# Simulation as a Method for Determining Inventory Classifications for Allocation

by

Braden Ball

B.S. Manufacturing Engineering Technology, Brigham Young University, 2008

Submitted to the MIT Sloan School of Management and the Engineering Systems Division in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Engineering Systems

In conjunction with the Leaders for Global Operations Program at the Massachusetts Institute of Technology

June 2012





Signature of Author _____
Engineering Systems Division, MIT Sloan School of Management
May 11, 2012

Certified by _____
Stephen Graves, Thesis Supervisor
Abraham J. Siegel Professor of Management Science,
Professor of Mechanical Engineering and Engineering Systems

Certified by _____
Chris Caplice, Thesis Supervisor
Executive Director, Center for Transportation and Logistics

Accepted by _____
Oli de Weck, Chair, Engineering Systems Education Committee
Associate Professor of Aeronautics and Astronautics and Engineering Systems

Accepted by _____
Maura Herson, Director, MBA Program
MIT Sloan School of Management

*This page intentionally left blank.*

# Simulation as a Method for Determining Inventory Classifications for Allocation

by

Braden Ball

Submitted to the MIT Sloan School of Management and the Engineering Systems Division on May 11, 2012 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Engineering Systems

## Abstract

Companies that utilize multiple facilities to satisfy customer demand are faced with the same basic question – where should inventory be held? This thesis presents a method for answering this question, specifically for a company that allocates multiple units across multiple facilities, where any facility can fulfill an order to any customer, though with differing shipping costs. The model presented is a simulation of the shipping costs of various allocation strategies across a range of allocated inventory quantities, where the strategies simulated include consolidating all inventory in a central facility, constraining inventory to regional hubs, and spreading inventory throughout the network. The simulated results are then compared to find the low cost allocation strategy at a given level of allocated inventory. With this comparison, product groupings with the same low cost allocation strategy are identified, and are defined as "Slow", "Medium-A", "Medium-B", and "Fast" products. These groups can then be used to manage the allocation process, where "Slow" inventory is held centrally, "Medium-A" inventory held regionally, and "Fast" inventory spread throughout the network. "Medium-B" items serve as a cost-mitigating flexible option, where they are spread throughout the network when possible but consolidated when necessary to avoid changing the allocation for "Fast" items. At a broad level, the model presented is applicable to any company that can fulfill demand to a single customer from multiple facilities.

Thesis Supervisor: Stephen Graves
Title: Abraham J. Siegel Professor of Management Science, Professor of Mechanical Engineering and Engineering Systems

Thesis Supervisor: Chris Caplice
Title: Executive Director, Center for Transportation and Logistics

*This page intentionally left blank.*

## Acknowledgments

I would like to thank Amazon.com for their sponsorship of this project, the Leaders for Global Operations (LGO) program, and for the opportunity they provided me to work and learn with them. It was truly an enlightening experience to work with a world-class organization. Special thanks to Larry Wilson, Russell Allgor, and Bill Campbell, whose collective experience and insight helped to make sure I was asking the right questions and working toward the right goals; and Steven Beach, who showed me how data can be used as a competitive edge and without whom I would likely still be drowning in a sea of SQL SELECT statements in the Amazon databases.

Special thanks also go to my MIT advisors, Professors Stephen Graves and Chris Caplice, who challenged my assumptions, guided my questions, and helped me to understand how to apply theoretical concepts learned in the classroom to real-world systems. Their mentorship and instruction is greatly appreciated.

Finally – and most importantly – I'm forever grateful to my wonderful wife Tiffany, and my kids Kayla (3 years) and Logan (8 months). The kid's smiles have made the challenges of LGO seem light, and I can't thank Tiffany enough for her unwavering support and encouragement through these past two years. She deserves much more recognition than this page could provide.

*This page intentionally left blank.*

# Table of Contents

## List of Figures

# 1 Introduction

The purpose of this thesis is to explain a method to determine inventory classifications for allocation in a fulfillment network. These classifications can then be used to better allocate inventory and reduce expected shipping costs in light of uncertain demand. The research presented was done as a part of an internship completed with Amazon.com, in conjunction with the Leaders for Global Operations program at the Massachusetts Institute of Technology. The following overview of the company, project problem statement and project goals are provided to supply the reader with the context necessary to understand the problem.

## 1.1 Amazon.com

While Amazon.com began in 1995 selling books that were packed and shipped from Jeff Bezos' garage, it has grown dramatically since those humble beginnings to a $34 billion company in 2010[1] that boasts "Earth's Biggest Selection" and seeks to be "Earth's most customer-centric company for three main customer sets: consumers, sellers, and enterprises"[2]. Amazon attributes this growth to its focus on the customer, and the influence of a positive customer experience on what it has described as its "Virtuous Cycle" or "Flywheel of Growth". Corporate lore holds that while on a cross-country drive, Bezos drew the following diagram[3] on a napkin, and it provided the impetus to start up the company. While the specifics of the story are unimportant, the diagram does provide a good explanation of how the company has grown and been successful by putting the customer experience first, and is used internally to help Amazonians understand the business model and the crucial role of the customer experience on the success of the firm[4].

---

[1] Amazon 10k, 2011 – http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-SECText&TEXT=aHR0cDovL2lyLmludC53ZXN0bGF3YnVzaW5lc3MuY29tL2RvY3VtZW50L3YxLzAwMDExOTMxMjUtMTEtMDE2MjUzL3htbA%3d%3d

[2] Amazon 10k, 2011

[3] http://seekingalpha.com/article/121955-amazon-s-wheel-of-growth

[4] Personal experience of the author

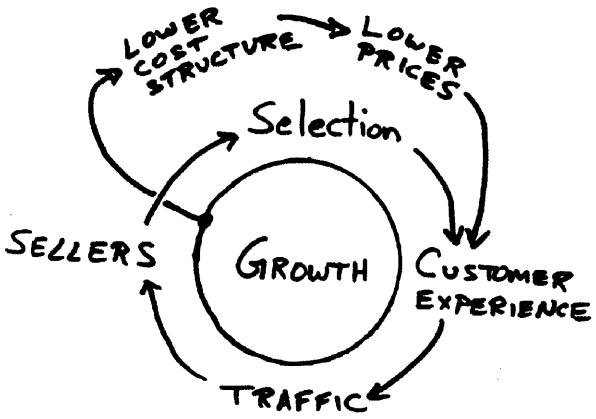


**Figure 1 - Amazon's Virtuous Cycle**

Growth is central to the cycle because it enhances the Customer Experience through both increased product selection and lower prices. Increased selection comes from adding more sellers and products to the site, which are attracted by the increased traffic at the Amazon.com marketplace that results from a good customer experience. In other words, a good experience at Amazon means a customer is more likely to return (and tell their friends), causing more traffic and attracting more sellers to the site to sell their products. This then improves the selection at Amazon, creating a better experience for customers who are able to find more of the products that they want to buy online. As this cycle continues to "spin", it gains momentum and propels the company to increasing growth, which is where the comparison to a flywheel comes from. However, the virtues of this cycle don't stop there, because with increased growth come economies of scale. This gives the company a lower cost structure that it passes on in the form of lower prices, further delighting customers and promoting the cycle. Following this strategy has resulted in millions of satisfied customers and dramatic sales growth over the life of the company, as shown by the following chart of sales growth throughout Amazon's history[5].

[5] Created using data from Amazon.com Form 10-K, years 2010, 2008, 2004, and 1999

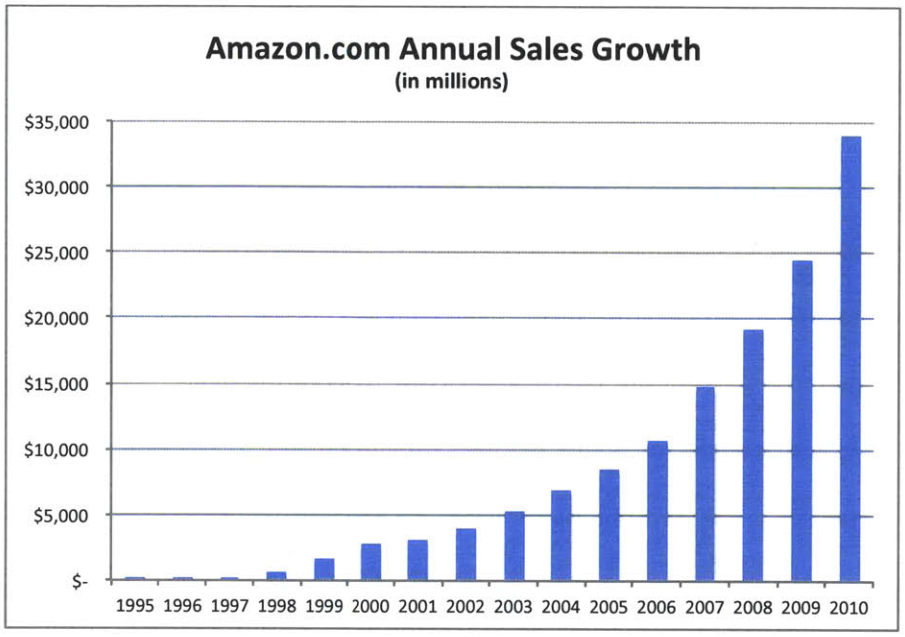


**Figure 2 - Sales Growth at Amazon.com**

While this cycle is relatively simple to explain and Amazon's results from using it are impressive, it is very difficult to execute. Increases in the product offering add complexity to the supply chain in both the variety and volume of disparate units that have to be managed. A quick search on Amazon.com will reveal that the current physical product offering includes everything from books to treadmills to ant farms[6]. This vast variety of inventory makes it very difficult to ensure that each item is ordered and stored efficiently. The difficulty is then compounded by the fact that every item might need to be shipped at a moment's notice to a customer anywhere in the United States, creating a significant operational challenge. Late or incorrect shipments degrade the customer experience and inhibit the flywheel, and as a result operational effectiveness and execution are core to a good customer experience and the success of the virtuous cycle.

In response to this challenge, Amazon has developed a Fulfillment Center (FC) Network that receives, stores, and ships products to customers. While at first glance one might confuse an FC with a

[6] Amazon.com search performed on 5/05/2012, keywords "books", "treadmill", & "ant farm". Example ASINs (Amazon Standard Identification Number) include 039480001X, B003TNJRLI, and B00000IRUG.

traditional warehouse, there are various differences that enable the FC network to fulfill its purpose of effectively satisfying customer demand.

Most notable is the way the high-level supply chain is arranged, which is quite different from how a traditional retailer operates. This comparison is illustrated in Figures 3 and 4. In a traditional retail supply chain, inventory from suppliers is received and held in a central warehouse (while large retailers may have multiple regional warehouses, the basic structure remains the same). These warehouses hold the inventory in bulk, and in accordance with the firm's inventory policy will ship to physical retail locations where customers in the surrounding area come to purchase products. In this model, inventory policy must be set such that demand variation is accounted for at a fairly low level – each retail store. Safety stock will be held at retail locations and/or warehouses to account for variation in customer demand and to provide whatever service level that management deems appropriate.

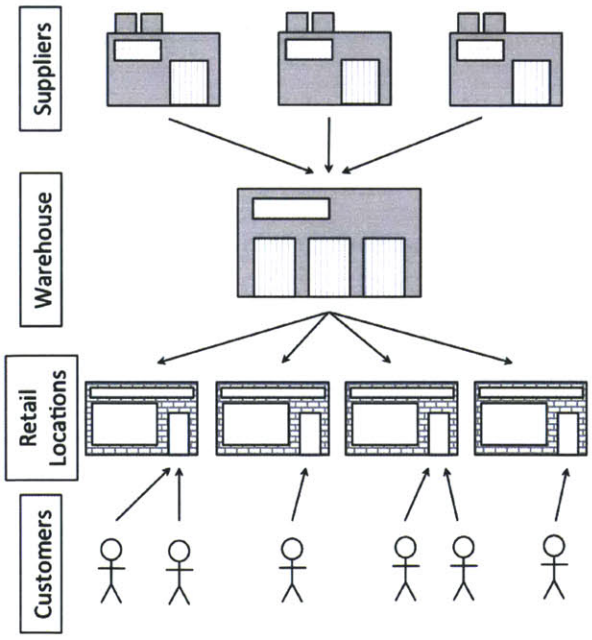


**Figure 3 - Traditional Retail Supply Chain Model**

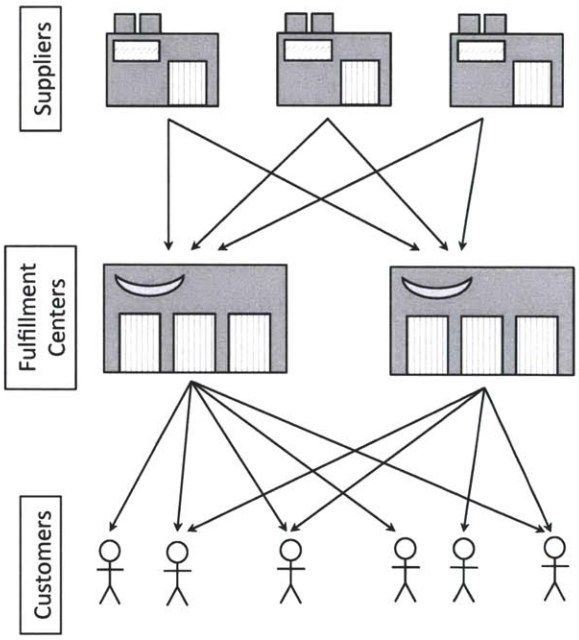


**Figure 4 - Amazon.com Supply Chain Model**

For Amazon the supply chain is markedly different. While many of the same suppliers will be used as in the typical retail example, they will not just ship into a few central warehouses. Instead, they will ship directly into the various FCs spread throughout the country (subject to the rules of the inventory

allocation process). These FCs will then break the bulk inventory into units that can be individually stored and shipped, and store them until demand for the unit is realized. When this happens, the product is boxed and shipped directly to the customer. While inventory will be shipped from the FC that minimizes the fulfillment cost of the order, all products aren't held in all FCs and it is possible to ship to any customer from any FC if necessary, regardless of the customer location. This model provides two important benefits:

1. Customers don't have to leave their homes to shop, making it easier and more convenient, thus improving the customer experience.
2. Demand variation does not need to be planned for at a regional level. Because shipments can go to any customer from any FC, demand variation can be pooled at a network-wide national level, reducing holding costs from inventory. This further improves the customer experience, as these savings are passed on to the customer in the form of lower prices.

While Amazon's supply chain model has enabled it to delight customers and grow in response to ever-increasing demand for an ever-increasing product line, it is not without challenges. Various inefficiencies that aren't present in traditional retail must be accounted for, one of the most notable being the extra cost that is frequently incurred to fulfill a customer order from a distant FC. As such, efficiently allocating inventory across FCs while planning for demand and its variation nationally becomes a significant operational challenge, which is the topic explored in this thesis.

## 1.2 Problem Statement

The capability (provided by the FC network) to pool regional demand variation and plan for inventory at a national level enables the company to maintain a high service level while keeping inventory levels relatively low. This strategy, however, inevitably results in some demand being fulfilled from an FC that does not provide the cheapest transportation option. Amazon will still fulfill these orders by the delivery date promised because of the capabilities of modern carriers, but may incur a higher shipping

cost to do so than if the products had been at an FC closer to the customer. While the difference in cost between the individual shipments is relatively small, with annual outbound shipping costs of $2.58 B[7], an improvement in outbound transportation costs that yields even a fraction of a percentage will result in considerable bottom-line savings.

In an effort to control these costs and lessen the impact from inefficient shipments, planned inventory purchases are classified into groups for allocation, with some groups being held in central locations and others being spread throughout the network. In theory, those groups held centrally represent slow-moving items with only a few total units in the network, making it hard to predict where they will ship to. Those groups spread throughout the network are of fast-moving items with many units held, which tend to ship closer to regional demand patterns. The assumption is that by spreading out fast-moving items, more often than not there will be a unit of inventory in an FC close to a customer. However, the method for determining these classifications has room for improvement, and it isn't possible to tell whether or not the groups are achieving their intended purpose. As a result, it requires significant skill and training to manage the nuances of the system, and there is a sizeable opportunity in reducing shipping costs by improving the allocation process.

## 1.3 Project Goals

The goal of the project is to improve the method by which these allocation groups are determined in order to better align with the objectives of the system, and to reduce inefficient shipments and their associated costs. This will include examining the current allocation tools and processes, and identifying ways in which the specification of product groupings can be changed to improve control of the process. An inventory allocation/shipment simulation will be used to show the potential benefits of different inventory allocation strategies and to gain insight into how product groupings should be defined and managed to improve network performance.

[7] Amazon.com Form 10-K, 2010

## 1.4 Summary

While Amazon is a very successful company that has produced very impressive results, opportunities for improvement always exist in systems as large and complex as its supply chain. As a result, this thesis aims to demonstrate how one specific opportunity – high fulfillment expense from cross-country shipments – can be approached in order to improve the process.

# 2 Amazon Fulfillment and Supply Chain Overview

As Amazon's supply chain differs some from that of traditional retailers, a brief explanation of the current FC network, allocation process, and outbound shipment process is beneficial, and this chapter will explore these different aspects of the supply chain.

## 2.1 Fulfillment Center Network

In order to supply such a deep product offering to customers all over the United States, Amazon has developed a network of specially designed warehouses that it refers to as the Fulfillment Center (FC) Network. This network is comprised of various types of facilities, specialized tightly by the size and weight of products that they will hold, and loosely by the geographic region they are designed to supply. For example, a pool table is much larger and heavier than a Blu-ray disc, and so will require a different material handling strategy. As a result, FCs are specialized to handle items based on their size and weight. A precise algorithm is used to make this distinction, the optimization of which is outside the scope of this research.

Additionally, each FC is located in a strategic area for fulfillment, with three main regions across the United States, the West, Midwest, and East. However, the definitions for geographic regions for fulfillment are fairly loose, because as previously explained every FC has the ability to ship to any customer, regardless of location. This allows the company to hold lower overall levels of inventory, but also results in increased shipping costs when a product is only available at an FC distant from the customer.

## 2.2 Overview of the Current Allocation Process

An interesting feature of the inventory allocation process is that it is almost exclusively managed centrally, while the network is very broad with many different FCs. Decisions about what product lines will be held in what quantities and at which locations are made at a network level, in an effort to ensure that the right product mix is held in each facility for the best network-wide performance. While this

prevents local optimization at the expense of network performance, it also presents a significant problem – knowing what quantities and variety of inventory should be held at each facility. To address this challenge, a process has been established to allocate inventory across the network.

At its core, the process is a simple Periodic Review, Order-Up-To-Level system, where inventory is purchased at regular intervals, up to a specified quantity based on the desired service level/in-stock percentage for the product[8]. Subtracting the current inventory quantity from the order-up-to level provides the order quantity for the period. In a simple form, this quantity can be calculated using the following equation[9], assuming Normal demand with mean $x$ and standard deviation $\sigma$:

$$S = x_{L+R} + k\sigma_{L+R}$$

Where $S$ is the order-up-to-level, $x_{L+R}$ is expected demand over the lead-time plus the review period, $k$ is a safety factor chosen by management to achieve a desired level of service, and $\sigma_{L+R}$ is the standard deviation of demand over the lead-time plus the order period.

Additionally[10]:

$$\sigma_{L+R} = \sqrt{E(L)var(D) + [E(D)]^2 var(L)}$$

Where $E(L)$ is the expected lead-time, $var(D)$ is the variance in demand over the unit time period, $E(D)$ is the expected demand over the unit time period, and $var(L)$ is the variance of the lead-time.

While these equations provide a basic framework to work from, they fail to address key differences between a traditional supply chain and that of Amazon. First, they assume a single expected lead-time (and lead time variation) from a supplier, which is a fair assumption for the traditional supply chain model (see Figure 3 - Traditional Retail Supply Chain Model). However, as shown previously, Amazon's

[8] Silver, et al., pp 240
[9] This equation is the simple form for finding the order-up-to-level in a periodic-review system. However, it does not represent the actual method by which Amazon calculates this value, which is omitted to protect confidential information. However, this form is sufficient for the purpose of this thesis.
[10] Silver, et al., pp 283

supply chain is different. For each ordering period, each supplier will ship directly to multiple FCs, with different corresponding lead times and lead-time variations for each supplier-FC connection. Second, the equations assume that the level $S$ (and the resulting order quantity) calculated will satisfy all of the demand considered; in the case of Amazon this would be national demand. However, the various lead-times for each supplier-FC connection result in different values of $S$ for each FC. This makes a "national" level for $S$ something that can't be calculated without combining the different lead times, and even if it were it would be of little value because of the need to manage inventory at an FC level.

To take these different elements into account, allocation factors were developed to use as a tool to allocate demand and the inventory required to fulfill it across FCs. To address the wide variety of products, the factors are assigned to product groups based on the product line, the product's size and weight classification, and the popularity or "velocity" of the product (the definition of which will be further explored in chapter 4.1), and ultimately determine which products will be held in which FCs and in what quantities. The factors represent a fraction of demand that an FC is planned to fulfill (the total of all factors for a given product group sums to one, and each factor is positive and less than or equal to one) and they serve as the primary lever for changing the inventory allocation in the network.

With the factors in place, the following basic steps[11] are followed at the end of each ordering period to determine the quantity of inventory that will be ordered into each FC:

1. Allocation factors are checked to determine which FCs will receive inventory for a given product group.
2. $S$ is calculated for each FC with a non-zero allocation, accounting for total system demand and the FC-specific vendor lead-time and its variation.

[11] While these steps represent the basic process of inventory allocation at Amazon, there are various specifics and exceptions that have been left out of the explanation, both for simplicity in explanation and to protect the confidential interests of Amazon.com

3. This value of $S$ is multiplied by the FC's allocation factor to determine the quantity of demand to be allocated to each FC, resulting in $S_{FC}$.
4. The FC order quantity is calculated by subtracting the current FC inventory from $S_{FC}$ and then rounding for whole-unit, case, and minimum order quantities.

It is not uncommon for the factors to be adjusted in order to ramp facilities up or down, to better match regional demand for products, or in response to capacity issues at various FCs.

## 2.3 Outbound Shipment Decision

With inventory allocated throughout the network, each individual unit would ideally ship from its respective FC to a customer in the surrounding region. However, as actual demand is realized and inventory is depleted this isn't always possible, and inefficient cross-country shipments will inevitably occur. In an effort to minimize these shipments, the outbound shipment decision of which FC will fulfill an order is determined by a "greedy" algorithm. In other words, the product will be fulfilled by the FC that both has the unit of inventory to ship and can ship the unit with the lowest cost, while satisfying the delivery date promised to the customer. This generally results in the shipment originating from the FC nearest to the customer with the inventory available.

Ideally then, shipments will always be fulfilled in-region or from a nearby region, and only fulfilled from a distant region when absolutely necessary. Unfortunately, due to the uncertain nature of demand in both its absolute quantity over the buying period and its point of origination (the customer location), regional stock-outs are not uncommon. In addition, multi-item orders (when two units from a single order aren't present in the same FC) will often require shipment from a far away FC. This will occur if the distant FC is the only one with all items of inventory in stock, as it is typically cheaper to ship multiple items in the same box from a distant FC than to create multiple shipments.

## 2.4 Summary

While Amazon's supply chain is very interesting, it is not 100% unique. Similar to others, the company uses strategically located facilities to provide efficient fulfillment and a high level of service to customers. As well, a simple-order-up-to model can be used to describe the ordering process, and the greedy algorithm for shipments is a pragmatic approach that can be easily replicated in our model.

# 3 Literature Review

The literature surrounding inventory placement strategies, ordering policies, and the handling of safety stock is extensive, and this review is not intended to be exhaustive. Rather, it is meant to provide a summary of various traditional and non-traditional approaches to inventory management, as well as how these approaches can be used to better understand inventory allocation in a fulfillment network like that of Amazon.

## 3.1 Risk Pooling

A large amount of research has been done regarding the benefits of risk pooling. A great example of this work is the influential paper by Eppen (1979), that illustrates how centralizing inventory rather than spreading it amongst multiple locations can reduce holding and penalty costs. Additionally, it shows that the magnitude of the savings is dependent on the correlation of demands from the disparate sources considered, and if these demands are uncorrelated, then the savings increases with the square root of the number of demands considered. This relationship has since been dubbed the "Square Root Law", and is often used by management to anticipate cost increases or decreases from adding or subtracting facilities from a network. Gerchak and He (2002) and Berman et al. (2011) are more recent examples that explore the benefits of risk pooling, specifically how the savings achieved are impacted by increased variation in expected demand and the demand distribution. Additionally, Sobel (2008) provides approachable, illustrative examples of how risk pooling can be used to manage various operational uncertainties in practice. While risk at Amazon isn't pooled in the traditional sense of centralizing inventory, it is virtually pooled by centralizing planning for demand at a national level (see Chapter 2). Because any location in the network can serve any demand in the United States, lower overall levels of inventory are needed to handle uncertainties in demand, yielding similar benefits to those explored in this literature.

Related to risk pooling through inventory consolidation is how planning for demand at a global level rather than a local level can facilitate more accurate forecasts. Intuitively, it will be easier for one to

forecast demand for a certain item over all of North America than to reliably predict demand for the same item independently in every town in North America. Zotteri et al. (2005) illustrates this principle by exploring the results of various forecasting models at varying levels of aggregation, and illustrates how using the appropriate level and type of aggregation for a given system can improve forecasting results. For Amazon, this relationship – it is easier to forecast demand nationally than locally – motivates demand planning at an aggregate national level, while the FC network enables order fulfillment with a high level of service at a local level.

## 3.2 Multi-Echelon Inventory Models

One way in which companies take advantage of the benefits of risk pooling without consolidating their entire inventory in a single facility is to set up their system with multiple echelons of inventory, such as a warehouse that feeds multiple retail locations. Many models have been developed to show how these systems can best be managed given a set of system characteristics. Examples of this work include Federgruen and Zipkin (1984), which uses a periodic review system where the central warehouse only orders and transships (it does not hold inventory); Svoronos and Zipkin (1988), which uses a continuous review system; Jackson (1988), which uses a periodic review system and allows the central warehouse to hold inventory and develops an approach for determining the appropriate level of safety stock at the central warehouse; and Graves (1996) that assumes a periodic review system at both the central warehouse and retail locations, but where warehouse stock is virtually allocated, or reserved, by the retail locations when they realize actual demand. More recently, Parker and Kapuscinski (2004) examine optimal policies for a two-echelon system where both echelons have finite capacity, and Axsäter and Marklund (2008) explore how to optimally manage a multi-echelon system where modern IT capabilities enable access to real-time information that can be efficiently incorporated into ordering decisions.

While this expansive body of work is very useful in understanding multi-echelon supply chains, multi-echelon inventory management is not currently the dominant strategy used at Amazon. As such these models have little direct application to the current FC network. It should be noted that there are

some efforts to implement a multi-echelon strategy in order to enable pull-based allocation in the FC network, but this practice is not yet widespread at Amazon and is out of the scope of this research. If Amazon ultimately decides to alter its supply-chain strategy along these lines, application of the large body of work in multi-echelon inventory placement will likely become quite valuable.

## 3.3 Inventory Allocation in Similar Systems

While Amazon's supply chain is non-traditional (see Figure 4 - Amazon.com Supply Chain Model), there is still literature that has been published exploring ways to allocate inventory in such a system. Xu (2005) outlines a way of determining the best strategy for stocking inventory in a multi-Unit multi-Location problem, where locations are not established in echelons. It shows how the problem can be formulated as a Markov Chain, and provides excellent examples of allocations in a 2-Unit 2-Location problem (with both the same and different lead times and with Compound Poisson demand), and a 2-Unit 3-Location Problem. The optimal allocation in each case is determined by finding the allocation with the lowest expected cost, given probabilities of demand from each region considered. It provides a way to examine allocation strategies based on different ranges of regional demand probabilities, and shows that the lowest-cost strategy in each scenario will be highly dependent on regional demand probabilities, variation in demand, and the required fill-rate of the system. While the model provides good intuition on a way to think about the problem, and a potentially viable strategy for allocating low-volume, slow-moving items, the Markov Chain quickly explodes when additional units and locations are added. As a result, it is an impractical way to model the allocation of many fast-moving items in Amazon's actual FC Network, which consists of greater than 20 FCs. However, the author suggests that this methodology may still prove beneficial if certain assumptions are made to decompose the problem as an approximation, which can then be used to formulate a more general model that could be applied to a more complex system.

Additionally, Benjaafar et al. (2008) demonstrate how a similar problem with multiple inventory locations and multiple sources of demand can be formulated as a mixed-integer linear program, providing

an exact method for finding an optimal allocation. However, this model does not address complications from a wide product line with both slow and fast-moving items, and relies on the assumption that there are no transshipments between facilities. This assumption is made because such a capability is only possible with a sophisticated IT system and responsive transportation processes that most companies simply don't have[12]. However, Amazon possesses the capability to transship. This flexibility makes it possible for Amazon to fulfill orders cheaper and faster, and it makes the model proposed by Benjaafar et al. (2008) inapplicable to Amazon's FC Network..

## 3.4 Product Classification in Inventory Management

Many companies group products to help them efficiently manage their inventory while ensuring that they commit enough resources to those items that generate the most sales volume and that they don't commit too many resources to those items that have little relative importance to the bottom line. Silver et al. (1998) provide a good overview of using an A-B-C Classification in this type of system. To summarize the strategy, Class A items are generally the 5-10% of SKUs that make up about 50% of sales volume, and so it is important to manage them closely to ensure that customers receive a high level of service and that sales aren't lost. For these items, it is not uncommon for management to make exceptions to standard process in order to fulfill orders, and planning parameters are generally reviewed more often than for other items. Class B items generally make up about 50% of SKUS, and don't require as much planning attention as Class A items. Finally, Class C items are the final 40% of items that don't contribute a large amount to either inventory costs or total dollar sales, and so management of these items should be kept as simple as possible.

While this type of product classification can help inform decisions at Amazon, the high number of SKUs and the demand distribution of those SKUs make it difficult to use. For example, a large retailer may stock 100,000 plus unique items[13], which would result in 5-10,000 Class A items that would need to

[12] Benjaafar et al. (2008)
[13] Roberts and Berg (2012)

be closely managed. However, at Amazon the number of items stocked is easily at least an order of magnitude higher[14], making this method impractical to use effectively.

## 3.5 Summary

Much work has been done related to the problem of efficiently allocating inventory in a network, but Amazon's operating strategy makes it difficult to directly apply this research to its FC network. While some models can be used to accurately describe a portion of the system, applying them broadly to the network is impractical due to either the complexity of the model or assumptions that don't apply.

[14] The precise number of unique items stocked and sold by Amazon.com is not publicly available, but the scope of the product offering can be estimated. To estimate this number, a search by department for "*" was performed on 5/5/2012 on Amazon.com, filtering for items sold by Amazon (excluding all third party offerings). This was done for every department except Books, Movies, Music, and Watches, which can't be limited by seller. Summing the outcomes resulted in over three million items. The number is inexact as some items are listed in multiple departments and some departments are excluded. Due to the exclusions, more items are likely offered, but it does demonstrate the breadth of products that are managed by Amazon.com, making A-B-C classification in the traditional sense impractical.

# 4 Product Groupings

As previously mentioned, in this system products purchased into the network are assigned to a group at the time of allocation. Different allocation strategies are then employed for each group, in an effort to get more products closer to consumers, while accounting for the reality that FCs sometimes run out of capacity making it necessary to change the allocation to satisfy capacity constraints.

## 4.1 Current Groupings

When products are ordered into the network they are given a classification, which determines the set of allocation factors that will be used. These factors are set according to two main strategies: spread the inventory through the network as close as possible to historic regional demand, or consolidate it in one or two central locations. The classifications are fairly straightforward, and depend on an item's product line, its size and weight, and its "velocity" (defined by the item's proximate weekly forecast).

The item's product line and size and weight are simple and understandable – it makes practical sense to group similar products together for allocation, and differentiating by size and weight is essential due to the material handling capabilities of various FCs in the network. Velocity however is a different type of classification. It does not depend on an item's physical characteristics, but rather on its anticipated popularity (the forecast). As a result, this classification is essential in determining which allocation strategy (spread or consolidate) will be used. Popular (Fast velocity) items will be spread throughout the network in an attempt to get them as close to the customer as possible, and less-popular (Slow velocity) items will be consolidated to hedge against cross-country shipments, as it is difficult to predict where they will ship in advance. To make the distinction, each item's weekly forecast is measured against a threshold set by management. If the forecast is above the threshold, the item will receive a Fast classification and be assigned the spread out allocation factors. If the forecast is below the threshold, it will be classified as Slow and assigned the consolidated allocation factors.

For example[15], consider two unique items: a baseball and a hockey puck. Both items will belong to the same product line, which we'll call "athletic equipment". Additionally, because they are both small, easy to handle items, they will both have the same size/weight classification. However, suppose these items are ordered into the network in the month of June, when (presumably) demand for baseballs in North America is much higher than demand for hockey pucks. For the purpose of the example, the weekly forecast for the baseball in question is 30 units, and that of the hockey puck 4 units. Additionally, suppose management's threshold for a "fast" item is that it must have a weekly forecast of at least 15 units. By these classifications, the incoming baseball inventory will be spread throughout the network according to the allocation factors for fast/small/athletic equipment, and that of the hockey puck consolidated according the allocation factors for slow/small/athletic equipment. Six months later, when winter sets in and hockey pucks become more popular and baseballs are no longer selling well, the velocities of each item would switch and hockey pucks would be spread throughout the network while baseballs would be consolidated.

Assigning allocation factors (and so allocation strategies) by these groups enables Amazon to place inventory throughout the network in-line with how efficiently it expects the items will ship – "faster" items, which by definition will have more inventory, are more likely to ship efficiently if spread out than "slower" items in light of uncertain demand. Additionally, it enables management to change the allocation strategies of only specific groups of units when necessary if an FC becomes over-capacitated.

## 4.2 Issues with Current Groupings

While grouping products in this way provides various planning and tactical benefits, it also presents some challenges. Specifically, the actual quantity of inventory purchased into the network is affected by other parameters in addition to the item's weekly forecast, namely the product's buying period (the amount of time between scheduled orders), and its service level. This quantity is important,

[15] The product line, velocity thresholds, and allocation factors were created for the purpose of the example explained. They are non-indicative of actual product lines, thresholds, or factors.

especially when allocating to spread out FCs according to historical regional demand percentages. These demand percentages are essentially the average or expected demand that will come from a given region. By the law of large numbers, the larger the number of units (trials), the higher the likelihood that actual shipments (results of the trial) will be in-line with the expectation. In other words, we need to know the quantity that will be allocated throughout the network before we can know how close to expect demand for those items to fall in-line with expected demand from each region/FC.

In addition to an item's weekly forecast, specific factors that affect the amount of inventory that will be purchased into the network include its buying period and service level. A longer buying period will result in a higher quantity of inventory allocated in the network, as will a higher service level. Without taking these factors into account, it becomes possible for items that might perform well if allocated as "fast" to be classified as "slow", e.g. the hockey puck from the previous example may have a monthly buying period and a very high service level, which would easily put the actual inventory quantity ordered into the network above the threshold of 15 units. Figure 5[16] shows the inventory Order-Up-To Level ($S$) for nine unique items, each with the same weekly forecast of 10 units. However, each item has a different buying period and service level, resulting in a different level of $S$ for each item, ranging from 28 to 89 units. By the current methodology, each of these items would receive the same allocation (because they have the same forecast) even though the actual quantities allocated to the network are very different. Because there are different quantities, the degree to which the regional demand for the item with 28 units will align with expected regional demand will be much different than that of the item with 89 units allocated. As a result, if both items were allocated with a consolidated strategy, the item with 89 units would likely result in more expensive shipments than would be achieved had it been spread throughout the network.

[16] Data represented is general, for illustrative purposes and not actual Amazon.com data.

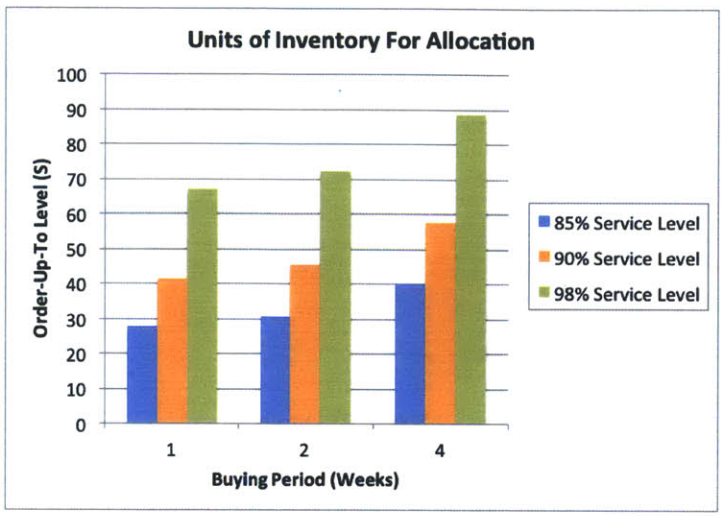


**Figure 5 - Inventory for Allocation**

While this problem might be of little concern if the vast majority of products were reordered on a weekly basis (in line with the weekly forecast) and with a single buying service level, in reality the buying periods and service levels for items within a single grouping can vary quite a lot. Figure 6 and Figure 7 illustrate this problem by ranking the various review periods (in days) and service levels[17] for items in a single product grouping[18].

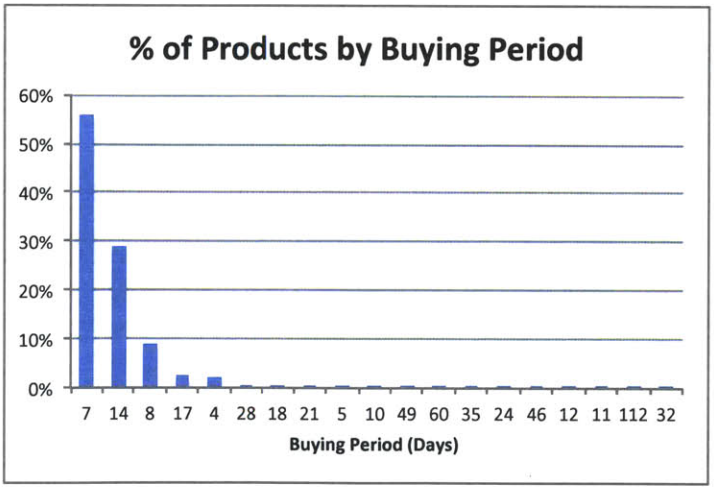


**Figure 6 - Percentage of Products by Buying Period**

[17] While the figure illustrates the variation in service levels, the actual service levels have been left out to protect confidential information.

[18] The figure shown represents a single product grouping. Some current groupings have more buying period variation and some have less, but the challenge illustrated remains the same.

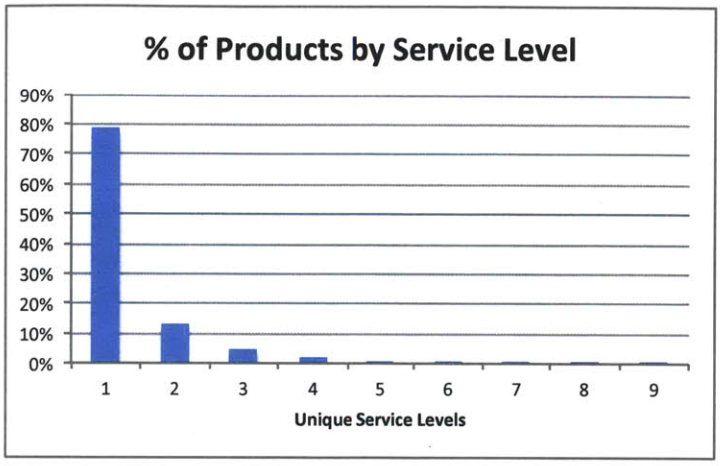


**Figure 7 - Percentage of Products by Service Level**

While it is true that 56% of products in this group are reviewed weekly, almost 30% are reviewed bi-weekly, and the remaining 15% reviewed according to different schedules entirely. As well, considerable variation still exists in the service levels within the product grouping. Ignoring these two factors makes the system difficult to manage efficiently, because when an allocation change is made it is impossible to know exactly how much inbound inventory will be affected. As a result, it is impossible to know with certainty precisely what will result from a given change, or what change should be made to create a desired outcome.

## 4.3 Changes for Proposed Groupings

Incorporating each item's buying period and service level in the velocity definition will resolve these differences in allocation quantity (and expected performance). Not only will this provide a better understanding of the quantity of items purchased into the network for allocation, but it also allows management to make a better decision if allocation factors are strategically changed (for example for a capacity concern). By knowing more precisely what quantity of inventory will be affected by an allocation factor change, management can understand what results to expect from a given change and make a better decision. For these reasons, buying period and service level will be incorporated in the velocity definition for the remainder of the thesis and in the simulation model.

## 4.4 Summary

While managing inventory according to product groupings makes it easier to handle Amazon's vast variety and quantity of products, it is important that these groups capture the factors most important to how the inventory is expected to ship once it is in the network. While the current method is functional, the thresholds are set according to the forecast in units/week, and so it allows for wide variation in the actual amount of inventory allocated into the network by a given strategy. By incorporating an item's buying period and service level into the definition of the set, it becomes possible to describe an item by how much inventory will be allocated to the network. This ability allows us to gain a better understanding of how that inventory will ship, given an allocation strategy and stochastic demand. As a result, the actual quantity of inventory allocated is used as an input to the simulation model described in the following chapter.

# 5 Simulation Model

While the Amazon FC network is an incredibly complex system that functions quite differently from a traditional supply chain, the uncertainties surrounding demand are fairly simple to understand. As a result, while it would be difficult to construct an optimization model to approach this problem that accurately characterizes the network behavior (especially transshipments between the 20+ FCs), it is fairly easy to construct a simulation to demonstrate the benefits of various allocation strategies across the spectrum of inventory levels for an item. For these reasons, simulation was used as the main analytical model for this project.

To protect confidential information, the data used for the model described in this chapter is fictitious, and some simplifying assumptions have been made. However, the framework explained is the same as the simulation built for the project, as are the relationships and opportunities discovered.

## 5.1 Overview

The key question this simulation attempts to answer is "if we are going to stock $x$ units of inventory in the network, how should they be allocated among the FCs?" To answer this question, we allocate a wide range of inventory quantities ($S$) exclusively by one of three practical strategies, and simulate demand for the products according to uncertain demand. Simulated shipments are then executed based on a greedy algorithm, from which anticipated shipping costs are calculated. This trial is repeated multiple times to generate an expected shipping cost per unit over time at each level of $S$. The simulation is repeated for each allocation strategy, after which it becomes a simple exercise of comparing the results of each allocation strategy to find which one has the lowest expected shipping cost at any level of $S$. In order to minimize costs, the lowest cost strategy should be used to allocate inventory in the actual network. Naturally, the results of the simulation depend upon the specific allocation strategies included. In this instance, we explore three specific strategies:

1. Spread inventory proportionally across the FCs according to historic local demand.
2. Hold inventory in three regional hubs.
3. Consolidate everything centrally in a single FC.

While we recognize that this method does not necessarily produce an "optimal" allocation strategy, it does provide a framework for improving the inventory allocation to reduce outbound shipping costs. As well, it is both easy to understand and actionable, which is important in generating sustained savings. Finally, even if an "optimal" solution were discovered, it would not necessarily continue to be optimal in the face of a dynamically changing network, requiring potentially complex updates to the model to chase an increasingly difficult optimal solution. In contrast, the simulation is relatively easy to update with a changing network, and allows us to better see the effects of demand uncertainty.

## 5.2 Model Components

### 5.2.1 Allocation Quantities

The first item that needs to be determined for the simulation is $S$, which is the order-up-to quantity that will be allocated to the network. For simplicity, this quantity was taken as two times an item's expected buying period demand. For example, if an item had a weekly forecast of 10 units and were purchased weekly, 20 units would be allocated.

### 5.2.2 Allocation Strategies

As mentioned previously, the three allocation strategies considered are to spread inventory in proportion with historic demand, allocating it to three regional hubs, or consolidating it in a single, central location. The figure below provides a simple representation of a network where these three strategies could be employed. The thick black line outlines the total area for demand, and is divided into five distinct geographic regions (A, B, C, D, E), with an FC located in each (1, 2, 3, 4, 5). Ideally, these FCs fulfill demand for the region in which they are located. However, they can ship to any region when necessary.

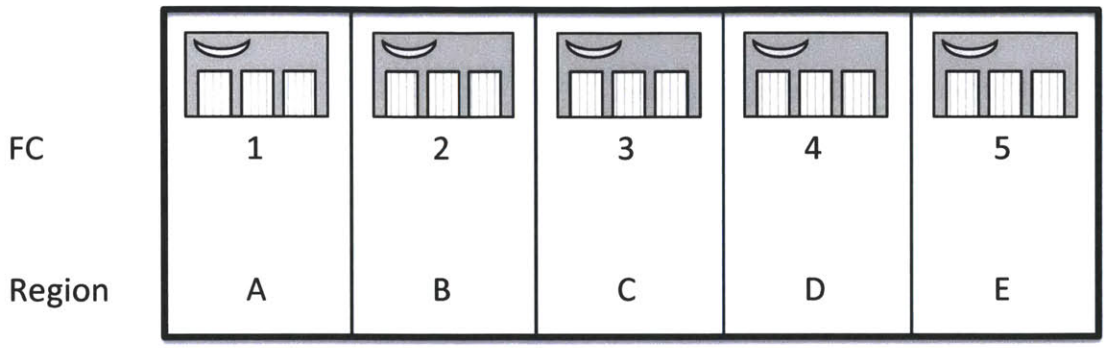


**Figure 8 - Simplified Representation of a Fulfillment Network**

Each of these regions has a corresponding percentage of regional demand, which can be calculated from historical data. For the purpose of demonstrating the simulation, the values in Figure 9 are used for demand percentages for each region[19].

| FC | Region | Example Historic Demand Percentages |
|---|---|---|
| 1 | A | 5% |
| 2 | B | 25% |
| 3 | C | 40% |
| 4 | D | 25% |
| 5 | E | 5% |

**Figure 9 - Example Demand Percentages**

With these demand percentages, each allocation strategy can be explicitly explained. For the "Spread" allocation, inventory will be divided amongst the five FCs according to their respective regional demand percentages. In the case of fractional units, inventory is rounded to a whole number, giving preference to FCs farthest from the center of the network. For example, Figure 10 shows how six units would be placed in the spread out allocation.

[19] Non-indicative of actual Amazon.com regional demand percentages

| FC | Example Historic Demand Percentages | Units Allocated (Calculated) | Units Allocated (actual) |
|---|---|---|---|
| 1 | 5% | 0.3 | 1 |
| 2 | 25% | 1.5 | 1 |
| 3 | 40% | 2.4 | 2 |
| 4 | 25% | 1.5 | 1 |
| 5 | 5% | 0.3 | 1 |

**Figure 10 - Placement of six units by the "Spread" allocation**

For the "Regional Hub" allocation, no inventory is allocated to regions A or E, and their demand is added to the nearest adjacent region. In this case, regions B and D would now receive 30% of the allocation each. Region C would remain the same at 40%, and the same basic process previously explained is employed for fractional units.

Finally, in the "Consolidate" allocation, all inventory is allocated to Region C because it is the most central FC. By allocating here we hedge against very long shipments, because the longest shipment possible is only two regions away.

### 5.2.3 National Demand

In order to create the simulation, the distribution for actual demand given an expectation (forecast) needs to be generated. For the purpose of this example, a modified normal distribution is used, with the mean equal to the forecast and the standard deviation equal to the mean. Negative values are truncated to zero. An example of this example distribution is included below, for an item with a weekly forecast of 10 units.

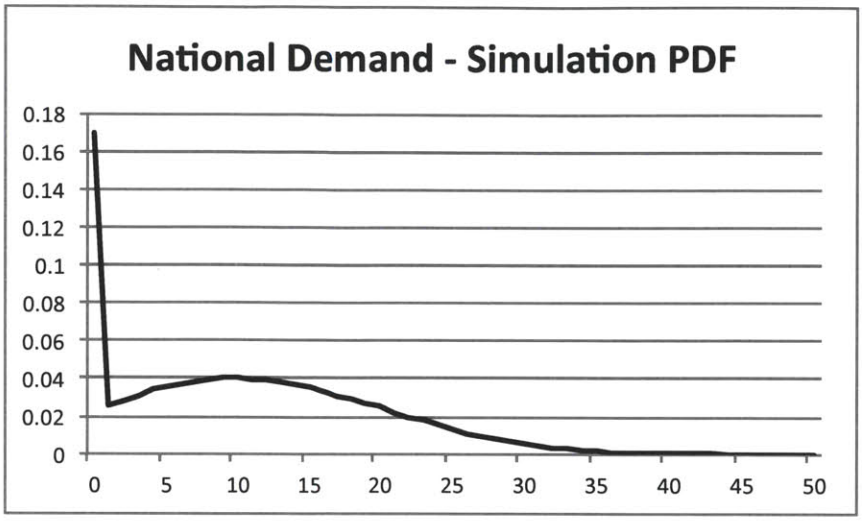


**Figure 11 - Example PDF of demand for simulation, mean = σ = 10**

### 5.2.4 Regions and Demand Percentages

With national demand established, it needs to be broken down into regional demand for FC fulfillment. Using the regional demand breakdown from 5.2.1 as the probability that a unit of demand will ship to a given region, we model the geographic distribution of demand as a multinomial random variable using these probabilities. For example, for each unit of demand, the probability that it is from Region A is 5%, the probability that it is from Region B is 25%, the probability it is from Region C is 40%, etc. As regional demand is calculated as a probability and not a fraction, it is possible (and not unlikely at low levels of demand) for all of demand to originate from a single region, as each unit demanded is treated as an independent event.

### 5.2.5 Sequencing Demand and Multiples

With demand assigned to each region, it is sequenced to simulate actual demand arrivals and fulfillment. It is possible for an FC to stock out before any in-region demand is realized, which results in additional inefficient shipments. By sequencing demand, we allow for this possibility and the simulation more accurately reflects the behavior of the system. Demand is sequenced in a process similar to that described for regional demand assignments, but instead of the probabilities for each region coming from historical trends, they are derived from the regional demand quantities. For example, if four units were

demanded nationally, and one came from region A, the probability that it would be the first shipment would be 25%. If the first unit of demand were to go to another region, then the second unit of demand would go to region A with probability of 33% (one out of three), and so on until each unit of simulated demand were sequenced.

Additionally, multi-item orders must be accounted for, where a multi-item order is one containing multiple units of a single item[20]. This matters to the simulation, because (as will be explained in 5.2.7) it is almost always better to ship two similar units in one box from one FC than in two boxes from two FCs, even if the consolidated shipment must originate from an FC distant to the customer. Each unit of demand becomes a multiple-item order with a set probability, derived from historical data. For the purpose of the example, the probability of a multiple-item order will be 50%. In addition, multiple-item orders are generated simultaneous to regional demand assignments to avoid inflating realized demand.

### 5.2.6 Assigning Demand to FCs ("Greedy)

Each unit of demand, now specified by quantity and region for shipment, must be fulfilled to the customer. This decision is made by a simple greedy algorithm – whichever FC that holds the requisite unit(s) of inventory with the lowest total ship cost will fulfill the order. In the case where multiple items are demanded and no FC has enough inventory to fulfill the order, the order will be split with the two lowest cost FC's handling the fulfillment.

### 5.2.7 Calculating Shipping Cost

With units of demand assigned to FCs, the shipping cost can be calculated for the simulation. For the example, the costs presented in Figure 12 were used for shipments from each FC to each region. Cells marked in green represent highly efficient shipments, while those marked in yellow are more expensive shipments that span multiple regions.

[20] While multi-item orders can also contain multiple distinct items, this complexity was considered out of scope for the project.

| | | Fulfillment Center | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Region | A | $ 3.00 | $ 3.50 | $ 6.00 | $ 6.75 | $ 7.00 |
| | B | $ 3.50 | $ 3.00 | $ 3.50 | $ 6.00 | $ 6.75 |
| | C | $ 6.00 | $ 3.50 | $ 3.00 | $ 3.50 | $ 6.00 |
| | D | $ 6.75 | $ 6.00 | $ 3.50 | $ 3.00 | $ 3.50 |
| | E | $ 7.00 | $ 6.75 | $ 6.00 | $ 3.50 | $ 3.00 |

**Figure 12 - Model Shipping Costs; FC-Region**

In addition to estimating the cost of shipping a single item from any FC to any region, it is also necessary to calculate the cost of shipping multiple items (in the same box) from a single FC to any region. To generate this cost, shipping cost can be thought of as having two parts: fixed costs $C_f$ (putting the item in a box) and variable costs $C_v$ (transporting the box to it's destination). For a single shipment, the following will represent the shipping cost $C$:

$$C = C_f + C_v$$

If we were to ship two items at the same time to the same customer, rather than shipping multiple boxes we could reduce costs by putting both items in the same box. As a result, we would only incur the fixed costs once for this shipment, and so the cost of this shipment would not be $2C$. Generally then, the cost of shipping $n$ units can be represented by the following equation:

$$C(n) = C_f + nC_v$$

While we recognize that this equation would not be accurate for a high number of items (100 units will require a much larger box and more handling considerations than two units, resulting in a higher $C_f$) we feel it is a fair assumption for low values of $n$, which is representative of what generally occurs.

Depending on the magnitude of $C_f$, the savings from putting multiple units in the same box can be quite substantial. For Amazon, this represents a significant savings opportunity that cannot be ignored

in any allocation decision, and the value of $C_f$ used was a parameter that had been derived from historical data. For the example simulation, $C_f$ is set to \$1.00.

With an estimate of shipping costs from any FC to any region, paired with the anticipated cost savings from shipping multiple units in the same box, the shipping cost output from the model can be then translated into a shipping cost per unit (CPU) to facilitate comparison. While these cost calculations result in a normalized output that does not represent actual anticipated shipping costs from a given allocation strategy, it does enable a fair comparison for the relative costs of various strategies, which is the goal of this research.

### 5.2.8 Running the Simulation

The model described in sections 5.2.1 through 5.2.7 represents a single inventory review period, for which inventory is allocated and then demand for that inventory simulated. Each review period is assumed to have independent national and regional demand, and in the case where simulated demand exceeds the allocated inventory quantity, a stock-out occurs and a simplifying assumption is made that excess demand is lost. Finally, inventory does not accumulate from one period to the next because of the nature of the allocation process, which allocates the order-up to quantity (which contains the FC's current inventory position) rather than in inbound inventory quantity.

This simulation is then run for 1000 simulated review periods in order to generate enough data to show an expected relative shipping cost for each allocation strategy at each level of demand. Running the simulation for a higher number of trials will result in more consistent results, though it becomes computationally time-consuming in Microsoft Excel, and 1000 iterations provides results with enough consistency to make a fair comparison. Fewer iterations, however, make it difficult to generate a repeatable expected cost at a given demand value.

## 5.3 Findings/Basic Outcomes

The outputs of this simulation demonstrate a relationship between the various allocation strategies that is both intuitively satisfying and can be used to make better allocation decisions – when both unconstrained and constrained by actual FC capacity. Figure 13 illustrates this comparison[21]. The x-axis of the chart represents the quantity of inventory that is allocated for the buying period, and the y-axis represents the expected shipping cost per unit for each allocation strategy at varying levels of allocated inventory. As mentioned previously, the strategies compared are to consolidate centrally, to hold inventory in the three regional hubs, or to spread inventory throughout the five possible FCs.

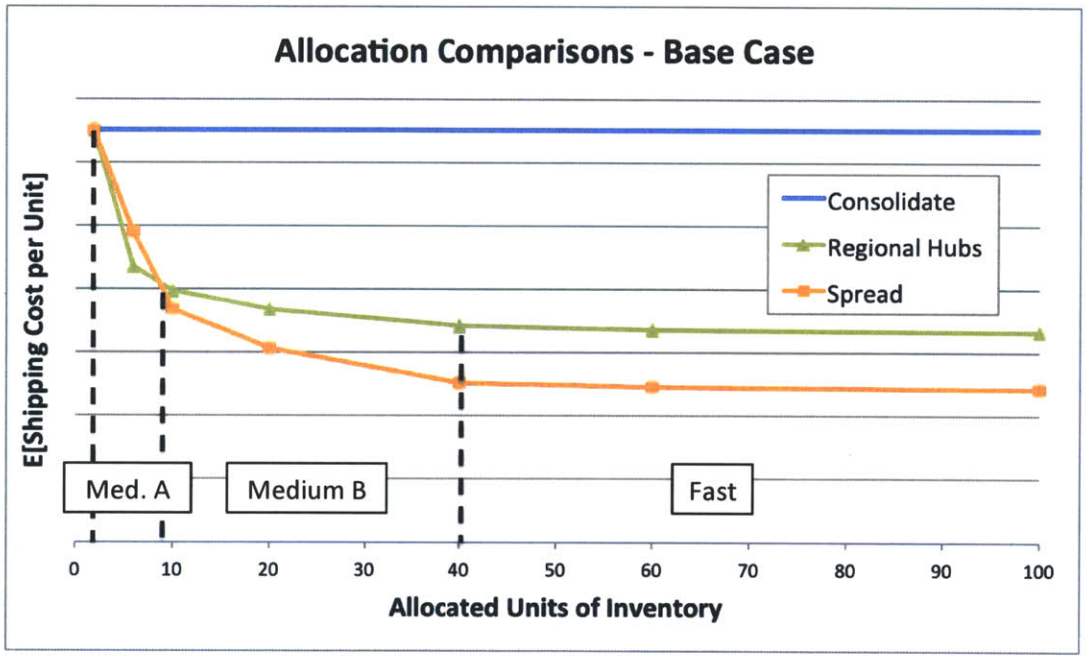


**Figure 13 - Simulation Output/Allocation Strategy Comparison: Base Case**

For the consolidated allocation, the shipping CPU is flat at all levels of inventory allocated. This is in-line with the expected shipping cost that can be easily calculated for a single location. The central location is used because, it is essentially the center of mass of uncertain demand, and so has the lowest

[21] Figures 13-16 show the output from the example simulation described for four different scenarios. They were not generated with actual Amazon.com data, nor do they necessarily reflect actual values for shipping costs or service levels. However, the relationships shown are in-line-with the outputs of the model with actual Amazon data, which is the key takeaway from the research.

expected shipping cost of all FCs in the network. While the CPU of this strategy is stable, it is the low-cost option only for very low levels of allocated inventory. In this instance in fact, the other two options are never more expensive than the Consolidated allocation. This is because of the way fractional units are allocated, and at very low numbers the allocation for the model is essentially the same for all three strategies. If this were not the case, a region on the chart would exist where consolidating inventory would be more cost effective than the Regional Hub or Spread allocations. This region of inventory would be classified as "slow".

As the quantity of inventory allocated increases, the shipping cost for the Regional Hub allocation quickly falls to levels far below the Consolidated allocation. However, it also levels out and begins to stabilize near a theoretical minimum CPU for the allocation. Placing inventory in the regional hubs saves so much money over the Single location allocation because it allows more units to be located close to the customer where very inexpensive shipments are possible. While this strategy will also incur expensive cross-country shipments, the increased costs of these shipments are far outweighed by the benefit garnered by having inventory closer to customers.

The Spread allocation follows a similar relationship to the Regional Hub allocation, though its initial cost reduction isn't as steep, due to the higher number of cross-country shipments incurred. However, while the initial cost reduction isn't as steep, it falls much lower at higher levels of inventory allocated. This is because at these levels this strategy provides even more inventory in locations closer to customers, facilitating many very cost-efficient shipments.

Comparing the three allocations in this way enables us to identify four unique and important regions, which can then be used to prescribe an allocation strategy for each region's respective inventory. These regions are "Slow"[22], "Medium-A", "Medium-B", and "Fast". Naturally, using the allocation

[22] The "Slow" region is where the Consolidate allocation is the lowest cost option. While this region does not exist in the example simulation (here the other allocations are always at least as good as the

strategy with the lowest expected CPU will be the preferred choice. If a "Slow" Region exists (as it will in some cases), it will provide a very straightforward recommendation – exclusively allocate slow inventory to the central FC. The "Medium-A" and "Fast" regions are similarly conducive to straightforward recommendations – allocating them with the Regional Hub and Spread allocations (respectively) will provide the lowest expected fulfillment cost.

However, "Medium-B" items provide a different strategic opportunity. The Spread allocation does provide the lowest expected CPU in this region, as it does for Fast items. However, the difference between the Spread and Regional Hub allocations for Medium-B items is not as great. In fact, it ranges from zero (where the CPUs for the strategies intersect), up to an approximate maximum difference where region four begins[23]. What this means is that if we were required to re-allocate some inventory in light of a capacitated network, and pull it away from the Spread allocation, Medium-B items would be the best items to move because they have a lower opportunity cost from using a non-preferred allocation than do Fast items. Functionally then, rather than moving a portion of all Fast and Medium-B items to satisfy a capacity constraint (which is in-line with the current practice), savings can be achieved by moving *all* Medium-B items before pulling any Fast items from the preferred allocation. This will allow the very popular items to consistently be located closer to the customer, enabling the firm to take fuller advantage of the resulting benefits.

## 5.4 Changes that Effect the Model

As with any model, the outputs of this allocation simulation model depend on the factors and assumptions that are incorporated into it. Among the most notable factors in the model that, when changed, create a material affect on the output are the probability of multiple shipments, the shipping cost function, the buying period and service level, and the variation in the demand distribution.

---

Consolidate allocation), it will exist in cases where demand is distributed differently than the symmetric distribution used for the example, shown in 5.2.1

[23] For Fast items, the difference between the Regional Hub and Spread allocations remains approximately constant. That is, the lines on the chart are very close to parallel.

### 5.4.1 Regional Demand Probabilities

Different regional demand probabilities will affect the output of the model. Most noticeably, if demand is skewed to one of the non-central regions, then a section of Slow items will be present for the allocation strategy comparison. As a result, it is quite important to use a good approximation of regional demand percentages in order for the outputs of the model to be applicable.

### 5.4.2 Probability of Multiple Shipments

When the probability of multiple item shipments is increased, the size of the Medium-A segment increases. In other words, it becomes more efficient to place items according to the Regional Hub allocation. The reason is that with more multiple shipments, it becomes more important to always have multiple items at each FC for shipment in order to take advantage of cost savings from consolidating units in a single box. This is most likely when inventory is held in fewer FCs. Figure 14 illustrates this relationship, where all units are ordered as a set of two units.

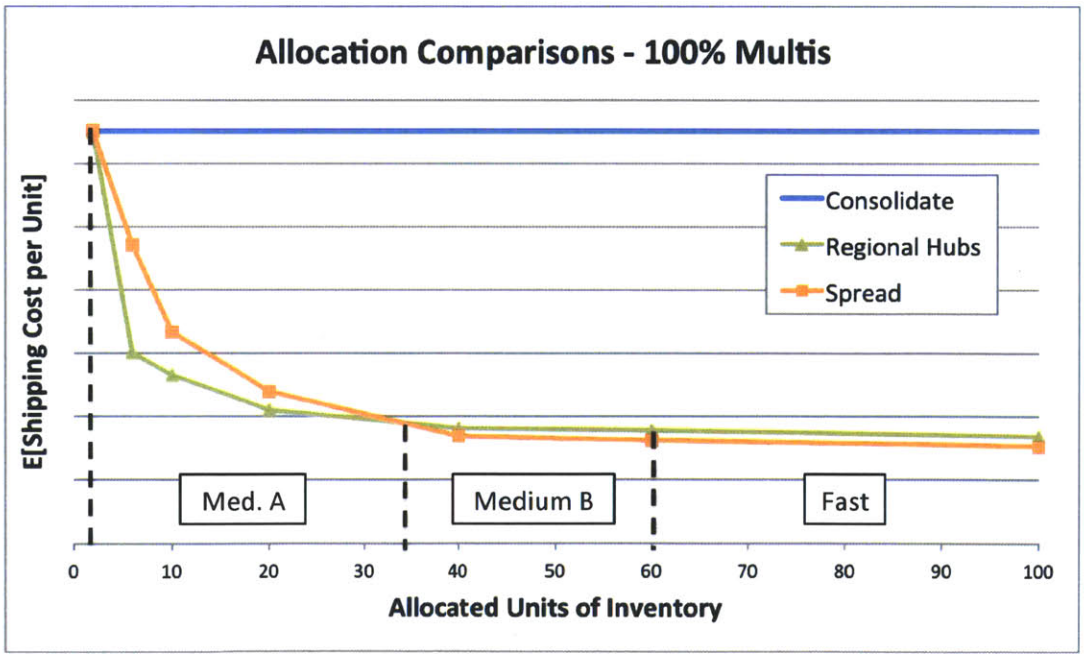


**Figure 14 - Simulation Output/Allocation Strategy Comparison: 100% Multis**

### 5.4.3 Shipping Cost Function

If it becomes more expensive to ship items cross-country, then the Medium-A segment will again increase in size. As well, the difference in expected costs of the Regional Hub and Spread allocations will become much smaller. Significantly higher costs for cross-country shipments mean they should be avoided as much as possible, which is exactly what the Regional Hub allocation enables. Additionally, much of the benefit of having inventory close to customers that is provided by the Spread allocation is lost due to very inefficient cross-country shipments, even at high levels of allocated inventory. This is a very real possibility, especially when it is necessary to fulfill a shipment quickly and it must be fulfilled by air because the only unit available is far from the customer. Figure 16 illustrates this relationship, where the shipping costs from Figure 15 were used. In comparison to the base case shipping cost example, there is much higher cost for shipments greater than two regions in distance in this scenario, representing the expense of air shipments.

| | | Fulfillment Center | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Region | A | 3 | 3.5 | 7 | 10 | 11 |
| | B | 3.5 | 3 | 3.5 | 7 | 10 |
| | C | 7 | 3.5 | 3 | 3.5 | 7 |
| | D | 10 | 7 | 3.5 | 3 | 3.5 |
| | E | 11 | 10 | 7 | 3.5 | 3 |

**Figure 15 - High-Cost Long Shipment Costs: FC-Region**

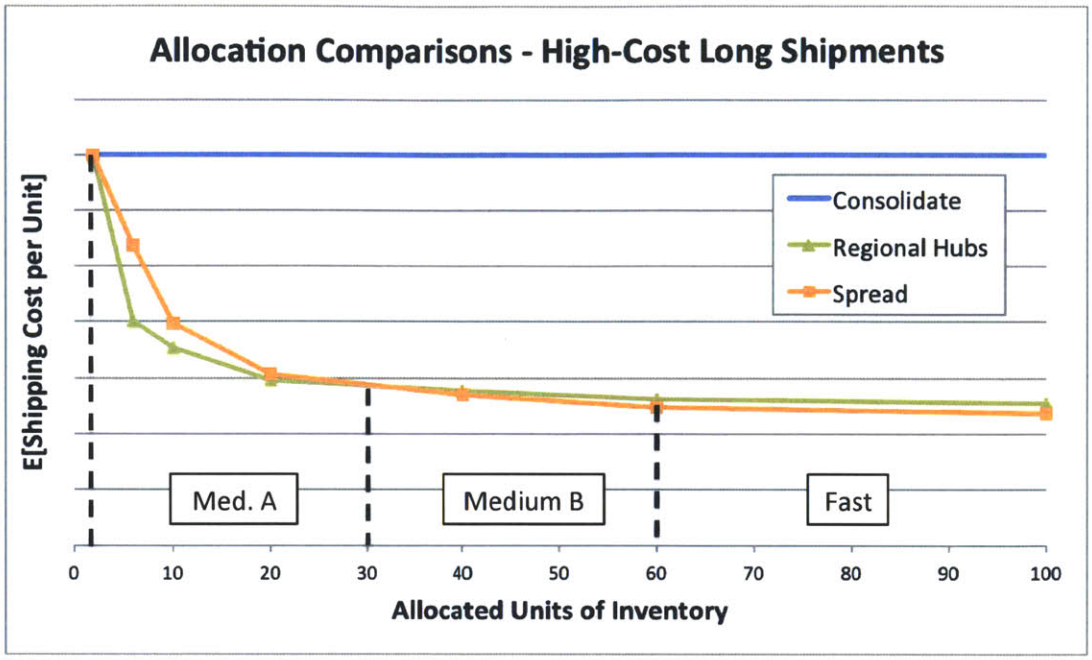


**Figure 16 - Simulation Output/Allocation Strategy Comparison: High-Cost Long Shipments**

### 5.4.4 Service Level

Changes to the service level will dramatically affect the output of the model, because it will directly impact the amount of inventory allocated to serve a given level of demand. If the service level is reduced, the Medium-A segment will increase in size and the thresholds for Medium-B and Fast items will also increase. This is because less inventory will be held to handle variation in demand, both in its quantity and location. As a result, consolidating more inventory regionally becomes the most cost-effective strategy to protect against the variation and subsequent expensive, long shipments. Conversely if the service level is increased, the increased safety stock results in the Medium-A section essentially disappearing and it becomes most cost-effective to spread everything out as much as possible. These effects can be seen in Figure 17 and Figure 18 where the model was run with $S$ covering one and three times the buying period demand, respectively.

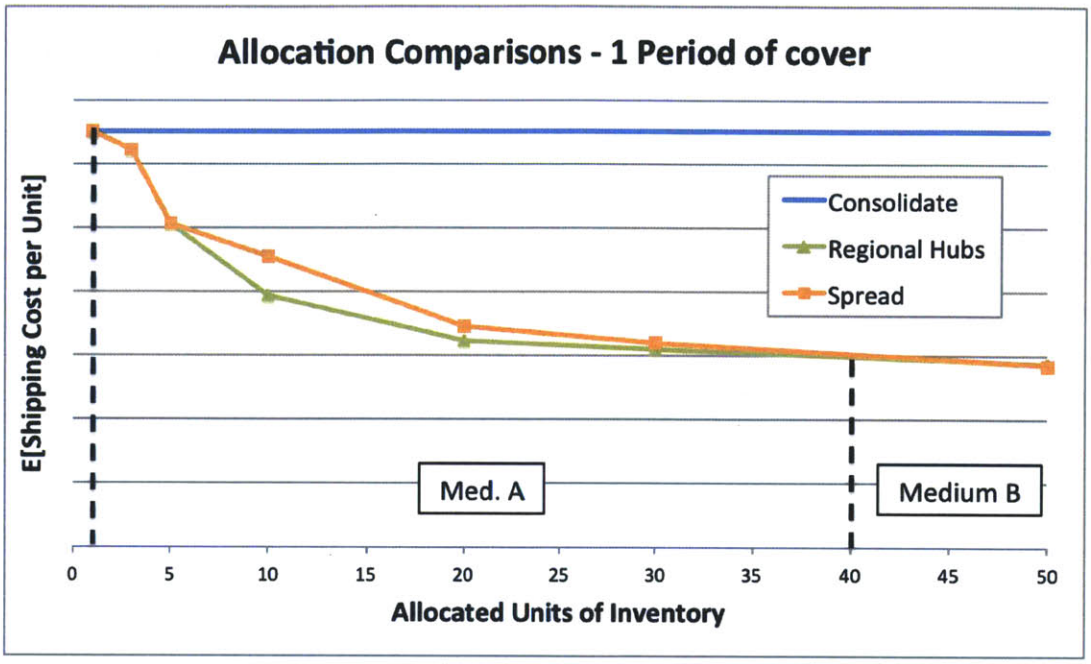


**Figure 17 - Simulation Output/Allocation Strategy Comparison: Low Service Level**

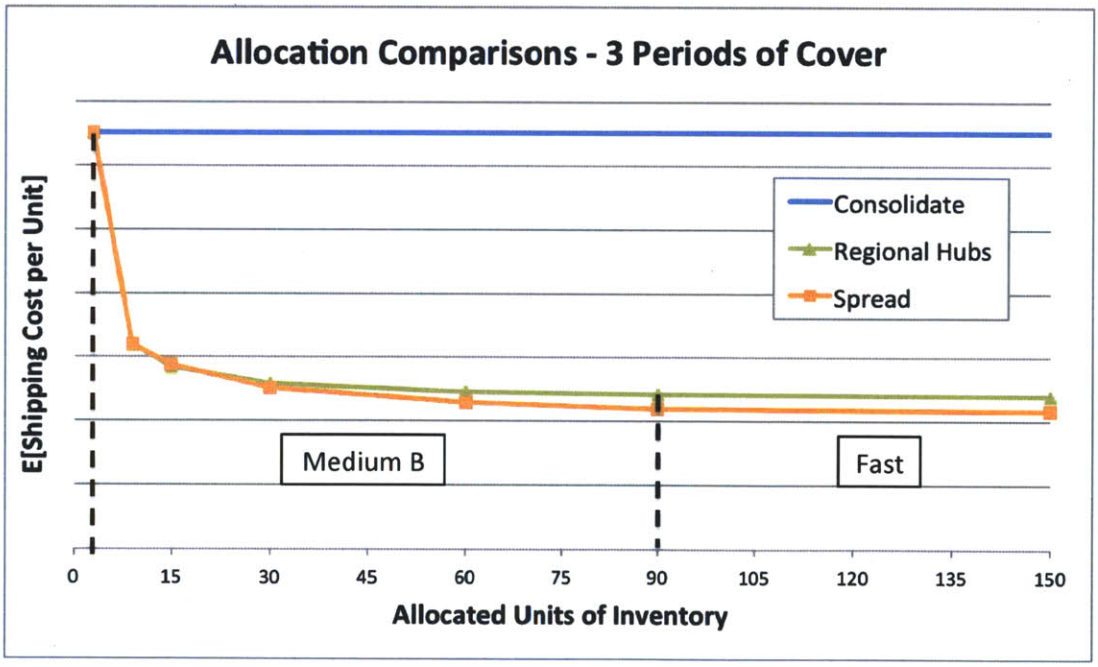


**Figure 18 - Simulation Output/Allocation Strategy Comparison: High Service Level**

### 5.4.5 Variation in the Demand Distribution

Increased variation in demand naturally makes efficient inventory planning and allocation even more difficult. Relating to this problem, it makes it more difficult to take advantage of being close to the

customer – which is provided by the Spread allocation – because you can't reliably predict where the customer is. However, increased variation does affect the model differently than those factors previously explained. To illustrate this relationship, a demand distribution with $\sigma = 2\mu$ was used for the simulation (in contrast to the base case, where $\sigma = \mu$). As shown in Figure 19, with higher variation the Medium-A segment becomes slightly larger than in the base case, and similar to the base case the Spread allocation becomes significantly better than the Regional hub allocation at higher levels of allocated inventory. However, the Medium-B segment is much larger with higher variation than in the base case. As a result, if demand is highly variable at all levels, the savings potential from using Medium-B items as a flexible option to react to capacity becomes even more valuable.

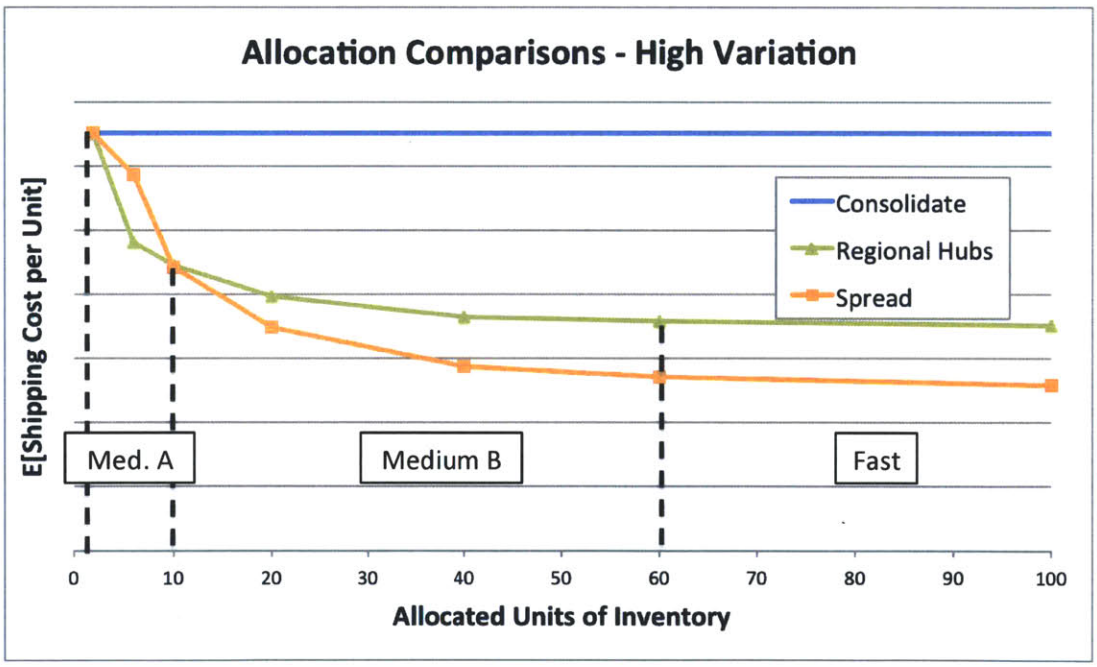


**Figure 19 - Simulation Output/Allocation Strategy Comparison: High Variation**

### 5.4.6 Comparison Between Different Scenarios

As shown in the previous sections, different model inputs result in different outputs for the different segments. The thresholds explained previously define these segments, a summary of which is

shown in Figure 20[24]. While the threshold values are different given different model inputs and assumptions, the way in which these segments should be used to effectively manage the supply chain is the same across scenarios.

| | Thresholds | | |
|---|---|---|---|
| | Medium A | Medium B | Fast |
| **Base** | 0 | 9 | 40 |
| **100% Multis** | 0 | 34 | 60 |
| **High-Cost Long Shipments** | 0 | 30 | 60 |
| **Low Service Level (1 Period of Cover)** | 0 | 40 | - |
| **High Service Level (3 Periods of Cover)** | 0 | 0 | 90 |
| **High Variation** | 0 | 5 | 30 |

**Figure 20 - Comparison in thresholds with different model inputs**

## 5.5 Summary

The simulation presented provides a method for classifying inventory for allocation, given various practical allocation strategies. While the outputs of the model depend on the assumptions used, it shows that there are some items that will be most efficiently fulfilled by consolidating either regionally or in a single FC, and many items should simply be spread as much as possible to align with expected regional demand.

Additionally – and perhaps most importantly given the physical constraints of the network – it provides a method for finding a "medium" band of items for allocation. These items can be used as a flexible option when an FC runs out of physical capacity. By pulling them back into a regional hub (where presumably there is more capacity), outlying FCs can more effectively use their capacity to satisfy demand for higher running items, which generate greater savings from the spread out allocation.

[24] The Fast threshold for low service level items could not be calculated because the simulation was not built to handle allocated inventory quantities this high at such a low service level, as it was deemed unrealistic. Is likely a very high level however, due to the fact that there is no safety stock to handle variation in demand and its region of origination.

# 6 Conclusions & Recommendations

## 6.1 Conclusions for the project

While Amazon has a non-standard supply chain, analytical methods can still be used to improve the process for inventory allocation. In this instance, simulation was used to evaluate a given set of practical allocation strategies. From this, product groups were then assigned to the best relative strategy, providing more control to the process and facilitating a lower expected cost of fulfillment.

Interestingly, while using the spread out allocation strategy for Medium-B items may provide the lowest expected cost for these items, if capacity is exhausted in outlying FCs these items should be the first pulled back into the regional hub allocation. The reason is because while the spread allocation does have the lowest expected cost, the cost difference between the spread allocation and the regional hub allocation is less than that of the Fast items. As a result, by moving Medium-B items to satisfy a capacity constraint, we give up less in expected cost savings than if Fast items are moved to fulfill a capacity constraint. As a result, the total expected loss would be less by using this strategy than if a portion of all fast and medium items were pulled back into regional hubs.

Given the fact that capacity constraints are the main driver for changing inventory allocations, this is a very pragmatic and useful result that can give the managers of this type of system more control over their process and provide significant cost savings.

## 6.2 Recommendations for other areas of research

Through the course of this project, other interesting and related problems have surfaced, which while out of the scope of this thesis, would nonetheless be interesting areas for further research. Principal among these would be to more fully analyze the system dynamics of the inventory allocation and fulfillment process. There are a multitude of variables that make this system difficult to control manually and prone to oscillations in inventory levels. Some of these dynamics are the result of seasonal effects,

but much of the variation can be attributed to the operation of the system. An example of these dynamics can be seen in Figure 21, which shows how changes to allocation factors for a product group at an FC relates to inventory levels and shipments of the product grouping. The y-axis in the figure is the percent of network allocation, inventory, and shipments for the product group at the FC, and the x-axis is the progression of time[25]. While it is natural (and present in the figure) for there to be a delay between the time an allocation factor is changed and the time that inventory and shipments arrive at the new target percentage, the cyclical nature of the allocation changes (and the response in inventory and shipments), indicates that there is likely an opportunity to improve the way in which the supply chain is managed to smooth out this relationship. A thorough analysis of the dynamics of the system would not only be interesting, but would undoubtedly yield improvements to the manner in which the supply chain is managed, resulting in both improved control and lower fulfillment expense.

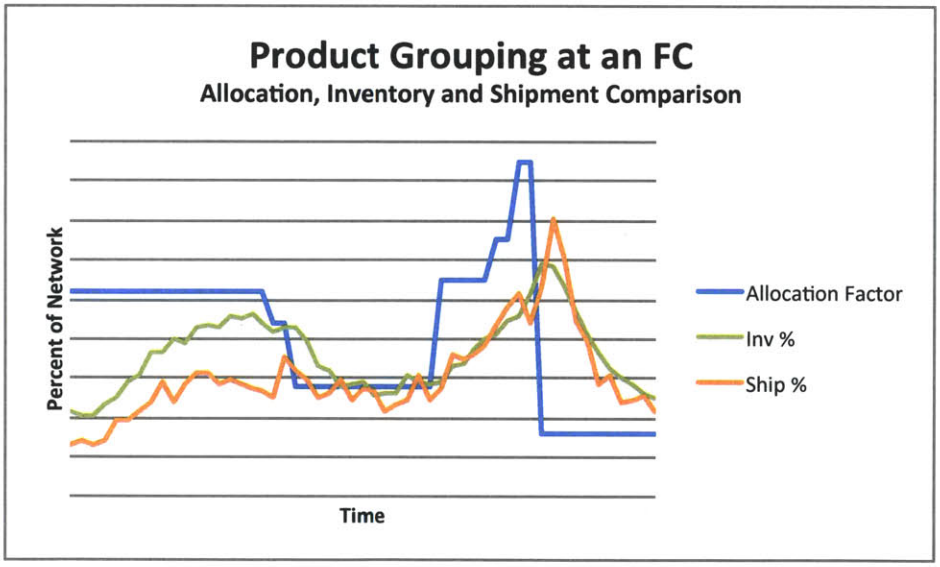


**Figure 21 - Cyclical relationship between allocation, inventory, and shipments**

Additionally, the model presented provides a way to compare relative allocation strategies given the state of the FC Network, but the FC network is constantly changing. While it is relatively easy to change the model to add or subtract facilities, this does not capture what should be done (from an

[25] The actual time period has been omitted from the figure to protect confidential information, but the peaks and valleys shown in the chart cannot exclusively be attributed to specific seasonal effects.

allocation perspective) as the firm transitions from a current state of the network to a future state. Analyzing this transition would likely yield further improvements in how the network should be managed as the company and its fulfillment capabilities continue to grow and evolve.

# 7 References


[1] Wingo, Scot. Amazon's Wheel of Growth. Seeking Alpha article 22 Feb. 2009. Retrieved 18 Jan. 2012. http://seekingalpha.com/article/121955-amazon-s-wheel-of-growth

[2] Anupindi, R; Chopra, S; Deshmukh, S; Van Mieghem, J; & Zemel, E. 2005. Managing Business Process Flows (Second Edition ed.). Pearson Prentice Hall

[3] Silver, E A; Pyke, D F; & Peterson, R. 1998. Inventory Management and Production Planning and Scheduling (Third Edition). John Wiley & Sons.

[4] Eppen, G D 1979. Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem. Management Science Volume 25, Issue 5 pp 498–501.

[5] Gerchak, Y; He, Q M 2002. On the Relation Between the Benefits of Risk Pooling and the Variability of Demand. IIE Transactions Volume 35, Issue 11 pp 1027-1031

[6] Berman, O; Krass, D; Tajbakhsh, M M 2011. On the Benefits of Risk Pooling in Inventory Management. Production and Operations Management Volume 20, Issue 1 pp 57-71.

[7] Sobel, M J 2008. Risk Pooling. Building Intuition: Insights from Basic Operations Management Models and Principles (Book). Series: International Series in Operations Research and Management Science. Volume 115, pp 155-174.

[8] Zotteri, G; Kalchshmidt, M; Caniato, F 2002. The Impact of Aggregation Level on Forecasting Performance. International Journal of Production Economics. Volume 93-4, pp479-491.

[9] Federgruen, A; Zipkin, P1984. Approximations of Dynamic, Multilocation Production and Inventory Problems. Management Science. Volume 30, Issue 1 pp 69-84

[10] Svoronos, A; Zipkin, P 1988. Estimating the Performance of Multi-Level Inventory Systems. Operations Research. Volume 36, Issue 1, pp 57-72

[11] Jackson, P L 1988. Stock Allocation in a Two-Echelon Distribution System or "What To Do Until Your Ship Comes In". Management Science. Volume 34, Issue 7 pp 880-895.

[12] Graves, S C 1996. A Multiechelon Inventory Model with Fixed Replenishment Intervals. Management Science. Volume 42, Issue 1, pp 1-18

[13] Parker, R; Kapuscinski, R 2004. Optimal Policies for a Capacitated Two-Echelon Inventory System. Operations Research. Volume 52, Issue 5 pp739-755.

[14] Axsäter, S; Marklund J 2008. Optimal Position-Based Warehouse Ordering in Divergent Two-Echelon Inventory Systems. Operations Research. Volume 56, Issue 4 pp 976-991

[15] Xu, P J 2005. Order Fulfillment in Online Retailing: What Goes Where. Phd Dissertation at the Massachusetts Institute of Technology, Operations Research.

[16] Benjaafar, S; Li, Y; Xu, D.; Elhedhli, S. 2008. Demand Allocation in Systems with Multiple Inventory Locations and Multiple Demand Sources. Manufacturing & Service Operations Management Vol 10, No. 1 pp 43-60.

[17] Roberts, B; Berg, N 2012. Walmart: Key Insights and Practical Lessons from the World's Largest Retailer. Kogan Page Ltd. pp 76.