynomial time. It remains to show that C contains an exact cover for X iff the n orders can be satisfied from the two warehouses with at most n shipments.

First, suppose we have an exact cover for X, $C' \subseteq C$. Then the set of $m/3 = q$ orders, where each order corresponds to a subset in C', can be satisfied from Warehouse 1. Every element in Warehouse 1 is in exactly one order in C'. The rest of the $n - q$ order can be satisfied from Warehouse 2 by definition. We then have one shipment for each order and n shipments total.

Conversely, suppose we have a solution where n orders are satisfied by the two warehouses with n shipments. Then, each order is satisfied from one warehouse. There are exact $m/3$ orders satisfied Warehouse 1. Then those $q = m/3$ orders form an exact cover of the $3q$ elements in Warehouse 1. The collection of subsets $C' \subseteq C$ corresponds to the q orders from Warehouse 1 is an exact cover for X. ■

Corollary 3.2.2. *The problem of minimizing the number of shipments by assigning 3-SKU orders to 2 warehouses is NP-hard.*

Proof. From Proposition 3.2.2, we see that a special case of assigning 3-SKU orders to 2 warehouses, the 3-SKU 2 warehouses problem, is NP-complete. By restriction, the general 3-SKU 2 warehouses problem is also NP-complete. Therefore, the optimization problem is NP-hard. ■

Corollary 3.2.3. *The problem of minimizing the number of shipments by assigning orders to warehouses is NP-hard.*

We have shown that special cases of the re-evaluation problem are NP-hard. Therefore, the general problem of assigning orders to warehouses such that the number of shipments is minimized is NP-hard. We show that we need efficient and easy to implement heuristics to solve the re-evaluation problem.

3.2.4 Literature Review

There are two clusters of literature that are most relevant to our problem: network design problems and local search algorithms, a wide class of improvement algorithms.

The literature on network design problems is directly related to the second formulation of the re-evaluation problem. Magnanti and Wong [MW84] provides a survey of models and classic solution methods of the general network design problem up to 1985. They show that the problem is very flexible and contains a number of well known network optimization problems as special cases: minimal spanning trees, shortest paths, vehicle routing problems, facility location problems, etc. Minoux [Min89] surveys the models and solution methods of the variants of the general problems. He discusses the general models using minimum cost multicommodity flows, models of tree-like networks, models using nonsimultaneous single-commodity or multicommodity flows. Balakrishnan, Magmanti, and Mirchandani [BMM97] list an annotated bibliographies on network design since 1985. The focus of the survey is on uncapacitated network design, capacitated network design, network loading, and network restoration problems.

The general form of a network design problem is the multicommodity network design problem. Let n be the number of nodes and m be the number of edges in a graph, then we have the following [BMM97]:

$$
\begin{aligned}
\min \quad & \sum_{i \in \mathcal{I}} c^i x^i + fy & \\
\text{s.t.} \quad & \mathcal{N} x^i = b^i & \forall i \in \mathcal{I} \\
& x^i \leq k^i y & \forall i \in \mathcal{I} \\
& \sum_{i \in \mathcal{I}} x^i \leq Ky & \\
& x^i \geq 0 & \forall i \in \mathcal{I} \\
& y \in Y &
\end{aligned}
$$

where $\mathcal{I}$ is a set of commodities, $c^i \in \mathbb{R}^m$ is a vector of edge cost per unit of flow, $f \in \mathbb{R}^m$ is a vector of edge design or installation cost, $\mathcal{N}$ is a node-edge incidence matrix, $b^i \in \mathbb{R}^n$ is a vector of node supplies or demand, k^i is a vector of edge capacity for commodity i, and K is the edge capacity for all commodities. There are two types of decision variables: each element in the vector $x^i \in \mathbb{R}^m$ models the continuous choice of routing commodity i flow on edges , and each element of $y \in \mathbb{R}^m$ models the discrete design choice of installing edges. The total cost is the sum of installation and routing costs.

The first constraint is a flow balance constraint. The second constraint ensures no flow is routed on the arcs that are not installed and edge capacity on each commodity is satisfied. The third constraint imposes the total capacity on each edge for all commodities. Our problem is a fixed charge uncapacitated multicommodity flow problem, a special case of

the general network design problem. Specifically, our model has the following properties: it only has installation costs but no routing costs, it has no bundling capacity constraint on all commodities, the graph topology is bipartite, and each commodity has multiple origins and multiple destinations. The current literature has made the greatest progress in solving uncapacitated problems. Most progress, however, has been made on single-origin single-destination problem, which is the simplest case of the uncapacitated problem. Our model has a simple cost structure and topology and has no side constraints, but has multiple origin and multiple destination.

Many of the special cases of the fixed charge design problem, e.g., the Steiner Tree problem, are known to be difficult to solve or NP-hard in complexity. The fixed charge problem then is also NP-hard [MW84]. In addition to the theoretical arguments, substantial empirical evidence also confirms the difficulty of the problem on large-scale instances [e.g., BG73, Won85]. Balakrishnan, Magnanti, and Wong[BMW89] develop a dual-ascent algorithm for the fixed-charge network design problem. They test problems with up to 500 integer and 1.98 million continuous variables and constraints. The procedure shows promising results of 1 to 4% of optimality. Holmberg and Hellstrand [HH98] show a Lagrangian heuristic within a branch-and-bound framework to find the exact optimal solution. Judging by the size of the instance solved in the current literature, none are close to the scale of our problem.

There is a vast amount of literature on local search or neighborhood search. Neighborhood search is the inspiration of our proposed heuristics, and the neighborhood of our heuristics is exponentially large. We refer the readers to an extensive survey by Aarts and Lestra [EL97]. Ahuja, Ergun, Orlin, and Punnen [AEOP02] provide a comprehensive survey on very large-scale neighborhood search techniques. Neighborhood search algorithms are a wide class of improvement algorithms that try to improve iteratively by searching the "neighborhood" of the current solution. Our first formulation shows the proximity to set partitioning problems. We can view our heuristics as a type of local search procedure for partitioning problem, which involves transferring or exchanging elements between clusters. In our problem, the clusters are customer orders and the elements are inventory units. We refer readers to Thompson and Orlin [TO89] for more discussion of the literature.

We can also view our heuristics as a network flow based improvement algorithm, which use network flow techniques to identify improving neighborhoods. In our heuristics, by solving a transportation problem, we are able to identify cyclic exchanges of items among customer orders. Some of these algorithms in the literature characterize improvement moves by negative cost disjoint cycles in a so called "improvement graph". For example, Thompson and Psaraftis [TP93] apply the technique to vehicle routing problems. They try to reduce the total cost of a set of routes by transferring demand among routes cyclically. Their results show the local search method as either comparable to or better than the vehicle

routing heuristic algorithms. Ahuja, Orlin and Sharma [AOS01] apply the technique to the capacitated minimum spanning tree problem. Using variants of shortest path label-correcting algorithms, they are able to identify cycles that exchange nodes among multiple subtrees simultaneously. Their results can improve the best available solution for most of the benchmark instances by as much as 18%. Others effectively apply the search method to minimum makespan parallel machine scheduling problem [FNS04], and capacitated facility location problem [AOP+04].

An application close in flavor to ours is the paper by Talluri [Tal96] on daily airline fleet assignment problem. The fleet assignment problem can be modeled as an integer multicommodity flow problem subject to side constraints and each commodity is a fleet. The problem assigns fleet types to flight legs. After a fleet assignment problem is solved during the planning stage, often the airlines need to change the assignment to accommodate updated demand forecast, disruptions to the schedule, or breakdowns. He develops a swapping procedure to identify and change the fleet types on flights from a given solution. The exchanges are identified by finding negative cost cycles in a related network.

3.3 Complex Network Properties

In solving the problem, we understand that specially tailored heuristics are more likely to outperform any general heuristics. To develop any solution procedure tailored to the problem structure, we must examine the problem data carefully. In this section, we summarize the important characteristics of the customer orders and the real-time assignments in the e-tailing setting.

To facilitate the presentation, we introduce the following definitions.

Definition 3.3.1.

A *single order* is a customer order that consists of exactly one unit.

A *multi order* is a customer order that consists of more than one unit, may have more than one SKU or multiple units of one SKU.

A *split order* is a customer order split over warehouses in the real-time assignment, i.e., multi orders that require more than one shipment.

A *single shipment* is a one-unit shipment of a split order, that is, a shipment of a single unit that is part of a multi order.

A *double shipment* is a two-unit shipment of a split order.

Recall that once customer orders are placed, the orders are assigned to one or more warehouses and entered to a picking queue. We take a snapshot of all the orders that are

waiting to be picked at a random time. We examine a few such data sets in the off season from a large e-tailer and one data set from the peak season, as illustrated in Table 3.1. The off-season data is taken at random days during a period of five months in 2004 and the peak-season is a random day in December, 2004. The term *"Total orders"* represents the

	Off-Peak Season				Peak Season
	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5
Total orders	869K	925K	918K	956K	1.55M
Total SKUs	411K	385K	388K	406K	526K
Single orders	64%	65%	66%	65%	56%
Multi orders	36%	35%	34%	35%	44%
Split orders	3.9%	3.9%	3.7%	3.6%	6.4%

Table 3.1: Snapshot Data

total number of customer orders that have not yet been picked. The term *"Total SKUs"* is the number of unique SKUs among the total number of orders. *"Single orders"*, *"Multi order"*, and *"Split orders"* are the percentages of single, multi, and split orders among "Total orders". Overall, the snapshot data is very consistent from day to day during the off-peak season. There are close to 1 million orders with 2 to 3 million units in the not-yet-picked queue for the snapshot data. Furthermore, the size of orders tends to follow a geometric distribution.

Comparing with the peak season snapshot data, the off-peak season has less orders. Also, the percentage of multi orders and split orders are less in the off-peak season. The real-time assignments in the snapshot data split about 10% of the multi orders in the off-peak season and 18% in the peak season. Overall, the number of shipments in each split order is two or three shipments with few exceptions. There is at least one single shipment in more than 80% of the split orders. Over 90% of the split orders have at least one single shipment or one double shipment. We will discuss the implication of these numbers in the later section. In particular, we exploit the abundance of the single and double shipments in our heuristics.

To investigate whether the problem can be decomposed into a number of smaller problems, we examined the connectivity of the order-SKU graph constructed for the snapshot of the not-yet-picked queue . There is one node for each SKU and we connect two SKU nodes when an order that includes both SKUs exists. We find that there exists one very large component in the graph, containing the majority of the SKU's. Furthermore, any removal of small subsets of SKU's does not change the connectivity of the graph. Therefore, we do not see a clear way to decompose the problem by considering a limited number of orders or SKUs.

Our exploration of the order-SKU network in the e-tailing setting echoes the theory of random graphs. Solomonoff and Rapoport [SR51] and independently Erdős and

Rényi [ER59] study a very simple model of network called Poisson random graphs. In a graph with n nodes, a random edge is independently chosen with probability of p to connect any two nodes in the graph. They show that in the limit of a large n, the degrees of the nodes in the graph is Poisson distributed. They also demonstrate that when the value of p is high, a large percentage of the nodes are joined together in a single *giant component.* Recently, Newman [New03] reviews the development in complex networks. The empirical studies of networks, such as the Internet, social networks, and biological networks, show some common properties among the different complex networks. In particular, as reproduced by the random graphs, the network exhibits the "small-world effect", or most pairs of nodes in most real-world networks seem to be connected by a short path through the network. Our order-SKU network is not as simple as the Poisson random graphs: our edges are correlated. Our network, however, seems to have the small-world effect.

3.4 Heuristic Approach

In solving the re-evaluation optimization problem, we already have an initial feasible solution, i.e., the real-time assignments. It seems natural to focus on improvement algorithms, by which we iteratively create better solutions. The focus on improvement algorithms is also driven by two major practical concerns: **i)** Improvement algorithms generate a feasible solution at every iteration. After each iteration, we can implement the recommended changes to the incumbent assignments to get an improved order assignment. This facilitates the implementation of this solution approach greatly, since we always have a feasible solution, even if there were a sudden termination of the algorithm. **ii)** Retrieving data in the setting of a large e-tailer is time consuming, because of the large scale of the information systems and data storages in place. A wide class of improvement algorithms is local search algorithms which searches for the "neighborhood" of the current solution at each iteration. Local search algorithms allow the users to retrieve only a small amount of data pertaining to the neighborhood for each iteration. Therefore, we can improve overall running time by solving the problem and retrieving data in parallel.

One key idea of the heuristics is using the single orders to "fix" the split orders. The motivation is twofold. First, single orders always entail a single shipment and therefore are very flexible in their assignment. Also, we know that the special case of our network design problem where all orders are single orders is an easy transportation problem. Second, the vast majority of split orders in the real-time assignment include a single shipment. By re-assigning a single order from warehouse A to warehouse B, we free up a unit of inventory at warehouse A that might be used to avoid a split order. Our Example 3.1.1 illustrates such an instance.

We start with the real-time assignments and iteratively improve upon it by reducing

the number of shipments at each iteration. Clearly, the number of splits within an order in iteration $t+1$ is never more than the number of splits in iteration t. We observe in all data sets that the majority of the orders in the real-time assignments are already of one shipment. Typically, the percentage of the multi-item orders with no splits is around 30% of all orders. We do not consider those orders in the heuristics. In summary, the input data of the heuristics include all single orders, split orders in the initial assignment as well as all "free" or unassigned inventory at the time the snapshot was taken.

Note that we can treat all unassigned inventory as single orders. We can assume that they are assigned to a dummy customer and have an infinitely long promise-to-ship date. We also assume that all on-order inventory will arrive to the warehouse at the expected due date. That is, we assume that on-order inventory will arrive at the date promised by the supplier.

Our heuristics consist of two distinct parts. We name the first part as **Order Swap** as we consider split orders one at a time and examine possible swaps. The second part is **SKU Exchange** as we consider one SKU at a time and examine possible cyclic exchanges. In our implementation, we start with Order Swap and then proceed to SKU Exchange on the remaining split orders. We view Order Swap as a fast and extremely simple greedy algorithm to exploit the abundance of unassigned inventory and single orders. To incur incremental benefits, we employ the efficient but more time consuming SKU Exchange heuristic. *For the rest of the section, we refer to single orders to include single orders as well as uncommitted inventory.*

In the remainder of the section, we discuss the details of the heuristics and present their worst-case analysis.

3.4.1 Order Swap

Order Swap exploits the flexibility of single orders and the abundance of unassigned inventory. We start with the split orders in the initial assignment. Let j be the index for split orders, and let k be the index for warehouses. For each order j, we examine each warehouse k. If k has sufficient single orders to swap with j such that the entire order of j can be fulfilled at k, then we complete the swap. We terminate examining order j if a swap involving j has occurred or all warehouses have been considered. The heuristic is named for considering one order at a time. Figure (3-7) specifies the general implementation. We illustrate the Order Swap heuristic with the following example.

Example 3.4.1. We consider performing the Order Swap procedure on O1 as illustrated in Figure 3-8. O1 has one unit of SKU X and Y currently assigned to warehouse 1 and 2. O2 and O3 are the single orders at warehouse 3. There are no single orders at warehouse 1 or 2, so we consider warehouse 3 next. The entire order of O1 is currently not assigned to

algorithm Order Swap:
input: an initial feasible solution

1. for each split order $j = 1, ..., J$ **do**

1.1. for each warehouse $k = 1, ..., K$ **do**

1.1.1. let I_{jk} be the set of items in order j currently *not* assigned to k.
1.1.2. if k's single orders contain I_{jk},
then make swap, **go to** $j + 1$.
else go to $k + 1$.

Figure 3-7: Order Swap Algorithm

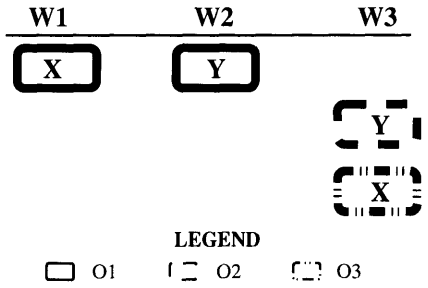

Figure 3-8: Order Swap Example 3.4.1.

warehouse 3: $I_{13} = \{XY\}$. The set of single orders at warehouse 3 contains $\{XY\}$ and we can make the swap. The assignments after the swap is illustrated in Figure 3-9

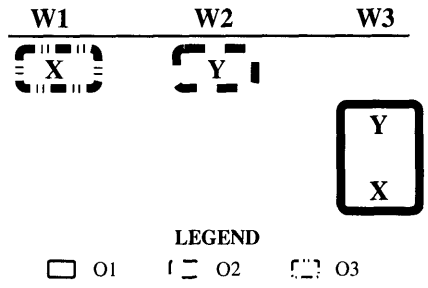

Figure 3-9: Order Swap Example 3.4.1– After a Swap.

Order Swap with time dimension

The actual problem we need to solve has a time dimension. First, we need to consider the promise-to-ship dates quoted by the e-tailer when customer orders occur. Second, we need to differentiate whether the unit of inventory assigned to the customer order is physically in the warehouse or on order. Recall the re-evaluation problem solves on the snapshot data taken on customer orders that are already placed but have not yet been picked. We then

create T time periods with respect to the time the snapshot was taken. In general, time period 0 is the snapshot time, time period $t < T$ is t days in the future of the snapshot date, and time period T is T days or more in the future of the snapshot date. We denote

v_i promise-to-ship date of item i in the real-time assignment,

u_i time at which the inventory committed to item i arrives to the warehouse.

To ensure that customer service level promised at the real time is not violated, we add the following constraint in Order Swap: it is feasible to swap item i_1 with i_2 of the same SKU iff $u_{i_1} \leq v_{i_2}$ and $u_{i2} \leq v_{i_1}$. That is, a swap is feasible if the e-tailer can still ship the orders within the promised dates after the swap.

3.4.2 SKU Exchange

The second key idea of the heuristics is to consider one SKU at a time. The main motivation is the special case of the general re-evaluation problem, where all orders are single orders. This special case can be formulated as a transportation problem and we know that polynomial algorithms can solve those problems optimally. In SKU Exchange, we start with a sequence of SKUs. For each SKU in the sequence, we can construct and solve a transportation problem that attempts to reduce the number of split orders. After solving each transportation problem, we update the affected orders, and continue with the next SKU. We terminate at the end of the SKU sequence. The transportation problem allocates the supply of the SKU at the supply nodes (the warehouses) to the demand nodes (orders that include the SKU). Figure (3-10) is a general implementation.

algorithm SKU Exchange:
input: an initial feasible solution

1. **generate** a sequence of SKU $A = \{1, 2, ..., n\}$.
2. **for** each SKU $a = 1, ..., n$ **do**,

   2.1. **construct** a transportation problem for SKU a.
   2.2. **solve** the transportation problem a.
   2.3. **update** orders affected by a.

Figure 3-10: SKU Exchange Algorithm

SKU Exchange of single shipments

We only consider SKUs that have single orders or uncommitted inventory, as well as split orders with single shipments that consist of SKU a. We start with the following example to illustrate the transportation problem before we summarize the details.

Example 3.4.2. We consider a batch of orders with the real-time assignment listed in Figure (3-11). We construct the corresponding maximization transportation problem for

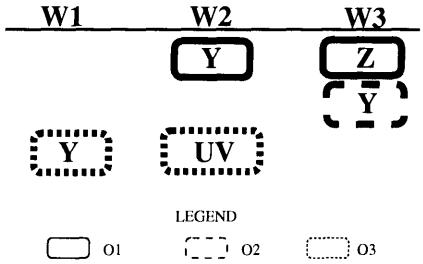

Figure 3-11: Real-Time Assignments – Example 3.4.2

SKU Y in Figure 3-12. Each warehouse represents a supply node, and each order with

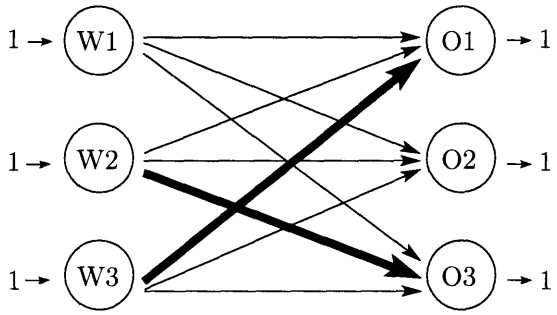

Figure 3-12: Transportation Problem for SKU Y – Example 3.4.2

a single shipment of SKU Y represents a demand node. From Figure (3-12), we see that the supply at each supply node is the available units of SKU Y at the warehouse for re-assignment, and the demand at each demand node is the number of units of SKU Y in the order. A unit of flow from supply node k to demand node j signifies that warehouse k ships a unit of SKU Y to fill order j's requirement.

Definition 3.4.1. Let the set of *profitable warehouses* of shipment y in order j be $P_j(y)$ such that, $\forall k \in P_j(y)$, while maintaining other shipments, shipping the shipment y from warehouse k reduces a split in order j.

That is, there will be one less shipment if warehouse k supplies the SKU Y for order j. The arc profit for arc $(k,j), \forall k \in P_j(Y)$ is 1, signifying that a unit flow on this arc results in one less shipment. The arc cost is zero for all other arcs. In Figure 3-12, $P_1(Y) = \{3\}, P_2(Y) = \emptyset, P_3(Y) = \{2\}$, and only arcs (2, 3) and (3, 1) (the dark arcs) have a profit of 1.

By inspection, we see that the optimal solution is to send one unit of flow along arcs, $(1,2),(2,3)$ and $(3,1)$. The optimal solution corresponds to the results in Figure (3-13). We reduce the number of shipments in the three orders from 5 to 3.

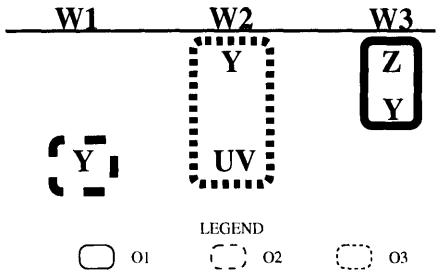

Figure 3-13: Re-Evaluation Reduces No. of Shipments from 5 to 3 – Example 3.4.2

Figure (3-14) is an augmenting cycle with respect to the initial solution: O3 → W1 → O2 → W3 → O1 → W2 → O3. Starting from the initial solution, the augmentation

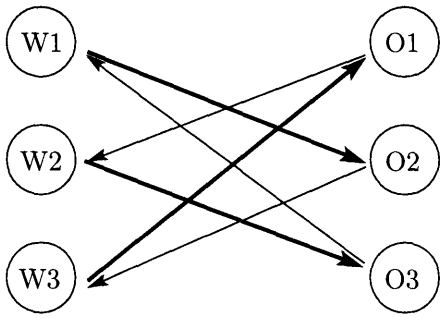

Figure 3-14: Augmenting Cycle – Example 3.4.2

increases one unit of flow on the forward arcs (warehouse to order) and decrease one unit of flow on the backward (order to warehouse) arcs. After we augment the cycle from the initial solution, we reach the optimal solution of the transportation problem of Y. We can interpret the augmenting cycle in the context of the problem: a backward arc (j, k) represents un-assigning the inventory unit of Y that's currently assigned to order j into warehouse k; a forward arc (k, j) represents assigning an inventory unit of Y in warehouse k to order j. We now can see the cyclic exchanges that are required to implement the solution: we un-assign the unit of Y at W1 to O3, and re-assign it to O2; we un-assign the inventory from W3 to O2, and re-assign it to O1; we un-assign the inventory from W2 to O1, and re-assign it to O3. That is, by implementing the cyclic exchange of SKU Y according to O3 → O2 → O1 → O3, we arrive at the solution in Figure (3-13). The term *SKU Exchange* is named in view of the one or more cyclic exchanges in each transportation problem.

With the discussion of Example 3.4.2, we are now ready for a general specification of SKU Exchange. Let's first define an admissible order and shipment.

Definition 3.4.2. Split order j is an *admissible order* of SKU a if

1) order j has SKU a in a single shipment, and

2) order j has only one single shipment of SKU a.

Then, that single shipment is an *admissible shipment* of SKU a.

That is, each admissible order of SKU a only has one admissible shipment. Let the notation $\{X, YZ\}$ represent an order having three items, one unit of each SKU X, Y, Z. The unit of X is currently committed at warehouse 1, and the units of Y and Z are committed at warehouse 2. Then order $\{X, YZ\}$ is an admissible order of X. In addition, order $\{X, YX\}$ is also an admissible order of X but order $\{X, X\}$ is not an admissible order of X.

We construct a maximization transportation problem for each relevant SKU according to Figure (3-15). Below we describe the details of the transportation problem for SKU a.

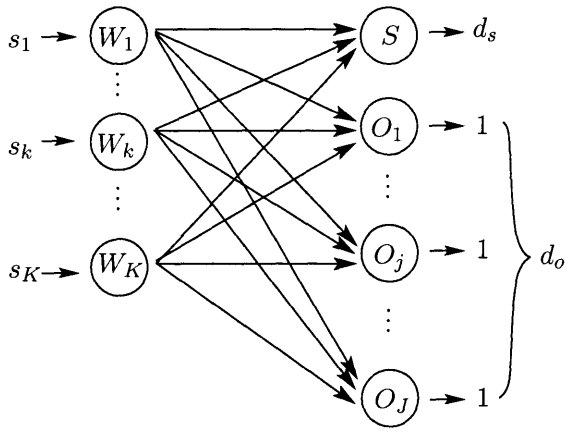

Figure 3-15: Transportation Problem of a SKU

Supply: Each supply node represents a warehouse. The supply available at warehouse W_k is s_k, representing the number of single orders (including uncommitted inventory) of a as well as the number of admissible shipments of a in admissible orders, all currently assigned to warehouse W_k.

Demand: Demand node S represents the single orders, and d_s is the total number of single orders of SKU a. All other demand nodes represent an admissible order of a, which is d_o in sum.

Arcs: We permit arcs from every supply node to every demand node.

Costs: The cost of every arc to demand node S is zero, since there is no reduction in the number of shipments from any re-assignment of a single order. The profit of an arc to an admissible order O_j from its profitable warehouse $W_k \in P_j(a)$ is 1, since this would reduce one shipment. The cost of an arc to O_j from all other warehouses is zero.

Observation. Note that if the supply or demand of every node is exactly one, we have an assignment problem.

As we mentioned earlier in the section, one major practical advantage of the heuristics is the ability to solve and implement one cyclic exchange at a time while maintaining a feasible solution at each iteration. While the major objective is to minimize the number of shipments, we also want to do so without implementing any unnecessary exchanges. That is, we have another objective of minimizing the number of exchanges from the initial real-time solution. This objective is easy to add on to the current transportation problem by modifying some arc costs. As denoted before, let j be a demand node and k be a supply node. Arc (k, j) has an additional profit of $n\epsilon$ if there are n committed inventory units currently in warehouse k and whose associated demand node is j. The value of ϵ is specified as $0 < \epsilon < 1$. To ensure the objective of minimizing number of exchanges is a secondary objective, we set the value of ϵ as $\epsilon < \frac{1}{m}$, where $m = \sum_{k=1}^{K} s_k$ is the total number of supply units in the transportation problem.

In a realistic setting, we want to minimize the total transportation costs, instead of the number of shipments. Recall we approximate the transportation costs by the number of shipments, because the items we consider tend to have small weight and thus have a significant fixed cost in the shipment. Note that we can naturally extend the transportation problem to incorporate actual transportation costs. Let's illustrate the cost modification with a previous example, Example 3.4.2. Consider arc (k, j), let $c_j(k)$ be the transportation cost of shipping a unit of SKU Y from warehouse k in order j. For example, if $j = 1$, then $c_1(3)$ is the transportation of shipping one package containing one unit of SKU Y and Z from warehouse 3. Let warehouse k_j be the current location of SKU Y in order j. Then, we set the profit of arc (k, j) as $c_{kj} = c_j(k_j) - c_j(k)$.

SKU Exchange with double shipments

In the real data sets we have examined, a large percentage of the split orders in the real-time assignment (over 85%) have at least one single shipment. A larger percentage (94%) of the split orders have at least one single or double shipment. To include more orders in the SKU Exchange heuristic, we can also incorporate all orders with double shipments. We start the discussion with the following example.

Example 3.4.3. We consider a batch of orders with real-time assignment in Figure 3-16.

Here the only change from Figure (3-11) is that the first order has a double shipment of YX instead of a single shipment of Y. Similarly, we implement a SKU Y transportation problem for the batch of orders, which all have a single or double shipment of Y. The resulting transportation problem, as shown in Figure (3-17), resembles Figure (3-12). Again,

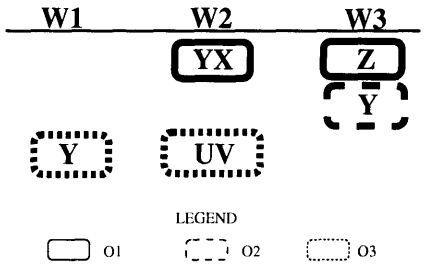

Figure 3-16: Real-Time Assignments – Example 3.4.3

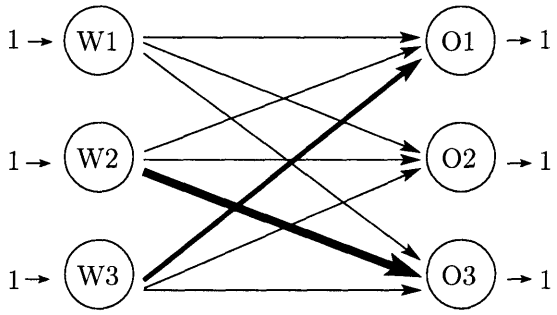

Figure 3-17: Transportation Problem for SKU Y – Example 3.4.3

the darkness of the arcs indicates the relative amount of arc profit. The only difference is the arc cost of arc $(3,1)$. In Figure 3-12, one unit of flow on arc $(3,1)$ indicates that we supply one unit of SKU Y in warehouse 3 to order 1, which reduces one split. However, in the current example, such unit of flow on the arc does not reduce one split. The split can be reduced only if the unit of SKU X can also be supplied from warehouse 3. Therefore, in our current example, we set the arc cost of $(3,1)$ to be, p_{31}, the probability that SKU X can also be shipped from warehouse 3. There are many ways one could estimate such probabilities. For instance, we could estimate the probabilities based on the expected outcome of solving the transportation problem for SKU X. For the sake of convenience, we simply set $p_{31} = 0.5$.

In summary, to include double shipments in the heuristics, we augment the definition of admissible orders and shipments.

Definition 3.4.3. Split order j is an *admissible order* of SKU a if

1) order j has SKU a in a single shipment or a double shipment, and

2) order j has only either either single or double shipment which has only one unit of SKU a.

Then, that shipment is an admissible shipment of SKU a.

For example, the order $\{Y, XZ\}$ is an admissible order of X but $\{X, XZ\}$ is not an admissible order of X.

The transportation problem is still defined as in Figure 3-15 with some slight augmentations in arc costs.

Costs: The profit of an arc to an admissible order O_j from its profitable warehouse $W_k \in P_j(a)$ is 1 if the admissible shipment of O_j is a single shipment; 0.5 if the admissible shipment of O_j is a double shipment.

There are other ways to include orders with double shipments. For example, we can treat each double shipment as a "special" SKU, and formulate a transportation problem for such "special" SKU. We find that such a heuristic is not very efficient in comparison, because the number of double shipments can be very large and the number of orders having a particular double shipment is very small.

SKU Exchange with time dimension

The actual transportation problem that we need to solve is a bit more complex. We augment the transportation problem with a time dimension. Recall we have T time periods with respect to the snapshot date. We outline the general idea in the following and discuss the implementation details in the next section "Implementation".

Figure 3-18 is a condensed representation of a transportation problem, where each square block contains a collection of nodes. Figure 3-19 is a full-scale representation of a transportation problem. In both representations, arcs are drawn from blocks to blocks, which indicates that every node in a block has arcs going to every node in a connected block.

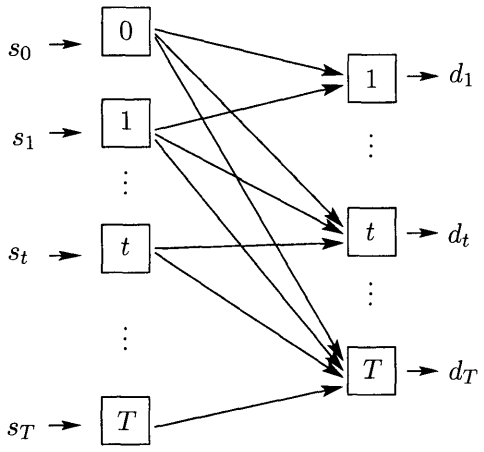

Figure 3-18: Condensed Representation of a Transportation Problem

Supply: We have $T + 1$ supply blocks, where each block contains a supply node for each warehouse as in the problem with no time dimension in Figure 3-15. Each supply node of warehouse k in time block t has a supply of s_{kt}. The supply available at all

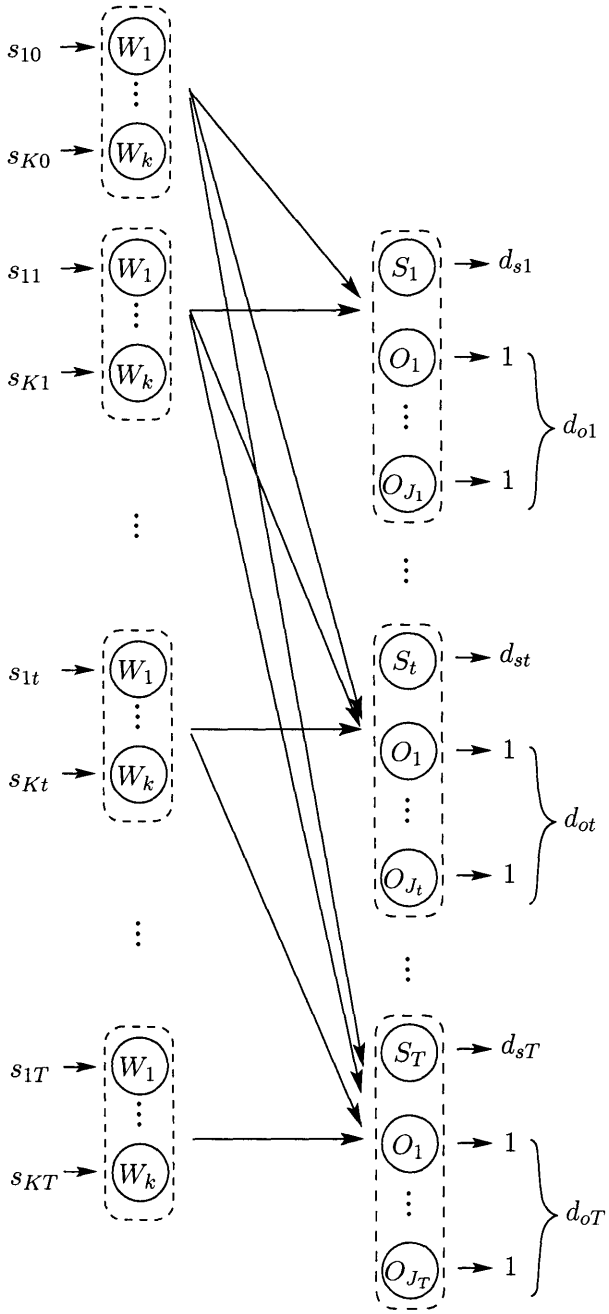

Figure 3-19: Transportation Problem of One SKU (with Simplified Arcs)

warehouses for the snapshot time block, $s_0 = \sum_{k=1}^{K} s_{k0}$, reflects the on-hand inventory, whereas the supply for future time blocks, $s_t = \sum_{k=1}^{K} s_{kt}$, $t > 0$, is the on-order inventory that will arrive during the time block.

Demand: We have T demand blocks, one for each promise-to-ship date category. Each demand block contains order nodes as in Figure 3-15. Demand block t has a single-order node S_t with d_{st} number of single orders whose promise-to-ship date is t. Each admissible shipment of SKU a with promise-to-ship date t has a node in demand block t.

Arcs: We permit arcs from the nodes with nonzero supply in supply block t_2 to the demand nodes in demand block t_1, $\forall\ t_1 \geq t_2$.

3.4.3 Worst-Case Analysis

In this section, we analyze the closeness of our heuristics to the optimal solution. We perform worst-case analysis. First recall the following notations in the problem of minimizing number of shipments.

K	number of warehouses
J	number of customer orders
u_j	number of items (not SKUs) in order j

We also define a constant for each order H_j as the maximum number of shipments order j can possibly have: the minimum of number of warehouses or number of items,

$$H_j = \min(K, u_j).$$

We denote H accordingly as the ratio of maximum number of shipments over the minimum number of shipments the J customer orders can possibly have,

$$H = \frac{\sum_{\forall j} H_j}{J}.$$

Let's recall the definition of approximation algorithms.

Definition 3.4.4. A *ρ-approximation algorithm* for a problem $\mathcal{P}$ is a polynomial-time algorithm that returns a feasible solution to $\mathcal{P}$ of cost within a factor of ρ of the optimal cost. If $\mathcal{P}$ is a minimization problem, then $ALG \leq \rho \cdot OPT$, where ALG is the cost of an algorithm and OPT is the cost of the optimal.

That is, we want to find a polynomial-time algorithm that is "close" to the optimal. Closeness is measured by a guaranteed factor of the optimal. Essentially, ρ is the constant

upper bound of $\frac{ALG}{OPT}$.

Remark. In the problem of minimizing the number of shipments given orders and warehouse supplies, any algorithmic solution is at most $H \cdot OPT$.

In other words, this problem has an obvious worst-case constant factor bound for any algorithm.

In the real data, we observe certain properties in the customer orders. Let Z be a random variable representing the size of an order. We find that Z is geometrically distributed with the probability of an order having size k, $p_k = q(1-q)^{k-1}$, $k = 1, 2, \ldots$, where $0 \leq q \leq 1$ is the parameter of Z. We can find the constant factor H for such set of orders.

Proposition 3.4.1. *In the problem of minimizing number of shipments given J orders whose sizes are geometrically distributed with parameter q and supplies in K warehouses, the worst-case bound for any algorithmic solution, given J is large, is*

$$\frac{1-(1-q)^K}{q} \leq \frac{1-e^{-K/q}}{q}.$$

Proof. Given H is defined as $\frac{\sum_{\forall j} \min(K,\ u_j)}{J}$, we can write H as

$$H = \sum_{k=1}^{\infty} \frac{n_k\ \min(K,k)}{J},$$

where n_k is the number of order with $u_j = k$. We let $p_k = \frac{n_k}{J}$ where p_k is the percentage of orders having order size k. We then have

$$H = \sum_{k=1}^{\infty} p_k \min(K,k).$$

Given J is very large and the orders are geometrically distributed in size, p_k can be seen as the PMF of Z. We have

$$\begin{aligned}
E[Z] &= \sum_{k=1}^{\infty} k\ p_k \\
&= \sum_{k=1}^{K} k p_k + \sum_{k=K+1}^{\infty} (K + (k-K))\, p_k \\
&= H + \sum_{k=K+1}^{\infty} (k-K) p_k, \quad \text{where } p_k = q(1-q)^{k-1} \\
&= H + \sum_{j=1}^{\infty} j\ (1-q)^K p_j \\
&= H + (1-q)^K E[Z].
\end{aligned}$$

Therefore, since $E[Z] = \frac{1}{q}$,

$$H = E[Z]\left(1 - (1-q)^K\right) = \frac{1-(1-q)^K}{q} \leq \frac{1-e^{-K/q}}{q} \qquad \blacksquare$$

We see that the bound is small for large values of q, which is obvious since a large q indicates that there is a large proportion of single orders.

To consider only the multi orders, we denote Z' as the size of a multi order. We start with a Lemma.

Lemma 3.4.1. *Let X be a discrete random variable with PMF p_k, $k = 1, 2, \ldots$ and X' be a related random variable such that $P(X' = 1) = 0$ and $P(X' = k) = c\ p_k$, $\forall k \geq 2$ and constant c. Then, we have*

$$E[X'] = E[X] + 1$$

iff X is geometrically distributed.

Proof. Suppose X is geometrically distributed and we define the expectation of X' as

$$\begin{aligned} E[X'] &= \sum_{x=0}^{\infty} P(X' > x) \\ &= P(X' > 0) + \sum_{x=1}^{\infty} P(X > x | X > 1) \\ &= 1 + \sum_{y=0}^{\infty} P(X > y+1 | X > 1) \\ &= 1 + \sum_{y=0}^{\infty} P(X > y) \\ &= 1 + E[X]. \end{aligned}$$

The above equation is a result of the memoryless property of geometric distribution. Since the geometric distribution is the only discrete distribution that has as the property, X is geometrically distributed if the $E[X'] = 1 + E[X]$. $\blacksquare$

Corollary 3.4.1. *In the problem of minimizing number of shipments in multi orders given orders whose sizes are geometrically distributed with parameter q and supplies in K warehouses, the worst-case bound for any algorithmic solution, given the number of orders is large, is $1 + \frac{1-e^{-(K-1)/q}}{q}$.*

Proof. We have $E[Z'] = E[Z] + 1$, and $H = \sum_{k=2}^{\infty} \frac{p_k}{1-p_1} \min(K, k)$ because we are only

considering multi-item orders, where p_k is the PMF of Z. Then,

$$E[Z'] = H + \frac{1}{1-p_1}\left((1-q)^K E[Z]\right).$$

As a result,

$$H = 1 + \frac{1}{q} - \frac{(1-q)^K}{q(1-q)} = 1 + \frac{1-(1-q)^{K-1}}{q} \leq 1 + \frac{1-e^{-(K-1)/q}}{q} \qquad \blacksquare$$

We will show the following sections that Order Swap and SKU Exchange can perform arbitrarily bad. That is, this worst-case bound is tight for both heuristics. However, the typical real data we observe does not exhibit the worst-case characteristics.

Order Swap

Order Swap performs well when there is a large amount of single orders or unassigned inventory, especially at the warehouses where the inventory of a large subset of SKUs are carried. This heuristic aims entirely at reducing all split orders to one shipment. However, it ignores reducing split orders with more than $n > 2$ shipments to orders with $2, \ldots, n-1$ shipments. In other words, the worst case of the heuristics can be extremely poor. We will show with the following tight example that Order Swap can perform arbitrarily bad. Following the previous convention, let $\{A, B, C\}$ be an order with a single shipment of A in warehouse 1, a single shipment of B in 2, a single shipment of C in 3.

Example 3.4.4. There are 3 warehouses and 3 SKUs A, B, and C. Suppose that we have three orders ABC, where each order has one unit of each SKU. Let their initial assignments be {A, B, C}, {B, C, A}, and {C, A, B}.

Since there are no single orders in any warehouses, Order Swap does not improve upon the initial solution. However, the optimal solution has 3 shipments ($OPT = 3$) for the three orders: assigning one order of ABC to each warehouse. The heuristic solution has 9 shipments, equivalent to $3 \cdot OPT$. We also have $H = K = 3$. Therefore, the worst-case bound is tight for Order Swap, $\frac{ALG}{OPT} = H$. Since Order Swap relies on the abundance of single orders, it is obvious that Order Swap can perform arbitrarily bad in the absence of single orders.

Now let's suppose that there is ample amount of single orders. That is, let's suppose that if warehouse k stocks SKU i, then at any time it has infinite amount of uncommitted inventory of SKU i; if warehouse k does not stock SKU i, then at any time it has zero amount of SKU i. The problem of minimizing the number of shipments is still NP-hard. It basically reduces to minimizing the number of shipments for each order, which is a set cover problem.

Proposition 3.4.2. *Assume that there is an infinite amount of inventory at the warehouses for the SKUs that are in stock. That is, let I be the set of SKUs and $I_k \subset I$ be the subset of SKUs carried by warehouse k. Then, the supply of SKUs in I_k at warehouse k is infinitely large. The Order Swap heuristic solution is at most $\frac{H}{2} \cdot OPT$ for multi orders.*

Proof. The supply constraint in $\mathcal{MIP}$ is $\sum_j x^i_{jk} = s^i_k$, which is the supply at warehouse k of SKU i. If $s^i_k \to \infty$, the constraint is eliminated. If warehouse k does not stock SKU i, $s^i_k = 0$, then $x^i_{jk} = 0$ for all orders. Therefore, the supply constraint can be eliminated entirely. The problem can be decoupled by orders. We consider orders individually, and we want to find the largest ratio of $\frac{ALG}{OPT}$.

For an order j, we claim that the worst-case bound is not tight. For the worst-case bound H_j to be tight, in the worst case, the optimal solution has one shipment and the heuristics have H_j shipments for each order. However, if the optimal solution has one shipment so does the heuristic solution. Because if the optimal has one shipment at warehouse k, then warehouse k must stock all of the SKUs in order j. Since there is infinitely amount of uncommitted inventory for the SKUs in order j, then the heuristic solution can also ship the order from warehouse k. Therefore, H is not tight. In the worst case, the optimal solution has to have more than one shipment for each order. As a result, the largest bound that can be constructed is $\frac{H}{2} \cdot OPT$: the heuristic solution having H shipments and the optimal solution having 2 shipments. ■

We show with the following example that this bound is tight.

Example 3.4.5. Let the set of SKUs be $\{A, B, C\}$. Suppose we stock $I_1 = \{A, B\}$ in warehouse 1, $I_2 = \{B, C\}$ in 2, and $I_3 = \{A, C\}$ in 3. That is, warehouse 1 only stocks SKU A and B but not C. Suppose there are n identical order of ABC (one unit of each SKU) and their initial assignments have three shipments for each order.

Since no warehouse can ship all three SKUs in the order, we cannot improve upon the initial solution using Order Swap. Therefore, the heuristic solution has $3n$ shipments. In the optimal solution, we can ship two out of three SKUs from one warehouse. Therefore, the optimal solution has $2n$ shipments. The heuristic solution is $\frac{3}{2} \cdot OPT$. Since $H = 3$, we show that the bound of $\frac{H}{2}$ is tight.

Corollary 3.4.2. *Given infinite supply for stocked SKUs and orders are geometrically distributed, the worst bound for multi orders is* $\frac{1}{2} + \frac{1-(1-q)^{K-1}}{2q} \leq \frac{1}{2} + \frac{1-e^{-(K-1)/q}}{2q}$.

Recall q is the parameter for order size Z. For instance, with values of $q = 0.5, K = 5$, the bound is 1.44.

We have shown that the Order Swap heuristic can perform arbitrarily bad. In particular, the worst cases take place when single orders are scarce and the orders are large in size. However, Order Swap performs well in practice because

- majority of the split orders have two shipments in the real-time assignment,
- ample amount of single orders and uncommitted inventory.

SKU Exchange

The SKU Exchange heuristics complements the Order Swap heuristics. SKU Exchange considers reducing split orders with $n > 2$ shipments to $2, 3, ..., n-1$ shipments, whereas Order Swap does not. The Order Swap heuristics also complements the SKU Exchange. For a split order considered in SKU Exchange, each shipment of the order can only be re-assigned to a warehouse where the order occupies in the real-time assignment. We show with the following tight example that the $H \cdot OPT$ bound is tight.

Example 3.4.6. Suppose we have three warehouses and two SKUs A, B. There is only one order O1 of AB. The initial assignment for O1 is {A, B, }. That is, there are two shipments with A in warehouse 1 and B in warehouse 2. W1 only stocks SKU A, W2 only stocks SKU B, and W3 stocks both SKU A, B.

When applying SKU Exchange heuristic to the example, we cannot improve upon the initial solution. Since we only consider moving SKU A to W2 and SKU B to W1, but only W3 can satisfy O1. Here $H = 2$. The heuristic solution has 2 shipments and the optimal solution has 1 shipment. Therefore, the bound of $H \cdot OPT$ is tight.

Even given infinity amount of supply for the stocked SKUs, the above example still shows that $H \cdot OPT$ is tight. The SKU Exchange can perform arbitrarily bad. The SKU Exchange procedure performs poorly when the warehouses currently occupied by the order do not carry the entire or large portion of the SKUs in the order. Order Swap complements that shortcoming by considering shipping the order from warehouses other than it currently occupies. However, SKU Exchange replies much less on the abundance of single orders or uncommitted inventory than Order Swap.

3.4.4 Generate Exchanges

Here we analyze the running time of the transportation problem in SKU Exchange and argue why solving a transportation problem is preferred than solving an assignment problem. In addition, the transportation problem only gives the number of extra shipments reduced but does not give the specific exchanges. We discuss the details of finding the implementable exchanges. We start with some notations. Here is the input data to the transportation

problem for each SKU a.

n_p	no. of admissible orders of SKU a
n_g	no. of single orders with SKU a
K	total no. of warehouses
T	no. of time periods
n_1	no. of supply nodes in transportation problem a
n_2	no. of demand nodes in transportation problem a

Since we have one demand node in each time period for *all* single orders, and one demand node for *each* admissible orders, we have

$$n_2 \leq n_p + T.$$

Clearly, the maximum number of supply nodes is in the order of KT. Therefore,

$$n_1 \leq KT.$$

Let n be the number of nodes in a graph. Here $n = n_1 + n_2$, then,

$$n \leq n_p + T(K+1)$$

In addition, we can set the maximum arc cost as $C = 1$, and the maximum arc capacity as $U \leq n_p + n_g$. There are many efficient minimum cost flow algorithms that can be applied to the transportation problem. To state as an example, we can apply the Primal-Dual Algorithm with a running time of $O(\min(nU, nC))$ [AMO93].

Remark. We have a running time of $O(n_p + TK)$ for the transportation problems.

Solving the transportation problem alone does not give the exact inventory exchanges for orders. In practice, we need to find the exact exchanges. We perform a simple procedure to extract the exact exchanges from the transportation problem solution.

Example 3.4.7. We have a transportation problem for SKU a in Figure 3-20, where only the nonzero-cost arcs are indicated. We want to find the exact exchanges after solving the transportation problem.

A dark arc (k, j) in Figure 3-20 has a profit of 1, indicating that warehouse k is a profitable warehouse of order j. A light arc (k, j) in the figure has a profit of $\epsilon \ll 1$, indicating that the unit of SKU a in order j is currently assigned at warehouse k. The zero-cost arcs are not drawn in the figure. As we defined perviously, nodes in supply block t_1 have arcs to nodes in demand block t_2 only if $t_1 \leq t_2$.

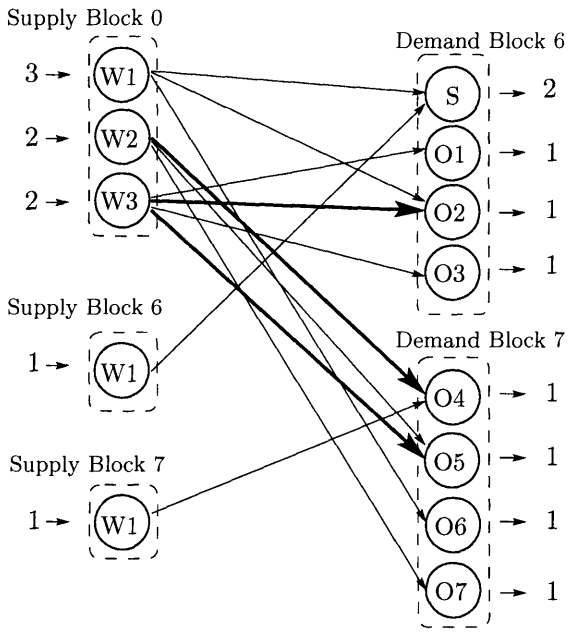

Figure 3-20: Example 3.4.7, Only Zero-Cost Arcs

We see that the supply of warehouse 1 in time block 0 is $s_{10} = 3$. From the light arcs, we know that s_{10} includes the inventory currently assigned to demand node S_1, O2, O6, where S_1 is a single order included in the node S. We can arrive at the same conclusion for the other supply nodes. If we push one unit of flow on each of the light arcs, we obtain the current solution. Figure 3-21 is an optimal solution of the transportation problem, where each drawn arc has one unit of flow and each dotted arc has a cost of 0.

To find the exact exchanges from the optimal solution, we need to find augmenting cycles to reach the optimal solution from the current solution, e.g., in Example 3.4.2. We can also do that for the optimal solution by inspecting Figure 3-21. Table 3.2 is another representation of the initial and optimal solution. The bolded orders represent orders whose assignments have not been changed by solving the transportation problem. For example,

(k,t)	Supply	Initial Soln.	Optimal Soln.
(1,0)	3	$\mathbf{S_1}$, O2, **O6**	$\mathbf{S_1}$, O1, **O6**
(2,0)	2	O5, O7	O3, O4
(3,0)	2	O1, O3	O2, O5
(1,6)	1	$\mathbf{S_2}$	$\mathbf{S_2}$
(1,7)	1	O4	O7

Table 3.2: Example 3.4.7, Changes in the Initial and Optimal Solution.

for supply $(1,0)$ in warehouse 1 time block 0, there are three unit of inventory of SKU a. In the initial solution, the three units were assigned to order O2, O6 and S_1. In the optimal solution, they are assigned to order S_1, O1, and O6. We can see that the assignment of

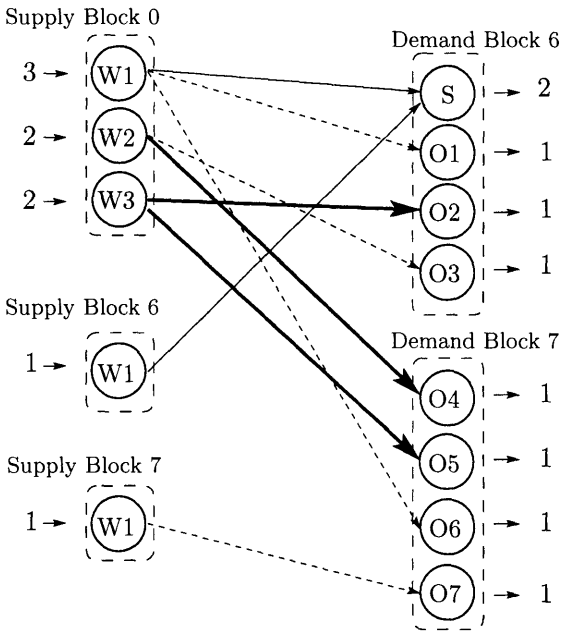

Figure 3-21: Example 3.4.7, an Optimal Solution

O6 and S_1 have not been changed. However, the unit that was assigned to O2 is now assigned to O1. Another example is supply $(2,0)$. The two unit of supply were assigned to O5 and O7 but now are assigned to O3 and O4. *We use the notation $O_i \rightarrow O_j$ to represent un-assigning the unit of SKU a from order O_i and re-assigning it to O_j.* Then, the re-assignment of either O5 $\rightarrow$ O3, O7 $\rightarrow$ O4 or O5 $\rightarrow$ O4, O7 $\rightarrow$ O3 is a result of the optimal solution.

By examining each supply (k,t), we can find the exact exchanges. One way to realize the exchange is: O2 $\rightarrow$ O1 $\rightarrow$ O2, O7 $\rightarrow$ O3 $\rightarrow$ O5 $\rightarrow$ O4 $\rightarrow$ O7. In short, we have a swap between O2 and O1 and a cyclic exchange between O7, O3, O5, O4.

We summarize the process in Figure 3-22. Let $I_1(k,t)$ be the list of orders assigned to the supply in (k,t) in the initial solution, and $I_2(k,t)$ in the optimal solution. Note that the procedure always return cyclic exchanges, because every order is included in the initial as well as the optimal solution.

This process takes $O(n_p+n_g)$ running time, and we have the exact exchanges thereafter. That is, we examine each unit of inventory once, and for each unit we look up if the unit can be assigned to itself (O(1) time if the information is stored in a hash table).

Remark. The running time of solving the transportation problem and finding the exact exchanges thereafter is of $O(n_p + n_g + TK)$.

The transportation problem has the flavor of an assignment problem. Certainly, when all demand and supplies are of one unit, the transportation problem is an assignment problem. An alternative algorithm is to formulate a bipartite weighted matching problem, which is

algorithm Find Exchanges:
input: I_1, I_2

1. for each (k,t) **do,**

1.1 remove order j if $j \in I_1(k,t), j \in I_2(k,t)$.

1.2 while $I_1(k,t) \neq \emptyset, I_2(k,t) \neq \emptyset$ **do,**

1.2.1 assign $i \rightarrow j$ s.t. $i \in I_1(k,t), j \in I_2(k,t)$,
1.2.2 remove i, j from $I_1(k,t), I_2(k,t)$.

Figure 3-22: Finding Exact Exchanges

the assignment problem. In this case, $n_1 = n_2 = n_p + n_g$, $C = 1$, and $U = 1$. Every unit of inventory of SKU a is listed as a supply node as well as a demand node. We would have the exact exchange after solving the assignment problem. The running time of the Hungarian Algorithm, which is the a direct implementation of the Primal-Dual Algorithm for the minimum cost flow problem, is $O(n_1 S(n,m,C))$ [AMO93], where $S(n,m,C)$ is the running time of solving a shortest path problem with number of nodes n and arcs m. The shortest path problem is on the order of $O(n^2)$.

Remark. The corresponding assignment problem has a running time complexity of $O(n_1^3) = O((n_p + n_g)^3)$.

Observation. In the case that the number of warehouses K and the number of time buckets T are small, the transportation problem performs significantly faster than the assignment problem.

3.5 Implementation

In this section, we describe in details the data and heuristic requirement in the implementation.

3.5.1 Data and Parameters

At the placement of a customer order, the e-tailer quotes a promise-to-ship date to the customer. Since items in the order may be shipped from different warehouses and/or in different times, we assume that each item i in customer order j is given a promise-to-ship date v_i in the real-time assignment. In addition, a unit of inventory is assigned to item i in real time and its inventory status is given as z_i. Specifically,

$$z_i = \begin{cases} 0, & \text{if the inventory assigned to item } i \text{ is physically available} \\ 1, & \text{if the inventory assigned to item } i \text{ is on order} \end{cases}$$

For example, $z_i = 1$ if a customer orders a book that's not yet been published.

In summary, we require the vector $\boldsymbol{A_i} = (\boldsymbol{a_i}, \boldsymbol{j_i}, \boldsymbol{k_i}, \boldsymbol{v_i}, \boldsymbol{z_i})$ for each item i as the real-time assignment in the re-evaluation problem. The elements of vector A_i represent the SKU a_i, order j_i, assigned warehouse k_i, promise-to-ship date v_i, and assigned inventory status z_i. We view the A_i's as the input data to the heuristic procedure.

In addition to the input data, we make the following assumptions on the operations and services.

Inventory replenishment We define $\boldsymbol{u_i}$ as the time at which the inventory assigned to item i would arrive at warehouse. Given no additional inventory replenishment information, we set $u_i = z_i v_i$. That is, on order inventory would arrive on the promise-to-ship date. We consider u_i to be the supply date and v_i to be the demand date of item i.

Order promise-to-ship date Since the promise-to-date is a promise by items or shipments, we need to define the promise-to-ship date for the order: let $\boldsymbol{V_j}$ be the earliest promise-to-ship dates of all items in the order, $V_j = \min_{i:j_i=j} v_i$.

Time dimension We create T time periods with respect to the snapshot date. We define time period 0 as the time at which the snapshot was taken. Time period t, $0 < t < T$, is t days in the future of the snapshot date, and time period T is T days or more in the future of the snapshot date. As a result, we set the range of the promise-to-ship date v_i to be $1, 2, ..., T$. Since we may be solving the re-evaluation problem periodically, those orders will eventually be in the specific day category. Accordingly, the range of u_i is $0, 1, ..., T$, and $u_i = 0$ represents the assigned inventory is physically in the warehouse in the real-time assignment. The value of T is an adjustable parameter in the heuristics.

Split Order j is *not* a split order if 1) all items in the order are to be shipped from one warehouse, and 2) all items can be shipped before the promise-to-ship date of the order, $u_i \leq V_j$, $\forall i : j_i = j$.

Shipments The number of shipments in an order j is defined by the following. Items in the same warehouse and can be shipped before the order promise-to-ship date, $u_i \leq V_j$, are in the same shipment; items in the same warehouse and having the same value promise-to-ship date of $u_i > V_j$ are in the same shipment.

3.5.2 Order Swap

The Order Swap heuristic follows the general algorithm listed in Figure 3-7. In the following, we specify the detailed rules of swapping.

Recall it is feasible to swap item i_1 and i_2 of the same SKU iff $u_{i_1} \leq v_{i_2}$ and $u_{i_2} \leq v_{i_1}$. That is, the customer service level promised at the real time cannot be violated after the swap. To be more specific about the swap of item i_1 in split order j and i_1 of single order at k, we examine such different scenarios. For item i_2, because of the relationship of $u_{i_2} = z_{i_2} v_{i_2}$, where $z_{i_2} \in \{0, 1\}$, we have the following scenarios:

1) if $z_{i_2} = 0$, then $u_{i_2} = 0 < v_{i_2}$

2) if $z_{i_2} = 1$, then $u_{i_2} = v_{i_2} > 0$

However, for item i_1, because of $v_{i_1} = V_j$ by construction, we have scenarios:

1) if $z_{i_1} = 0$, then $u_{i_1} = 0 < V_j$

2) if $z_{i_1} = 1$ and $v_{i_1} = V_j$, then $u_{i_1} = V_j > 0$

3) if $z_{i_1} = 1$ and $v_{i_1} > V_j$, then $u_{i_1} > V_j > 0$

Considering all the scenarios, we have

Remark. In Order Swap, it is profitable to swap item i_1 in a split order j with item i_2 of single order in warehouse k, for all i_1 in j, if i_1 and i_2 satisfy the following condition:

1) if $u_{i_1} = 0 < V_j$, then $u_{i_2} = 0$ or $u_{i_2} = v_{i_2} \leq V_j$

2) if $u_{i_1} = V_j > 0$, then $u_{i_2} = 0, v_{i_2} \leq V_j$ or $u_{i_2} = v_{i_2} = V_j$

3) if $u_{i_1} > V_j > 0$, then $u_{i_2} = 0$ and $v_{i_2} \leq v_{i_2}$

If there are many such item as i_2 in the above discussion, we may start with the items that have later supply date u_{i_2}.

3.5.3 SKU Exchange

Recall that in SKU Exchange, we solve a transportation problem for each SKU. We have a demand node for each admissible split order. For each such order j, we denote C_j to be the collection of shipments in order j. Each shipment is defined by the warehouse assigned at the real time as well as the promise-to-ship date, $C_j = \{(t_1, w_1), ..., (t_n, w_n)\}$.

For each SKU a in an admissible shipment (t, w), we represent the order in transportation problem a as a demand node in demand block t.

Remark. For a demand node in block t representing an admissible shipment (t, w) in order j, profitable arcs are from warehouse w_1 in supply block t_1 such that $t_1 \leq t$ and

1) if $t_1 \leq V_j$, then $(t_1, w_1) \in C_j \backslash (t, w)$

2) if $t_1 < V_j$, then $(V_j, w_1) \in C_j \backslash (t, w)$

We set the profit to be $c_1 = 1$ if the shipment (t, w) is a single shipment and $c_2 = 0.5$ if a double shipment. In addition, we set the profit from supply node (t, w) to be ϵ to minimize the number of exchanges w. r. t. to the real-time assignment.

3.6 Computations

We implement the heuristics on several real data sets from a global e-tailer. To examine the sub-optimality of the heuristics on the data, we would like to benchmark the heuristic solutions with the optimal solutions. However, the entire re-evaluation problem is too large to be solved optimally in Cplex due to the lack of computer memory. Instead, we extract a much smaller test data set to benchmark the heuristic solutions. In this section, we first discuss the results from the test data sets, and then present the benefits in the real data sets. Here all results from Order Swap utilize the version with a time dimension; all results from SKU Exchange consider the single and double shipments with a time dimension.

3.6.1 On Test Data

The snapshot data of all not-yet-picked orders include all orders that were placed before the snapshot time. We extract a subset of orders which were placed on the day of the snapshot date. The resulting orders are the reduced test data. The reduced test data sets generally have 110-120K orders, 100K SKUs, and 7 warehouses. The reduced re-evaluation problem can be solved with *no time dimension* in Cplex within minutes [AS04]. That is, we ignore the promise-to-ship date for each order in the test data.

Table (3.3) displays the summary of a few test data sets. We observe that overall the data sets are very similar. The columns *"Single orders", "Multi orders", "Split orders"* are as defined in §3.3. The column *"Split"* represents the number of splits or extra shipments in the data, which is the number of shipments minus the number of orders.

Data Set	Single orders	Multi orders	Split orders	Split
A	56.7K	62.7K	10.5K	11.3K
B	55.2K	65.0K	10.2K	10.9K
C	52.4K	61.8K	9.6K	10.2K
D	45.8K	54.8K	8.1K	8.6K

Table 3.3: Test Data

We implemented the heuristics on the test data sets and Table (3.4) displays the results. In Table (3.4), *Algorithm 0* is the optimal solution from Cplex, *Algorithm 1* is the Order Swap procedure, and *Algorithm 2* is the combined Order Swap and SKU Exchange procedure. The column *"Shipments"* represents the number of shipments reduced from the real-time assignments. Since we minimize the number of shipments in the re-evaluation

Data Set	Algorithms	Shipments	(%) of Opt.	(%) of Splits	Time (sec)
A	0	6074	100	54	300
	1	5466	89.9	48.5	12
	2	5862	96.5	52.0	188
B	0	5836	100	53.3	300
	1	5175	88.6	47.3	12
	2	5669	97.1	51.8	150
C	0	5566	100	54.4	300
	1	4984	89.5	48.9	8
	2	5407	97.1	52.8	122
D	0	4435	100	51.3	300
	1	3974	89.6	46	8
	2	4371	98.6	50.6	162

Table 3.4: Heuristic Results on Test data

problem, it is equivalent to maximize the number of shipments reduced from the real-time assignments. The column *(%) of Opt.* is the reduced shipments in the heuristic solution as a percentage of the reduced shipments in the optimal solution. The column *(%) of Splits* is the reduced shipments in the heuristic solution as a percentage of the total extra shipments in the real-time assignments. The running time of the optimal is approximately 300 seconds.

We implement the heuristics on UNIX machines with 1.5 GHz processors and 1GB RAM. We extract the necessary data using a text processor Perl. For each SKU in the sequence, we solve each transportation problem in Cplex. We then update the affected orders back in Perl. Overall, the Order Swap runs within seconds, and the combined procedure of Order Swap and SKU Exchange terminates within a few minutes. We did not optimize the running time. The running time can be reduced further by streamlining the implementations.

Comparing with the optimal solution, we see that the heuristic procedure Order Swap performs well by itself. This phenomenon can be explained by the large amount of unassigned inventory for many SKUs in the system. However, to reap additional benefits, we need to implement the SKU Exchange heuristic procedure after the Order Swap procedure.

In the SKU Exchange results, we start with a random sequence of SKUs. For each SKUs, we solve a transportation problem. To investigate the impact of how SKUs are sequenced, we perform tests on running the SKU Exchange heuristics based on sorted SKU sequence, e.g., sorting SKUs by the size of their transportation problems (size of nodes or supplies), sorting SKUs by the amount of uncommitted inventory. We find that the sequence of SKUs have little impact on the heuristic results.

In the Order Swap results, we examine the (order, warehouse) pair randomly. We test the heuristics by using different random sequence of orders and warehouses. Again, we find insignificant evidence that the sequence of orders or warehouses affect the heuristic results.

3.6.2 On the Entire Data

Having seen that the heuristics perform well on the test data sets comparing to the optimal solutions, we implement the procedures on the entire data sets. Table (3.5) presents the data summary. Each data set of *A,B,C,D* includes all the not-yet-picked orders from a

Data Set	Single orders	Multi orders	Split orders	Splits
A	600K	314K	34K	38K
B	618K	338K	34K	38K
C	624K	338K	35K	38K
D	625K	329K	33K	36K
E	875K	680K	99K	112K

Table 3.5: Entire Data (Not-Yet-Picked Orders)

snapshot, which is a randomly chosen day during a five month off-season period in 2004. The data set *E* is from a randomly chosen day in the peak season of 2004. We use $T = 12$ time buckets.

We list the heuristic solution on the entire data sets in Table (3.6). The columns are

Data Set	Algorithms	Shipments	(%) of Splits	Time (sec)
A	1	11,028	29	43
	2	15,643	40.9	732
B	1	13,058	34.2	32
	2	19,579	51.3	893
C	1	13,795	35.9	28
	2	20,074	52.2	820
D	1	12,937	35.6	26
	2	19,055	52.4	862
E	1	37,862	34	89
	2	55,408	49.6	2931

Table 3.6: Heuristic Results on Entire Data

defined as in Table (3.4). Since we do not have the optimal solution for the entire data set, we eliminated the rows correspond to *Algorithm 0*. Notice that the number of extra shipments we can reduce from the heuristics ranges from 15K to 20K consistently. Also, note that, for the off-peak data, the number of extra shipments reduced ranges from 40% to 50% of the total number of extra shipments in the real-time assignments. These numbers also resemble those in Table (3.4). Even though we have no optimal solution here to benchmark the heuristic solution, we claim that the data sets are well-behaved and the entire-data solutions resembles the test-data solutions. The number of reduced shipments is more significant for data set *E*, the peak season data. Again, we made no attempts to optimize the running times.

3.6.3 Summary

As the not-yet-picked queue corresponds to orders for one or two days, we expect that we can re-solve the re-evaluation problem in one or two days. Suppose that the off-peak season data resembles our random snapshot data of `A, B, C, D`. We estimate that we can reduce 15K to 20K of extra shipments from the real-time assignments for each re-evaluation problem. Suppose that each extra shipment reduced saves approximately \$1 to \$2. We conjecture that there is a significant opportunity for cost reduction by solving the re-evaluation problems. We estimate that the cost reduction can range from \$2.7 million to \$14.6 million per year.

Our heuristic is relatively easy to implement, as each iteration translates into a series of swaps or cyclic exchanges among a limited set of orders. We can feed these exchanges into the e-tailer's existing order-management systems, and as such, are optimistic that implementation is possible.

We conclude that there is an opportunity to reduce the transportation costs for an e-tailer by means of a re-evaluation of its real-time fulfillment decisions. We have developed a heuristic to do this re-evaluation and shown with preliminary testing that it results in better decisions by utilizing more resources and more information.

3.7 Bounds and Extensions

In this section, we show the lower bounds of the general re-evaluation problem. Recall the following problem is the LP relaxation of the $\mathcal{MIP}$ with variable transportation costs.

$$
\begin{array}{lll}
(\mathcal{LP}) \quad \min & f\sum_{j,k} y_{jk} + \sum_{i}\sum_{j,k} c_{jk}\, x^i_{jk} & \\
\text{s.t.} & \sum_{j\in J_i} x^i_{jk} = s^i_k, & \forall i \in I,\ k \in K_i \\
& \sum_{k\in K(i)} x^i_{jk} = d^i_j, & \forall i \in I,\ j \in J_i \\
& 0 \le x^i_{jk} \le d^i_j\, y_{jk}, & \forall i \in I,\ j \in J_i,\ k \in K_i \\
& y_{jk} \ge 0, & \forall j,\ k
\end{array}
$$

Recall that all uncommitted inventory can be treated as a single order from a dummy customer and infinitely long promise-to-ship date. Therefore, in the first constraint, the equality indicate that all supplies in the warehouses are "used" to satisfy customer orders.

The following problem is the dual of $\mathcal{LP}$ and p^i_k, q^i_j, and r^i_{jk} are the dual variables

associated with the first three constraints in $\mathcal{LP}$, respectively.

$$
\begin{array}{lll}
(\mathcal{D}) \quad \max & \sum_{i\in I}\sum_{k\in K_i} s_k^i\ p_k^i + \sum_{i\in I}\sum_{j\in J_i} d_j^i\ q_j^i & \\
\text{s.t.} & p_k^i + q_j^i - r_{jk}^i \le c_{jk}, & \forall\ i\in I, j\in J_i, k\in K_i \\
& \sum_i d_j^i\ r_{jk}^i \le f, & \forall\ j,k \\
& r_{jk}^i \ge 0, & \forall\ i\in I, j\in J_i, k\in K_i
\end{array}
$$

We can decompose $\mathcal{D}$ by SKUs. That is, given the r_{jk}^i's satisfying

$$\sum_i d_j^i\ r_{jk}^i \le f, \quad r_{jk}^i \ge 0, \quad \forall i\in I, j\in J_i, k\in K_i$$

for each SKU i, we have the following sub-problem $\mathcal{SB}_i$:

$$
\begin{array}{lll}
S_i(r^i) = \max & s_k^i\ p_k^i + d_j^i q_j^i & \\
\text{s.t.} & p_k^i + q_j^i \le r_{jk}^i + c_{jk}, & \forall\ j\in J_i, k\in K_i,
\end{array}
$$

where r^i is the a vector of r_{jk}^i's. We can take the dual of subproblem $\mathcal{SB}_i$. Let α_{jk} be the dual variable associated with the constraint and we denote the dual of $\mathcal{SB}_i$ as $\mathcal{SD}_i$.

$$
\begin{array}{lll}
T_i(r^i) = \min & \sum_{j\in J_i, k\in K_i} \left(c_{jk} + r_{jk}^i\right)\alpha_{jk} & \\
\text{s.t.} & \sum_{j\in J_i} \alpha_{jk} = s_k^i, & \forall\ k\in K_i \\
& \sum_{k\in K_i} \alpha_{jk} = d_j^i, & \forall\ j\in J_i \\
& \alpha_{jk} \ge 0, & \forall\ j\in J_i, k\in K_i
\end{array}
$$

Therefore, for a set of feasible r_{jk}^i's, we can solve $|I|$ transportation problems, and we arrive at a feasible solution to $\mathcal{D}$. Notice that we employ transportation problem in our SKU Exchange heuristics.

Let $Z_{\mathcal{MIP}}$ be the optimal objective value of $\mathcal{MIP}$, $Z_{\mathcal{LP}}$ be the optimal objective value of $\mathcal{LP}$, and $Z_{\mathcal{D}}$ be the optimal objective value of $\mathcal{D}$. Then,

$$Z_{\mathcal{LP}} \le Z_{\mathcal{MIP}}$$

because $\mathcal{MIP}$ is a more constrained minimization problem. Also,

$$Z_{\mathcal{D}} = Z_{\mathcal{LP}}$$

because of strong duality. For a set of feasible values of r^i's,

$$\sum_i T_i(r^i) = \sum_i S_i(r^i)$$

because of strong duality and $\sum_i S_i(r^i)$ is the objective value of a feasible solution of $\mathcal{D}$. Therefore, for a set of feasible value of r^i's,

$$\sum_i T_i(r^i) \leq Z_{\mathcal{D}} = Z_{\mathcal{LP}} \leq Z_{\mathcal{MIP}}.$$

That is, solving $|I|$ transportation problems give a solution lower bound on the objective value of $\mathcal{MIP}$. To find the best lower bounds on the $\mathcal{MIP}$, we can use sub-gradient methods to solve the following $\mathcal{Q}$.

$$\begin{array}{lll} \max & \sum_{i \in I} T_i(r^i) & \\ \text{s.t.} & \sum_i d_j^i \, r_{jk}^i \leq f, & \forall j \in J, k \in K \\ & r_{jk}^i \geq 0, & \forall i \in I, j \in J_i, k \in K_i \end{array}$$

We have shown that we can exploit the special structure of the dual of the LP relax to generate lower bounds on the optimal solution. This Dual-ascent methods have been proven effective for a number of difficult problems. Erlenkoter [Erl78] uses a dual-ascent procedure for uncapacitated facility location problem. Wong [Won84] uses the method for the Steiner tree problem. Balakrishnan, Magnanti, and Wong [BMW89] use it for a large-scale uncapacitated network design problem where each commodity has a single origin and destination. Their results are guaranteed to be within 1 to 4% of optimality. Raghavan [Rag94] studies the procedure on network design problem with connectivity requirements. His algorithm also solves optimally the special cases of the k-edge-disjoint path and k-node-disjoint path problems. The solutions from the dual-ascent procedure are within 4% of the optimal for typical telecommunication applications and within 1% of the optimality for Steiner tree problems. Future research needs to explore the dual-ascent method on our model as well.

By construction, the re-evaluation optimization problem is based on a snapshot of not-yet-picked orders at a random time. We employ effective heuristics to solve the problem. Naturally, we need to solve this problem on a rolling horizon basis. Immediate future research should address how often to solve the problem. Considering the heuristics developed in this chapter, we could solve the less time consuming and simple Order Swap procedure very frequently during a day, and solve the more time consuming SKU Exchange procedure daily.

Recall our heuristics are improvement algorithms based on the initial feasible solution, the real-time assignment. The effectiveness of our heuristics certainly depends on the real-

time assignment. We should explore the impact of the real-time assignment on the sub-optimality of the heuristics.

Finally, we take the promise-to-ship or the delivery promise as given in the model. We believe e-tailers can further improve costs by optimizing the delivery date quotation.

Chapter 4

Inventory Allocation for Low-Demand SKUs

4.1 Introduction

A large e-tailer strategically stocks inventory for SKUs with low demand for several reasons. One motivation is to provide a wide range of selections, since such SKUs actually constitute a significant portion of the total SKUs. The second incentive, of course, is to provide faster customer fulfillment service. The third motivation is to gain a competitive advantage from other online retailer. Suppose that an e-tailer only drop-ships the low-demand SKUs, its drop-shipper who serves many online retailers, may choose to satisfy a competitor's demand. For many of these SKUs, the e-tailer may only stock a handful of inventory units across all warehouses.

Inventory planning for low-demand SKUs is challenging because the discrete effect is much more pronounced while the current inventory models often assume all variables are continuous. We illustrate the discrete effect with the following example.

Example 4.1.1. 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 is a Poisson distribution with rate d, which is the expected demand in leadtime. We want to stock enough inventory in the system so that the fill rate (prob. of a customer served 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). According to Table 4.1,

d	Global Planning	Regional Planning
0.5	2	4
10	15	17

Table 4.1: Example of the Discrete Effect

if the demand in leadtime is low $d = 0.5$, then we stock 2 units of system inventory to reach the desired fill rate in global planning but we stock 4 units in the system in regional planning. Because of the discrete effect, we stock twice as much inventory if we employ a different inventory planning process. Whereas, for high-demand SKUs $d = 10$, the relative difference is not as extreme.

Efficient inventory planning for low-demand SKUs is also important in the retailing setting. We observe the demand follows the Pareto Law (law of the vital few and trivial many): the majority of the SKUs have low sales volume. Table 4.2 displays the percentage of SKUs by sale volume. The data is based on a six-week demand data in 2003 from a large

Percentage of SKUs	Fast	Medium	Slow
Books	2.6	5.3	92.1
Music	2.5	4.5	93.0
DVD	3.1	3.9	93.0

Table 4.2: Histogram of SKUs by Sale Volume

e-tailer. Notice that the "Slow" or low-demand SKUs are more than 90% of the SKUs in the books, music, and DVD product category. The "Slow" SKUs typically have 0.2-0.8 units of average weekly demand. Therefore, the impact of poor inventory planning on low-demand SKUs is very significant.

Here we focus on the effect of inventory allocation on outbound transportation costs. We assume that an 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 simultaneously. Specifically, the e-tailer will utilize its entire inventory, regardless of location, to serve demand. Given we stock certain units of inventory in the system, we intend to discover how best to allocate inventory to warehouses by considering outbound transportation costs from the warehouses to customers. The focus on transportation costs is a result of the system control policy, and we refer the reader to § 4.2 for the discussion.

We envision the inventory planning process for low-demand SKUs takes two stages. In the first stage, the managers decide the total system inventory units to stock according to the system fill rate and costs. In the second stage, given the total system inventory, she decides the inventory allocation according to the minimum transportation cost. Given the minimum transportation cost, she may want to increase or decrease the total system inventory. The process can iterate back and forth between the two stages.

4.1.1 Literature Review

A loosely related cluster of literature is on risk pooling. Typically, they consider a distribution system with 1 depot and n retailers, where the depot supplies the retailers. In so called

"joint order effect", the depot does not carry inventory but is used to pool risk over the outside-supplier leadtime. In "depot effect", the depot holds inventory and uses it between system replenishments to rebalance retailer inventories which have become unbalanced due to different demand. Some sample papers are: Schwarz, et al. [SDB84], Jonsson and Silver [JS87], Jackson [Jac88], McGavin, et al. [MSW93]. They discuss how to allocate the inventory arriving at the depot to the individual retailers. Our model is different. Because all warehouses face the same demand in our model, in a sense, the global execution policy already balances the individual warehouse inventory.

A more tightly related cluster of literature addresses lateral transshipment. A significant number of papers have considered a single-item, multi-location, periodic review inventory systems with lateral transshipments. The main objective often is to define the optimal policy for reordering at each location and the optimal policy for transferring among the locations. Some notable early papers are: Gross [Gro63], Krishnan and Rao [KR65], Das [Das75], and Karmarkar and Patel [KP77]. Robinson [Rob90] extends Krishnan and Rao's work to a multiperiod model using a stochastic dynamic program. At the beginning of each period, the inventory position at each of n locations are reviewed, additional inventory is ordered from outside suppliers, and then demand is observed. Before demand is satisfied, transshipment can take place among the locations. The leadtimes for ordering and transshipment are assumed to be zero. He shows that the optimal policy is a basestock policy.

Many early publications on lateral transshipment focus on two-location systems, since it is the first step to understanding multi-location problems. Tagaras [Tag89] treats a two-location problem by minimizing the expected cost per period. Inventory level is reviewed at the beginning of each period, and both locations employ an order-up-to policy. When the demand exceeds the order-up-to level at one location but not at the other, then lateral transshipment takes place before demand must be satisfied. The outside and pooling replenishment times for the two locations' supply are assumed to be zero. Tagaras and Cohen [TC92] extend this model by adding nonzero constant supply leadtime. They consider transshipment policies that are based on on-hand inventory or inventory position at each location. Archibald, et al. [AST97] characterize the optimal policy by considering lateral transshipments as well as emergency orders. Rudi, Kapur, and Pyke [RKP01], recently, consider a decentralized model. They find transshipment prices that lead to joint-profit maximization. Recently, Hu, Duenyas, and Kapuscinski [HDK04] first address the issue of capacity uncertainty. They characterize the optimal ordering and transshipment policy for two locations facing capacity uncertainty. At the beginning of each period, production quantities at the two locations are determined, then the production and demand uncertainties are revealed, transshipment decisions are then made, and demand is satisfied and unsatisfied demand is lost. Contrary to earlier results, they characterize the optimal transshipment policy as a rationing policy.

Other papers consider continuous review inventory systems with lateral transshipments. Dada [Dad92] analyzes a two-echelon one-warehouse-N-service center system where lateral transshipments are allowed on the lower-echelon when a stock-out occurs. He develops exact solutions when each location stocks only one unit, and approximates the general case. Lee [Lee87] develops a model for repairable items where lateral transshipments between identical bases are allowed when one base is out of stock. He approximates the expected number of backorders and lateral transshipments when a high service level is imposed. Axsater [Axs90a] extends Lee's model by focusing on the demand at a service center or local site. He models the demand rate depending on the inventory situation: demand met by on-hand order, transshipment, or backorders. Building on Axsater's work, Alfredsson and Verrijdt [AV99] model a two-echelon system where the demand at the local sites can be met by on-hand inventory, lateral shipment, direct delivery from the central warehouse, and direct delivery from the plant. Their simulations indicate that the leadtime distribution does not appear to affect the service performance. Using an approximate technique related to Axsater's, Grahovac and Chakravarty allow a transshipment not only when there is a stock-out, but at arbitrarily chosen levels of inventory levels. Recently, Axsater [Axs03] develops an effective heuristic decision rule for transshipment for a multi-location problem where each location employs a (R, Q) ordering policy and faces compound Poisson demand.

Almost all papers assume that the lateral transshipment leadtimes are instantaneous but with additional cost. This is consistent with our assumptions. We assume that if a warehouse is out of stock, its demand can be satisfied by on-hand units in other warehouses, which is equivalent to an instantaneous transshipment from other warehouses with additional transportation cost. This instantaneous transshipment assumption was not realistic for individual retailers as in the transshipment literature, but for retailers with good IT infrastructure, this assumption is valid. 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. This is the main difference between our model and those in the literature: our "transshipment policy" in effect allows the disaggregate model to act exactly like the aggregate model on the system level.

Our methodology is related to one approach of control policy performance evaluation: match every supply unit with a demand unit. That is, for an arbitrary supply unit, we keeps track of how it enters the systems, traverses through the system, and exits the systems. Svoronos and Zipkin [SZ88] first present this idea of matching supply units with demand units. Axsater [Axs90b, Axs93b, Axs97] later develops this idea into an evaluation method. Using this idea, Muharremoglu and Tsitsiklis [MT03] show the optimality of state dependent basestock policies for multi-echelon systems with Markov modulated demand. They are able to decompose the problem into single-unit single-customer problems.

Finally, there is a cluster of literature that considers inventory models for low-demand

SKUs. Some examples are: Sherbrooke [She68], Simon [Sim71], Shanker [Sha81], Graves [Gra85], and Svoronos and Zipkin [SZ88]. We also refer the reader to the section on $(S-1, S)$ policy in § 2.1.1. One-for-one replenishment policies are consistently used for low-demand SKUs. Here we also deploy this inventory policy.

4.2 2-Unit 2-Location (2U2L) Problem

We start with the simplest but non-trivial model: a 2-Unit 2-Location Problem for a single item. Suppose the e-tailer decides to stock two units of inventory in two warehouses in the system, A and B. We intend on find the allocation scenario that minimizes the expected outbound transportation costs of shipping from warehouses to customers. We start with the following assumptions.

A-1 Demand arrival to the system is Poisson with rate λ.

A-2 The demand process is split into two independent processes, 1, and 2. With probability α_1, a demand arrival is from region 1; with probability $\alpha_2 = 1 - \alpha_1$, a demand arrival is from region 2.

A-3 Region 1 is closer to A than B and region 2 is closer to B than A. As in Figure 4-1, c_1 (c_2) is the cost of shipping an order of region 1 (2) from the closest warehouse. We pay

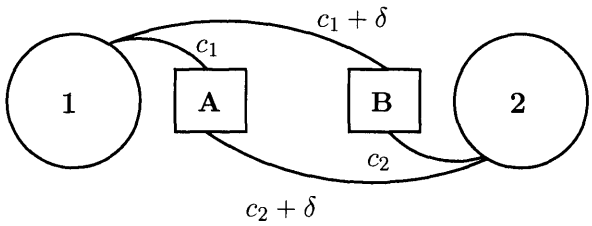

Figure 4-1: 2-Unit 2-Location Problem

a penalty of δ for a demand being served from the further warehouse. The costs $c_1(c_2)$ can be seen as the expected values of a random order in region 1 (2) being served from its closest warehouse. The penalty δ can be seen as the expected premium of a random order being served from a further warehouse.

A-4 The replenishment lead time for each warehouse is the same constant L.

A-5 Inventory policy is one-for-one replenishment: a replenishment is triggered as soon as a demand arrives.

A-6 Demand is backlogged when there is no on-hand inventory in the system.

We denote stocking scenario $(\mathbf{i},\mathbf{j})$ as stocking i units of inventory at location A and j units in B. Our allocation options are then limited to stocking one unit at each warehouse or stocking two units at one of the warehouses, in symbols (1,1), (2,0), (0,2). In the context of online retailing, the e-tailer can utilize all of its warehouse or fulfillment centers to serve the customer demand. It also has the technological capability to manage and control the inventory globally. 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 replenishment. We then have the following assumptions on how the system operates for all stocking scenarios.

A-6 If a customer arrives and its closest warehouse has on-hand inventory, then its closest warehouse serves the demand and triggers a replenishment.

A-7 If a customer arrives and only one warehouse has on-hand units in the system (e.g., its closest warehouse does not have inventory on-hand and the other warehouse does have inventory on hand), then the warehouse with the on-hand unit serves demand and triggers a replenishment.

A-8 If a customer arrives and the system has no on-hand units, then the warehouse with the next arriving unassigned replenishment is assigned to serve this demand and trigger a replenishment.

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

To facilitate the discussion, we denote the following inventory notation:

$IP_k(t)$	inventory position of location k at time t
$IP(t)$	inventory position of the system at t
$IL_k(t)$	inventory level (on-hand inventory) at location k at t
$IL(t)$	inventory level of the system at t
I	steady state net inventory of the system

where inventory position is on-hand and on-order inventory minus backorders, and net inventory is on-hand minus backorders.

As a result of our assumptions, we see that the system inventory position is always 2, $IP(t) = 2,\ \forall t$. In addition, we can model the system as an $M/G/\infty$ queue with service

time L, and the steady state system net inventory is a Poisson distribution:

$$Pr\{I = j\} = e^{-\lambda L}\frac{(\lambda L)^{j-2}}{(j-2)!}, \quad j = 2, 1, 0, -1, -2, \ldots$$

By assumption, every demand is matched with the next available unit and replenishment is triggered as soon as a demand arrives. We see that the system inventory level and the customer waiting times are the same as in the aggregate model where all inventory in the system is aggregated into one warehouse. In other words, those costs at the system level, e.g., 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 at which demand are served. Therefore, we examine how outbound transportation costs influence inventory allocations among the warehouses in the disaggregate model.

Also by assumption, each stock scenario (i,j) has $IP_A(t) = i, IP_B(t) = j, \forall t$. Given the values of α_1, α_2, we would like to determine the scenario with minimum expected transportation penalty cost per order. We denote $\underline{C}$ to be the expected outbound transportation cost of an order given that it is served by its closest warehouse. Clearly, we have

$$\underline{C} = \alpha_1 c_1 + \alpha_2 c_2.$$

We denote $\boldsymbol{C(i,j)}$ as the expected transportation cost of an order of scenario (i,j). For the remainder of the section, we derive the scenario costs $C(i,j)$.

4.2.1 Scenario (2,0) and (0,2)

In scenario (2,0), we stock two units of inventory in warehouse A only. Since all demand is served by A, we derive $C(2,0)$ by conditioning on the type of a random demand.

$$C(2,0) = \alpha_1 c_1 + \alpha_2 (c_2 + \delta) = \underline{C} + \alpha_2 \delta$$

Similarly, we have

$$C(0,2) = \alpha_1 (c_1 + \delta) + \alpha_2 c_2 = \underline{C} + \alpha_1 \delta \tag{4.1}$$

4.2.2 Scenario (1,1)

By assumption, the inventory position at each location is always 1, $IP_A(t) = IP_B(t) = 1$ for all t in scenario (1,1). The cost derivation is more complicated.

We introduce additional notation to facilitate the discussion.

Π_{kj}	$Pr\{$a demand is from region j and is served by location $k\}$, $j \in \{1,2\}, k \in \{A,B\}$
P_k	$Pr\{$a demand is served by location $k\}$
q	$Pr\{$zero demand in L time periods$\}$

By definition, we have

$$\begin{aligned} \Pi_{A1} + \Pi_{B1} = \alpha_1, \qquad & \Pi_{A2} + \Pi_{B2} = \alpha_2 \\ P_A + P_B \;\; = \;\; & 1 \end{aligned}$$

Since we assume Poisson demand, we have

$$q \;\; = \;\; e^{-\lambda L}$$

Therefore,

$$\begin{aligned} C(1,1) \;\; &= \;\; \Pi_{A1}c_1 + \Pi_{A2}(c_2+\delta) + \Pi_{B1}(c_1+\delta) + \Pi_{B2}c_2 \\ &= \;\; \underline{C} + (\Pi_{A2} + \Pi_{B1})\,\delta \end{aligned} \tag{4.2}$$

In the remainder of the section, we introduce a methodology to derive the probabilities of Π_{A2} and Π_{B1}.

Observation. Given $IP_A(t) = IP_B(t) = 1$ for all t and 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.

To see that, we consider two cases: if there is an on-hand unit at a warehouse, obviously that unit is not assigned to any demand; if there is no on-hand units at a warehouse, then the inventory position equals to on-order units minus the backorders, then there is one unit of on-order replenishment that's not yet been assigned while the others are assigned to backorders. Furthermore, the unassigned unit is the most recent on-order unit, which resulted from the replenishment of the most recent demand served by the warehouse.

In scenario (1,1), only five relative positions of two unassigned units in A and B can occur, as illustrated in Figure 4-2. Each box in Figure 4-2 represents one unit of unassigned inventory, the units below the bars are on-hand units, and the units above are on order and their proximity to the bar represents their time until reaching the warehouse. Light-color units are at location A and the dark-color units are at location B. The three cases are:

i) both units are on hand;

ii) the unit in B is on order, and the unit in A is on hand;

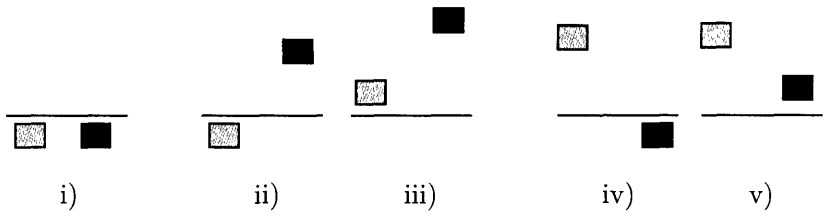

Figure 4-2: 2U2L, Positions of Unassigned Units

iii) both units are on order, and the unit destined for A will arrive before the unit in B destined for B;

iv) the unit in A is on order, and the unit in B is on hand;

v) both units are on order, and the unit destined for B will arrive before the unit in A destined for A.

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

We denote an epoch to be the time of a demand arrival. In addition, in a *type-A epoch*, a demand is served by A and in a *type B epoch* a demand is served by B. We define a Markov Chain with states that are defined on the demand epochs. State A defines a type-A epoch, and state B defines a type-B epoch. The Markov Chain is as illustrated in Figure 4-3.

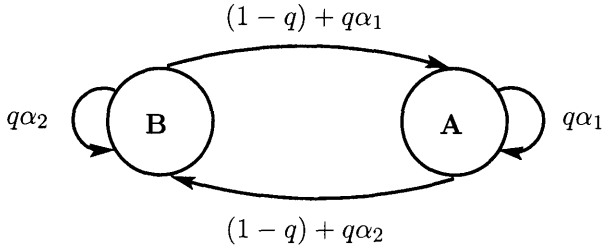

Figure 4-3: 2U2L Markov Chain

To explain the transitions, we represent the Markov Chain in a different form as in Figure 4-4. 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 a type-A epoch. 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 solid-line transitions represent the next demand arriving before $t_k + L$, $t_{k+1} < t_k + L$. The dotted-transitions represents the next demand arriving after $t + L$, $t_{k+1} > t_k + L$. The state of the system at t_k is of case iv) or v) in Figure 4-2 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 $k + 1^{\text{st}}$ demand would be served by B and the system transition to state B. If $t_{k+1} > t + L$ (with probability q), the state

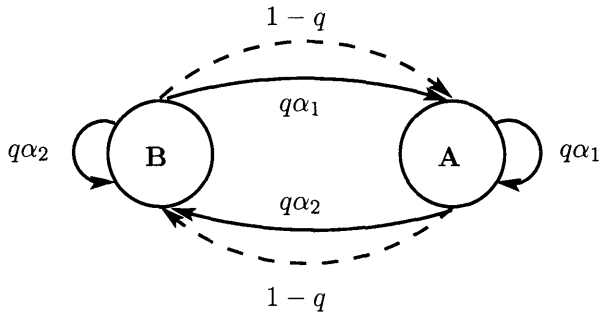

Figure 4-4: 2U2L Markov Chain

of the system at $t_k + L$ would be of case i) in Figure 4-2. Then, with probability α_1, the system transitions to state A and with probability α_2, the system transitions to state B.

Let $\boldsymbol{\pi_A, \pi_B}$ be the steady state probabilities of the Markov Chain. We see that all epochs are marked by either state. The value of π_A represents the steady state probability of an epoch being type A. This is equivalent to the steady probability of a demand being served by location A. Therefore, we have

$$\pi_a = P_A, \quad \pi_b = P_B.$$

This Markov Chain is the embedded chain of a Semi-Markov process. Each transition duration represents a demand inter-arrival. This process is not Markov, since the duration of a transition depends on the current as well as the next state. We then find the steady state probabilities of the embedded Markov Chain to be

$$\pi_B = \frac{1 - q\alpha_1}{2 - q}, \quad \pi_A = \frac{1 - q\alpha_2}{2 - q}$$

Proposition 4.2.1. *Let $q = e^{-\lambda L}$, then the probability of a demand being served at A in the 2-Unit 2-Location Problem is $P_A = \frac{1 - q\alpha_2}{2 - q}$ and served at B is $P_B = \frac{1 - q\alpha_1}{2 - q}$. When $\lambda L = 0$, $P_A = \alpha_1$, $P_B = \alpha_2$, and as $\lambda L \to \infty$, $\lim_{\lambda L \to \infty} P_A = P_B = \frac{1}{2}$. In addition, if $\alpha_2 \geq \frac{1}{2}$, then*

- *P_A is concave and P_B is convex in λL,*
- *$\alpha_1 \leq P_A \leq \frac{1}{2} \leq P_B \leq \alpha_2$, $\forall \lambda L \geq 0$.*

If $\alpha_2 \leq \frac{1}{2}$, then

- *P_A is convex and P_B is concave in λL,*
- *$\alpha_2 \leq P_B \leq \frac{1}{2} \leq P_A \leq \alpha_1$, $\forall \lambda L \geq 0$.*

Proof. We show the properties for P_A for $\alpha_2 \geq \frac{1}{2}$ only, since the same reasoning applies to P_B.

Let P'_A and P''_A be the first and second derivative of P_A as a function of λL. We have

$$P''_A = \frac{q(2+q)(2\alpha_2 - 1)}{(q-2)^3}.$$

Since $0 \leq q \leq 1$ $\forall \lambda L \geq 0$, $P''_A \leq 0$ or P_A is concave in λL iff $\alpha_2 \geq \frac{1}{2}$. We also have

$$P'_A = \frac{q(2\alpha_2 - 1)}{(q-2)^2},$$

which is always nonnegative if $\alpha_2 \geq \frac{1}{2}$. Since $P_A(\lambda L = 0) = \alpha_1$, $\lim_{\lambda L \to \infty} P_A = \frac{1}{2}$, and $P'_A \geq 0$, $\forall \lambda L \geq 0$, $\alpha_1 \leq P_A \leq \frac{1}{2}$, $\forall \lambda L \geq 0$. ■

To find the probabilities of Π_{A2}, Π_{B1}, we only need to examine the dotted-line transitions in Figure 4-4, as these transitions include the events where demand is served from a further warehouse. To derive Π_{A2}, we examine the dotted transition from state B to state A where the next demand occur within L time units. Since the next arriving demand has to be served by A, the probability that the next arriving demand is from region 2 is α_2. Then, we have

$$\begin{aligned} \Pi_{A2} &= \pi_B(1-q)\alpha_2 \\ \Pi_{B1} &= \pi_A(1-q)\alpha_1 \end{aligned}$$

As a result, from Equation 4.2 we have

$$C(1,1) = \underline{C} + \frac{1-q}{2-q}\left(1 - 2q\alpha_1\alpha_2\right)\delta$$

Alternatively, we can derive the value of Π_{A2} by conditioning on the state of the system at a demand arrival:

$$\Pi_{A2} = Pr\{I = 2\}Pr\{A2|I = 2\} + Pr\{I < 2\}Pr\{A2|I < 2\},$$

where $Pr\{A2|I = 2\}$ is the probability of a demand from region 2 and it is served by A given the system has two units on hand at time of its arrival. Clearly, $Pr\{A2|I = 2\} = 0$ since each demand is always matched with its closest location when there are on-hand units. To derive $Pr\{A2|I < 2\}$, we condition on the next arriving unit being from A or B. Let Q_A be the probability that the on-hand unit is A or the next arriving unit is A given $I < 2$.

Then,

$$\Pi_{A2} = Pr\{I < 2\} Q_A \alpha_2.$$

Since $Pr\{I < 2\} = 1 - q$, we have $Q_A = \pi_B$.

Proposition 4.2.2. *Given that the system has less than two units on hand ($I < 2$), the steady state probability that the on-hand unit is A or the next arriving unit is A is* $\dfrac{1 - q\alpha_1}{2 - q}$.

4.2.3 Comparison

Given the relative magnitude of demand process 1 and 2 (i.e., α_1, α_2), we want to find the minimum cost scenario. We first start with a useful Lemma.

Lemma 4.2.1. *Let $f : \mathbb{R}^1 \to \mathbb{R}^1$ be a quadratic function in $x \in \mathbb{R}^1$ with coefficients a, b, c, $f = ax^2 + bx + c$. In addition, $a \geq 0$ and the discriminant $b^2 - 4ac > 0$, that is, f is convex and has two real roots. Let the roots be $s_1 = \dfrac{-b - \sqrt{b^2 - 4ac}}{2a}$ and $s_2 = \dfrac{-b + \sqrt{b^2 - 4ac}}{2a}$.*

1) If f satisfies the two following conditions:

i) either $b \geq 0,\ c \leq 0$ or $b \leq 0,\ c \leq 0$

ii) $a + b + c \geq 0$,

then $f \leq 0\ \forall x \in (0, s_2)$ and $f \geq 0\ \forall x \in (s_2, 1)$.

2) If f satisfies the two following conditions:

i) $b \leq 0,\ c \geq 0$,

ii) $a + b + c \leq 0$

then $f \geq 0\ \forall x \in (0, s_1)$ and $f \leq 0\ \forall x \in (s_1, 1)$.

Proof. If $a \geq 0$, then f is convex in x and $s_1 \leq s_2$.

1) If $b \geq 0,\ c \leq 0$ or $b \leq 0,\ c \leq 0$, then by Descartes' Sign Rule, f has exactly one positive and negative root. The positive root $s_2 \leq 1$ iff

$$\begin{aligned} & \sqrt{b^2 - 4ac} && \leq && 2a + b \\ \Rightarrow \quad & b^2 - 4ac && \leq && (2a + b)^2 = 4a^2 + 4ab + b^2 \\ \Rightarrow \quad & 0 && \leq && 4a(a + b + c). \end{aligned}$$

Since $s_1 \leq 0,\ s_2 \leq 1$, $f \leq 0$ for $x \in (0, s_2)$ and $f \geq 0$ for $x \in (s_2, 1)$.

2) If $b \leq 0,\ c \geq 0$, then by Descartes' Sign Rule, f has two positive roots. The roots $s_1 \leq 1$ and $s_2 \geq 1$ iff

$$\begin{array}{ccccc} \sqrt{b^2-4ac} \geq -(2a+b) & \text{and} & \sqrt{b^2-4ac} \geq (2a+b) \\ \Rightarrow & b^2-4ac & \geq & (2a+b)^2 \end{array}$$

Therefore, $s_1 \leq 1$ and $s_2 \geq 1$ iff $0 \geq 4a(a+b+c) \Rightarrow a+b+c \leq 0$. Therefore, $f \geq 0$ for $x \in (0, s_1)$ and $f \leq 0$ for $x \in (s_1, 1)$. ■

We first compare the two scenarios of (1,1) and (2,0) by examining their transportation penalty costs. If scenario (1,1) has a larger expected penalty cost per order, then we prefer scenario (2,0) and vice versa. Let the ratio r be the ratio of penalty costs,

$$\begin{aligned} r &= \frac{C(1,1)-\underline{C}}{C(2,0)-\underline{C}} \\ &= \frac{\frac{1-q}{2-q}(1-2q\alpha_1\alpha_2)}{\alpha_2} \\ &= \frac{1-q}{2-q}\left(\frac{1}{\alpha_2}-2q+2q\alpha_2\right) \end{aligned}$$

and we have the following property.

Theorem 4.2.1. *Let* $q = e^{-\lambda L}$ *and*

$$\alpha = \frac{2+q-2q^2-\sqrt{2-q}\sqrt{2-4q^3+4q^2-q}}{4q(1-q)}. \tag{4.3}$$

If $\alpha_2 \in (\alpha, 1-\alpha)$, *then we prefer scenario (1,1); if* $\alpha_2 \leq \alpha$, *then we prefer scenario (2,0); if* $\alpha_2 \geq 1-\alpha$, *then we prefer scenario (0,2).*

Proof. Let the function $f = r - 1 = 0$. Then, $f = \frac{1-q}{2-q}\left(1-2q\alpha_2+2q(\alpha_2)^2\right)-\alpha_2$. Note that f is a quadratic function of α_2, and the coefficients are: $a = 2\theta q,\ b = -1-2\theta q,\ c = \theta$ where $\theta = \frac{1-q}{2-q}$. Since $0 \leq q \leq 1$, $\theta \geq 0$. Therefore, $a \geq 0,\ b \leq 0,\ c \geq 0$.

By inspection, the discriminant

$$b^2-4ac = 1+4\theta^2q^2+4\theta q(1-2\theta).$$

Since $\theta \leq \frac{1}{2}$, the discriminant is positive.

In addition, $a+b+c = \theta - 1 < 0$.

Let $\alpha = \frac{-b-\sqrt{b^2-4ac}}{2a}$. Then, according to Lemma 4.2.1 Part 2), $f \geq 0$ for $\alpha_2 \in (0, \alpha)$ and $f \leq 0$ for $\alpha_2 \in (\alpha, 1)$. Therefore, for $\alpha_2 \in (\alpha, 1)$ we prefer the scenario (1,1) over (2,0). Since scenario (2,0) is symmetric to scenario (2,0), we draw the symmetric conclusion: for $\alpha_2 \in (0, 1-\alpha)$, we prefer the scenario (1,1) over (0,2). ■

Let **FR** be the fill rate, or, the probability of a demand served by an on-hand inventory unit. We can derive the fill rate from the aggregate model. Because of PASTA, the long run probability of a demand served by an on-hand unit is the equivalent of the time average of the system having on-hand units.

$$\begin{aligned} FR &= Pr\{I = 2\} + Pr\{I = 1\} \\ &= q(1 + \lambda L). \end{aligned}$$

We provide a few numerical examples in Table 4.3. For a given value of expected demand

λL	Fill Rate	Range of α_2 to Choose (1,1)
0.2	0.98	(0.13, 0.87)
0.4	0.94	(0.20, 0.80)
0.5	0.91	(0.22, 0.78)
0.75	0.83	(0.28, 0.72)
1	0.73	(0.33, 0.68)
1.5	0.56	(0.39, 0.61)

Table 4.3: 2U2L Numerical Results

in leadtime, λL, we compute the fill rate for the system and the range of α_2 for which we prefer scenario (1,1). For α_2 smaller than the range in the table, we prefer (2,0), and for α_2 larger we prefer (0,2). We only present the values of λL for which the fill rate is not extremely poor.

Remark. We note that for balanced split of demand in region 1 and 2, we tend to prefer scenario (1,1), and for extreme split of demand, we prefer scenario (2,0) or (0,2).

Remark. With higher fill rate, the range of demand splits to choose (1,1) becomes larger. This is because with high fill rate, customers are more likely to see (1,1) when they arrive to the system, and thus be served from the closest warehouse.

In this section, we have established the simplest model of 2U2L. The allocation result is intuitively satisfying. In the next few sections, we will relax some of the assumptions and build upon the 2U2L model.

4.3 2-Unit 2-Location with Different Leadtimes

Here we consider a natural extension of the 2U2L problem where the two locations have different supply leadtimes. This is a more realistic assumption. Often, warehouses may be located far or close to their supply injection point. We are interested in how the results in the 2-Unit 2-Location problem may change by relaxing the equal leadtime assumption.

While we maintain all the assumptions in the 2U2L problem, we assume that $\boldsymbol{L_A, L_B}$ are the supply leadtimes to location A, B. In addition, $\boldsymbol{L} = L_A - L_B \geq 0$. The expected

transportation costs of scenario (2,0) and (0,2) stay the same as in the 2-Unit 2-Location problem. We derive the cost of scenario (1,1) as in the following.

4.3.1 Scenario (1,1)

Recall in the 2U2L problem, the expected transportation cost of scenario (1,1) is $C(1,1) = \underline{C} + (\Pi_{A2} + \Pi_{B1})\delta$. We need to derive the Π's under the different leadtime assumption. Let

$$q_A = e^{-\lambda L_A}, \quad q_B = e^{-\lambda L_B}, \quad q = e^{-\lambda L}.$$

We construct a similar Markov Chain with more states: we define state A as before but break down state B into states B and AB. Figure 4-5 illustrates the three states: A, B, AB. Again, we examine the relative positions of the two unassigned units in location A and B

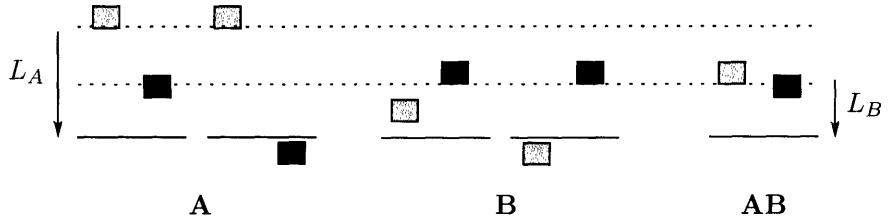

Figure 4-5: 2-Unit 2-Location with Different Leadtimes

at a demand epoch. Each box in the figure represents an unassigned inventory unit, with the darker box from location B and the lighter one from location A.

State A represents a type-A epoch where a demand has just been served by and thus triggers a replenishment at A. Therefore, the unassigned unit in A has L_A time periods to reach A, and the unassigned unit in B is either on hand or on order but will reach B before the unit on order at A reaches A.

State B represents a type-B epoch where a demand has just been served by and thus triggers a replenishment at B. In addition, the unit in A is either on hand or on order but has $\tau < L_B$ time units to reach the warehouse. The unit at B, of course, will reach B in L_B time units.

State AB represents a time where exactly L time periods ago, a demand (not necessarily the last demand) has just been served by and thus triggers a replenishment at A. Therefore, the unit in A has L_B time periods to reach A. The unit at B, of course, will reach B in $\tau \leq L_B$ time periods. Note that state AB does not necessarily represent a demand epoch.

Figure 4-6 is the embedded Markov Chain for the process. To be consistent with the

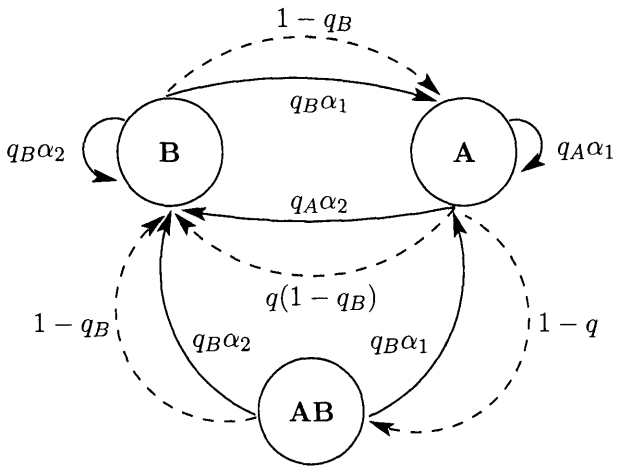

Figure 4-6: 2-Unit 2-Location with Different Leadtimes - Markov Chain

2U2L problem, we again denote the transitions by solid or dotted lines. The solid-line transitions reach the system state with two units on hand, whereas the dotted-line transitions do not. The transitions out of state B is the same as in the 2U2L problem. The transitions out of state A require more discussion. Let a type-A epoch occur at t. We first consider the solid-line transitions: the next demand arrives after $t + L_A$. Then, with probability α_1 the system transitions to A and with α_2 to B. Next, we consider the dotted-line transitions. *If the next demand arrives before $t + L$ (with probability $1 - q$), then the system stays at state A until $t + L$, at which it transitions to state AB.* Otherwise (with probability q), if the next demand arrives within $(t + L, t + L_A)$, the system transitions to state B.

Notice that by construction, the process is a Semi-Markov process. In particular, let's examine the transitions out of A. Let $\boldsymbol{U_{i\_j}}$ be the interval between transition from state i to j. Then, U_{A_A} is distributed as a conditional Poisson inter-arrival given the interval is greater than L_A; U_{A_B} is distributed as a conditional Poisson inter-arrival given the interval is greater than L; U_{A_AB} is a constant of L. Clearly, the U_{i_j}'s are only dependent on the current and next state.

Suppose the system transitions to state AB at time t. By definition, at time t, the unit at A has L_B time periods to reach its warehouse, and the unit at B has L_B or less to reach B. The system transitions out of AB at the next demand arrival after t. Since state AB does not necessarily mark a demand epoch, because we have Poisson arrival, from t to the next arrival after t is still exponentially distributed with rate λ. Furthermore, if the next demand arrives before $t + L_B$, the system transitions to B; otherwise, with probability α_1, the system transitions to state A and with α_2 to state B.

Observation. Note that not all demand epochs are marked by a transition into a state: there may be many demand arrivals during the transition from state A to state AB.

We derive the time averages of being in a state to find the Π's. We denote

p_j	time average probability of being in state j
π_j	steady state probability of state j of the Markov Chain
μ_j	expected time in state j before a transition
P_{ij}	transition probability from state i to j

We derive the followings based on the Semi-Markov balance equations $p_j = \frac{\pi_j \mu_j}{\sum_i \pi_i \mu_i}$ or $\frac{p_j}{\mu_j} = \sum_i \frac{p_j}{\mu_j} P_{ij}$.

$$\begin{aligned} \frac{p_A}{\mu_A} &= \frac{p_A}{\mu_A} q_A \alpha_1 + p_B \lambda (1 - q_B \alpha_2) + p_{AB} \lambda q_B \alpha_1 \\ p_B \lambda &= \frac{p_A}{\mu_A} q (1 - q_B \alpha_1) + p_B \lambda q_B \alpha_2 + p_{AB} \lambda (1 - q_B \alpha_1) \\ p_{AB} \lambda &= \frac{p_A}{\mu_A} (1 - q) \end{aligned}$$

$\Rightarrow$

$$p_B \lambda (1 - q_B \alpha_2) = \frac{p_A}{\mu_A} (1 - q_B \alpha_1). \tag{4.4}$$

In addition,

$$p_A + p_B + p_{AB} = 1. \tag{4.5}$$

We derive the expected time in state A, μ_A, by conditioning on whether the next state is AB or not. If the next state is AB, then $\mu_A = L$. If the next state is not, which indicates that the next demand arrival takes place more than L time periods from the start of state A, $\mu_A = E[X_L]$ where

X_L	the conditional time until next arrival given it arrives after L time periods.

Therefore, we have

$$\begin{aligned} \mu_A &= (1 - q) L + q E[X_L] \\ &= (1 - q) L + q \left(L + \frac{1}{\lambda} \right) \\ &= L + q \frac{1}{\lambda} \end{aligned} \tag{4.6}$$

With Equations (4.4), (4.5), (4.6), we can solve for p_A, p_B, p_{AB}:

$$\begin{aligned} p_A &= \frac{(\lambda L + q)(1 - q_B\alpha_2)}{2 - q_B + \lambda L(1 - q_B\alpha_2)}, \\ p_B &= \frac{1 - q_B\alpha_1}{2 - q_B + \lambda L(1 - q_B\alpha_2)}, \\ p_{AB} &= \frac{(1 - q)(1 - q_B\alpha_2)}{2 - q_B + \lambda L(1 - q_B\alpha_2)}. \end{aligned}$$

Using PASTA, we then can derive the Π_{A2}, Π_{B1} by conditioning on the state of the system at the time of a demand arrival.

$$\Pi_{A2} = p_B(1 - q_B)\alpha_2$$

The derivation of Π_{B1} requires more discussion. First, if a demand arrives when the system is in state AB, then with probability $(1 - q_B)$ the demand has to be served by location B. Next, consider a demand arrives and the system is in state A. Note that the duration of being in state A is at least L time periods. Also, any demand arriving during the first L time periods would have to be served by B, since the unit in A is on order and the unit in B is "ahead". By renewal theory, the time average of the system in the first L time periods is $p_A\frac{L}{\mu_A}$, which is the probability of a demand arriving then. If a demand arriving after the first L time periods in A, then with probability $(1 - q_B)$ the demand would have to be served by B. Therefore, we have

$$\begin{aligned} \Pi_{B1} &= p_A\frac{L}{\mu_A}\alpha_1 + p_A\frac{\mu_A - L}{\mu_A}(1 - q_B)\alpha_1 + p_{AB}(1 - q_B)\alpha_1 \\ &= \frac{p_A}{\lambda L + q}\left(\lambda L + q(1 - q_B)\right)\alpha_1 + p_{AB}(1 - q_B)\alpha_1 \end{aligned}$$

Let the penalty cost function be $g(1,1) = \frac{C(1,1) - \underline{C}}{\delta}$, then

$$\begin{aligned} g(1,1) &= \Pi_{A2} + \Pi_{B1} \\ &= \frac{a\alpha_2^2 + b\alpha_2 + c}{2 - q_B + \lambda L(1 - q_B\alpha_2)} \end{aligned}$$

where

$$a = 2q_B(1 - q_B + \lambda L),\ b = 2q_B(q_B - 1) - \lambda L(1 + q_B),\ c = 1 - q_B + \lambda L.$$

Following the same reasoning in the derivation of Π_{A2}, Π_{B1}, we derive the probabilities

of an order being served from location A or B, P_A, P_B.

$$\begin{aligned} P_A &= p_B(1-q_B\alpha_2) + p_{AB}q_B\alpha_1 + p_A\frac{\mu_A - L}{\mu_A}q_B\alpha_1 \\ &= p_B(1-q_B\alpha_2) + p_{AB}q_B\alpha_1 + \frac{p_A}{\lambda\mu_A}q_A\alpha_1 \\ &= \frac{1-q_B\alpha_2}{2-q_B+\lambda L(1-q_B\alpha_2)} \end{aligned}$$

$$\begin{aligned} P_B &= p_Bq_B\alpha_2 + p_{AB}(1-q_B\alpha_1) + \frac{p_A}{\lambda\mu_A}(\lambda L + q - q_A\alpha_1) \\ &= \frac{(1-q_B\alpha_1)+\lambda L(1-q_B\alpha_2)}{2-q_B+\lambda L(1-q_B\alpha_2)} \end{aligned}$$

Proposition 4.3.1. *Let $q = e^{-\lambda L}, q_B = e^{-\lambda L_B}$, then the probability of a demand served by location A in the 2-Unit 2-Location problem with different leadtimes is $P_A = \dfrac{1-q_B\alpha_2}{2-q_B+\lambda L(1-q_B\alpha_2)}$ and served at B is $P_B = \dfrac{(1-q_B\alpha_1)+\lambda L(1-q_B\alpha_2)}{2-q_B+\lambda L(1-q_B\alpha_2)}$. In addition,*

- *P_A is nonincreasing in λL, for $\lambda L \geq 0$;*
- *P_B is nondecreasing in λL, for $\lambda L \geq 0$;*
- *If $\lambda L \geq q_B$, then $P_B \geq P_A$ for all $\alpha_2 \in [0,1]$;*
- *If $\lambda L \leq q_B$, then $P_B \leq P_A$ for $\alpha_2 \in [0, \tilde{\alpha_2}]$ and $P_B \geq P_A$ for $\alpha_2 \in [\tilde{\alpha_2}, 1]$, where $\tilde{\alpha_2} = \dfrac{q_B - \lambda L}{q_B(2-\lambda L)}$.*

Proof. Given the numerator of P_A is nonnegative and the denominator of P_A is increasing in $\lambda L \geq 0$, then P_A is nonincreasing in λL. Since $P_B = 1 - P_A$, then P_B is nondecreasing in λL.

Let $\tilde{\alpha_2}$ be the value of α_2 at which $P_A = P_B$. Since the denominator of P_A and P_B is strictly positive, we can find $\tilde{\alpha_2}$ by equating $1 - q_B\alpha_2 = (1-q_B\alpha_1) + \lambda L(1-q_B\alpha_2)$. Then

$$\tilde{\alpha_2} = \frac{q_B - \lambda L}{q_B(2-\lambda L)}.$$

For $\tilde{\alpha_2}$ to be within the range of $[0,1]$, we examine the following three cases: i) $\lambda L \leq q_B$, ii) $q_B \leq \lambda L \leq 2$, iii) $\lambda L \geq 2$.

If $\lambda L \leq q_B$, then $\lambda L \leq 2$ and $\tilde{\alpha_2} \geq 0$. In addition, $\tilde{\alpha_2} \leq 1$ for $\lambda L \geq 0$.

If $q_B \leq \lambda L \leq 2$, then $\tilde{\alpha_2} \leq 0$.

If $\lambda L \geq 2$, then $\lambda L > q_B$ and $\tilde{\alpha_2} \geq 0$. However, for $\tilde{\alpha_2} \leq 1$ only if $\lambda L \leq \dfrac{q_B}{q_B - 1} < 0$.

Therefore, P_A and P_B only intersects when $\lambda L \leq q_B$. Since P_A is nonincreasing and P_B is nondecreasing, we have $P_B \leq P_A$ for $\alpha_2 \in [0, \tilde{\alpha_2}]$ and $P_B \geq P_A$ for $\alpha_2 \in [\tilde{\alpha_2}, 1]$.

If $\lambda L \geq q_B$, then

$$\begin{aligned} P_B &\geq \frac{(1-q_B\alpha_1)+q_B(1-q_B\alpha_2)}{2-q_B+\lambda L(1-q_B\alpha_2)} \\ &\geq \frac{(1-q_B)+q_B(1-q_B\alpha_2)}{2-q_B+\lambda L(1-q_B\alpha_2)} \\ &\geq \frac{(1-q_B)(1-q_B\alpha_2)+q_B(1-q_B\alpha_2)}{2-q_B+\lambda L(1-q_B\alpha_2)} = P_A \end{aligned}$$

for $\alpha_2 \in [0,1]$. ■

Observation. Note that the value of P_A is smaller than the value of P_A in scenario (1,1) in the 2U2L problem with the same leadtime of L_B. Moreover, a demand is always more likely to be served by B when the difference between leadtimes is large (i.e. $\lambda L \geq q_B$). If the difference between leadtimes is small, a demand is more likely to be served by A when α_2 is small and more likely to be served by B when α_2 is large.

4.3.2 Fill Rates

In the 2U2L problem, we compute the fill rate from the aggregate model. However, in the 2-Unit 2-Location problem with different leadtimes, the warehouses are no longer identical. The fill rate may depend on how the demand is split in the two regions, and, therefore, cannot be derived from the aggregate model. We derive the fill rates from additional analysis in this section.

The fill rate for scenario (2,0) is simply derived as

$$FR(2,0) = q_A(1+\lambda L_A),$$

and similarly, we have

$$FR(0,2) = q_B(1+\lambda L_B).$$

The fill rate of scenario (1,1) is more complicated. We devote the rest of the section on the derivation.

We *assume* $L \leq L_B$ *in the derivation*, while we can derive the case of $L > L_B$ similarly. First, we present a more detailed Markov Chain than in Figure 4-6, as illustrated in Figure 4-7. *Same as in the Figure 4-6 Markov Chain, we define each state on a demand epoch.* Specifically, we split each type-A or type-B epoch into two types: one with an on-hand unit and one without. Figure 4-8 is an illustration. State A is split into A′ and A″, B into B′ and B″. The states A′ and B′ are special cases of A and B where the unit ahead has not yet reached the warehouse. For instance, the state A′ is defined on a type-A demand epoch where the unit in B is on order. The states A″ and B″ are defined such that the unit

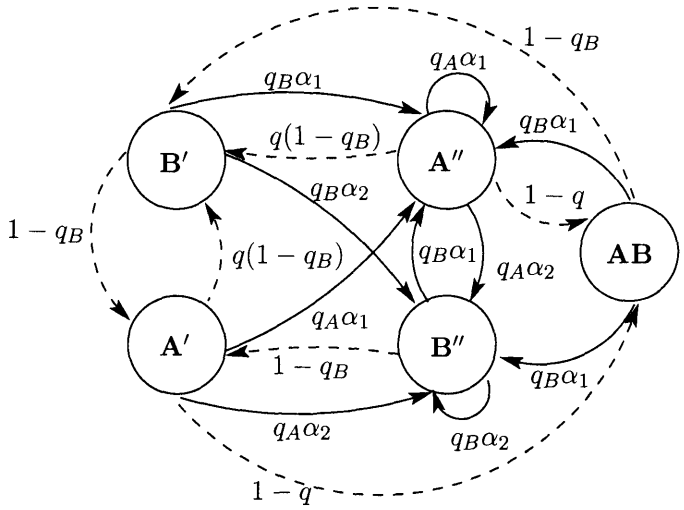

Figure 4-7: 2-Unit 2-Location with Different Leadtimes - Fill Rate of Scenario (1,1)

ahead is on hand. *By definition, the transitions out of state A' and A'' are the same as the*

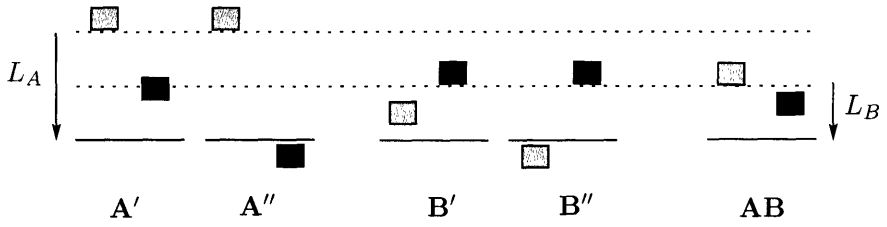

Figure 4-8: 2-Unit 2-Location Different Leadtime - Fill Rate

transitions out of state A, and the transitions out of state B' and B'' are the same as the transitions out of state B. Therefore, here we omit the discussion on the transitions.

We derive the time average of being in a state. By definition, $p_{A'} + p_{A''} = p_A$ and $p_{B'} + p_{B''} = p_B$. Deriving from the balance equations, we have

$$\begin{aligned} p_{B'} &= (1-q_B)\frac{p_A}{\lambda L+q} & p_{B''} &= q_B\alpha_2\left(\frac{p_A}{\lambda L+q}+p_B\right) \\ p_{A'} &= (1-q_B)(\lambda L+q)p_B & p_{A''} &= q_B\alpha_1\left(p_A+p_B(\lambda L+q)\right) \end{aligned}$$

Furthermore, we denote the conditional probability v_j as

v_j prob. of a demand served by an on-hand unit,
given the system is in state j at the time of its arrival

We then can write the fill rate of scenario (1,1) as:

$$FR(1,1) = p_{B''}v_{B''} + +p_{B'}v_{B'} + p_{A''}v_{A''} + p_{A'}v_{A'} + p_{AB}v_{AB}.$$

Clearly, $v_{B''} = 1$, since one unit is always on hand during state B''. The others require more discussion.

First, we derive $v_{B'}$. Suppose the system transitions to state B' at time t as illustrated in Figure 4-9. Let each black dot on the time line represent a demand arrival. At t, the

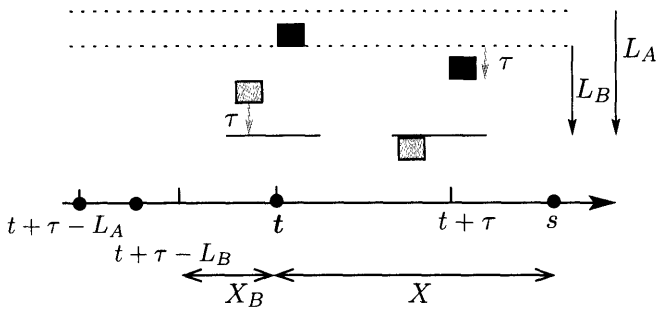

Figure 4-9: System Reaches State B' at Time t

unit in A has $\tau < L_B$ time periods to reach A and the unit in B has L_B time periods to reach B. Then, $v_{B'}$ is the probability of the next demand arriving after $t + \tau$, such as at time s. Let

$$X \quad \text{the time until the next arrival,}$$

and X is exponentially distributed with rate λ. Note that the demand associated with the unit in A at t must have arrived at $t + \tau - L_A$. That is, the system is at state A' or A'' at $t + \tau - L_A$. Then, regardless of whether the system transitions to state AB, the demand arriving at t must be the *first* demand arriving after $t + \tau - L_B$. Let

$$X_B \quad \text{a conditional inter-arrival given it's less than } L_B,$$

then at $t + \tau - L_B$, the time until the next arrival at t is distributed as X_B given we have $\tau < L_B$. We can represent the time during $(t + \tau - L_B, t)$ as X_B. We have

$$\begin{aligned} v_{B'} &= Pr\{X \geq \tau\} = Pr\{X \geq L_B - X_B\} \\ &= \int_{x_B=0}^{L_B} \frac{\lambda e^{-\lambda x_B}}{1 - e^{-\lambda L_B}} \int_{x=L_B-x_B}^{\infty} \lambda e^{-\lambda x} \mathrm{d}x \; \mathrm{d}x_B \\ &= \frac{\lambda L_B q_B}{1 - q_B} \end{aligned} \tag{4.7}$$

Next, we derive $v_{A'}$. Suppose that the system transitions to A' at t. We examine two types of demand arrivals during A': **1)** the first demand arrival after t, and **2)** the other demand arrivals. Note that type 2 demand arrival, or having more than one demand arrivals in A', can only occur during the transition from state A' to state AB. Given a

demand arriving to state A′, the probability that it is of type 1 demand is

$$\frac{1/\lambda}{\mu_{A'}} = \frac{1}{\lambda L + q},$$

since $\mu_{A'} = \mu_A$. Because we assume $L \leq L_B$, the type 2 demand would never be served by an on-hand unit: the replenishment associated with its previous demand would never reach B during A′. Therefore, we focus on the type 1 demand.

Suppose that the system transitions to state A′ at time t as illustrated in Figure 4-10. At time t, the unit in A has L_A time units to reach A and the unit in B has $\tau < L_B$ to

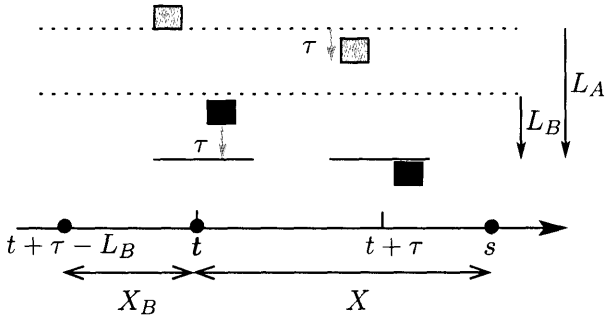

Figure 4-10: System Reaches State A′ at Time t

reach B. Then, the probability of a type 1 demand arrival being served by an on-hand unit is the probability of the next arrival (after t) arriving after $t+\tau$, such as at time s. The demand arrival, whose replenishment is the unit in B at t, must have arrived at $t+\tau-L_B$. In addition, there is no demand arrival during $(t+\tau-L_B, t)$: otherwise, the system could have reached state A′ earlier. Then, X_B represents the time between $(t+\tau-L_B, t)$. We can then derive the probability of a type 1 demand being served by an on-hand unit as:

$$\begin{aligned}(v_{A'} \mid \text{type 1 demand}) &= Pr\{X \geq \tau\} = Pr\{X \geq L_B - X_B\} \\ &= \frac{\lambda L_B q_B}{1 - q_B}.\end{aligned}$$

As a result, we derive the following:

$$\begin{aligned}v_{A'} &= Pr\{\text{type 1 demand}\}\ (v_{A'} \mid \text{type 1 demand}) \\ &\quad + Pr\{\text{type 2 demand}\}\ (v_{A'} \mid \text{type 2 demand}) \\ &= \frac{1}{\lambda L + q}\frac{\lambda L_B q_B}{1 - q_B} + \left(1 - \frac{1}{\lambda L + q}\right)(0) \\ &= \frac{1}{\lambda L + q}\frac{\lambda L_B q_B}{1 - q_B}\end{aligned} \tag{4.8}$$

We can derive $v_{A''}$ using the same logic:

$$\begin{aligned} v_{A''} &= Pr\{\text{type 1 demand}\}\ (v_{A''} \mid \text{type 1 demand}) \\ &\quad + Pr\{\text{type 2 demand}\}\ (v_{A''} \mid \text{type 2 demand}) \\ &= \frac{1}{\lambda L + q}(1) + \left(1 - \frac{1}{\lambda L + q}\right)(0) \\ &= \frac{1}{\lambda L + q} \end{aligned} \tag{4.9}$$

Finally, we proceed to derive v_{AB}. Suppose that the system transitions to state AB at time t as in Figure 4-11. This implies that the system reaches state A′ or A″ at $t - L$ and

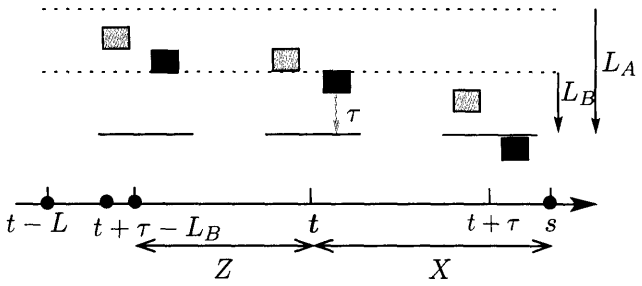

Figure 4-11: System Reaches State AB at Time t

there must be at least one demand arrival during $(t - L, t)$. Also, because we assume that $L \leq L_B$, the unit at B is always on order at t. At time t, the unit in A has L_B time periods to reach A and the one in B has $\tau < L_B$ time periods to reach B.

Since the unit in B reaches B at $t + \tau$, then its associated demand must have arrived at $t + \tau - L_B$ and it is the *last* demand before t. Let Z be the time between $(t + \tau - L_B, t)$, that is, let $Z = L_B - \tau$. To derive the distribution of τ, we must derive the distribution of Z. First, let

$N(t)$ the number of Poisson arrivals during t time periods.

Then, we derive the pdf of Z as:

$$\begin{aligned} f(Z = z) \cdot \delta &= Pr\{N(t - L, t - z) \geq 0,\ N(t - z, t - z + \delta) = 1, \\ &\qquad N(t - z + \delta, t) = 0 \mid N(t - L, t) \geq 1\} \\ &= \frac{1 \cdot \lambda\delta \cdot e^{-\lambda z}}{1 - q} \end{aligned}$$

Therefore, we have $f_Z(z) = \dfrac{\lambda e^{-\lambda z}}{1 - q}$ and Z is the conditional inter-arrival given it's less than

L. Consequently, we derive v_{AB} as:

$$\begin{aligned} v_{AB} &= Pr\{X \geq \tau\} = Pr\{X \geq L_B - Z\} \\ &= \int_{z=0}^{L} \frac{\lambda e^{-\lambda z}}{1-q} \int_{x=L_B-z}^{\infty} \lambda e^{-\lambda x} \, \mathrm{d}x\mathrm{d}z \\ &= \frac{\lambda}{1-q} \int_{z=0}^{L} e^{-\lambda L_B} \mathrm{d}z \\ &= \frac{\lambda L q_B}{1-q} \end{aligned} \tag{4.10}$$

We derive the fill rate of scenario (1,1) given $L \leq L_B$ as:

$$\begin{aligned} FR(1,1) &= q_B \frac{2-q_B}{2-q_B+\lambda L(1-q_B\alpha_2)} \left(1+\lambda L_B + \lambda L \frac{1-q_B\alpha_2}{2-q_B}\right), \quad L \leq L_B \\ &= q_B \left(1+\lambda L_B - \frac{(\lambda L)(\lambda L_B)}{\lambda L + \frac{2-q_B}{1-q_B\alpha_2}}\right) \end{aligned} \tag{4.11}$$

Note that f is increasing in α_2 for $\alpha_2 \in (0,1)$.

Observation. The fill rate for the 2U2L problem is a special case of $FR(1,1)$ with $L = 0$, $q_B(1+\lambda L_B)$, which is consistent with our analysis.

4.3.3 Comparison

We first compare the scenarios (1,1) and (2,0).

Proposition 4.3.2. *Let $L = L_A - L_B \geq 0$ and $q_B = e^{-\lambda L_B}$. We prefer scenario (2,0) if $\alpha_2 \in (0, \gamma_1)$, prefer (1,1) if $\alpha_2 \in (\gamma_1, 1)$, where*

$$\gamma_1 = \frac{2+q_B-2q_B^2+\lambda L(2+q_B) - \sqrt{2-q_B}\sqrt{4q_B^2(1+\lambda L-q_B)+(2-q_B)(1+\lambda L)^2}}{4q_B(1-q_B+\lambda L)}. \tag{4.12}$$

Proof. Let $f = g(1,1) - g(2,0)$, and

$$\begin{aligned} f &= g(1,1) - \alpha_2 \\ &= \frac{a\alpha_2^2 + b\alpha_2 + c}{2-q_B+\lambda L(1-q_B\alpha_2)} \end{aligned}$$

where

$$a = 2q_B(1-q_B+\lambda L), \; b = q_B(-1+2q_B-\lambda L) - 2 - 2\lambda L, \; c = 1-q_B+\lambda L.$$

Since the denominator $2-q_B+\lambda L(1-q_B\alpha_2) > 0$, finding the roots of f is equivalent to finding the roots of the nominator. Since $0 \leq q_B \leq 1$, $a \geq 0$ and $c \geq 0$. Also, since

$q_B(2q_B - 1) - 2 \leq 0$, $b \leq 0$.

The discriminant is

$$b^2 - 4ac = \frac{1}{2 - q_B + \lambda L(1 - q_B\alpha_2)}(2 - q_B)\left(4q_B^2(1 + \lambda L - q_B) + (2 - q_B)(1 + \lambda L)^2\right).$$

Clearly, the discriminant is nonnegative. Since $2 - q_B > 0$ and $1 + \lambda L > 0$, the discriminant is strictly positive.

In addition,

$$a + b + c = \frac{2 - q_B + \lambda L(1 - q_B\alpha_2)}{\lambda L(q_B - 1) - 1} < 0$$

Therefore, according to Lemma 4.2.1 Part 2), $f \geq 0$ for $\alpha_2 \in (0, \gamma_1)$ and $f \leq 0$ for $\alpha_2 \in (\gamma_1, 1)$, where $\gamma_1 = \dfrac{-b - \sqrt{b^2 - 4ac}}{2a}$. ■

Observation. Comparing with the solution in the 2U2L problem, we see that the two solutions are the same when $L = 0$.

Proposition 4.3.3. *Let $L = L_A - L_B \geq 0$ and $q_B = e^{-\lambda L_B}$. We prefer scenario (1,1) if $\alpha_2 \in (0, \gamma_2)$, prefer (0,2) if $\alpha_2 \in (\gamma_2, 1)$, where*

$$\gamma_2 = \frac{3q_B - 2q_B^2 - 2 + \sqrt{2 - q_B}\sqrt{(2 - q_B) + 4q_B^2(1 - q_B)}}{4q_B(1 - q_B)}. \tag{4.13}$$

Further more, γ_2 is the equivalent of the indifference point between scenario (1,1) and (0,2) in the 2U2L problem with the same leadtime L_B.

Proof. Let $f = g(1,1) - g(0,2)$, and $f = \dfrac{a\alpha_2^2 + b\alpha_2 + c}{2 - q_B + \lambda L(1 - q_B\alpha_2)}$ where

$$a = 2q_B(1 - q_B),\ b = 2 + q_B(2q_B - 3),\ c = -1.$$

Since the denominator $2 - q_B + \lambda L(1 - q_B\alpha_2) > 0$, thus, finding the roots of f is equivalent to finding the roots of the nominator which we denote as $\hat{f}$.

Next we consider the 2U2L problem with the same leadtime of L_B at both locations. Let $h = g(1,1) - g(0,2)$, and h is quadratic in α_2, where the coefficients are

$$a = \frac{2q_B(1 - q_B)}{2 - q_B},\ b = \frac{2 + 2q_B^2 - 3q_B}{2 - q_B},\ c = \frac{-1}{2 - q_B}.$$

Since the denominator $2 - q > 0$, thus, finding the roots of h is equivalent to finding roots of $\hat{h}$ whose coefficients are defined as

$$a = 2q_B(1 - q_B),\ b = 2 + 2q_B^2 - 3q_B,\ c = -1.$$

We see that $\hat{f} = \hat{h}$. Therefore, in both problems, we prefer scenario (1,1) for $\alpha_2 \in (0, \gamma_2)$, prefer (0,2) for $\alpha_2 \in (\gamma_2, 1)$, where γ_2 is defined as above. ■

Corollary 4.3.1. *We prefer (2,0) for $\alpha_2 \in (0, \gamma_2)$, prefer (1,1) for $\alpha_2 \in (\gamma_2, \gamma_1)$, prefer(0,2) for $\alpha_2 \in (\gamma_1, 1)$, where γ_1, γ_2 are defined in Equation (4.12), (4.13).*

Table 4.4 shows some numerical examples. For each value of expected demand during

		Fill Rate of Scenario			Range of α_2 to Choose
λL_A	λL_B	(2,0)	(1,1)	(0,2)	(1,1)
0.2	0.1	0.982	(0.988, 0.995)	0.995	(0.13, 0.92)
0.4	0.2	0.938	(0.959, 0.978)	0.982	(0.20, 0.87)
0.5	0.25	0.910	(0.940, 0.965)	0.973	(0.22, 0.85)
0.8	0.4	0.809	(0.876, 0.914)	0.938	(0.28, 0.80)
1	0.5	0.736	(0.830, 0.872)	0.910	(0.32, 0.77)
1.5	0.75	0.558	(0.710, 0.754)	0.827	(0.37, 0.72)

Table 4.4: 2-Unit 2-Location with Different Leadtimes Numerical Results

location A's leadtime, λL_A, and expected demand during location B's leadtime, λL_B, we present the fill rates for the three scenarios. In particular, the fill rate range for scenario (1,1) is for any value of $\alpha_2 \in [0, 1]$. We also present the range of α_2 for which we prefer scenario (1,1). If α_2 is smaller than that range, we prefer scenario (2,0) and if larger, we prefer scenario (0,2).

4.4 2-Unit 2-Location with Compound Poisson Demand

To model more demand variability in the 2-Unit 2-Location problem, we examine the 2-Unit 2-Location problem with Compound Poisson demand. We assume that demand arrive according to a Poisson process with rate λ. We assume at the i^{th} arrival, there are Y_i units in the order from either demand region 1 or 2, where Y_i is a discrete distribution with $Pr\{Y_i \geq 1\} = 1$. We assume the Y_i's are *i.i.d.* In addition, we execute the system with the following policy:

- If $Y_i = 1$, then we following the same execution policy as in the Poisson demand model.
- If $Y_i \geq 2$, the first two orders are matched with the two unassigned units in the system, and we order emergency orders for the remaining units in the order from the closest location (still with L unit leadtime).

We denote

$$p = Pr\{Y_i = 1\}.$$

The main purpose of this model is to examine the impact on the 2U2L model where there is more demand variability. For the sake of comparison, we only examine the cost of single orders, that is, orders for which $Y_i = 1$. We do this to simplify the presentation. We note that, by assumption, the cost to serve the multi orders is a constant that is independent of how the demand is split between the regions – this cost is just the cost of shipping one unit from each warehouse, plus the cost of the emergency shipment (if needed) for the remainder of the order. Therefore, the costs expressed below are expected transportation cost of a random single-item order.

We still have the same cost of $\underline{C} = \alpha_1 c_1 + \alpha_2 c_2$, $C(2,0) = \underline{C} + \alpha_2 \delta$, and $C(0,2) = \underline{C} + \alpha_1 \delta$.

The (1,1) scenario is quiet different from the 2U2L problem. To derive a similar Markov Chain, we need to add a state 0 as in Figure 4-12. We define a type-0 demand epoch where a demand arrival orders more than one unit. In particular, two units in the order are served by the two units in the system, and we trigger a replenishment at each warehouse at a type-0 demand epoch. That is, both replenishment units have exactly L time periods to reach their warehouses at the type-0 demand epoch. We define state 0 on a type-0 demand epoch. State A and B are as defined in the 2U2L problem. The steady state probabilities

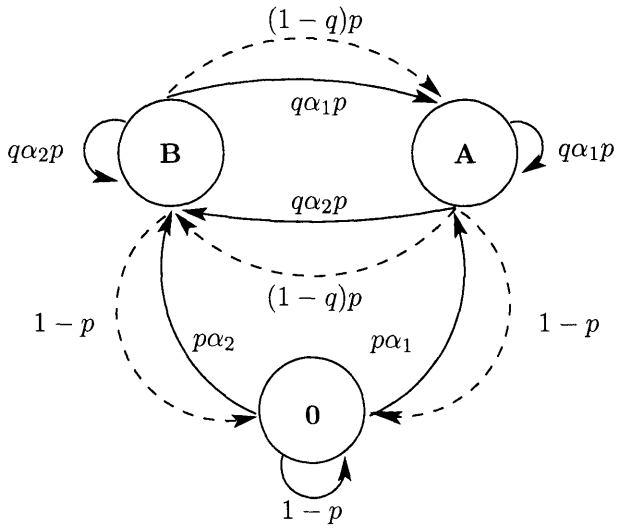

Figure 4-12: 2-Unit 2-Location Compound Poisson Markov Chain

of the Markov Chain are

$$\pi_0 = 1 - p, \quad \pi_B = p\,\frac{\alpha_2 + \alpha_1 p(1-q)}{1 + p(1-q)}, \quad \pi_A = p\,\frac{\alpha_1 + \alpha_2 p(1-q)}{1 + p(1-q)},$$

and again we can derive the Π's as from the dotted-line transition. The probability of a demand from region 2 and served by A is $\Pi_{A2} = \pi_b(1-q)p\alpha_2$, and the probability of a demand from region 1 and served by B is $\Pi_{B1} = \pi_a(1-q)p\alpha_1$.

Since the cost $C(1,1)$ is the expected cost of an order given it is a single order, we utilizes the conditional probabilities of the Π's, Π_{A2}/p, Π_{B1}/p.

$$C(1,1) = \underline{C} + \left(\frac{\Pi_{A2}}{p} + \frac{\Pi_{B1}}{p}\right)\delta$$

Let $g(1,1) = \dfrac{C(1,1)-\underline{C}}{\delta}$, then $g(1,1) = \dfrac{a\alpha_2^2 + b\alpha_2 + c}{1+p(1-q)}$ where

$$a = 2p(1-q)(1-p(1-q)), \quad b = 2p(1-q)(p(1-q)-1), \quad c = p(1-q)$$

Proposition 4.4.1. *Let* $q = e^{-\lambda L}$ *and* $\alpha = \dfrac{-b-\sqrt{b^2-4ac}}{2a}$ *where*

$$a = 2p(1-q)(1-p(1-q)), \quad b = 2p^2(1-q)^2 - 3p(1-q) - 1, \quad c = p(1-q)$$

If $\alpha_2 \in (\alpha, 1-\alpha)$, *then we prefer scenario (1,1); if* $\alpha_2 \leq \alpha$, *then we prefer scenario (2,0); if* $\alpha_2 \geq 1-\alpha$, *then we prefer scenario (0,2).*

We omit the proof here since it is very close to the proof in the 2U2L problem. Note that we do not need the full distribution of Y_i, but just the probability of a demand being a single order, p. We provide a few numerical examples in Table 4.5 for $p = 0.25$, $p = 0.5$, and $p = 0.75$. For each value of expected demand in leadtime, λL, we show the range of

	Range of α_2 to Choose (1,1)		
λL	$p = 0.25$	$p = 0.5$	$p = 0.75$
0.2	(0.04, 0.96)	(0.07, 0.93)	(0.10, 0.90)
0.5	(0.08, 0.92)	(0.13, 0.87)	(0.18, 0.82)
0.75	(0.10, 0.90)	(0.17, 0.83)	(0.22, 0.78)
1	(0.11, 0.89)	(0.19, 0.81)	(0.26, 0.74)
1.5	(0.13, 0.87)	(0.22, 0.78)	(0.30, 0.70)

Table 4.5: 2U2L with Compound Poisson Demand Numerical Results

α_2 for which we prefer scenario (1,1). For values of α_2 smaller than the range, we prefer scenario (2,0), and larger than the range, we prefer scenario (0,2).

4.5 2-Unit 3-Location (2U3L) Problem

We extend the 2U2L problem to consider three locations. Suppose there are three warehouses in the system: location A is close to the West Coast, C close to the East Coast, , and B in the center, as illustrated in Figure 4-13. We assume $\omega \geq 1$, and $\omega\delta$ is the penalty premium of shipping a region 1 demand from warehouse C or a region 3 demand from warehouse A. By assumption, for example, we prefer to ship a region 1 demand from warehouse

B than C. Suppose we stock two units of inventory in the system for an item, how should we best allocate the units so that the expected transportation cost is minimized? Consistent

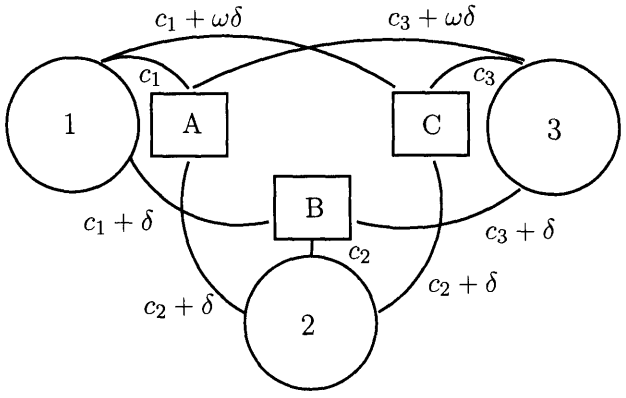

Figure 4-13: 2-Unit 3-Location Problem

with our previous definition, we define scenario (i, j, k) to be $IP_A = i, IP_B = j, IP_C = k$. Again, given the demand split $(\alpha_1, \alpha_2, \alpha_3)$, we want to compare the minimum cost scenario. We are mainly interested in given α_2, whether we should stock two units at location B, (0,2,0), to stock only one unit at B, (1,1,0) or (0,1,1), or to stock zero unit at B, (1,0,1), (2,0,0), (0,0,2).

Clearly, we have

$$\underline{C} = \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3$$

4.5.1 Scenario (0,2,0)

We stock two units of inventory in warehouse B only, and we have

$$C(0,2,0) = \alpha_1(c_1 + \delta) + \alpha_2 c_2 + \alpha_3(c_3 + \delta) = \underline{C} + (\alpha_1 + \alpha_3)\delta$$

Let $\boldsymbol{g(i,j,k)} = \dfrac{C(i,j,k) - \underline{C}}{\delta}$, and let $\boldsymbol{a_{ijk}, b_{ijk}, c_{ijk}}$ be the coefficients of $g(i,j,k)$ as a quadratic function of α_2. Then $g(0,2,0) = 1 - \alpha_2$ and $a_{020} = 0, b_{020} = -1, c_{020} = 1$.

The same reasoning applies to scenario (2,0,0) and (0,0,2). Let $\boldsymbol{\kappa} = \dfrac{\alpha_1}{1-\alpha_2}$ and $\boldsymbol{1 - \kappa} = \dfrac{\alpha_3}{1-\alpha_2}$.

$$\begin{aligned} C(2,0,0) &= \underline{C} + (\alpha_2 + \omega\alpha_3)\delta \\ &= \underline{C} + (\alpha_2 + \omega(1-\kappa)(1-\alpha_2))\delta \\ C(0,0,2) &= \underline{C} + (\alpha_2 + \omega\alpha_1)\delta \\ &= \underline{C} + (\alpha_2 + \omega\kappa(1-\alpha_2))\delta \end{aligned}$$

We have $a_{200} = 0, b_{200} = 1-\omega(1-\kappa), c_{200} = \omega(1-\kappa)$ and $a_{002} = 0, b_{002} = 1-\omega\kappa, c_{002} = \omega\kappa$.

4.5.2 Scenario (1,0,1)

We stock one unit each at location A and C. In addition to the assumptions in the 2U2L Problem, we assume that for a demand from region 2, it is equally likely to be served by location A and C if both have on-hand inventory.

$$\begin{aligned} C(1,0,1) &= \underline{C} + (\Pi_{B1} + \Pi_{B3} + \omega(\Pi_{A3} + \Pi_{C1}))\,\delta \\ &= \underline{C} + (\alpha_2 + \omega(\Pi_{A3} + \Pi_{C1}))\,\delta \end{aligned}$$

We pay a premium of δ for serving a region 2 demand since there is no inventory stocked at warehouse B. We pay a premium of $\omega\delta$ for serving a region 1 and region 3 demand from a further warehouse.

Similar to Scenario (1,1) in the 2-Unit 2-Location problem, we find the probabilities of Π_{A3}, Π_{C1} by examining the Markov Chain in Figure 4-14. The steady state probabilities

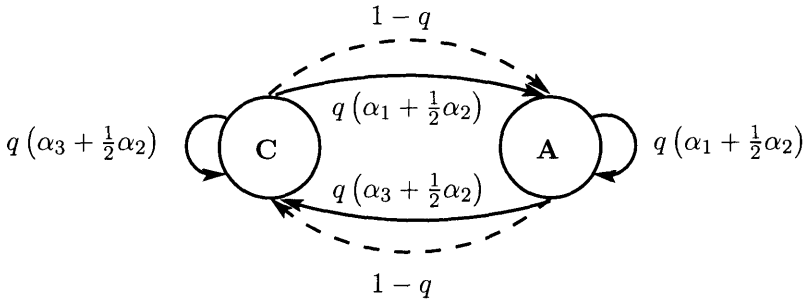

Figure 4-14: 2-Unit 3-Location Markov Chain, Scenario (1,0,1)

are

$$\pi_C = \frac{1 - q\left(\alpha_1 + \frac{1}{2}\alpha_2\right)}{2 - q}, \quad \pi_A = \frac{1 - q\left(\alpha_3 + \frac{1}{2}\alpha_2\right)}{2 - q}$$

We then have the probability of a demand from region 3 and served by A $\Pi_{A3} = \pi_C(1-q)\alpha_3$, and the probability of a demand from region 1 and served by C $\Pi_{C1} = \pi_A(1-q)\alpha_1$. Moreover, the probability of a demand from region 2 and served by a further warehouse is α_2. We then have

$$C(1,0,1) = \underline{C} + \left(\alpha_2 + \omega\theta(1-\alpha_2)\left(1 - 2q\kappa' + 2q\kappa'\alpha_2 - \frac{1}{2}q\alpha_2\right)\right)\delta,$$

where $\boldsymbol{\theta} = \dfrac{1-q}{2-q}$, $\boldsymbol{\kappa'} = \kappa(1-\kappa)$, and $\kappa = \alpha_1/(1-\alpha_2)$, $(1-\kappa) = \alpha_3/(1-\alpha_2)$. Also, $g(1,0,1)$ is a quadratic function in α_2, and

$$a_{101} = \omega\theta q\left(\frac{1}{2} - 2\kappa'\right), \quad b_{101} = 1 - \omega\theta\left(1 + q\left(\frac{1}{2} - 4\kappa'\right)\right), \quad c_{101} = \omega\theta(1 - 2q\kappa'). \quad (4.14)$$

Since $0 \leq \kappa \leq 1, \kappa' \leq \frac{1}{4}$, $a_{101} \geq 0$, $\forall \lambda L \geq 0$. Therefore, $g(1,0,1)$ is convex in α_2. For $\alpha_1 = \alpha_3$, i.e., $\kappa' = \frac{1}{4}$, $g(1,0,1)$ is linear in α_2 and

$$a_{101} = 0, \quad b_{101} = 1 - \frac{\omega}{2}(1-q), \quad c_{101} = \frac{\omega}{2}(1-q), \quad \text{if } \alpha_1 = \alpha_3. \tag{4.15}$$

4.5.3 Scenario (1,1,0)

We consider the scenario where we stock one unit in location A and B. We can write the cost as

$$C(1,1,0) = \underline{C} + (\Pi_{A2} + \Pi_{B1} + \omega\Pi_{A3} + \Pi_{B3})\,\delta$$

To derive the Π's, we construct the Markov Chain in Figure 4-15. Since we assume $\omega \geq 1$, to demand stream 3, location B is the closer warehouse with inventory. The steady state

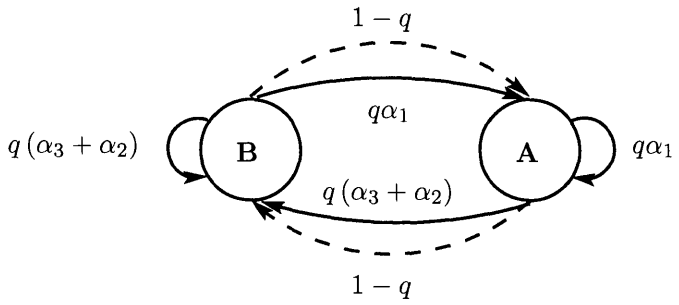

Figure 4-15: 2-Unit 3-Location Markov Chain, Scenario (1,1,0)

probabilities are

$$\pi_A = \frac{1 - q(\alpha_2 + \alpha_3)}{2 - q}, \quad \pi_B = \frac{1 - q\alpha_1}{2 - q}$$

The Π's are derived as

$$\Pi_{A2} = \pi_B(1-q)\alpha_2 \qquad \Pi_{A3} = \pi_B(1-q)\alpha_3$$
$$\Pi_{B1} = \pi_A(1-q)\alpha_1 \qquad \Pi_{B3} = \pi_A(1-q)\alpha_3 + q\alpha_3$$

The cost is then

$$C(1,1,0) = \underline{C} + (q\alpha_3 + \theta\left((1 - q(\alpha_2 + \alpha_3))\,(\alpha_1 + \alpha_3) + (1 - q\alpha_1)(\omega\alpha_3 + \alpha_2)\right))\,\delta$$

where $\theta = \frac{1-q}{2-q}$. Also, $g(1,1,0)$ is a quadratic function in α_2 where

$$\begin{aligned} a_{110} &= \theta q\left(2\kappa - \omega\kappa'\right), \\ b_{110} &= \theta(\omega-1)(2q\kappa' - (1-\kappa)) - \theta 2q\kappa^2 - \frac{1-\kappa}{2-q}, \\ c_{110} &= \frac{1-q\kappa}{2-q}\left(1+\omega(1-q)(1-\kappa)\right). \end{aligned} \tag{4.16}$$

For $\alpha_1 = \alpha_3$, we have

$$a_{110} = \theta q(1-\frac{\omega}{4}),\ b_{110} = \frac{2q^2+2\omega q-\omega q^2-3q-\omega}{2(2-q)},\ c_{110} = \frac{1}{2}+\frac{\omega}{4}(1-q),\ \text{if } \alpha_1 = \alpha_3. \tag{4.17}$$

The cost of scenario (0,1,1) is symmetric to (1,1,0), and we obtain $C(0,1,1)$ by substituting α_1 with α_3 and α_3 with α_1.

4.5.4 Comparison

Let $g = \min_{\forall ijk} g(i,j,k)$. For a given value of α_2, the minimum cost scenario is the scenario (i,j,k) s.t. $g(i,j,k) = g$. We devote the rest of the section on the special case of $\alpha_1 = \alpha_3$.

Special case of $\alpha_1 = \alpha_3$

In the special case of $\alpha_1 = \alpha_3$, we have the symmetry of α_1 and α_3. Therefore, the cost of scenario (i,j,k) is equivalent to (k,j,i).

Proposition 4.5.1. *If $\alpha_1 = \alpha_3$, then we always prefer scenario (1,0,1) over (2,0,0) or (0,0,2): $g(2,0,0) = g(0,0,2) \geq g(1,0,1)$ for $0 \leq \alpha_2 \leq 1$.*

Proof. Since $\alpha_2 \leq 1$, by definition,

$$\begin{aligned} g(2,0,0) &= (1-\frac{\omega}{2})\alpha_2 + \frac{\omega}{2} \\ &\geq (1-\frac{\omega}{2}(1-q))\alpha_2 + \frac{\omega}{2}(1-q) = g(2,0,0) + \frac{\omega}{2}q(\alpha_2 - 1) = g(1,0,1). \end{aligned}$$ ■

Therefore, to find the minimum cost scenario, we only consider scenarios (1,0,1), (1,1,0), and (0,2,0). In the remainder of the section, we derive the values of α_2 such that we are indifferent to any two scenarios among (1,0,1), (1,1,0), and (0,2,0).

Proposition 4.5.2. *Given $1 \leq \omega \leq 2$, if $\alpha_1 = \alpha_3$, we prefer scenario (1,0,1) over (0,2,0) for $\alpha_2 \in (0, \gamma_1)$ and prefer scenario (0,2,0) over (1,0,1) for $\alpha_2 \in (\gamma_1, 1)$, where*

$$\gamma_1 = \frac{1-\frac{\omega}{2}(1-q)}{2-\frac{\omega}{2}(1-q)}. \tag{4.18}$$

Proof. Let function $f = g(1,0,1) - g(0,2,0)$, and f is a linear function in α_2 given $\alpha_1 = \alpha_3$. Let γ_1 be such that $f(\gamma_1) = 0$. Then, $\gamma_1 = -\dfrac{c_{101} - c_{020}}{b_{101} - b_{020}} = \dfrac{1 - \frac{\omega}{2}(1-q)}{2 - \frac{\omega}{2}(1-q)}$. Since $\omega \leq 2$ and $0 \leq q \leq 1$, then $0 \leq \dfrac{\omega}{2}(1-q) \leq 1$. Then, $\gamma_1 \leq \dfrac{1}{2} < 1$. Also, f is an increasing function of α_2 since $b_{101} - b_{020} \geq 0$. Therefore, $f \leq 0$ for $0 \leq \alpha_2 \leq \gamma_1$ and $f \geq 0$ for $1 \geq \alpha_2 \geq \gamma_1$. ■

For $\omega = 1.4$, we have the numerical results in Table 4.6. For each value of expected

		Range of α_2 to Choose (1,0,1) Over (0,2,0)		
λL	Fill Rate	$\alpha_1 = \alpha_3$	$\alpha_1 = 3\alpha_3$	$\alpha_1 = 0$
0.2	0.98	(0, 0.47)	(0, 0.46)	(0, 0.45)
0.4	0.94	(0, 0.43)	(0, 0.43)	(0, 0.41)
0.5	0.91	(0, 0.42)	(0, 0.41)	(0, 0.39)
0.75	0.83	(0, 0.39)	(0, 0.38)	(0, 0.36)
1	0.73	(0, 0.36)	(0, 0.35)	(0, 0.33)
1.5	0.56	(0, 0.31)	(0, 0.31)	(0, 0.29)

Table 4.6: 2-Unit 3-Location Numerical Results, Scenario (1,0,1) and (0,2,0)

demand in leadtime, λL, we show the fill rate; we also present the range of α_2 for which we prefer scenario (1,0,1) over scenario (0,2,0) for three cases: i) the special case of $\alpha_1 = \alpha_3$ as in our previous discussion, ii) $\alpha_1 = 3\alpha_3$, and iii) $\alpha_1 = 0, \alpha_3 = 1$.

We compare scenario (1,1,0) and (0,2,0).

Proposition 4.5.3. *Given $1 \leq \omega \leq 2$, if $\alpha_1 = \alpha_3$, we prefer scenario (1,1,0) over (0,2,0) for $\alpha_2 \in (0, \gamma_2)$ and prefer scenario (0,2,0) over (1,1,0) for $\alpha_2 \in (\gamma_2, 1)$. The value γ_2 is defined as $\gamma_2 = \dfrac{-b + \sqrt{b^2 - 4ac}}{2a}$ where*

$$\begin{aligned} a &= \theta q(1 - \frac{\omega}{4}), \\ b &= \frac{4 - 5q + 2q^2 + \omega(2q - 1 - q^2)}{2(2-q)}, \\ c &= \frac{1}{2}\left(-1 + \frac{\omega}{2}(1-q)\right) \end{aligned}$$

Proof. Let function $f = g(1,1,0) - g(0,2,0)$, and f is a quadratic function of α_2 with coefficients defined above. Since $\omega \leq 2$, $a \geq 0$. In addition, $c \leq 0$. Since $\omega \leq 2$,

$$b = \frac{2q^2 - 5q + 4 - \omega(1-q)^2}{2(2-q)} \geq \frac{2-q}{2(2-q)} = \frac{1}{2} > 0.$$

Finally, $a + b + c = \theta \geq 0$, $\gamma_2 \leq 1$. Therefore, by Lemma 4.2.1 Part 1) $f \leq 0$ for $\alpha_2 \in (0, \gamma_2)$ and $f \geq 0$ for $\alpha_2 \in (\gamma_2, 1)$. ■

λL	Fill Rate	Range of α_2 to Choose (1,1,0) Over (0,2,0) $\alpha_1 = \alpha_3$	$\alpha_1 = 3\alpha_3$	$\alpha_1 = 1$
0.2	0.98	(0, 0.76)	(0, 0.84)	(0, 0.87)
0.4	0.94	(0, 0.65)	(0, 0.75)	(0, 0.80)
0.5	0.91	(0, 0.60)	(0, 0.71)	(0, 0.78)
0.75	0.83	(0, 0.52)	(0, 0.65)	(0, 0.72)
1	0.73	(0, 0.45)	(0, 0.59)	(0, 0.68)
1.5	0.56	(0, 0.36)	(0, 0.52)	(0, 0.61)

Table 4.7: 2-Unit 3-Location Numerical Results, Scenario (1,1,0) and (0,2,0)

For each value of expected demand in leadtime, λL, we show the fill rate; we also present the range of α_2 for which we prefer scenario (1,1,0) over scenario (0,2,0) for three cases: i) the special case of $\alpha_1 = \alpha_3$ as in our previous discussion, ii) $\alpha_1 = 3\alpha_3$, and iii) $\alpha_1 = 0, \alpha_3 = 1$.

Proposition 4.5.4. *Given* $1 \leq \omega \leq 2$, *if* $\alpha_1 = \alpha_3$, *we prefer scenario (1,0,1) over (1,1,0) for* $\alpha_2 \in (0, \gamma_3)$ *and prefer scenario (1,1,0) over (1,0,1) for* $\alpha_2 \in (\gamma_3, 1)$, *where* $\gamma_3 = \dfrac{-b - \sqrt{b^2 - 4ac}}{2a}$ *and*

$$a = \theta q(1 - \frac{\omega}{4}),\ b = \frac{2q^2 - q - 4 + \omega(1-q)}{2(2-q)},\ c = \frac{1}{2}\left(1 - \frac{\omega}{2}(1-q)\right).$$

Proof. Let function $f = g(1,1,0) - g(1,0,1)$, and f is a quadratic function of α_2 with coefficients defined above. Clearly, $a \geq 0$, since $\omega \leq 2$. By the same reason, $c \geq 0$. In addition,

$$b \leq \frac{2q^2 - q - 4 + 2(1-q)}{2(2-q)} = \frac{-(2-q)}{2(2-q)} = -\frac{1}{2} < 0.$$

Since we have

$$4ac = \frac{q(1-q)(4-\omega)(2+\omega(1-q))}{4(2-q)} \leq \frac{q}{8}(4-\omega)(2+\omega) \leq \frac{q}{8},$$

then

$$b^2 - 4ac \geq (q + \frac{1}{2})^2 - \frac{q}{8} > 0.$$

Furthermore, $a + b + c = -\dfrac{1}{2-q} \leq 0$.

Therefore, f satisfies the conditions in Lemma 4.2.1 Part 2). Let $\gamma_3 = \dfrac{-b - \sqrt{b^2 - 4ac}}{2a}$. Then, $f \geq 0$ for $\alpha_2 \in (0, \gamma_3)$ and $f \leq 0$ for $\alpha_2 \in (\gamma_3, 1)$. ∎

Table 4.8 has a few numerical examples. For each value of expected demand in leadtime,

		Range of α_2 to Choose (1,0,1) Over (1,1,0)		
λL	Fill Rate	$\alpha_1 = \alpha_3$	$\alpha_1 = 3\alpha_3$	$\alpha_1 = 1$
0.2	0.98	(0, 0.32)	(0, 0.17)	(0, 0)
0.4	0.94	(0, 0.32)	(0, 0.15)	(0, 0)
0.5	0.91	(0, 0.31)	(0, 0.14)	(0, 0)
0.75	0.83	(0, 0.30)	(0, 0.11)	(0, 0)
1	0.73	(0, 0.29)	(0, 0.09)	(0, 0)
1.5	0.56	(0, 0.27)	(0, 0.05)	(0, 0)

Table 4.8: 2-Unit 3-Location Numerical Results, Scenario (1,0,1) and (1,1,0)

λL, we show the fill rate; we also present the range of α_2 for which we prefer scenario (1,0,1) over scenario (1,1,0) for three cases: i) the special case of $\alpha_1 = \alpha_3$ as in our previous discussion, ii) $\alpha_1 = 3\alpha_3$, and iii) $\alpha_1 = 0, \alpha_3 = 1$.

Theorem 4.5.1. *Given $1 \leq \omega \leq 2$ and $\alpha_1 = \alpha_3$, we prefer scenario (1,0,1) if $\alpha_2 \in (0, \gamma_3)$,(1,1,0) or (0,1,1) if $\alpha_2 \in (\gamma_3, \gamma_2)$, (0,2,0) if $\alpha_2 \in (\gamma_2, 1)$, where the γ's are defined previously.*

Proof. Given α_2, we need to find a scenario (i, j, k) whose penalty $g(i, j, k)$ is the minimum among scenario (1,0,1), (1,1,0), and (0,2,0). The function $g(0, 2, 0)$ is a linear function in α_2 with negative slop. The function $g(1, 0, 1)$ is also linear in α_2 but with positive slop. In addition, the two function intersect at $\alpha_2 = \gamma_1 < 1$.

The function $g(1, 1, 0)$ is convex in α_2 given $1 \leq \omega \leq 2$, and it intersects with $g(0, 2, 0)$ at $\alpha_2 = \gamma_2 \leq 1$ and with $g(1, 0, 1)$ at $\alpha_2 = \gamma_3 \leq 1$. We claim that $g(1, 1, 0)$ is a non-increasing function in the interval $(0, 1)$ and given $\alpha_2 = \gamma_1$, $g(1, 1, 0) \leq g(0, 2, 0) = g(1, 0, 1)$. Therefore, for $\alpha_2 \in (\gamma_3, \gamma_2)$, $g(1, 1, 0) \leq \min(g(0, 2, 0), g(1, 0, 1))$. Then, we prefer scenario (1,0,1) if $\alpha_2 \in (0, \gamma_3)$,(1,1,0) or (0,1,1) if $\alpha_2 \in (\gamma_3, \gamma_2)$, (0,2,0) if $\alpha_2 \in (\gamma_2, 1)$.

Here we proceed to show that $g(1, 1, 0)$ is non-increasing in the interval (0,1). The minimum of $g(1, 1, 0)$ is achieved at $\alpha_2 = -\dfrac{b}{2a}$,

$$\begin{aligned} -\frac{b}{2a} &= \frac{3q - 2q^2 + \omega(1-q)^2}{(4-\omega)(1-q)q} \geq \frac{3q - 2q^2 + (1-q)^2}{(4-\omega)(1-q)q} = \frac{q(1-q)+1}{(4-\omega)(1-q)q} \\ &\geq \frac{1}{4-\omega}\left(1 + \frac{1}{q(1-q)}\right) \geq \frac{1}{3}(1+4) \geq 1 \end{aligned}$$

Since $g(1, 1, 0)$ is convex, it is non-increasing in $\alpha \leq -\dfrac{b}{a} \Rightarrow g(1, 1, 0)$ non-increasing in $\alpha \leq 1$.

We also need to show that at $\alpha_2 = \gamma_1$, $g(1,1,0) \leq g(0,2,0) = g(1,0,1)$. At $\alpha_2 = \gamma_1$,

$$g(1,1,0) - g(0,2,0) = \frac{q(2-\omega(1-q))(4-2q-\omega(1-q))}{(q-2)(4-\omega(1-q^2))}.$$

The nominator is positive given $1 \leq \omega \leq 2$. The denominator is negative since $q - 2 \leq 0$ and $4 - \omega(1-q^2) \geq 0$. Therefore, $g(1,1,0) \geq g(0,2,0)$. ■

We summarize the numerical examples in Table 4.9 and 4.10. In Table 4.9, we focus on examples where $\alpha_1 - \alpha_3$ and in Table 4.10, we focus on examples where $\alpha_1 = 3\alpha_3$. For each value of expected demand in leadtime, λL, we show the fill rate as well as the range of α_2 for which we prefer scenario (1,0,1), (1,1,0) or (0,2,0).

		Range of α_2 to Choose		
λL	Fill Rate	(1,0,1)	(1,1,0)	(0,2,0)
0.2	0.98	(0, 0.32)	(0.32, 0.76)	(0.76, 1)
0.4	0.94	(0, 0.32)	(0.32, 0.65)	(0.65, 1)
0.5	0.91	(0, 0.31)	(0.31, 0.60)	(0.60, 1)
0.75	0.83	(0, 0.30)	(0.30, 0.52)	(0.52, 1)
1	0.73	(0, 0.29)	(0.29, 0.45)	(0.45, 1)
1.5	0.56	(0, 0.27)	(0.27, 0.36)	(0.36, 1)

Table 4.9: 2-Unit 3-Location Numerical Results, $\alpha_1 = \alpha_3$

		Range of α_2 to Choose			
λL	Fill Rate	(2,0,0)	(1,0,1)	(1,1,0)	(0,2,0)
0.2	0.98		(0, 0.17)	(0.17, 0.84)	(0.84, 1)
0.4	0.94		(0, 0.15)	(0.15, 0.75)	(0.75, 1)
0.5	0.91		(0, 0.13)	(0.13, 0.71)	(0.71, 1)
0.75	0.83	(0, 0.15)		(0.15, 0.65)	(0.65, 1)
1	0.73	(0, 0.20)		(0.20, 0.59)	(0.59, 1)
1.5	0.56	(0, 0.27)		(0.27, 0.52)	(0.52, 1)

Table 4.10: 2-Unit 3-Location Numerical Results, $\alpha_1 = 3\alpha_3$

4.6 Summary

In this chapter, we examine inventory planning for low-demand SKUs. We investigate a few simple cases where given the amount of system inventory, we find the stocking allocation with the minimum expected outbound transportation costs. Here we summarize our findings.

In the 2-Unit 2-Location problem, we have demand region 1 and 2 where 1 is closer to warehouse A and 2 is closer to warehouse B. We find that

- a balanced split of demand in region 1 and 2 results in a minimum cost of balanced stocking allocation, namely (1,1). We give the specific range of demand splits such that either scenario (2,0), (1,1), or (0,2) has the minimum cost, for a given expected demand in leadtime λL.

- with higher fill rate (i.e., small λL), the range of demand splits for which scenario (1,1) has the minimum cost increases.

- in scenario (1,1), the probability of a demand served from warehouse A, P_A, is within the range of $[\min(\alpha_1, \frac{1}{2}), \max(\alpha_1, \frac{1}{2})]$, and the probability of a demand served from warehouse B, P_B, is within the range of $[\min(\alpha_2, \frac{1}{2}), \max(\alpha_2, \frac{1}{2})]$. That is, a demand is always more likely to be served by a warehouse that's close to a larger demand region.

In the 2-Unit 2-Location problem with different leadtimes, we add an assumption of $L_A = L + L_B,\ L \geq 0$. While other observations are consistent with the 2U2L problem, we find that

- the range of α_2 for which scenario (1,1) is the minimum-cost scenario is larger than the range in the 2U2L problem with a leadtime of L_A but smaller than the range in the 2U2L problem with a leadtime of L_B.

- if the difference of expected demand in leadtime between A and B is large enough (i.e., $\lambda L \geq q_B$), then a demand is always more likely to be served by location B. If the difference is small, then a demand is more likely to be served by A when α_2 is small and by B when α_2 is large.

In the 2-Unit 2-Location problem with Compound Poisson demand, we find that

- the larger the demand variance (i.e., smaller value of p), the larger the range of α_2 for which scenario (1,1) is the minimum-cost scenario.

In the 2-Unit 3-Location problem, we have two locations (A, C), one on each coasts, and location B in the middle of US. We find that

- for a reasonable demand fill rate (e.g., 0.9), we suppose the shipping premium across the coasts is 40% more than the premium from the middle of the country to either coast. If the demand region closest to B has more than 40% of the total demand, then we prefer stocking the two units of inventory in B, regardless of the demand split among region 1 and 3.

Overall, our results are intuitively satisfying. We are also able to provide specific guidelines for choosing the best inventory allocation for these simple cases. An important assumption in the model is on constant leadtimes. We believe the methodology can also be extended for stochastic leadtimes but with an assumption on no order crossing.

For problems with large units of system inventory and/or large number of locations, the state space explodes in our current methodology. However, we believe that the 2U2L methodology can be used as an approximation to a Multi-Unit 2-Location problem. In the approximation, we still have states A and B associated with the two locations, but we approximate the transition probabilities. Then, this approximation of the Multi-Unit 2-Location problem can be used as a basis for a general Multi-Unit Multi-Location problem. That is, any Multi-Unit Multi-Location problem may be decomposed into a Multi-Unit 2-Location problem. We can then extend the current methodology to analyze a more general problem.

Chapter 5

Conclusion

In this dissertation, we present three problems motivated by the customer fulfillment process in online retailing. These three problems underscore 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, tactical or operational.

In Chapter 2, we solve a multi-item two-stage serial inventory model with stochastic demand and space constraints. We are able to show with real data the large scale of the system as well as the effect of delay in demand fulfillment on inventory planning. In Chapter 3, we generate near-optimal heuristics to reduce the number of shipments in shipping from warehouses to customers. We show that there is a significant cost saving by exploiting the order-to-delivery window. The problem again illustrates the challenge due to the large scale of the system. In Chapter 4, we present a methodology and solve a few simple models of inventory allocation for low-demand SKUs. The results are intuitively satisfying. We also present specific guidelines to allocate inventory in the system based on outbound transportation costs. By allocating inventory efficiently, e-tailers are able to have high customer service level and low costs.

The three problems in this dissertation covers a few of the important issues in online retailing. Here are some additional considerations.

Combining marketing-operations In the second problem, we minimize the total number of shipments shipping from warehouses to customers. That is, the e-tailer prefers customers ordering items that are available in a single warehouse, in particular, in the warehouse that's closest to the customer. E-tailers often suggest a group of sale items or items that may be of interest to their customers. However, this marketing function is often separate from the operations function. That is, e-tailers often do not consider the availability of inventory in warehouses when making marketing decisions. By combining the marketing and operations function – by suggesting the group of

items that available at the closest warehouse to the customer, e-tailers may be able to reduce costs.

Impact of multi-item orders All problems in this dissertation demonstrate that some customer orders have multiple items create a substantial challenge to online e-tailers. First, the items in a customer order need to be synchronized in terms of location and timing. Second, e-tailers also need to consider the demand correlation of items when making inventory stocking decisions. Having an effective inventory planning process that balances the trade-off of costs and service level is crucial. For instance, a relevant question is given the warehouses are limited in storage space, what subset of SKUs should be stock in what warehouses?

Appendix A

Appendix

A.1 Single-Staged Exact (R, T) Model for Poisson Demand

Following the literature [e.g., HW63], we write the exact expected total cost per unit time for Poisson demand as:

$$
\begin{aligned}
C(R,T) &= \frac{a}{T}P(1|T) && \text{(A.1)}\\
&\quad + h\left(R - d(l+\frac{T}{2})\right) + \frac{h}{T}\int_l^{l+T}\sum_{x=R}^{\infty}(x-R)p(x|t)dt && \text{(A.2)}\\
&\quad + \frac{b}{T}\sum_{x=R}^{\infty}(x-R)p(x|l+T) - \frac{b}{T}\sum_{x=R}^{\infty}(x-R)p(x|l) && \text{(A.3)}
\end{aligned}
$$

where $p(x|\tau)$ is the PMF of demand during time τ, and $P(x|\tau) = \sum_{k=x}^{\infty} p(k|\tau)$ is the right-hand CDF of demand during time τ.

The first cost term (A.1) is the fixed ordering cost. We charge a per review period iff there is nonzero demand during the review period. The second term (A.2) is the holding cost term. The expected holding cost per period is

$$h\int_l^{l+T} E[I(t)]\mathrm{d}t = h\int_l^{l+T} E[IL(t)] + E[B(t)]\mathrm{d}t.$$

The expected holding cost per time period is the expected holding cost per period multiple by $1/T$. The third term (A.3) is the backorder cost term. The expected number of backorders incurred in a review period is the expected number of backorders incurred between time l and $l+T$:

$$E[B(l+T)] - E[B(l)].$$

A.2 Exact Two-Stage Serial Model

Here we derive an exact two-stage model to serve as a benchmark for the approximate two-stage model in § 2.2.2. In the exact model, we relax assumptions A-4, A-7, A-8, and A-9. Essentially, we build the exact two-stage model based on the exact single-stage (R, T) model but with the nested policy assumption A-6.

Echelon-2 by itself is a single stage. Therefore, the development of the echelon-2 setup cost and holding cost is the same as in the exact single-stage model with order-up-to level R_2 and review period T_2. We only derive the echelon-1 setup, holding, and backorder cost below.

Recall that a cycle is the time between successive echelon-2 replenishment.

A.2.1 Echelon-1 Holding Cost

We derive the holding cost based on the following equation:

$$\frac{h_1}{nT_1}\sum_{j=0}^{n-1}\int_{l+jT_1}^{l+(j+1)T_1} E[IL(t)] + E[B(t)]\ \mathrm{d}t.$$

The value $\int_{l+jT_1}^{l+(j+1)T_1} E[IL(t)] + E[B(t)]\ \mathrm{d}t$ is the expected on-hand inventory during the $(j+1)^{\text{st}}$ stage 1 replenishment in a cycle, and we devote the rest of the section on it's derivation.

In stage 1, we order up to R_1 if stage 2 has sufficient stock to satisfy the replenishment. Otherwise, we exhaust the stage 2 inventory. Therefore, the stage 1 inventory position is always less or equal to R_1. We define

$\mathbf{v_k}$ shortfall of echelon-2 inventory from R_1 at the kth review in a cycle

.

Given we order up to R_2 at time $t = 0$, the inventory level at the kth stage-1 review time is $IL_2(l_2 + (k-1)T_1)$. The shortfall at the kth review time is

$$\begin{aligned} v_k &= \max\{0, R_1 - IL_2(l_2 + (k-1)T_1)\} \\ &= \max\left\{0, R_1 - \Big(R_2 - D(l_2 + (k-1)T_1)\Big)\right\}. \end{aligned} \tag{A.4}$$

That is, we start the kth stage-1 review with inventory position of $R_1 - v_k = \min\{R_1, R_2 - D(l_2 + (k-1)T_1)\}$.

To derive the expected on-hand inventory in a stage-1 replenishment, we first derive the expected net inventory. The net inventory at the beginning of the kth stage-1 replenishment

cycle is

$$R_1 - v_k - D(l_1),$$

since a replenishment review took place l_1 time periods ago. The net inventory at the end of the kth replenishment is

$$R_1 - v_k - D(l_1 + T_1).$$

Then, the expected net inventory per unit time of the kth stage-1 replenishment cycle is:

$$\begin{aligned} & \frac{1}{2}\left(E[R_1 - v_k - D(l_1)] + E[R_1 - v_k - D(l_1 + T_1]\right) \\ = \; & R_1 - \bar{v}_k - dl_1 - dT_1/2, \end{aligned}$$

where $\bar{v}_k$ is the expected value of shortfall. The we write $\bar{v}_k$ as

$$\begin{aligned} \bar{v}_k &= E[D(l_2 + (k-1)T_1 - (R_2 - R_1)]^+ \\ &= \sum_{R_2 - R_1}^{\infty} x - (R_2 - R_1)\; p(x|l_2 + (k-1)T_1). \end{aligned}$$

The expected on-hand inventory is the sum of expected net inventory and backorders. We define

$\mathbf{I_k}$	expected on-hand inventory
	during the kth stage-1 replenishment cycle

Then, we have

$$I_k = (R_1\bar{v}_k - dl_1 - dT_1/2)\,T_1 + \int_{l_1}^{l_1+T_1} E\left[D(t) + v_k - R_1\right]^+ \mathrm{d}t, \tag{A.5}$$

where $E\left[D(t) + v(k) - R_1\right]^+$ is the expected number of backorders at time t. We can expand $E\left[D(t) + v(k) - R_1\right]^+$ as:

$$\begin{aligned} &= E[\max\{D(t) - R_1, D(t) - R_2 + D(l_2 + (k-1)T_1)\}] \\ &= E[D(t) - R_1]^+\; P(D(l_2 + (k-1)T_1) \le R_2 - R_1) \\ &\quad + E[D(t) + D(l_2 + (k-1)T_1) - R_2]^+\; P(D(l_2 + (k-1)T_1) > R_2 - R_1) \\ &= \left(\sum_{R_1}^{\infty}(x - R_1)p(x|t)\right)\left(\sum_{y=0}^{R_2-R_1} p(y|l_2 + (k-1)T_1)\right) \\ &\quad + \left(\sum_{z=R_2}^{\infty}(z - R_2)p(z|l_2 + (k-1)T_1 + t)\right)\left(1 - \sum_{y=0}^{R_2-R_1} p(y|l_2 + (k-1)T_1)\right) \end{aligned}$$

Notice that if v_k is zero for all k, then the cost term I is the same as in the exact single-stage model. The long-run average holding cost of stage 1 is

$$\frac{h_1}{T_2}\sum_{k=1}^{n} I_k. \tag{A.6}$$

A.2.2 Backorder Cost

We first derive the expected number of backorders in a cycle or echelon-2 cycle. Denote

$$\boldsymbol{B} \qquad \text{number of backorders in a cycle.}$$

Then, the backorder cost per time unit is $\frac{bE[B]}{T_2}$. To calculate the number of backorders in a stage-1 cycle, we suppose that a review takes place at time t. The next review then take place at time $t + T_1$. The number of backorder during $(t + l_1, t + l_1 + T)$ is the difference between the number of backorder during $(t, t + l_1)$ and during $(t, t + l_1 + T_1)$.

By assumption, we have n stage-1 inventory *reviews* in a cycle. However, we may have less than n stage-1 *replenishments* in a cycle. For instance, if at the kth ($k < n$) review, stage-2 has insufficient stock to raise the stage-1 inventory position to R_1, then stage-2 has insufficient stock in the remaining stage-1 reviews. In this example, we have at most k stage-1 cycles in a cycle. We derive B by conditioning on the number of stage-1 replenishments in a cycle. We let

$$E[B] \;=\; \sum_{k=1}^{n} E[B_k],$$

where B_k is the number of backorders in a stage-1 cycle where there are k stage-1 replenishment cycles. In other words, stage-2 has sufficient stocks at the $(k-1)$st stage-1 review and insufficient stock at the kth stage-1 review to raise the inventory position to R_1. Essentially, the first $(k-1)$ replenishments are normal replenishments and the last is exhaustive.

We denote

$$D(t_1, t_2) \qquad \text{random demand during } [t_1, t_2].$$

Also, denote $q_k (1 \le k \le n)$ as the probability of a cycle having k stage-1 replenishments (the first $k-1$ replenishment are normal and the kth replenishment is exhaustive):

$$\boldsymbol{q_k} = \begin{cases} Pr\{D(0, l_2) \ge R_2 - R_1\}, & k = 1 \\ Pr\{D(0, l_2 + (k-1)T_1) \ge R_2 - R_1 > D(0, l_2 + (k-2)T_1)\}, & 1 < k \le n \end{cases}$$

We also denote $q_{n'}$ as the probability of a cycle having n stage-1 normal replenishments:

$$\boldsymbol{q_{n'}} = Pr\{R_2 - R_1 > D(0, l_2 + (n-1)T_1)\}.$$

We can also represent B as:

$$E[B] = \sum_{k=1}^{n}\sum_{j=1}^{k} E[B_{jk}] + \sum_{j=1}^{n} E[B_{jn'}]$$

where B_{jk} is the number of backorders in the jth stage-1 replenishment where there are $k-1$ normal replenishments in a cycle; $B_{jn'}$ is the number of backorders in the jth stage-1 replenishment where there are n normal replenishments in a cycle.

To compute the value of B_k, we start with a few examples for $n > 2$. First, we examine B_1. Suppose there is only one replenishment in the cycle. That is, at the first stage-1 review, stage-1 exhausts the stage-2 inventory. We assume that we order up to R_2 at $t = 0$. Let random variable $\boldsymbol{X} = D(0, l_2)$, $\boldsymbol{Y} = D(l_2, l_2 + l_1)$, and $\boldsymbol{Z} = D(l_2, l_2 + l_1 + T_2)$. By definition, X and Y are independent, and X and Z are independent. We derive the expectation of B_1 as:

$$\begin{aligned}
\boldsymbol{E[B_1]} &= q_1 E[B_1|q_1] \\
&= q_1 \left(E[D(0, l_2 + l_1 + T_2) - R_2|q_1]^+ - E[D(0, l_2 + l_1) - R_2|q_1]^+\right) \\
&= \sum_{\substack{x \geq R_2 - R_1 \\ x+z \geq R_2}} (x + z - R_2) p(x, z) - \sum_{\substack{x \geq R_2 - R_1 \\ x+y \geq R_2}} (x + y - R_2) p(x, y) \\
&= \sum_{x=R_2-R_1}^{\infty} \sum_{z=R_2-x}^{\infty} (x + z - R_2) p(x|l_2) p(z|l_1 + T_2) \\
&\quad - \sum_{x=R_2-R_1}^{\infty} \sum_{y=R_2-x}^{\infty} (x + y - R_2) p(x|l_2) p(y|l_1) \qquad \text{(A.7)}
\end{aligned}$$

Next, we derive B_2 where there are two replenishments in a cycle. We derive the backorders in each stage-1 replenishments separately. Let $\boldsymbol{X} = D(0, l_2)$, $\boldsymbol{Y} = D(l_2, l_2 + T_1)$, $\boldsymbol{Z} = D(l_2 + T_1, l_2 + l_1 + T_1)$, $\boldsymbol{W} + \boldsymbol{V} = \boldsymbol{Y}$ where $W = D(l_2, l_2 + l_1)$ and $V = D(l_2 + l_1, l_2 + T_1)$.

Assume that $T_1 \geq l_1$.

$$
\begin{aligned}
E[B_{12}] &= q_2 \left(E[D(l_2, l_2 + l_1 + T_1) - R_1 | q_2]^+ - E[D(l_2, l_2 + l_1) - R_1 | q_2]^+ \right) \\
&= \sum_{\substack{x < R_2 - R_1 \\ x+y \geq R_2 - R_1 \\ y+z \geq R_1}} (y + z - R_1) p(x, y, z) - \sum_{\substack{x < R_2 - R_1 \\ x+w+v \geq R_2 - R_1 \\ w \geq R_1}} (w - R_1) p(x, w, v) \\
&= \sum_{x=0}^{R_2 - R_1 - 1} \sum_{y = R_2 - R_1 - x}^{\infty} \sum_{z = R_1 - y}^{\infty} (y + z - R_1) p(x|l_2) p(y|T_1) p(z|l_1) \\
&\quad - \sum_{x=0}^{R_2 - R_1 - 1} \sum_{w = R_1}^{\infty} \sum_{v = R_2 - R_1 - x - w}^{\infty} (w - R_1) p(x|l_2) p(w|l_1) p(v|T_1 - l_1)
\end{aligned}
$$

To derive B_{22}, we denote $\boldsymbol{U} = D(l_2 + T_1, l_2 + l_1 + T_2)$.

$$
\begin{aligned}
E[B_{22}] &= q_2 \left(E[D(0, l_2 + l_1 + T_2) - R_2 | q_2]^+ - E[D(0, l_2 + l_1 + T_1) - R_2 | q_2]^+ \right) \\
&= \sum_{\substack{x < R_2 - R_1 \\ x+y \geq R_2 - R_1 \\ x+y+u \geq R_2}} (x + y + u - R_2) p(x, y, u) - \sum_{\substack{x < R_2 - R_1 \\ x+y \geq R_2 - R_1 \\ x+y+z \geq R_2}} (x + y + z - R_2) p(x, y, z) \\
&= \sum_{x=0}^{R_2 - R_1 - 1} \sum_{y = R_2 - R_1 - x}^{\infty} \sum_{u = R_2 - x - y}^{\infty} (x + y + u - R_2) p(x|l_2) p(y|T_1) p(u|T_2 - T_1 + l_1) \\
&\quad - \sum_{x=0}^{R_2 - R_1 - 1} \sum_{y = R_2 - R_1 - x}^{\infty} \sum_{z = R_2 - x - y}^{\infty} (x + y + z - R_2) p(x|l_2) p(y|T_1) p(z|l_1)
\end{aligned}
$$

To generalize the computation of expected backorders during the jth replenishment where there are $(k-1)$ replenishments in a cycle $E[B_{jk}]$, $1 < k \leq n$, we denote

$$
f(t_1, t_2, t_3, t_4, r) = q_k E \left[D(t_1, t_2) - r | q_k \right]^+ ,
$$

where $q_k = Pr\{D(0, t_4) \geq R_2 - R_1 > D(0, t_3)\}$. Then, by definition,

$$
\boldsymbol{E[B_{jk}]} = \begin{cases} f\left(\tau_1, \tau_1 + l_1 + T_1, \tau_3, \tau_3 + T_1, R_1\right) - f\left(\tau_1, \tau_1 + l_1, \tau_3, \tau_3 + T_1, R_1\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } j < k,\ 1 < k \leq n \\ f\left(0, \tau_2, \tau_3, \tau_3 + T_1, R_2\right) - f\left(0, \tau_1 + l_1, \tau_3, \tau_3 + T_1, R_2\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } j = k,\ 1 < k \leq n \end{cases}
$$

where $\tau_1 = l_2 + (j-1)T_1$, $\tau_3 = l_2 + (k-2)T_1$, $\tau_2 = l_2 + l_1 + T_2$. Similarly, to generalize the computation of $E[B_{jn'}]$, we define

$$
g(t_1, t_2, t_3) = Pr\{R_2 - R_1 > D(0, t_3)\} E \left[D(t_1, t_2) - R_1 | R_2 - R_1 > D(0, t_3) \right]^+ .
$$

Then, by definition,

$$\boldsymbol{E[B_{jn'}]} = g(\tau_1, \tau_1 + l_1 + T_1, \tau_3) - g(\tau_1, \tau_1 + l_1, \tau_3),$$

where $\tau_1 = l_2 + (j-1)T_1$ and $\tau_3 = l_2 + (k-1)T_1$. Depending on the relative value of t_1, t_2, t_3, t_4, we can compute the functions f, g as in the examples of $E[B_1]$, $E[B_{12}]$, $E[B_{22}]$.

Since each function of f, g involves multiple discrete conditional distributions, the computations of f, g functions are complicated. In our computation of benchmarking the approximate model, we assume that in the exact model

$$E[B_{jk}] \quad = \quad q_k E[B_{jk}|q_k] \cong q_k E[B_{jk}].$$

We then can replace the function $f(t_1, t_2, t_3, t_4, r)$ by

$$\hat{f}(s_1, s_2, s_3, s_4, r) \quad = \quad q_k E[D(s_1, s_2) - r]^+.$$

As an example, here is the expected value of backorders where there is one replenishment in a cycle in the computation:

$$\begin{aligned} E[B_1] \quad = \quad & \left(\sum_{x=R_2-R_1}^{\infty} p(x|l_2) \right) \sum_{x+z=R_2}^{\infty} (x+z-R_2) p(x+z|l_2+l_1+T_2) \\ & - \left(\sum_{x=R_2-R_1}^{\infty} p(x|l_2) \right) \sum_{x+y=R_2}^{\infty} (x+y-R_2) p(x+y|l_2+l_1) \end{aligned}$$

Comparing with Equation (A.7), we see that we overestimate the backorders slightly in the computation . In summary, we use the following expected backorders per time unit in the computation:

$$\begin{aligned} \frac{b}{T_2} \quad & \Bigg(q_1 \hat{f}(l_2 + l_1, l_2 + l_1 + T_2, R_2) \\ & + \sum_{k=2}^{n} q_k \Big((k-1) \hat{f}(l_1, l_1 + T_1, R_1) + \hat{f}(l_2 + l_1 + (k-1)T_1, l_2 + l_1 + T_2, R_2) \Big) \\ & + q_{n'} n \hat{f}(l_1, l_1 + T_1, R_1) \Bigg) \end{aligned}$$

A.2.3 Echelon-1 Setup Cost

It costs a_1 for every stage-1 replenishment. Following the ideas in the backorder cost derivation, we derive he long-run average echelon-1 setup cost as

$$\frac{a_1}{T_2}\left(\sum_{k=1}^{n} kq_k + nq_{n'}\right) \tag{A.8}$$

Bibliography

[AEOP02] R.K. Ahuja, O. Ergun, J.B. Orlin, and A.P. Punnen. A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics*, 123:75–102, 2002.

[AMO93] R.K. Ahuja, T. Magnanti, and J.B. Orlin. *Network Flows: Theory, algorithms, and applications.* Prentice Hall, Inc., 1993.

[AOP+04] R.K. Ahuja, J.B. Orlin, S. Pallottino, M.P. Scaparra, and M.G. Scutella. A multi-exchange heuristic for the single-source capacitated facility location problem. 50(6):749–760, 2004.

[AOS01] R.K. Ahuja, J.B. Orlin, and D. Sharma. New neighborhood search structures for the capacitated minimum spanning tree problem. *Mathematical Programming*, 91:71–97, 2001.

[AS04] R. Allgor and D. Stratila. Personal communications. 2004.

[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.

[Axs90a] S. Axsater. Modeling emergency lateral transshipments in inventory systems. *Management Science*, 36(11):1329–1338, 1990.

[Axs90b] S. Axsater. Simple solution procedures for a class of two-echelon inventory problems. *Operations Research*, 38:64–69, 1990.

[Axs93a] S. Axsater. Continuous review policies for multi-level inventory systems with stochastic demand. In S. Graves, A. Rinnooy Kan, and P. Zipkin, editors, *Handbook in Operations Research and Management Science*, volume 4. North Holland, Amsterdam, 1993.

[Axs93b] S. Axsater. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Operations Research*, 41(4):777–785, 1993.

[Axs97] S. Axsater. Simple evaluation of echelon stock (R, Q)-policies for two-level inventory systems. *IIE Transactions*, 29:661–669, 1997.

[Axs00] S. Axsater. Exact analysis of continuous review (R, Q)-policies in two-echelon inventory systems with compound poisson demand. *Operations Research*, 48(5):686–696, 2000.

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

[Ber99] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, 1999.

[BG73] J. Billheimer and P. Gray. Network design with fixed and variable cost elements. *Transportation Science*, pages 49–74, 1973.

[BMM97] A. Balakrishnan, T. Magnanti, and P. Mirchandani. Network design. In F. Maffioli M. Dell'Amico and S. Martello, editors, *Annotated Bibliographies in Combinatorial Optimization.* John Wiley & Sons, 1997.

[BMW89] A. Balakrishnan, T.L. Magnanti, and R.T. Wong. A dual-ascent procedure for large-scale uncapacitated network design. *Operations Research*, 37(5):716–740, 1989.

[BS00] E. Brynjolfsson and M.D. Smith. Frictionless commerce? a comparison of internet and conventional retailers. *Management Science*, 46(4):563–585, 2000.

[CH00] K.L. Cheung and W. Hausman. An exact performance evaluation for the supplier in a two-echelon inventory system. *Operations Research*, 48(4):646–653, 2000.

[Che99] F. Chen. 94%-effective policies for a two-stage serial inventory system with stochastic demand. *Management Science*, 45(12):1679–1696, 1999.

[Che00] F. Chen. Optimal policies for multi-echelon inventory problems with batch ordering. *Operations Research*, 48(3):376–389, 2000.

[Che01] F. Chen. Market segmentation, advanced demand information, and supply chain performance. *Manufacturing & Service Operations Management*, 3(1):53–67, 2001.

[CS60] A. Clark and H. Scarf. Optimal policies for a multi-echelon inventory problem. *Management Science*, (6):475–490, 1960.

[CS62] W. Clark and H. Scarf. Approximate solutions to a simple multi-echelon inventory problem. In S. Karlin K. J. Arrow and H. Scarf, editors, *Studies in Applied Probability and Management Science*, pages 88–110. Stanford University Press, Stanford, CA, 1962.

[CS01] F. Chen and J.S. Song. Optimal policies for multiechelon inventory problems with Markov-modulated demand. *Operations Research*, 49(2):226–234, 2001.

[CZ94a] F. Chen and Y.S. Zheng. Evaluating echelon stock (R, nQ) policies in serial production/inventory systems with stochastic demand. *Management Science*, 40:1262–1275, 1994.

[CZ94b] F. Chen and Y.S. Zheng. Lower bounds for multi-echelon stochastic inventory systems. *Management Science*, 40:1426–1443, 1994.

[CZ97] F. Chen and Y.S. Zheng. One-warehouse multi-retailer systems with centralized stock information. *Operations Research*, 45(2):275–287, 1997.

[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.

[DG85] M. DeBodt and S.C. Graves. Continuous-review policies for a multi-echelon inventory problem with stochastic demand. *Management Science*, 31(10):1286–1299, 1985.

[DL03] L.X. Dong and H. Lee. Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Operations Research*, 51(6):969–980, 2003.

[DS81] B. Deuermeyer and L. Schwarz. A model for the analysis of system service level in warehouse/retailer distribution systems: The identical retailer case. In L. Schwarz, editor, *Studies in the Management Sciences: The Multi-Level Production Inventory Control Systems*, volume 16, pages 163–193. Amsterdam, 1981.

[EL97] Aarts E. and J.K. Lestra. *Local Search in Combinatorial Optimization.* Wiley, New York, 1997.

[ER59] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[Erl78] D. Erlenkotter. A dual-based procedure for uncapacitated facility location. *Operations Research*, 26:992–1009, 1978.

[ES81] G. Eppen and L. Schrage. Centralized ordering policies in a multiwarehouse system with leadtimes and random demand. In L. Schwarz, editor, *Multi-Level Production Inventory Control Systems: Theory and Practice*, pages 51–69. North Holland, 1981.

[Fed93] A. Federgruen. Centralized planning models for multi-echelon inventory systems under uncertainty. In S. Graves, A. Rinnooy Kan, and P. Zipkin, editors, *Handbook in Operations Research and Management Science*, volume 4. North Holland, Amsterdam, 1993.

[FNS04] A. Frangioni, E. Necciari, and M.G. Scutella. A multi-exchange neighborhood for minimum makespan parallel machine scheduling problems. *Journal of Combinatorial Optimization*, 8:195–220, 2004.

[FZ84a] A. Federgruen and P. Zipkin. Allocation policies and cost approximation for multi-location inventory systems. *Naval Res. Logist. Quart*, 31:97–131, 1984.

[FZ84b] A. Federgruen and P. Zipkin. Approximation of dynamic, multi-location production and inventory problems. *Management Science*, 30:69–84, 1984.

[FZ84c] A. Federgruen and P. Zipkin. Computational issues in an infinite-horizon, multiechelon inventory model. *Operations Research*, 32(4):818–836, 1984.

[GJ79] M.R. Garey and D.S. Johnson. *Computers and Intractability, a guide to the theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[GO01] G. Gallego and O. Ozer. Integrating replenishment decisions with advance demand information. *Management Science*, 47(10):1344–1360, 2001.

[GO03] G. Gallego and O. Ozer. Optimal replenishment policies for multiechelon inventory problem with advance demand information. *Manufacturing & Service Operations Management*, 5(2):157–175, 2003.

[Gra85] S.C. Graves. A multi-echelon inventory model for a reparable item with one-for-one replenishment. *Management Science*, 31:1247–1256, 1985.

[Gra96] S.C. Graves. A multiechelon inventory model with fixed replenishment intervals. *Management Science*, 42(1):1–18, 1996.

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

[GZ99] G. Gallego and P. Zipkin. Stock positioning and performance estimation in serial production-transportation systems. *Manufacturing & Service Operations Management*, 1:77–88, 1999.

[HDK04] X. Hu, I. Duenyas, and R. Kapuscinski. Optimal joint inventory and transshipment control under uncertain capacity. 2004.

[HH98] K. Holmberg and J. Hellstrand. Solving the uncapacitated network design problem by a Lagrangian heuristic and branch-and-bound. *Operations Research*, 46(2):247–259, 1998.

[Hol81] I. Holyer. The NP-Completeness of edge-colouring. 10(4):718–720, 1981.

[HW63] G. Hadley and T.M. Whitin. *Analysis of Inventory System.* Prentice-Hall, Englewood Cliffs, N.J., 1963.

[HZ95] R. Hariharan and P. Zipkin. Customer-order information, leadtimes, and inventories. *Management Science*, 41(10):1599–1607, 1995.

[Jac88] P. Jackson. Stock allocation in a two-echelon distribution system of 'What to do until your ship comes in'. *Management Science*, 34:880–895, 1988.

[JS87] H. Johnsson and E.A. Silver. Analysis of a two-echelon inventory system with complete redistribution. *Management Science*, 34(7):880–895, 1987.

[JW02] M.E. Johnson and S. Whang. E-business and supply chain management: An overview and framework. *Production and Operations Management*, 11(4):413–423, 2002.

[Kee99] R.L. Keeney. The value fo internet commerce to the customer. *Management Science*, 45(4):533–542, 1999.

[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.

[Lov72] S. Love. A facilities in series inventory model with nested schedules. *Management Science*, 18(5):327–338, 1972.

[Min89] M. Minoux. Network synthesis and optinum network design problems: Model, solution methods and applications. *Networks*, 19:313–360, 1989.

[MSW93] E.J. McGavin, L.B. Schwarz, and J. E. Ward. Two-interval inventory-allocation policies in a one-warehouse n-identical-retailer distribution system. *Management Science*, 39(9):1092–1107, 1993.

[MT03] A. Muharremoglu and J. Tsitsiklis. A single-unit decomposition approach to multi-echelon inventory systems. Forthcoming, 2003.

[Muc73] J.A. Muckstadt. A model for a multi-item , multi-echelon, multi-indenture inventory system. *Management Science*, 20:472–481, 1973.

[MW84] T.L. Magnanti and R.T. Wong. Network design and transportation planning: Models and algorithms. *Transportation Science*, 18:1–55, 1984.

[Nad82] E. Naddor. *Inventory Systems.* Robert E. Krieger Publishing Co., Malabar, Florida, 1982.

[New03] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[NR04a] S. Netessine and N. Rudi. Supply chain choice on the internet. Working paper, University of Pennsylvania, 2004.

[NR04b] S. Netessine and N. Rudi. Supply chain structures on the internet and the role of marketing-operations interaction. In D. Simchi-Levi, S. D. Wu, and M. Shen, editors, *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era.* Kluwer, 2004.

[Rag94] S. Raghavan. *Formulations and algorithms for network design problems with connectivity requirements.* PhD thesis, Massachusetts Institute of Technology, 1994.

[Rao03] U.S. Rao. Properties of the periodic review (R, T) inventory control policy for stationary, stochastic demand. *Manufacturing & Service Operations Management*, 5(1):37–53, 2003.

[RKP01] N. Rudi, S. Kapur, and D. Pyke. A two-location inventory model with transshipment and local decision making. *Management Science*, 47(12):1668–1680, 2001.

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

[Ros89] K. Rosling. Optimal inventory policies for assembly systems under random demand. *Operations Research*, 37:565–579, 1989.

[Rou86] R. Roundy. A 98% effective lot sizing rule for multi-product, multi-stage production inventory systems. *Math. OR.*, 11:699–727, 1986.

[Sch73] L.B. Schwarz. A simple continuous review deterministic one-warehouse n-retailer inventory problem. *Management Science*, 19(5):555–566, 1973.

[SDB84] L.B. Schwarz, B.L. Deuermeyer, and R.B. Badinelli. Fill-rate optimization in a one-warehouse, n-identical retailer distribution system. *Management Science*, 31(4):488–498, 1984.

[Sha81] K. Shanker. Exact analysis of a two-echelon inventory system for recoverable items under batch inspection policy. *Naval Res. Logist. Quart.*, 4:579–601, 1981.

[She68] C. Sherbrooke. METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16:122–141, 1968.

[She86] C.C. Sherbrooke. VARI-METRIC: Improved approximation for multi-indenture, multi-echelon availability models. *Operations Research*, 34:311–319, 1986.

[Sim71] R.M. Simon. Stationary properties of a two-echelon inventory model for low demand items. *Operations Research*, 19:761–777, 1971.

[SN85] C. Schmidt and S. Nahmias. Optimal policy for a two-stage assembly system under random demand. *Operations Research*, 33(5):1130–1145, 1985.

[SR51] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin and Mathematical Biophysics*, 13:107–117, 1951.

[SS03] K.H. Shang and J.S. Song. Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Science*, 49(5):618–638, 2003.

[ST03] J. Swaminathan and S. Tayur. Models for supply chains in e-business. *Management Science*, 49(10):1387–1406, 2003.

[SZ88] A. Svoronos and P. Zipkin. Estimating the performance of multi-level inventory systems. *Operations Research*, 36:57–72, 1988.

[SZ91] A. Svoronos and P. Zipkin. Evaluation of one-for-one replenishment policies for multiechelon inventory systems. *Management Science*, 37(1):68–83, 1991.

[SZ03] J.S. Song and P. Zipkin. Supply chain operations: Assembly-to-order systems. In S. Graves and T. de Kok, editors, *Supply Chain Management*, volume 30 of *Handbooks in Operations Research and Management Sciences*, chapter 11. North Holland, 2003.

[Tag89] G. Tagaras. Effects of pooling on the optimization and service levels of two-location inventory systems. *IIE Transactions*, 21(3):250–257, 1989.

[Tal96] K.T. Talluri. Swapping applications in a daily airline fleet assignment. *Transportation Science*, 30:237–248, 1996.

[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.

[TD02] G. Torkzadeh and G. Dhillion. Measuring factors that influence the success of internet commerce. *Information Systems Research*, 13(2), 2002.

[TO89] P.M. Thompson and J.B. Orlin. The theory of cylic transfers. Working paper, Operations Research Center, MIT, Cambridge MA, 1989.

[TP93] P.M. Thompson and H.N. Psaraftis. Cyclic transfer algorithms for multivehicle routing and scheduling problems. *Operations Research*, 41(5):935–946, 1993.

[UPS05] www.ups.com, 2005.

[WCW73] M. Wagner W. Crowston and J.F. Williams. Economic lot size determination in multi-stage assembly systems. *Management Science*, 19(5):517–527, 1973.

[WCZ00] Y.Z. Wang, M. Cohen, and Y.S. Zheng. A two-echelon repairable inventory system with stocking-center-dependent depot replensishment lead times. *Management Science*, 46(11):1441–1453, 2000.

[Won84] R. Wong. A dual ascent approach for steiner tree problems on a directed graph. *Mathematical Programming*, 84:271–287, 1984.

[Won85] R.T. Wong. Probabilistic analysis of an optimal network problem heuristic. *Networks*, 15:347–363, 1985.

[Zhe92] Y.S. Zheng. On properties of stochastic inventory systems. *Management Science*, 38(1):87–103, 1992.

[Zip00] P. Zipkin. *Foundations of Inventory Management.* McGraw-Hill, New York, 2000.