Inbound Supply Chain Optimization with Ship-Mode Variation in a Fixed-Capacity Fulfillment Center

by

Jeffrey Birenbaum
B.S. Mechanical Engineering, George Washington University, 2014

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

Master of Business Administration
and
Master of Science in Civil and Environmental Engineering

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

June 2018

Signature of Author________ Signature redacted ___________
Jeffrey P. Birenbaum, Student
MIT Sloan School of Management, MIT Civil and Environmental Engineering
May 11, 2018

Certified by________________ Signature redacted ___________
Stephen C. Graves, Thesis Supervisor
Abraham J. Siegel Professor of Management, MIT Sloan School of Management

Certified by________ Signature redacted ___________
David Simchi-Levi, Thesis Supervisor
Professor, MIT School of Engineering

Accepted by_____________Signature redacted__________
Jesse Kroll
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Accepted by_________________Signature redacted___
Maura Herson
Director of MIT Sloan MBA Program, MIT Sloan School of Management

Inbound Supply Chain Optimization with Ship-Mode Variation in a Fixed-Capacity Fulfillment Center

by

Jeffrey Birenbaum

Submitted to the MIT Sloan School of Management and MIT School of Engineering Civil and Environmental Engineering Department on May 11th, 2018 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Civil and Environmental Engineering

Abstract

Amazon.com's retail business in North America is growing rapidly, with sales increasing by 25% each of the past two years. In order to scale with expected continued sales growth, Amazon has been investing heavily in its inbound supply chain, where product is received and allocated to various nodes, with cross-dock facilities, Amazon Robotics fulfillment centers and traditional fulfillment centers constituting a multi-echelon distribution network. In an Amazon Robotics fulfillment center, robotic drives retrieve and deliver portable inventory pods, where product is stowed and picked at fixed stations.

Currently, approximately 65% of associate hours within the inbound department are utilized in the direct process of stow, while the other 35% of associate hours in the inbound department are utilized in support of the stow process in tasks such as corrugate removal and product container management. As a result, there is a continued emphasis on improving the efficiency of the non-value added tasks utilized in support of the stow process in order to utilize as many hours as possible in the value-added stow process.

This thesis proposes a linear optimization-based analysis framework and capital allocation model that can be utilized to determine the investment viability for different automation systems and process improvements, which could improve efficiency and reduce overall cost in the Amazon Robotics fulfillment centers. This is especially the case within those fulfillment centers that are labor constrained. Labor constraints within a fulfillment center result in artificial limits set within Amazon's inventory placement algorithm, changing the origin of the shipment of product(s) to customers, which results in additional outbound transportation cost. This study will uncover unrealized cost-improvement areas by suggesting an inbound conveyance solution that can improve upon the current human-powered inbound system, and provides further areas of investigation for additional improvement. Implementation of the selected automation solution reduces inbound department hours by ~3% with a payback period of ~0.93 years for the fulfillment center in question, while improving labor-constrained fulfillment center capacity by as much as 1%, and suggests further areas of investigation that can improve overall cost within the inbound supply chain by over 10%.

Thesis Advisor: Stephen C. Graves
Title: Abraham J. Siegel Professor of Management, MIT Sloan School of Management

Thesis Advisor: David Simchi-Levi
Title: Professor, Department of Civil & Environmental Engineering

This Page Is Intentionally Left Blank

Acknowledgements

This thesis is the result of a broad and intense effort to uncover improvements in Amazon's inbound supply chain, and it's a direct result of the help and encouragement of colleagues and friends that I can call my experience at Amazon a success.

I owe a great deal of gratitude to my supervisor at Amazon, Matt Martindale. You allowed me to explore Amazon Fulfillment in search of where I could make the largest contributions, and that resulted in six months of learning and experimenting that I wouldn't trade away for anything. I would also like to thank Andrew Bignell and Ben Mullican for the endless support, mentorship, and friendship during my time at Amazon and afterwards. Both of you helped steer me towards success (or away from trouble) and provided a guiding voice or ear when I needed it most. Another thanks to Brian Urkiel, Brian Donato, and Greg Zielinski; as prior Sloan graduates and important sponsors of Amazon's relationship with LGO and Sloan during my project, your support of me and the program allowed my dreams of calling myself an Amazonian to be realized.

I also owe a great debt to Stephen Graves and David Simchi-Levi for agreeing to take me on as an advisee and for being engaged in the thesis creation process and the LGO program. Professor Graves continuously challenged me to improve my research and thesis and provided crucial feedback that helped me achieve more through this process, and Professor Simchi-Levi provided early encouragement and direction when it was needed most.

To colleagues I was lucky to call friends by the end of my time at Amazon, thanks to Josh Nahley, John McEwen, Olivia Ramirez, and April Gibbs for making me feel at home in a new city far from home. Same goes for the inbound team at MSP1: Aaron Meyers, Nathan Ray, Jay Erwin, John Russell and Mike Andrea were critical to the research effort and furthermore to making me feel a part of the MSP1 team.

Though my time was mostly spent at MSP1, the support from Amazonians in Seattle (and elsewhere) was unparalleled and I cannot thank Linnea Ottervik-Clearfield, Tyler Park, and countless others for patiently answering all of my questions.

To the LGOs that played a major part in this process; namely Drew Walker as the other off-cycle Amazon LGO fellow and Aaron Small as the predecessor to my project, I can't thank each of you enough for the friendship and advice during the research project.

And finally, to my parents, grandparents and other supporters who have supported all of my adventures over the years. My continuing curiosity and motivation to succeed comes from you!

This Page Is Intentionally Left Blank

Table of Contents

List of Figures

List of Tables

List of Equations

Glossary

Amnesty	Amnesty product is product that is not located in its rightful place within the fulfillment center environment; either not in its correct container or inventory location. Amnesty product can result from container overflow, missed scans, or an improper pick, for instance.
ASIN	Amazon Standard Identification Number, which Amazon uses to identify unique products within in the entire network.
Customer Promise	Amazon considers its promised delivery date to a customer a customer promise; essentially, Amazon promises that it will deliver a product on time, every time.
Cross-Dock (IXD)	A cross-dock is a high-volume inbound facility that receives customer orders and dis-aggregates them, allowing inventory to be efficiently received and spread across a distribution network.
Direct Labor Hours	All labor hours utilized in direct process paths, including stow, stow-to-prime, and vendor receive.
Down Stacking	The process of "destroying" a complete pallet, or removing all containers from the pallet such that all that remains is the pallet and the containers are located elsewhere.
Fast-Track (Prime)	A fast-track ASIN is an item that is available under the Amazon Prime service, or another similar, expedited delivery option.
Fulfillment Center (FC)	A fulfillment center is the primary location in which inventory is stored in and shipped to customers. An Amazon Robotics fulfillment center utilizes the Kiva (now Amazon Robotics) inventory management system to improve capacity and efficiency of the overall inventory storage system.
Effective Hourly Labor Cost Rate	The effective cost that Amazon incurs per labor hour. The cost includes the average wage rate, as well as insurance and other variable labor costs.
Inbound Throughput-Per-Labor-Hour (TPH)	The primary performance metric utilized when determining inbound performance, which is calculated by aggregating all of the labor hours utilized and units processed through the inbound department.
Indirect Labor Hours	All labor hours utilized in support of the direct process of stow: this includes all manual pallet and container moves and quality / inventory support associates.
Lane	A transshipment lane is a transportation arc that connects an IXD-FC pair or FC-FC pair. Trailers flow through these lanes to deliver product and other items (such as reusable totes and pallets) back and forth between different facilities.

Load Balancing	A process utilized to move network charge (or shippable customer orders / volume of sales) from location to location to balance the amount of work according to the capacity by location
Origin-Destination (OD) Pair	An OD pair is a connection between an origin site and a destination site serviced by a transportation arc. A lane is the connection between these sites.
Process Path	Job functions within a fulfillment center performed by Amazon associates are considered process paths, and process paths are aggregated at various levels to show building performance. If an associate is "in path", the associate is performing a standard function within the fulfillment center, such is pick, pack, stow, or count.
Robotic Storage Platform (RSP)	The area within an Amazon Robotics sortable fulfillment center in which products are stowed, picked, and counted.
Ship-Mode	The format in which freight is loaded onto a truck, with two primary options: floor-load and palletized.
Transportation Arc	A connection between an origin-destination pair where trailers move inventory from the origin to the destination (or vice versa).
Transshipment	A trailer that is sent from one internal location to another internal location; in Amazon's case, usually from a cross-dock to a fulfillment center.
Velocity	Velocity characterizes the inventory turns associated with a product. A high-velocity product turns (sells) very quickly, while a low-velocity product turns very slowly.
Vendor Receive	Product received directly from the vendor (generally) in a fulfillment center, which requires product to be manually received into another container by an associate.
Vertical Reciprocating Conveyor (VRC)	A vertical lift that is utilized to move pallets and carts to and from different floors.

1 Introduction

1.1 Amazon Customer Fulfillment

Amazon.com, established in July of 1994 by Jeff Bezos in Bellevue, Washington, is a retail and technology firm that had over $130 billion in revenue in 2016 [1]. In Amazon's 23-year history, Amazon has expanded from selling books and CDs to being the world's largest web services provider and one of the world's premier electronics developers, with devices like the Kindle and Amazon Echo. While they have entered several diverse markets, Amazon's primary business remains its retail business, where two-day shipping (via the Amazon Prime membership service) and the world's largest selection have resulted in significant advantages for the firm in comparison to its competitors.

Recently, this primary business has been expanding, with new services opening, including PrimeNow, an urban 2-hour delivery service, and Prime Pantry, a grocer and household products provider. This is in addition to its traditional fulfillment business; Amazon continues to open additional fulfillment centers in the United States (and abroad) at a staggering pace, with over 10 robotics fulfillment centers now open in the United States alone since its acquisition of Kiva Systems (now Amazon Robotics) in 2012.

1.2 Problem Statement

Amazon has several different versions of its fulfillment centers, which are characterized by the technology that supports order fulfillment and by the type of inventory that the fulfillment center stocks. In an Amazon Robotics (AR) Sortable fulfillment center, small-and-medium-sized product is stored in mixed-ASIN pods and retrieved by Amazon Robotics drives from the Robotic Storage Platform (RSP), which operate under a complex set of rules designed to ensure that Amazon associates always have a storage location in front of them.

Together, this system constitutes many of the processes that traditional order fulfillment is composed of, including the functions of storing (stowing) the product, picking a product, and maintaining inventory accuracy (count). These stowers, pickers, and counters operate at fixed stations where inventory pods are delivered to them, in comparison to traditional warehouse operations where employees manually stow and retrieve product using forklifts and carts.

Figure 1: Amazon Robotics Inventory System

In the inbound area of an Amazon Robotics fulfillment center, Amazon associates stow product into the RSP, at which point the product is considered to be in a prime-eligible sellable (fast-track) location. In the current state, approximately 65% of labor in the inbound area is related to the direct function of stowing product into the RSP while the remaining 35% of labor is utilized in support of stow. As a result, optimizing the productivity of the direct function of stow, while minimizing the usage of indirect labor in support of stow becomes the requisite objective in order to improve the efficiency of the current generation of Amazon Robotics fulfillment centers.

This thesis explores three areas of opportunity that might yield improvement for the Amazon Robotics fulfillment center's inbound department:

1. Upstream (cross-dock network) processing changes or improvements that could result in less variability of the inputs, resulting in lower inbound processing cost. The primary sources of variability are in ship-mode, defined as the format of the product that is shipped to the fulfillment center (i.e. pallet or floor-loaded), and container-mix, which is defined as the type of container that holds the product (i.e. tote or case).
2. Process selection within the fulfillment center, as it pertains to the different processing options that buildings have in regards to processing inbound product.
3. Inbound automation selection within the fulfillment center, and the approach that should be utilized to justify capital investment within a warehouse operation that suffers from competing inbound constraints (i.e. station constraints or labor constraints).

1.3 Project Hypothesis

The hypothesis of this thesis is that using ship-mode specific automation or processes to offset ship-mode and container-mix-based inefficiencies can result in improved performance for indirect labor in the inbound area, as well as improvements in stow rates (defined as units per labor hour, or UPH). Furthermore, labor constraints and their effects on both building productivity and supply chain continuity should be considered in an effort to make better capital decisions.

1.4 Project Approach

The first part of the project involved a deep dive into the various process paths that Amazon associates are placed in within the inbound department of an Amazon Robotic sortable fulfillment center. Through data collection and analysis, it was hypothesized that the variation of performance witnessed within the fulfillment center would be heavily correlated to the arrival ship-mode of product from the cross-dock facilities.

As a result, the search for methods to improve indirect performance at the Amazon Robotics sortable fulfillment center began by investigating possible improvements in two separate work streams. The first work stream involved a comprehensive supply chain analysis, which focused on theorizing ways to reduce the variation in ship-mode of incoming transshipments from the cross-dock facilities. The second work stream focused on improving the processing capabilities of the fulfillment center by utilizing ship-mode specific automation to reduce certain ship-mode-based effects on performance.

Two conclusions were generated from these analyses. First, it was determined that the costs within the inbound supply chain severely limit advantageous opportunities to control transshipment ship-mode variation before the fulfillment center. Second, and because of this, investing in ship-mode specific automation for current buildings through a simple retrofit, and investigating design improvements for future buildings, yields the greatest potential in improving fulfillment center performance in the current state.

1.5 Thesis Overview

The thesis begins with a background on Amazon's inbound supply chain. Each stage in the multi-echelon system is outlined, with a special emphasis on the implications of the different containers, ship-mode choices, and labor requirements. Areas of theorized improvement are introduced in each section, as relevant. This section is outlined chronologically, outlining first the upstream sources (the cross-dock network) and then the primary area of investigation (the Amazon Robotics fulfillment center).

This is followed by an exploration into theoretical supply chain improvements, which ultimately yield no significant opportunities due to the significant constraints within the current system, which will be discussed. As a result, a review of the potential automation solutions and process improvements, highlighting areas of indirect labor expense within the inbound portion of the fulfillment center, is completed.

This overview includes a comprehensive review of a linear optimization model, which was used to calculate theoretical improvements in various different cost buckets of the inbound department, as outlined in section 2.6. There is also a focus on the relevancy of capital investment within a multi-constrained system, and the importance of utilizing automation to alleviate constraints within a multi-echelon distribution environment. A framework, named the Integrated Supply Chain Capital Allocation Framework, is introduced in an effort to discuss this.

Next, a savings calculation methodology is provided for the selected automation solution and the implications of this on a labor-constrained scenario are evaluated. This conversation provides a framework and corresponding case study for thinking about capital investment in a labor-constrained scenario, and a corresponding review is conducted to understand how other industries could benefit from a similar framework.

The thesis then concludes by presenting future areas of improvement for the inbound supply chain and a discussion on the impact of the proposed inbound conveyance solution from a qualitative standpoint. Further projects that could benefit the inbound supply chain at Amazon are also discussed.

2. Amazon.com's Inbound Supply Chain

2.1 Introduction

Increasing customer demand has required Amazon to scale at an extraordinary pace. As a result, Amazon has looked for ways to improve its inbound supply chain in order to improve product availability, reduce cost to the customer, and continue its mission to be the most customer-obsessed corporation on the planet. These attributes define the inbound supply chain's purpose; to receive product accurately and efficiently.

Amazon's inbound supply chain begins with receiving a shipment from a vendor. There are two main classifications for product that flow through the Amazon network: those include "Fulfillment by Amazon" and Amazon Retail. "Fulfillment by Amazon", also known as FBA, is a service where individual sellers and small businesses pay Amazon a fee to utilize Amazon's fulfillment network to sell product directly to customers. Amazon receives, picks, and packs this product and provides customer service for exceptions and returns. Amazon Retail is Amazon-owned inventory that the company sources through a variety of vendors, and usually represents items that have a certain baseline demand. Amazon Retail also sources private label items and has several services that produce make-to-order products, such as books and CDs.

Vendors selling within the Amazon marketplace can have their product received by Amazon through two shipping methods; the "We Pay" method or the "They Pay" method. This point in the supply chain is named the "1^{st} leg transportation arc", where product changes hands from the vendor to Amazon.

The "We Pay" method allows vendors to take advantage of Amazon's low shipping rates, and Amazon prefers to send these orders to the nearest fulfillment center or cross-dock (relative to the origin of the product) to consolidate the product with other inbound shipments. The "They Pay" method allows vendors to utilize their preferred method of shipping, but Amazon determines the destinations of the product, and usually chooses multiple (non-distance limited) destinations to improve inventory placement. In both cases, placement selection is made by an inventory placement algorithm.

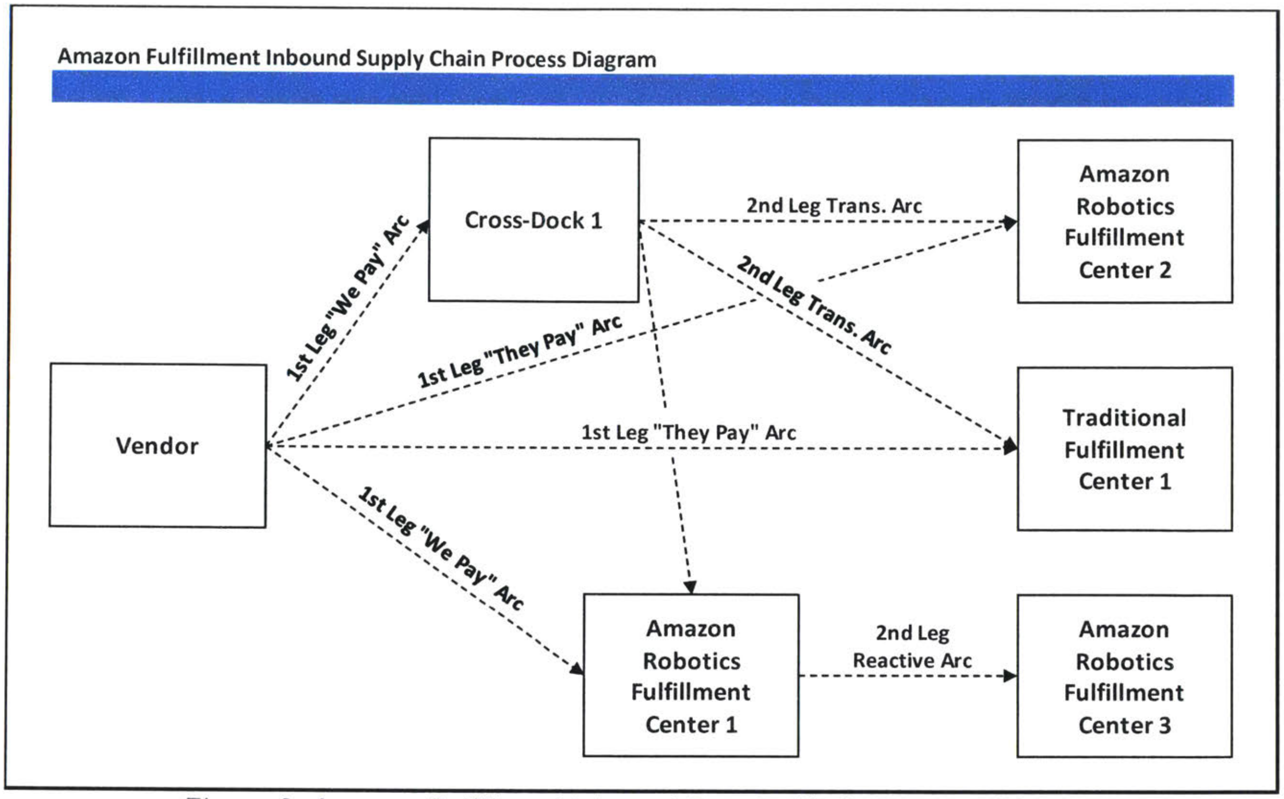


Figure 2: Amazon Fulfillment Inbound Supply Chain Process Diagram

In cases where the first placement was a cross-dock facility, that product will be efficiently received and will receive a second placement from the algorithm, at which point that product will be transshipped, or transported between two internal locations (likely in full truckloads) in the supply chain. In cases where the first placement was a fulfillment center, that product may be stowed or transshipped through a "reactive" arc. In the case of a reactive shipment, the product has already been ordered by a customer and the lowest cost method to ship that product to the final destination and have it arrive on time is to ship the item internally from one fulfillment center to another before sending the product into the outbound portion of the supply chain. In this case, the product is picked and shipped in a full truckload from the original inventory location to a second fulfillment center, where the product is stowed, picked, packed and shipped to its final destination. This only occurs if customer promise can still be met; otherwise, the more expensive shipping option will be chosen.

Various decisions occur at each of the stages of the inbound supply chain that determine the total supply chain cost of the product as it flows throughout the inbound system. Those decisions will be outlined in the following sections of the thesis.

The purpose of section two is to outline the upstream inputs that determine what the fulfillment center receives (in terms of container and ship-mode) and how those inputs are determined. This section will substantiate my claim that making changes to the upstream processes in the current value stream at this time will be difficult due to several system constraints and costs. These constraints and their associated costs ultimately outweigh any of the savings possible within the inbound department of the fulfillment center.

2.2 Distribution System Overview and Concepts: Literary Review

As discussed above, Amazon's inbound supply chain acts as a hub-and-spoke system, with a large amount of vendors ("spokes") providing products to a few strategically located cross-dock facilities ("hubs") who then disaggregate the product to improve product placement at a larger number of fulfillment centers ("spokes") [2].

This system provides several advantages in comparison with other distribution systems, including an overall reduction in costs, centralization of material handling and sortation, and an overall consolidation of logistical flows [3].

The cross-dock is conveniently named for the process that is called "cross-docking", where product is unloaded from inbound freight, sorted, and then re-loaded into outbound freight destined for the next location [4]. The goal of cross-docking is to reduce inventory levels and improve lead times through the "seam-less flow of products through the distribution network" [5].

Given the amount and volume of products that Amazon manages within its supply chain, it can be assumed that Amazon's primary goal of utilizing a hub-and-spokes network can be to reduce the combination of transportation costs within it's entire (end-to-end) supply chain [6]. This has been confirmed by other theses in the past: Olufemi Oti (MIT LGO '13) estimated that utilizing hubs within the inbound supply chain had the potential to reduce costs significantly, for Amazon in particular [7]. Since then, Amazon has opened at least four additional cross-dock facilities (hubs) and Amazon's utilization of more cost-effective full truckload (TL) shipments has increased.

As a result, any project within the fulfillment center must not require any major change to the distribution network (i.e. inputs to the fulfillment center) without considering the impacts to the cross-dock facility and corresponding transportation legs as those costs will likely outweigh the fulfillment center improvements. Proof of this is provided in the following sections.

2.3 Common Terms

A few common terms are particularly relevant to introduce in order to understand the complexity of the inbound supply chain in the following sections (in addition to those referenced in the glossary on page ten):

Amazon Standard Yellow Tote: A yellow tote is a standard container that is utilized within the inbound supply chain in various capacities. This plastic tote is designed for human usage with handles, and is stackable in two positions to allow for either (1) the stacking of full totes for easy and safe movement or (2) the stacking of empty totes for recycling purposes. The liquid usable cube of the yellow tote is ~1.50 ft^3 out of a total envelope of ~2.25 ft^3.

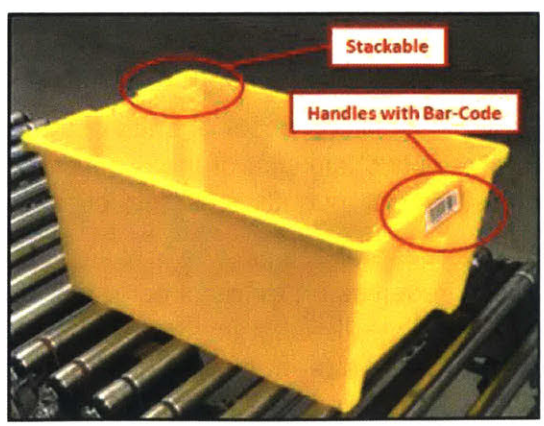


Figure 3: Amazon Yellow Tote

Juice Cart: A juice cart, or cart for short, is a container that is utilized to transport vendor-received product to the stow stations. All product on a juice cart is assumed to be formatted in a way that it is ready to be stowed immediately upon receiving this cart at a stow station. Juice carts are characterized by a large amount of items being easily contained and rolled from point to point.

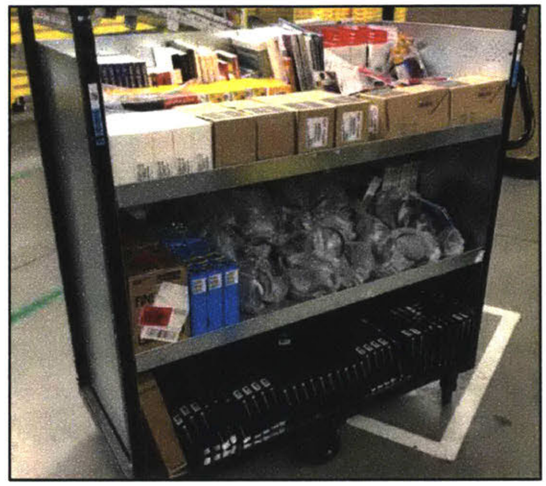


Figure 4: Juice Cart

Case: A case is a cardboard / corrugate-based container that can arrive in a variety of sizes, with either a single ASIN or multiple different ASINs present within the case. Typically, the container is determined by the vendor. The number of products within a case also varies significantly, from cases which arrive with a single product inside of them to cases that arrive with several hundred products in them. Dimensions vary from 4" × 4" × 1" cases to cases that have single dimensions exceeding 30".

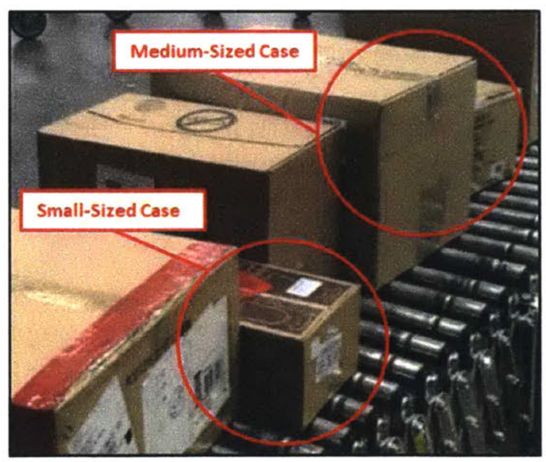


Figure 5: Corrugate Cases

Single-ASIN Pallet: A pallet composed entirely of a single product type. For example, this could be a pallet of Amazon Echo Dots or a pallet of Smartwater® cases. The average pallet size is 48" × 40" × 6", though pallets arriving from vendors can vary in size. The average height of product on a Single-ASIN pallet can also vary significantly. These pallets arrive at the fulfillment center in a transhippment from the cross-dock.

Figure 6: Single-ASIN Pallet

Mixed-ASIN Pallet: A pallet composed of multiple different sized cases with multiple different ASINs represented. These pallets are down stacked and rebuilt at various stages in the supply chain. Down stacking is the process of removing containers from a complete pallet, such that the pallet is made empty. These pallets can be built on standard wooden pallets, or they can be built inside of black plastic pallets. Black plastic pallets are recyclable pallets that are built out of six pieces (four walls, one bottom pallet, and one top) and utilized to allow pallets to be double-stacked within trailers. These pallets may arrive in a transshipment from the cross-dock or be built at the fulfillment center from the contents of a floor load.

Figure 7: Mixed-ASIN Pallet

Tote Pallet: A tote pallet is composed completed of totes. A maximum of 25 totes can be placed safely on a pallet, with most pallets composed of between 20 and 25 totes. In a trailer, a pallet of 20 totes and a pallet of 25 totes can be stacked together to maximize the utilization of space within that pallet location in the trailer. These pallets arrive at the fulfillment center in a transhippment from the cross-dock or another fulfillment center.

Figure 8: Tote Pallet

Floor Load: A full-truck-load ship-mode that contains loose-loaded, floor-stacked cases (and potentially totes). If totes are present, the trailer is designated a hybrid load and otherwise, it is considered to be a fluid load. In a hybrid load, totes can only be stacked to a maximum height of five totes, and then the remaining space above must be covered with cases.

Trailer Types: Fluid Load (Cases) and Hybrid Load (Cases and Totes)

Figure 9: Floor-Loaded Trailer

Palletized Load: A trailer ship-mode that contains palletized cases, totes, and single-ASIN pallets. The maximum number of pallets that can fit onto a trailer is 60 pallets, with 30 pallet locations loaded two pallets high. A trailer can have 60 tote pallets or 60 pallets of cases, or some combination of these, depending on the properties of the origin and the destination. A trailer can also have less than 60 pallets for several reasons; including if a pallet is too tall to allow an additional pallet be stacked or if, operationally, 60 pallets are not available to be loaded at the time of departure of the trailer.

Trailer Types: Palletized Tote Load (50+ Tote Pallets) and Mixed Pallet Load (50+ Total Pallets, <50 Tote Pallets)

Figure 10: Palletized Trailer

Transportation Cost-Per-Unit

Transportation Cost-Per-Unit (Trans CPU) is the primary financial metric used to understand transportation performance. It is represented by total cost of a trailer divided by the number of units in a trailer, which can be aggregated at many different levels. It is usually compared on a lane-by-lane level and at a cross-dock level on a monthly basis. Lane cost is defined as the all-inclusive cost of shipment, including fuel surcharges, mileage surcharges, trucking company fees, and other miscellaneous charges.

$$Trans\ CPU = \frac{\sum Lane\ Cost \times Trailer\ Count}{\sum Unit\ Count}$$

Equation 1: Transportation Cost-Per-Unit

Variable Cost-Per-Unit

Similar to Trans CPU, Variable Cost-Per-Unit (VCPU) is the primary financial metric used to understand labor performance within the Amazon Fulfillment Network. It is represented by the total labor hours divided by the number of units processed multiplied by the labor rate, and can be aggregated at many different levels as well. Included here is all variable labor; this metric does not include the salaried workforce. In addition, this metric is rolled up at various levels, for instance; considering just the outbound department or the entire building (with hourly HR, security, and safety hourly employees considered).

$$VCPU = \frac{\sum Labor\ Hours}{\sum Unit\ Count} \times Labor\ Rate$$

Equation 2: Variable Cost-Per-Unit

Throughput-Per-Labor-Hour

Throughput-Per-Labor-Hour, abbreviated as TPH, is the primary performance metric utilized to compare performance building to building. A higher TPH indicates better performance. TPH can be used to compare performance in many different areas: TPH can describe a value stream (such as the inbound value stream) or for the entire building. For example, a 100 TPH indicates that one labor hour was utilized to process 100 units through a particular area, or through an entire department, depending on the metric. Generally, TPH is rolled up at the department and building level; during building performance reviews, inbound and outbound TPH will be the primary metrics that are focused on.

$$TPH = \frac{\sum Units\ Processed}{\sum Labor\ Hours\ Expended}$$

Equation 3: Throughput-Per-Labor-Hour

2.4 Amazon Cross-Dock Operations

The cross-dock operation is a component of the inbound supply chain that operates as a central hub and receive center for vendors to send their product to in order to have that product efficiently received and distributed to various inventory locations. The following section will outline the value proposition of the cross-dock (IXD) operation, and then outline the three main areas of the operation that have an impact on the fulfillment center: the "Receive" process, the "Sortation" process, and the "End-Of-Line" process.

There are seven primary cross-docks within the Amazon Fulfillment network as of July 2017. Each of these cross-docks are located in different parts of the country, and have various design differences, peak throughput levels, and connections to fulfillment centers. There are several primary goals of the cross-dock network:

1. Efficiently receive high volumes of product from vendors and format product to be easily transferred to a final inventory location.
2. Allocate product to a wide range of fulfillment centers so as to make product closer to the customer.
3. Combine small parcel and less-than-truckload (LTL) shipments to create more efficient full truckloads bound for final fulfillment center destinations.

Within the cross-dock, several processes (potentially) change the aggregate supply chain cost of a product as it travels to a final inventory location at the fulfillment center. These processes are described in the following table:

Amazon Cross-Dock Overview

Cross-Dock Operation	Description
Receive from Vendor	Based on the vendor, product will be either auto-received by the PID or manually-received at a designated station by an associate.
Each Sortation	Based on the ITS (Inventory Transfer Service) network optimization, a determination is made whether to sort the product into totes or leave the cases / pallets intact.
Palletization vs. Floor Loading	At each IXD, a determination is made as to which lanes will be floor loaded and which lanes will receive only palletized freight based on lane velocity and cost.

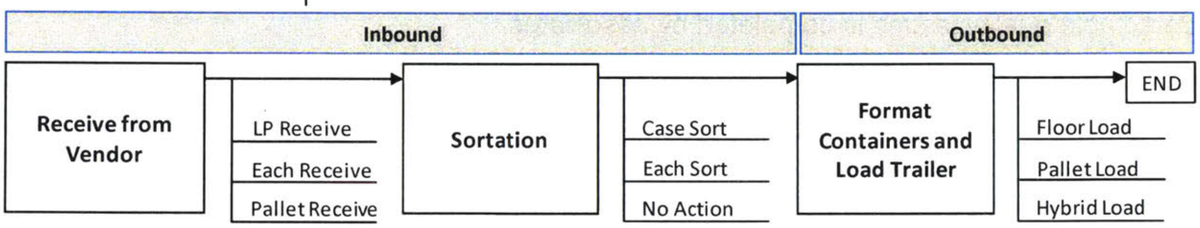


Table 1: Amazon Cross-Dock Overview[1]

Each of these primary processes also has a varying impact on fulfillment center performance, which will be discussed in depth in the "Amazon Robotics Fulfillment Center Operations" section.

2.4.1. Receive from Vendor

At the cross-dock, there are three primary ways that a product can be received. The three primary ways are through the License Plate (LP) Receive process, the Each Receive Process, and the Pallet Receive Process.

The first method (LP Receive) is completed by using a parcel identification unit (PID), where product is automatically received through an automated bar-code scan. Several automated quality checks are performed during this process, including a weight measurement and size measurement for verification purposes, but the main purpose of using a PID is to receive product from trusted vendors efficiently.

The second method (Each Receive) is completed through the manual receive process. In this process, Amazon associates (AA) open the cases of product and manually receive each product back into the case by using a bar-code scanner. In most cases, the ASINs are received back into the container (case) that the product arrived in, but sometimes, the system will trigger the product to be received into a yellow tote. This can happen if the original case is too heavy, or if the associate has difficulty returning all of the items to the case. Naturally, this process increases the cost of fulfilling the product, as associate hours need to be utilized to receive the product into Amazon inventory.

[1] A **Parcel Identification Device (PID)** is a device used to receive product efficiently and automatically. A PID uses a combination of cameras, scanners, and scales to receive the product and check the quality of the shipment against the shipping manifest.

The third method (Pallet Receive) is completed through a different manual receive process. In this process, Amazon associates unload a single-ASIN pallet from the trailer and receive the pallet by verifying the contents of the manifest for the pallet. The major difference here is that the pallet is not immediately broken down, and each ASIN is not scanned, such as in the Each Receive method. Given the efficiency of this process, the variable cost per unit (VCPU) of Pallet Receive is very low.

Ultimately, to reduce cost within this step in the inbound supply chain, the goal would be to LP Receive and Pallet Receive as much of the vendor product as possible. However, both of these processes do nothing to prepare the product to be easily stowed at a fulfillment center outside of acknowledging the existence of the inventory and reducing the manual receive volume within the fulfillment center; in both cases, the cases remain closed and no manual inspection for undetectable quality issues is completed by associates.

2.4.2. Sortation

As soon as a product is either Each Received, LP Received, or Pallet Received, a determination is made, via the Inventory Transfer Service (ITS) network optimization algorithm, on how the product will be routed and where the final inventory locations (fulfillment center) will be. Sortation is classified into two different processes: the Case Sort (Down stack) process or the Each Sort process.

In the Case Sort process, a determination is made on whether or not to break down a single-ASIN pallet (via the down stack process) into individual cases. For a low-velocity single-ASIN pallet, it usually makes sense to break the pallet down in order to ensure that the inventory position for the individual product at each fulfillment center does not exceed the usual required days of cover. In addition, it is likely that low-velocity ASINs have lower inventory positions within the network, so ensuring that the inventory is placed in as many geographic areas as possible, while maintaining a low inventory position network-wide, becomes the requisite goal to limit the amount of inventory space taken up by slow-moving ASINs.

In the Each Sort process, a determination is made on whether or not to sort product into totes. The value proposition here is that by sorting a case of a particular ASIN into totes, that particular ASIN can be spread across multiple different regions, improving the overall fast-track capabilities of an ASIN. One of the core value propositions of Amazon is the Prime service, and this sortation process allows a higher percentage of ASINs to be available via that service by affordably placing product within two days of the customer, improving customer satisfaction and increasing sales.

Ultimately, each of these processes has a cost associated with it, from the associate hours that are used to perform these case or ASIN-level manipulations. These costs, along with other relevant factors (tote recycle cost: the reverse supply chain of sending totes back to the IXD, maximum sortation capacity) is compared to the projected benefit in outbound transportation cost and projected ASIN sales associated with improved product placement and fast-track availability. Then, a decision is made by the ITS algorithm whether or not to perform the sortation function. Sortation is currently only considered for cases with over 18 units in them. Tote generation is primarily based on each sortation, and thus the amount of sortation occurring by origin-destination pair results in the tote percentage that the destination receives.

2.4.3. End-Of-Line Processes by Ship Mode

After the products pass through what constitutes the inbound portion of the cross-dock process, a variable amount of cases and totes with pre-determined destinations arrive to the ship-sorter. The ship-sorter has two primary functions within cross-docks that floor load and palletize outbound product:

1. Route cases and totes to FC-specific production lines in outbound.
2. Determine if the case or tote is to be palletized in an FC-specific production line or floor loaded in an FC-specific, floor-loaded-enabled dock door.

Essentially, the major differentiator within the outbound portion of the cross-dock is the internal decision made on whether to floor load or palletize containers in advance of sending them to the final destination (fulfillment center). The costs of each of these processes vary significantly. Associated rates and processing steps can be seen below in Figure 11:

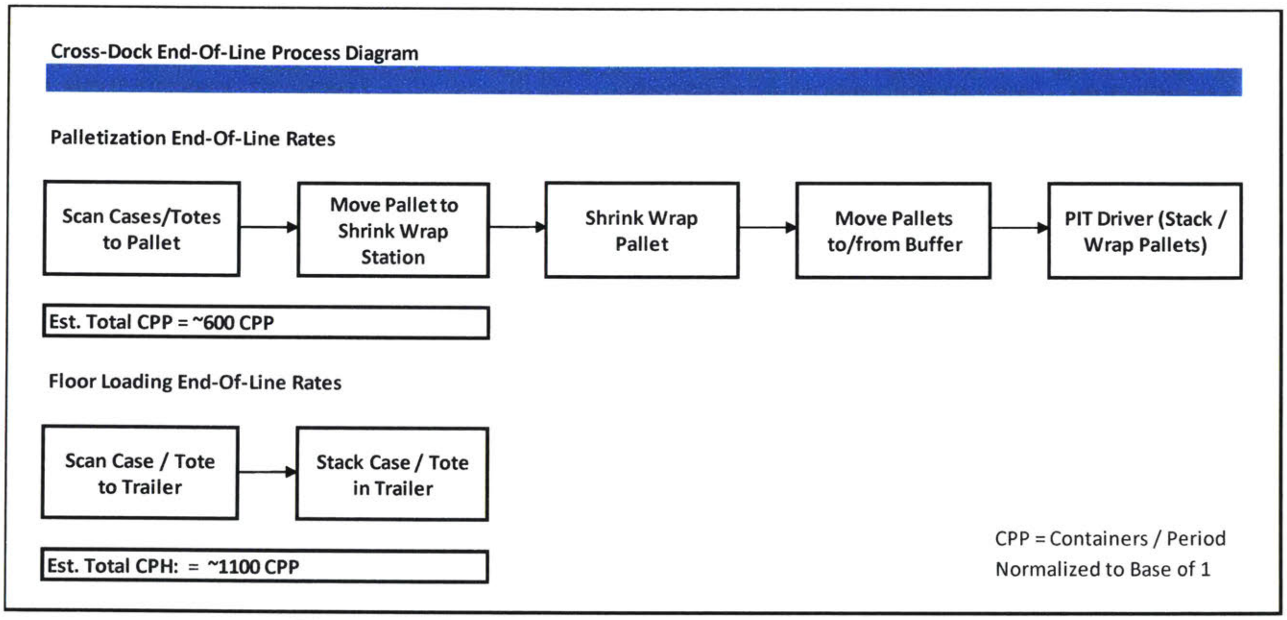


Figure 11: Cross-Dock End-Of-Line Process Diagram

A critical insight is that the end-of-line rate for palletization is nearly half that of floor loaded cases, on average. This makes the initial case to utilize floor loads as much as possible within the cross-dock in order to reduce labor cost and requirements. However, each cross-dock has a different number of floor load enabled dock doors, due in part to the different layouts of each facility and the supporting technology required to allow floor loading (separate fixed conveyance, routing). As a result, there is an upper limit on the number of floor loads that can be sent daily from the cross-docks to the fulfillment centers and the remainder of the containers must be palletized at the traditional end-of-line production lines.[2]

[2] **Floor Load Constraints:** There is an additional constraint as well; the number of potential floor loads is capped based on the tote generation of a cross-dock; only about 40% of the trailer cube within a trailer can be utilized by totes in a floor load.

2.4.4. Conclusions

There are a few main conclusions from the analysis undertaken at the cross-dock facility in relation to the problem we are trying to solve at the fulfillment center in regards to container and trailer ship-mode mix:

1. Floor loaded trailer labor cost associated with loading the trailer is less than the labor cost to create a pallet-loaded trailer.
2. Each cross-dock has a unique, and currently fixed, capacity of floor loads that can be produced each day. This leads to limitations on the locations that receive floor loads, and also leads to significant variability in floor load arrivals at the FCs. The highest velocity (volume) FCs will receive the most floor loads because those floor loads can be built and shipped quickly (allowing high utilization of limited floor load capacity at the IXDs).
3. Floor load production is based on several characteristics within the cross-dock facility, including the automation infrastructure, number of floor load capable doors, production of totes and cases within the receive and sortation processes, and the destination arc volumes. Each marginal floor load capable door that is added recovers less savings, so there is an optimal point at which adding floor loaded doors at a cross-dock no longer makes financial sense.

2.5 Amazon 2nd Leg Transportation

Connecting the cross-dock network and the fulfillment center network is 2nd Leg Transportation. The transportation leg is completely composed of full truck-loads (FTL) of received product that are transshipped from the cross-dock (origin) to the fulfillment center (destination). Following the common theme in the previous sections, there are several different ship-modes possible, as outlined in Figure 8 and 9 on page 20 and 21.

In section 2.4.3, the differences in labor required to load each ship-mode were outlined. However, trailers have three primary factors that determine the variable cost on a trailer-by-trailer basis from a transportation cost-per-unit standpoint, which are included in the logic of whether or not to floor load a particular OD pair. These can be summarized as the following:

1. *Lane Cost:* four-week rolling average cost (for continuous analysis) or the quarterly average cost (for a static analysis) of a trailer for an origin-destination pair, measured in dollars / truckload for a given lane.
2. *Cube Utilization (ASIN Cube)*: the air space (in cubic volume) within a trailer utilized by actual product.
3. *Unit Count:* the number of units per trailer.

2.5.1 Lane Cost

For the purposes of this thesis, lane cost is the quarterly average cost of a trailer for an OD pair and a data set from Q2 of 2017 is used in support of calculations. An increasing amount of the 2nd-leg transportation shipments are completed by Amazon directly, but this lane cost includes both those shipments and shipments completed by trucking vendors. Lane cost is a function of several characteristics of the OD pair. For instance, base cost per mile is influenced by the specific lane that the trailer travels in and the supply / demand attributed to that lane. Fuel surcharges are added based on the cost of fuel at that time and the amount of fuel required for the distance travelled. A good example of this variation in lane cost is that lane cost rises substantially across most OD pairs in the 4th quarter as shipments, primarily from demand increases for the holiday season. These factors result in variability seen on a lane-by-lane and time basis. Lane cost data for Q2 of 2017 shows the following relationship between arc length and total lane cost, with lane-by-lane variability apparent in the following figure:

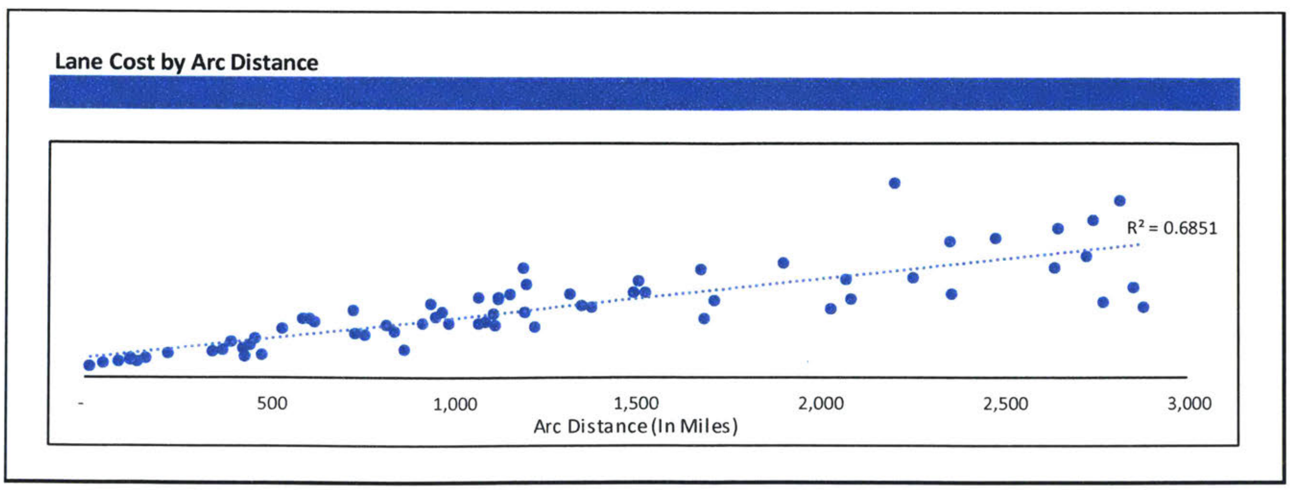


Figure 12: Lane Cost by Arc Distance

Lane cost is a critical piece of this analysis because it is the numerator of transportation CPU. Transportation CPU will be discussed further in section 2.5.3.

2.5.2 Trailer (Cube) Utilization

When utilizing FTLs, the requisite goal is to fill as much of the trailer as possible with actual product, in order to minimize the cost-per-unit of the trailer. Within the warehousing industry, two types of cube utilization are recognized as measurements of the success of "cubing out" trailers [8].

1. *Primary Cube:* The utilization of air space within a trailer by container envelope, assuming the container is utilizing 100% of the space it contains.
2. *Secondary Cube (Liquid Cube or ASIN Cube):* The utilization of air space within a trailer by actual product inside of the containers themselves. For the purposes of this thesis, cube utilization will be discussed in terms of ASIN Cube, unless otherwise noted.

To calculate the ASIN Cube of a trailer, the units (and their corresponding physical volume) within the trailer are aggregated and divided by the total available air space in the trailer. In the case of the standard 53' trailer, the assumed maximum air space available is 3960 ft^3:

$$ASIN\ Cube = \frac{\sum Volume_{ASIN}}{3960\ ft^3}$$

Equation 4: Trailer ASIN Cube Formula

In general, ASIN Cube is a good representation of overall trailer utilization. However, there are also weight limits. In the case of the data set utilized in this study, approximately 1000 trailers surpassed the designated weight threshold for the trailer, resulting in the utilization of that trailer being represented by weight utilization as opposed to ASIN Cube utilization. However, given the small size (<2%) of the weight-utilized data set, ASIN Cube is the primary metric focused on in this study.

In Section 2.2, we outlined the main trailer types that are utilized by the cross-dock facilities to ship product. Each of these four main trailer types has different ASIN Cube properties, related to product containerization and end-of-line choice (palletization vs. floor loading):

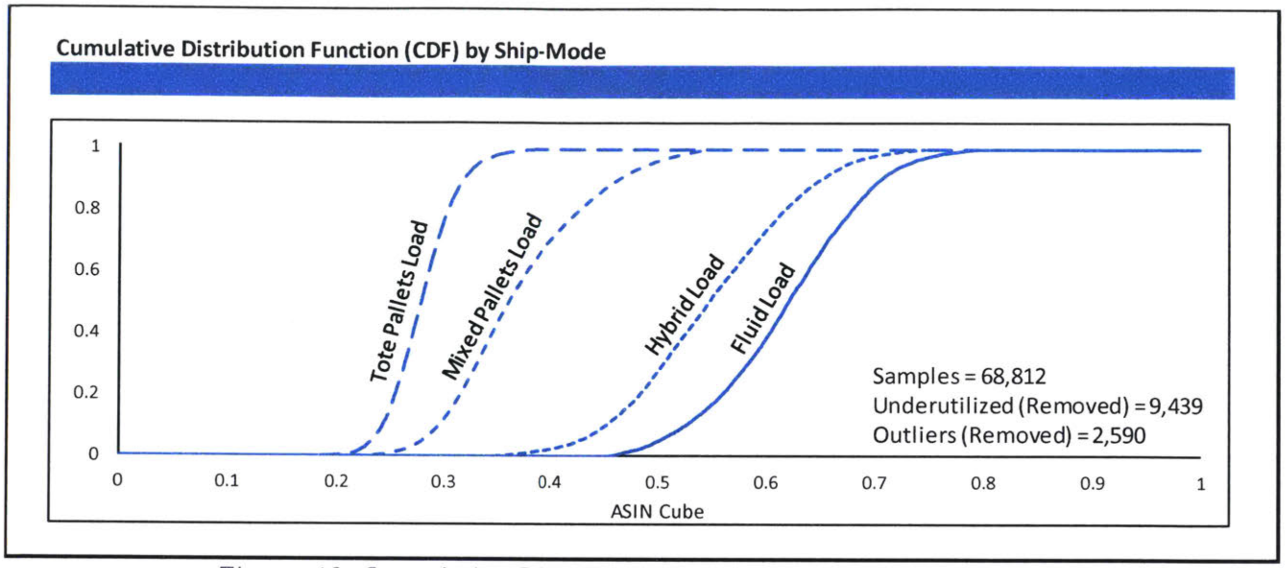


Figure 13: Cumulative Distribution Function (CDF) by Ship-Mode

The data set here was parsed to remove underutilized trailers (< 50 pallets) and outliers (± 2 standard deviations from the mean ASIN Cube by Ship-Mode) to remove abnormalities, and then normalized to remove actual ASIN cube metrics. Some of these abnormalities within the outliers are explained by various issues with ASIN properties, as the rollup of ASIN cube within Amazon's databases is based on known ASIN dimensional data, which is not always perfect.

The primary reason for removing underutilized trailers is to ensure skewness does not affect the ASIN Cube of palletized trailers. Often times, palletized loads can be sent without being completely loaded (and thus underutilized) if a "live load" is called. A live load is a reactive trailer that is called to drain an arc-specific buffer or ship expedited product due to a difference between expected arc volume and actual arc volume. The downside of live loads is that they have additional costs and only remain at the cross-dock for one hour prior to departure. As a result, if the buffer is not adequately sized for the live load such that it can be completely loaded in an hour, an underutilized trailer can be the resulting effect. This effect would skew the data set such that palletized trailer ASIN cube would be lower than should be expected if floor loads and pallet loads are treated the same, operationally.

2.5.3 Unit Count

Unit Count is defined as the number of units expected in an individual trailer. Unit Count is the denominator of transportation CPU.

Unit Count is primarily a function of four variables: (1) ASIN Cube, (2) Origin, (3) Ship-Mode, and (4) Unit Size. Given that ASIN Cube / Ship-Mode and Origin / Unit Size are collinear to some degree, ASIN Cube and Origin (both independent variables) are used as the primary factors that predict Unit Count. While Unit Count is easy to determine (a shipping manifest for each trailer will show the number of units on a trailer in an easily accessible portal), predicting the number of units in a trailer is more difficult without knowing some of the key characteristics identified above. Without being able to predict the number of units per trailer, estimating transportation CPU and the change in transportation CPU associated with changing Ship-Mode, for instance, becomes difficult.

ASIN Cube and Ship-Mode are collinear because of the fact that different containers have different cube utilization characteristics that drastically alter the ASIN Cube of a trailer, and a different container assortment is the defining characteristic of Ship-Mode. Origin and Unit Size are collinear because of the fact that different cross-docks receive a different assortment of ASINs.

Overall, as mentioned, the variables that best predicted unit count were ASIN Cube and Origin. The R^2 value with ASIN Cube and Origin as predictors was 60% with the residuals normally distributed (see Figure 14).

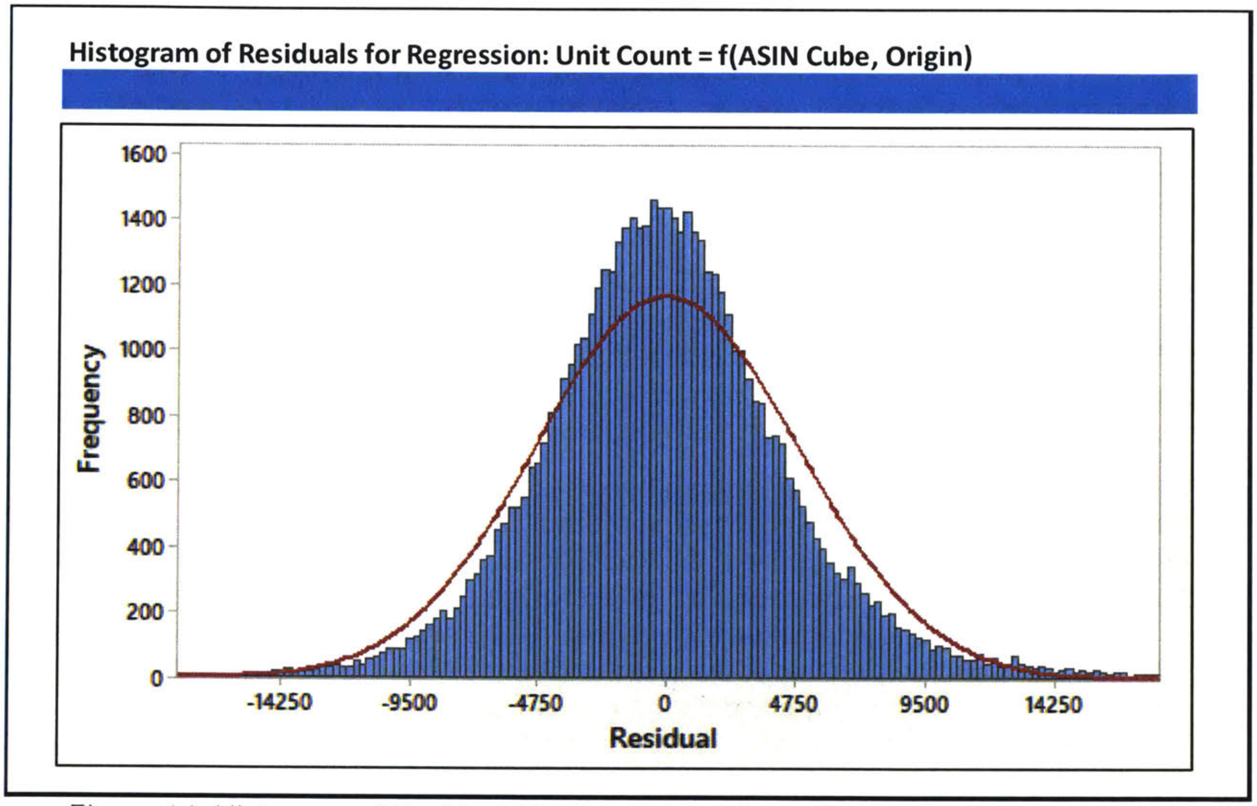


Figure 14: Histogram of Residuals for Regression: Unit Count = f(ASIN Cube, Origin)

2.5.4 Differentiating Characteristics by Ship Mode

As discussed, the primary differences in the ship-mode choice is variable ASIN Cube and unit count. Furthermore, the presence of pallets also changes the cost structure of a transshipment load. This all combines with the labor cost on the cross-dock side to create a differential cost structure based on load type. An overview (with normalized values) of this can be seen below:

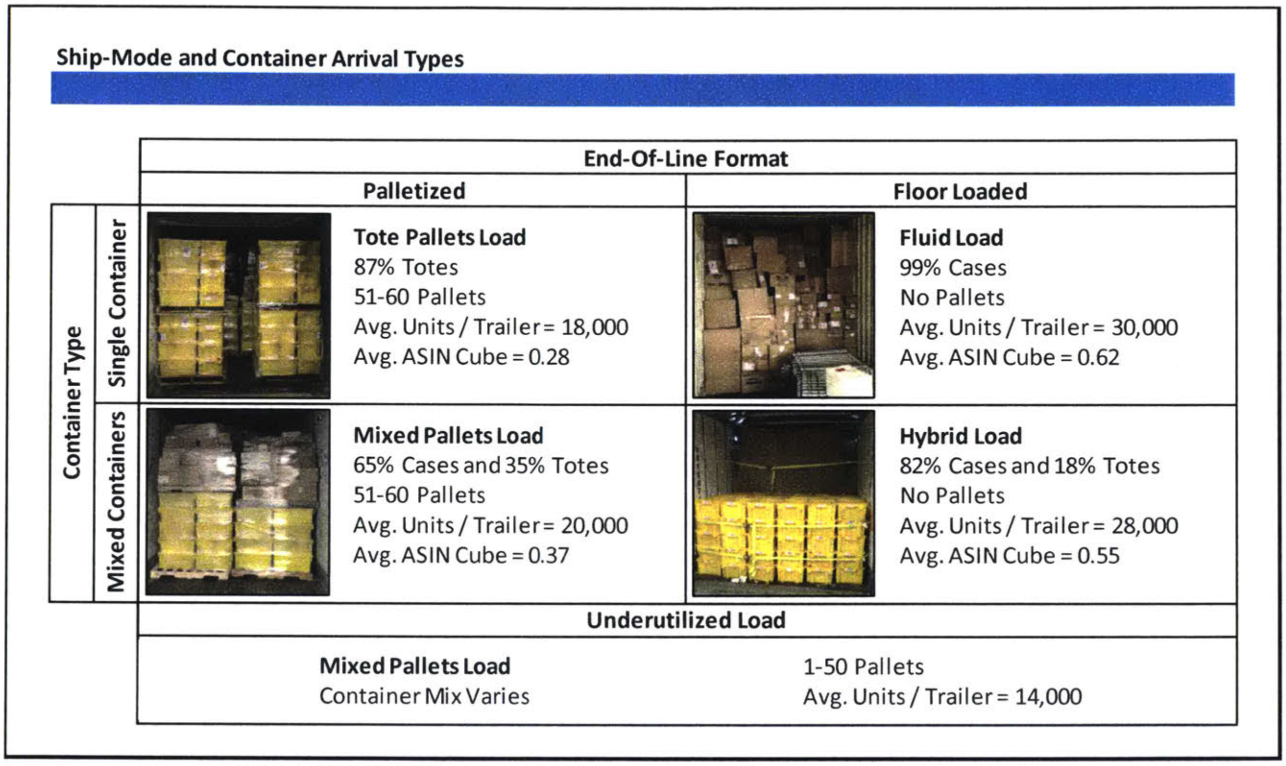


Figure 15: Ship-Mode and Container Arrival Types

2.5.5 Summary of 2nd Leg Transportation

There are a few main conclusions from the analysis focusing on the 2nd Leg Transportation area in relation to the problem we are trying to solve at the fulfillment center in regards to container and trailer ship-mode mix:

1. Floor loaded trailers are utilized nearly twice as well (by ASIN cube) and 50% better (by Unit Count) than pallet loads.
   a. One can think of this in the following way: the more utilized a trailer is with product, the fewer trailers needed to ship that product from the cross-dock to the fulfillment center. For instance, if 3 mixed pallet loads can be converted into 2 hybrid loads, the transportation CPU will be greatly decreased.
   b. Floor loaded trailers have a higher average item size. This can be thought of in the following sense – larger items are sorted less into totes for sizing and utilization reasons; given the disproportionate volume of cases that flows through floor loads vs. palletized loads, often times palletized loads contain smaller items from having more totes.
2. Each type (Ship-Mode) of trailer holds a different mix of containers. In some ways, this restricts the conversion of one trailer type to another; for instance, a tote pallet load

could never be converted to a fluid load without several repercussions in product allocation or labor costs.

2.5.6 Summary of Upstream Sources

A few facts have been established so far:

1. Ship-mode choice and container choice have a high influence on the product cost before the arrival of that product to the fulfillment center. Without divulging specific costs, transportation cost generally outweighs FC or cross-dock processing cost unless the 2nd leg transportation distance is short.
2. Several processes, established to reduce operational cost and burden and improve allocation (such as sortation), constrain free choice of ship-mode and product container.

These facts lend credibility to the thesis hypothesis; that fulfillment center based improvements (via automation utilization or process improvement) are the primary area of improvement that can be utilized to improve the inbound system in the current state.

However, a few ideas in regards to how some of these current constraints can be eased or improved upon will be proposed as the end of thesis.

2.6 Amazon Robotics Fulfillment Center Operations

The fulfillment center operations are the primary focus of this project. The specific portion of the operation focused on in this research project was the inbound process: the process that starts when a trailer arrives to the fulfillment center and ends when the product is stowed in a sellable location within the fulfillment center. The following figure outlines this process:

Amazon Robotics Fulfillment Center Overview

Cross-Dock Operation	Description
Dock Operations	Based on the arrival type of freight, product is unloaded from the trailers, received (if required), formatted, and placed in queues on the dock to be delivered to stow locations.
Material Flow to the Robotic Storage Platform	If the product is to be stowed into the Robotic Storage Platform, a series of touches with the product ultimately arriving at a stow station.
Stow Process	Product will be stowed into a sellable, prime-eligible location with the Robotic Storage Platform at a designated stow station by a stower.

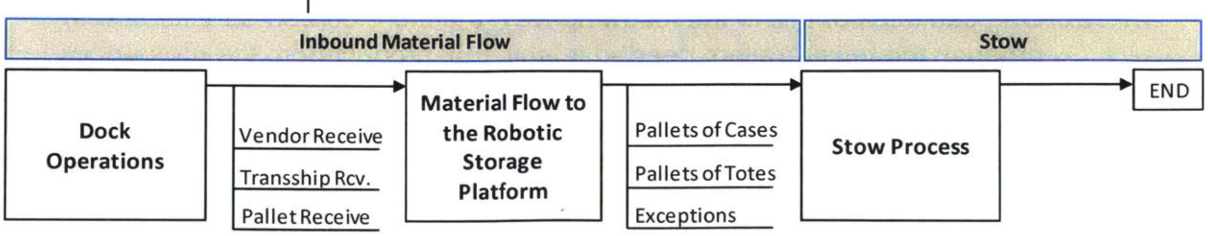


Table 2: Amazon Robotics Fulfillment Center Overview

An Amazon Robotics Sortable fulfillment center has either one or two receive areas, where both product arriving from other Amazon locations and vendors arrives at a variety of dock doors. Product, on pallets, is unloaded at each dock door and moved to a variety of queues on the dock. These pallets are placed in VRCs (large industrial elevators) which convey the pallets to identical floors; each having a large inventory storage field and dozens of stow, pick, and count stations. These pallets of product are placed in buffers once again when they are unloaded from the VRCs, and after a few more buffer placements, are moved to a stow station. Pallets are picked from buffers on a FIFO (first-in first-out) basis to ensure product doesn't sit outside of inventory locations too long and several buffers are utilized such that support associates are responsible for a smaller geographic area, which reduces the complexity of any one associate's decision making and task selection. Once a pallet arrives at a stow station, the containers on the pallet are unloaded onto a sled, which acts as a buffer for containers (as opposed to pallets). The product within these containers is stowed, and the pallets, containers, and any lose corrugate and plastic are collected in large containers and returned to the dock.

A fulfillment center has a designated number of floors, VRCs and stations based on its generation. For instance, earlier generation Amazon Robotics Sortable fulfillment centers had two receive areas, with four floors each, constituting eight total inventory storage fields. Current generation FCs have one receive location with four larger inventory storage fields. There are also several other variants, though the descriptions of the FCs above are largely appropriate for most FCs. When staffing the stow stations, one can think of the methods utilized as a focused "zone" approach. If you need to staff a certain number of stow stations, in order to gain the economies of scale associated with indirect support, one would want several consecutive stow stations staffed. For example, if a stow station was staffed individually away from others, the associate supporting that stower would be underutilized and would often be waiting for that stower to need support or moving to stowers farther away to support them. By placing eight stowers in a "zone", support associates could be fully utilized in supporting those stowers.

A current state analysis of the fulfillment center is shown on the following page. As mentioned in the abstract, approximately 35% of labor hours are focused in different indirect support areas. This project was focused on addressing this, and in particular, the areas listed under the general "Stow Support" bucket in Figure 17 that represent approximately 63% of total indirect hours and 22% of all inbound hours. Some of the jobs represented by the "Stow Support" bucket, including trainers and dock clerks, are not addressed in this study because they are highly variable or required. Other areas, such as "Stow Support", "Fluid Load Support", "Corrugate / Tote Removal", and "Dock Support" all involve manual tasks that are reminiscent of the "Muda" discussed within the Toyota Production System; they combine waiting, transportation, and excess motion in different proportions. Examples of the tasks include moving and stacking pallets, building queues of pallets, or loading containers into other queues. There are several ways to reduce this waste; namely through container standardization or the implementation of automation to reduce moving – however it is important that these improvements can be justified.

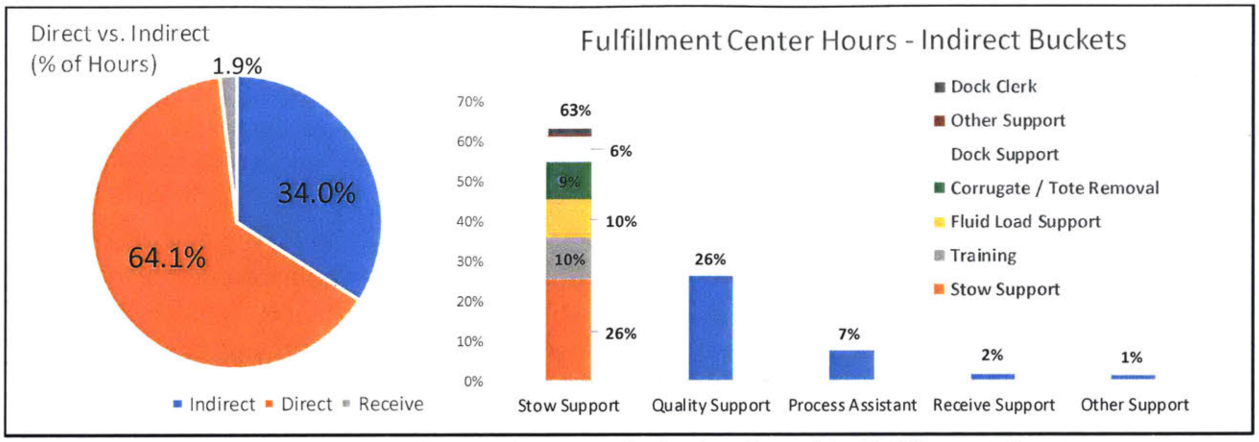


Figure 16: Current State, Amazon Robotics Sortable Fulfillment Center

Following the common themes of prior sections, the amount of support required in the "Stow Support" section results partially from major differences in operational and labor requirements from different arrival types (ship-modes) of freight. These differences in requirements are outlined in the following sections.

2.6.1 Dock Operations

Dock operations encompasses the group of processes that prepare product to be sent to the stow stations. The staffing and operational requirements are highly variable based on several characteristics, including: (1) the volume and ship-mode of arrivals and (2) the amount of product that is received from vendors versus the amount of product that is received from a transshipment. This is primarily because the processes for handling vendor receive product and floor loaded transshipped product is more tedious than the process for handling palletized transshipped product. A summary of these rates is listed in Table 3.

Process	**Vendor Receive**	**Pallet Transshipments**	**Floor-Load Transshipments**
Trailer Unload	Medium	High	Medium
Manual Receive	Low	Not Applicable	Not Applicable
Containerization / Palletization	Medium	Not Applicable	Medium
Queuing	High	High	High
Loading VRC	High	High	High
Overall Rate	**Low Rate**	**High Rate**	**Medium Rate**

Table 3: Rate Comparisons for Different Receive Types

In terms of the volume of arrivals, it makes sense that an increasing amount of transshipment product arriving corresponds to higher labor requirements on the dock, since there is a higher volume to process. However, the relationship between increasing arrivals and labor requirements isn't linear, if the type of arrivals aren't uniform. This is because the amount of floor loaded volume as a percentage of total volume increases as the amount of volume being processed at the fulfillment center increases (see Figure 18 on page 34). Given that floor loaded volume requires more labor to process (the product must be unloaded container by container and palletized), this increase in the percentage of floor loaded volume results in an increase in labor on the dock disproportionate to the increase in volume. I've named this phenomenon the "Increasing Volume Effect", which will be discussed in the following sections.

In addition, the amount of volume that is manually received from vendors corresponds to an increase in labor requirements. This is because each individual item must be received within the inventory management system and placed into a new container indicating the product has been received. As one can imagine, an additional individual product touch can add a significant amount of labor when tens of thousands of products are arriving each day through this method. However, in many cases, it still makes sense to have this process within the fulfillment center (as opposed to solely within the cross-dock), as was discussed in section 2.1. This decision is controlled by the network optimization algorithm that the Supply Chain Optimization Technologies (SCOT) team manages, and as a result, will not be the focus of this investigation.

2.6.2 Decant and Station Constraints

On the inbound dock, there exists the ability to "decant" cases into other containers. The process of decant involves essentially dumping the product from the case into another container. This allows the fulfillment center to centralize the processing tasks of opening cases and disposing of the corrugate. It also allows stowers to see less variation in stow container, which is attributed to higher stow rates; stowing from cases has been shown in several studies to be the slowest container to stow from, though teams differ on how much the effect of stowing from cases is. By increasing stow rates, a fulfillment center can stow the same amount of product using fewer stowers and stow stations. As a result, in a situation when every station in a fulfillment center is used, decant can result in additional fulfillment center inbound capacity and constitute the alleviation of one of two primary constraints that affect the inbound department (the other, labor constraints, which will be discussed in the next section).

The tradeoff is that you have to utilize a substantial amount of labor hours to decant the product itself, and the extra touch within the process also results in additional quality defects. As a result, the tradeoff is that the stow rate increase and reduction in labor hours associated with corrugate management must equal or be higher than the increase in labor hours required from decant and the amount of labor hours used to "clean" extra defects created, known as the decant cost of quality. If mechanical station capacity is an issue, there is further reason to decant as a result of the ability to stow more product at the same number of stations.

2.6.3 The "Increasing Volume Effect" and Labor Constraints

As the amount of volume being received by a fulfillment center increases, the percentage of volume that is floor-loaded volume increases. This causes an increase in the labor requirements on the dock – subject to certain constraints such as network floor-load capacity. This increase can be observed in the following figure:

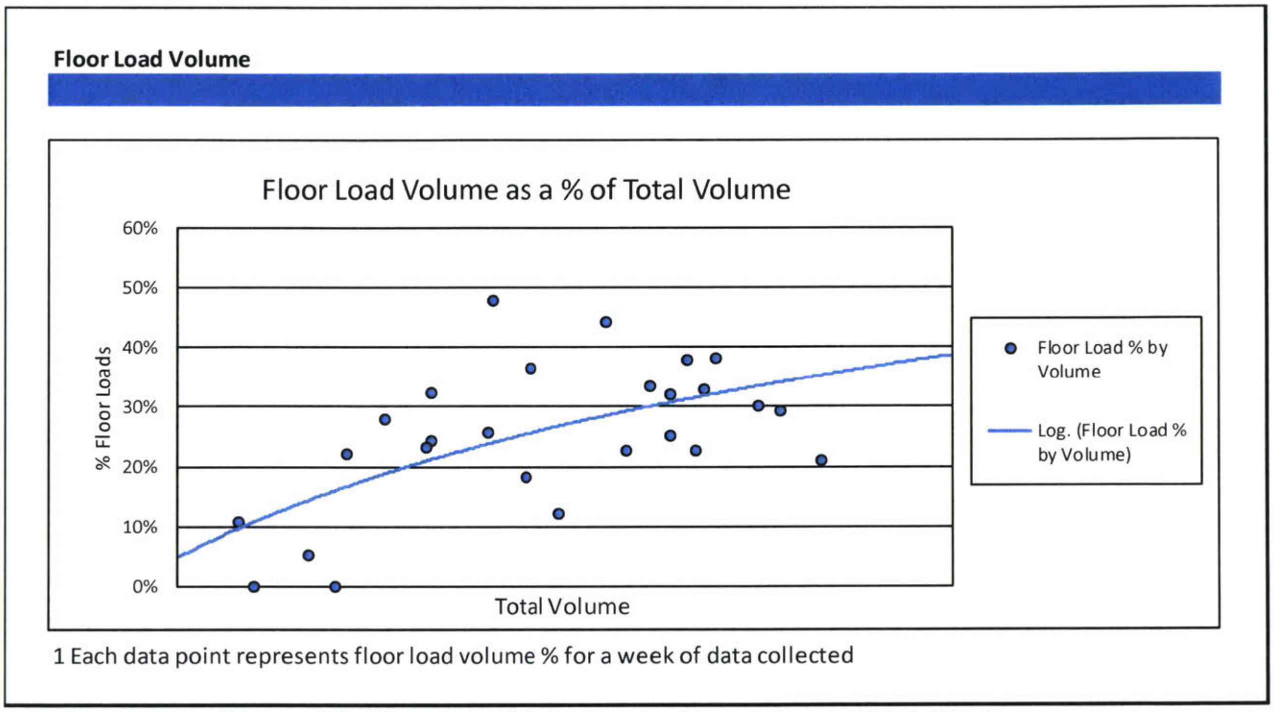


Figure 17: Floor Load Volume at a Percentage of Total Volume

The impacts of this increase are significant in two ways: (1) it causes a disproportionate increase in labor requirements during periods where hiring is occurring and (2) causes the amount of non-value-added labor as a percentage of total labor to increase.

The first issue has the potential to cause a significant problem for the fulfillment center, during a period when labor constraints are present. The Theory of Constraints, introduced in the novel "The Goal" by Eliyahu Goldratt, states that focusing on the limiting factor or weakest ring in the supply chain as a mechanism to improve the system overall [9]. In a fulfillment center, there are two possible constraints that the system can be exposed to: (1) mechanical constraints and (2) labor constraints. In the case of this thesis, and following the Theory of Constraints, it is assumed that if a labor constraint exists, a mechanical constraint cannot also exist at the same time.

Labor constraints arise when there is a local mismatch between the available labor force and the amount of labor required to fulfill a certain volume of product that has been allocated to a fulfillment center. It is generally easy to identify when this is the case; the labor plan will show a deficit in hiring to plan and the load balancing cost-per-unit (CPU) will sharply increase from normal. Load balancing CPU is the cost per unit of outbound customer orders that are shipped from sub-optimal fulfillment centers as a result of a fulfillment center not having the capacity or capability to ship a unit originally allocated (at the lowest cost) to that fulfillment center. A numerical example is shown below:

Load Balancing Cost Calculations

Optimized Scenario

	FC1	FC2
Volume	100,000	50,000
Trans CPU	$ 0.20	$ 0.20
Trans Cost	$ 20,000	$ 10,000

Load Balanced Units

Volume	-	-

Additional Units to Process

Volume	-	-
Trans CPU	$ -	$ -
LB Trans Cost	$ -	$ -

Total Cost	$ 30,000	
LB CPU	$ -	$ -

Standard Load Balancing (LB) Cost

	FC1	FC2
Volume	98,000	50,000
Trans CPU	$ 0.20	$ 0.20
Trans Cost	$ 19,600	$ 10,000

Load Balanced Units

Volume	2,000	-

Additional Units to Process

Volume	-	2,000
Trans CPU	$ -	$ 0.40
LB Trans Cost	$ -	$ 800

Total Cost	$ 30,400	
LB CPU	$ 0.20	$ -

High Load Balancing (LB) Cost

	FC1	FC2
Volume	90,000	50,000
Trans CPU	$ 0.20	$ 0.20
Trans Cost	$ 18,000	$ 10,000

Load Balanced Units

Volume	10,000	-

Additional Units to Process

Volume	-	10,000
Trans CPU	$ -	$ 0.60
LB Trans Cost	$ -	$ 6,000

Total Cost	$ 34,000	
LB CPU	$ 0.40	$ -

Given that Trans CPU was $0.20 from the optimal location, the Trans CPU from a sub-optimal location is higher, at $0.40 in a low load balancing situation and higher, at $0.60 in a high load balancing situation. The load balancing CPU is a calculation of the additional transportation cost of the load balanced units.

Figure 18: Load Balancing Cost-Per-Unit Calculations

As shown, in the high load balancing cost scenario, the 10,000 units that are shipped from FC2 (the sub-optimal ship location) cost $0.60 each to ship to the customer, as opposed to the $0.20 they would each cost to ship from FC1 (the optimal ship location) to the customer. This represents a load balancing cost of $0.40 per unit on 10,000 units and a $4,000 impact to the business. An example of this could be the need to air-ship units from a farther location instead of ground-ship them from the optimal location due to Customer Promise (on-time delivery).

I argue that this cost should be considered when justifying capital expenditure related to continuous improvement at locations that have experienced this constraint, which I have done in the following analysis for inbound conveyance. The insight here is that by reducing the amount of labor hours required in the inbound department through continuous improvement projects, those employees can be shifted to the outbound department to increase the capacity of outbound. By increasing the capacity of the outbound department, less units would need to be load balanced to a sub-optimal location, resulting in a decrease in load-balanced units and load balancing CPU.

At the time of my project, four of the ten fulfillment centers within the scope of my project were experiencing labor constraints during various periods. This is illustrated below:

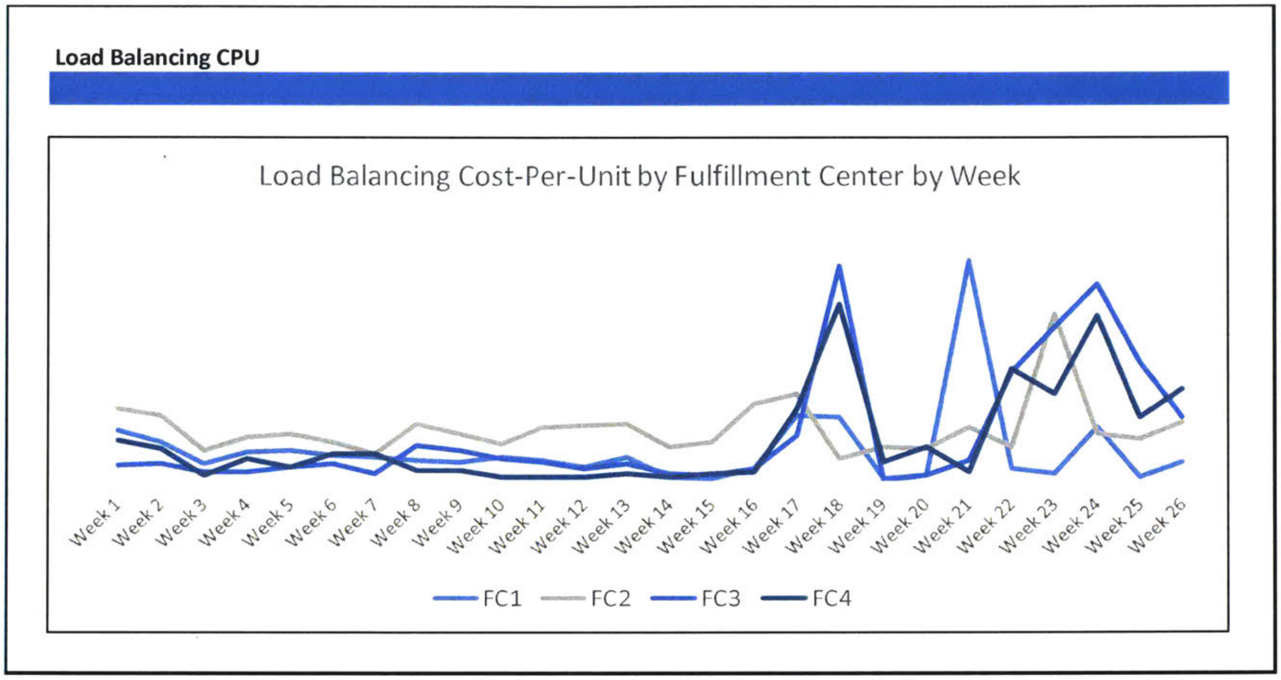


Figure 19: Load Balancing Cost-Per-Unit

Load balancing cost is essentially the increase in outbound transportation cost versus the minimized transportation cost according to the network optimization algorithm. That increase in transportation cost is then spread across the number of units that were displaced from their optimized shipment fulfillment center. Load balancing costs are calculated in the following way:

$$Load\ Balancing\ CPU\ (Daily) = \frac{\sum Realized\ Daily\ Trans\ Cost - \sum Optimized\ Daily\ Trans\ Cost}{\sum Load\ Balanced\ Units}$$

Equation 5: Load Balancing Cost-Per-Unit Equation

As shown, the network and sites considered have relatively low load-balancing costs between weeks 1 and 16. Following, in weeks 17 through 26, these four FCs experienced an increase in volume and a resulting inability to "solve" the labor plan, which results in a significant increase in load balancing costs CPU.

2.6.4 Material Flow to the Robotic Storage Platform

Once transshipment product and vendor product have been checked-in, palletized / containerized (if required), and placed in queues, the next part of the process is to move the containers to the robotic storage platform.

The first step of the process is to determine where product is needed. The robotic storage platform is rectangular, with stowers located at various places on the platform on several different floors. In the current fulfillment center design, the standard process to deliver product to stations is to utilize pallet-jacks and vertical reciprocating conveyors (VRCs, or elevators) to move the pallets or carts into place at a station. Information from the floors flows back to the dock to determine which floors have the smallest queues and those floors are prioritized for replenishment. Waste also flows back down to the dock to be discarded: waste, in this case, includes corrugate from opened containers, extra pallets, general trash, or totes.

The primary goal here is to ensure the queues are balanced in the areas that are being utilized to stow from, and that waste is removed and sent to the dock efficiently. This requires the associates (named water spiders) supporting this process to cycle the VRCs as quickly and as often as possible, while also ensuring they are providing the appropriate information to the dock so that their queues don't run low.

2.6.5 The Stow Process

Additional associates, also called water spiders, move pallets to stow stations and unload the pallets onto sleds, or smaller container-based queues that provide a constant buffer of material to stowers.

These associates are responsible for ensuring these buffers don't get too small; otherwise, the stower may run out of work which results in a lower overall stow rate. Stow rate is one of the primary metrics of performance given the stow process' overall importance in the system.

There are a few main barriers that these water spiders face. If they have to load cases onto the sleds, they also have to open those cases and remove any interior packaging within the case that is seen. On the other hand, if these water spiders are loading totes onto the sleds, all they have to do is place the standard container onto the sled. As a result, the container that associates have to deal with at this point in the process is critical to the takt time that these associates maintain. Based on takt times observed, associates who are handling cases take nearly 3-5 times the amount of time to complete an average case than the associate who handles totes. In addition, associates must then be utilized to remove the corrugate waste and send that waste back down to the dock, whereas totes are utilized by pickers at adjacent stations.

At this point, stowers then take the product and stow this product into the mobile inventory locations. The rate at which stowers are placing this product into the inventory locations is a major determinant of the performance of the inbound department; eliminating barriers for these associates is critical. Slower stow rates results in an increase in the requirements of stations and associates, which are the two constraints most often faced by a fulfillment center. This fixed-station constraint will be discussed further in section 4.1.

Associates are expected to stow at certain rates based on their position in the learning curve. A learning curve at Amazon represents the expectations that Amazon has of its associates in terms of improvement as they become more proficient at certain standardized tasks, such as pick, count, and stow. Associates progress through each level of the learning curve as they spend more hours in path (in stow, in this case). Currently, Amazon utilizes a five-stage learning curve for stow that looks as follows:

Learning Curve Level	Expectation to Mature Rate
Level 1	72%
Level 2	82%
Level 3	88%
Level 4	94%
Level 5	100%

Table 4: Learning Curve Expectations

2.6.6 Comments on Direct and Indirect Associate Roles

Several process paths, including the water spider and stower roles, have been discussed above. In general, these roles fall under two primary characterizations: direct or indirect. Direct process paths are tasks that add value for the customer. In the case of Amazon, these roles include the stower role and the receive role. These roles are critical to the customer experience of ordering an item, since an item cannot be ordered unless it is received and stowed to an inventory location. Indirect process paths are tasks that customers would consider wasteful – in the case of the fulfillment center; this includes waste disposal, pallet movement, truck unloading, problem solving (quality control), and other container management efforts. In general, the goal of the inbound process should be to minimize all of these efforts in order to reduce overall cost that the customer sees. This is the premise upon which this thesis is built.

2.6.7 Summary

Throughout the inbound value chain, there are several variables that affect overall performance, specifically for the indirect functions. The goal of this section was to introduce those various performance determinants, focusing specifically on areas not controllable by the fulfillment center. The chart below summarizes these:

FC Inbound Value Stream Performance

		Positive or Negative Effect on Performance	
		Positive	Negative
Impact on Performance	High	Amount of Totes Amount of Carts Amount of Palletized Totes	Amount of Cases Amount of Floor Loads
	Low	Amount of Floor Loaded Totes	Amount of Pallet Recycling Amount of Plastic / Trash Recycling

Figure 20: FC Inbound Performance Determinant Matrix

In the next section, proposals for improving the current, variable process will be explored. Section 3 will outline the network improvements theorized and section 4 will discuss the internal process improvements that were reviewed.

3. Network-Level Supply Chain Improvement Yields Little Possible Improvements

The first theorized area of improvement within the inbound department of the fulfillment center was to regulate the inputs to the system; i.e. increase tote quantity, reduce floor load quantity, or develop upstream process improvements that would eventually allow these things to occur. There were a few major barriers to each of the theorized improvements. The improvements studied are listed below:

1. *Load Balancing Floor Loads*: Divert floor loads away from sites that have labor constraints. Given that floor loads create additional labor requirements, diverting floor loads to sites with adequate or excess labor supply could help reduce the dock labor requirements at sites.
2. *All-Tote Arrivals:* Utilize a standard container (to start, the Amazon yellow tote) to standardize the stow-from container and reduce fulfillment center waste disposal of corrugate and (potentially) pallets.

As will be discussed in the following sections, neither of these suggestions are options due to the associated costs.

3.1 Load Balancing Floor Loads

Load balancing floor loads is defined as changing floor-loaded lanes based on the optimal site suggested by the Floor Loads Savings Calculator[3] to sub-optimal sites based on real-time constraint management.

In sections 2.3 and 2.4, the primary impacts of floor loads on the overall operations were discussed, showing a decrease in labor cost at the cross dock and an improvement in trailer utilization that improved the overall cost structure of the logistic system. Those primary impacts can be summarized as follows:

1. *Trailer Cube Utilization:* Floor-loading increases cube utilization, which lowers the corresponding transportation CPU and the number of trailers required, system-wide.
2. *Improved Cross-Dock Labor Efficiency:* The rate for loading floor-loaded trailers is higher than the rate for loading pallet trailers.

Figure 21 shows the impact that load balancing floor loads would have on the analyzed network, as a whole. This six-month study, involving 11 fulfillment centers and 6 cross-dock facilities (in total, 65 lanes) was built in the following way:

1. Aggregate six months of floor load data per lane, and calculate the average lane cost experienced over that six-month period.

[3] **Floor Load Savings Calculator:** a tool utilized by the cross-dock team to determine what sites should be prioritized for floor loads. Floor-loaded lanes are prioritized by a combination of increasing lane cost and velocity of product flowing through a particular lane.

2. Calculate the average cube utilization for those floor loads, and calculate the average cube utilization for a similar trailer with pallets instead, holding the tote-to-case ratio constant[4].
3. Based on the percentage decrease in cube utilization seen between floor loads and pallet loads of that lane, calculate the increase in lane cost as a result of having trailers that are under-utilized.
4. Calculate the labor savings at the fulfillment center, as seen on the dock by reducing the manual palletization required.
5. Calculate the labor costs at the cross-dock, as seen on the dock as a result of a reduction in rate seen from moving from floor loading to palletization.
6. Calculate the increase in pallet cost seen as a result of increasing the number of pallets procured per trailer from zero (floor load) to the pallet load requirements.
7. Sum all of these increases and decreased in cost to understand the overall per-load impact to cost seen.

An example of this calculation for a single lane, IXD1 to FC1, for a specific period is shown below with several characteristics normalized:

Parameter	Value	Notes
(1) Fluid Loads	33 Trailers	
(2) Hybrid Loads	24 Trailers	
(3) Total Floor Loads	57 Trailers	Fluid Loads + Hybrid Loads
(4) Totes	14,043 Totes	
(5) Cases	121,554 Cases	
(6) Exp. Containers Per Trailer	2379 Containers	(Totes + Cases) / Trailers
(7) Totes / Case Ratio	0.1	Totes / (Totes + Cases)
(8) Lane Cost	$2,000 / Trailer	
(9) Pallet Load ASIN Cube[5]	0.43	Based on (7), exp. ASIN Cube based on data set
(10) Floor Load ASIN Cube	0.59	Based on (7), exp. ASIN Cube based on data set
(11) Trailer Savings	+ $728	Lane Cost * [(10) – (9)] / (9) – this represents the trailers reduced by using these floor loads
(12) Pallet Savings	+ $140	Avg. Pallet Reduction * Pallet Cost
(13) Labor Savings (Cross-Dock)	+ $490	@ IXD (Hours / Pallet Load – Hours / Floor Load) * Wage Rate
(14) Labor Cost (FC)	- $426	@ FC (Hours / Floor Load – Hours / Pallet Load) * Wage Rate
(15) Savings / Trailer	$932	
(16) Total Savings / Cost	$53,124	Total Savings by Floor Loading Trailers IXD1 – FC1 without considering Load Balancing costs
(17) Labor Constraint Cost (FC)	- $757	Load Balancing CPU * Outbound Throughput / Hour * (Labor Cost (FC) / Wage Rate) @ Max Load Balancing Costs
(18) Savings / Trailer	$175	
(19) Total Savings / Cost	$9,975	Total Savings by Floor Loading Trailers IXD1 – FC1

Table 5: Floor-Load Load Balancing Study

[4] **Tote-to-Case Ratio:** Given that totes and cases have different cube utilization, air space utilization, and ASIN contents on average, holding the tote-to-case ratio constant allows for the impact of pallet loads specifically to be analyzed rather than biasing the data with tote and case impacts.

[5] **Pallet Load ASIN Cube to Fluid Load ASIN Cube:** Utilizing a constant tote-to-case ratio to maintain the same container distribution (since determination of loading with or without pallets doesn't change container selection), steps 9, 10, and 11 calculate the reduction in trailers required on a per floor load basis. For instance, if 1.5 pallet load trailers are needed for 1 floor load trailer, the trailer savings is 0.5 trailers per floor load utilized.

In the case of this lane, despite high load balancing costs[6] (of $757 per trailer), from a cost standpoint, it still makes sense to send floor loads in this lane because the transportation savings are so significant. This drives several key insights:

- Trailer savings drive the utilization of floor loads; given that this analysis utilizes the highest recorded load balancing costs, generally the trailer savings outweighs this.
- Labor savings at the cross-dock and labor costs at the fulfillment center are roughly the same depending on a number of factors, including containers per trailer and the cross-dock in question (since cross-docks have different configurations for loading trailers).
- Pallet savings should not be discounted; while they are the least significant, they have an impact on the suggested ship-mode in marginal cases.

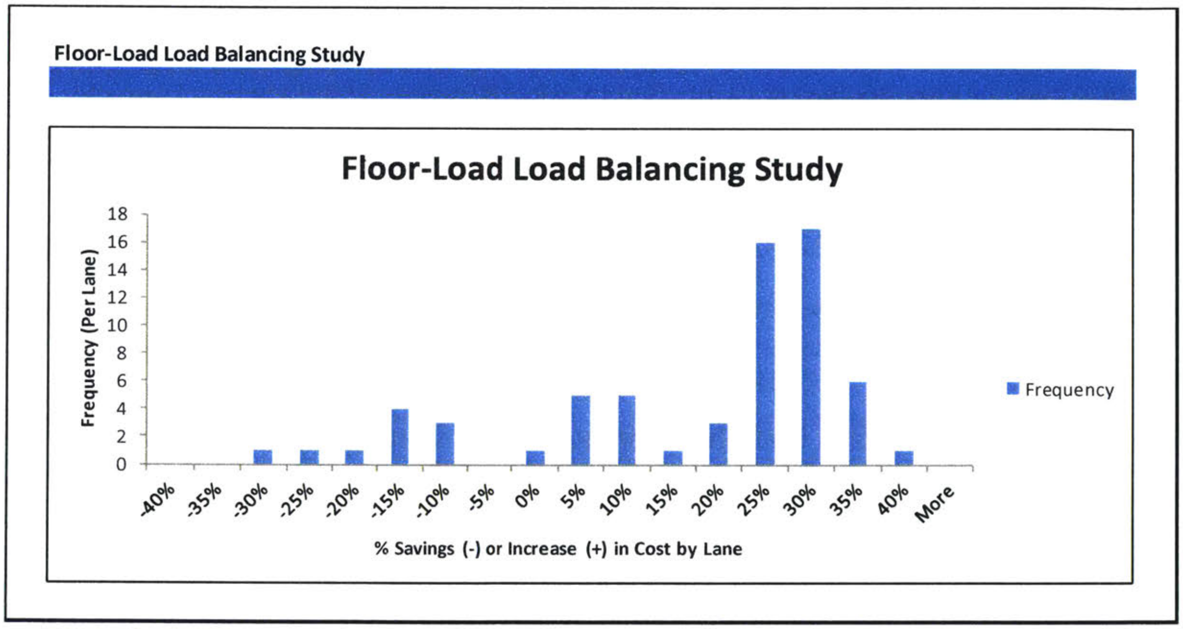


Figure 21: Floor-Load Load Balancing Study

As shown above, only 10 of the 65 lanes experience a decrease in cost as a result of changing from all floor loads to all palletized loads. All 10 of these lanes had a lower cost using palletized loads only because of the load balancing costs saved at shipment destinations (similar to the example provided on the previous page), where labor hours were that much more valuable at the fulfillment center because of labor constraints.

Based on the difficulty of predicting when exactly labor constraints will occur, implementing this strategy proactively would be difficult to undertake with the expectation of significant savings. The chosen strategy was to build awareness within the organization that there are certain lanes and certain times that would benefit from a different strategy based on real-time constraint analysis.

---

[6] **Load balancing costs:** the cost to Amazon that is incurred by shipping a product from a sub-optimal site due to capacity constraints, resulting in additional outbound transportation cost (due to longer distance logistics / upgraded shipping requirement). In this thesis, associate hours that could be used to pick, pack, and ship product are used to unload floor loads instead, and at a labor-capacity-constrained fulfillment center, this results in the aforementioned load balancing costs.

3.2 All-Tote Arrivals

Stow rates and indirect rates would improve if a fulfillment center only received totes, due to a reduction in corrugate management and material handling. As a result, a study was undertaken to determine if this would be beneficial to the inbound network as a whole. This study was completed with a network team, where I provided the FC analysis and stakeholders from within the organization provided the transportation and cross-dock analysis.

In the case of this project, the incremental costs associated with utilizing only totes within the inbound supply chain proved to be too high. There are three primary reasons for this:

1. *Transportation CPU:* Given the utilization difficulties associated with totes that was shown in Figure 14, the additional trailers required and cost incurred to ship all products in totes from the cross-docks would be insurmountable.
2. *Tote Recycle (Reverse Logistics) Cost:* Container recycling cost is a substantial cost that proved insurmountable due to the volume and frequency of trailers that would be required to ensure adequate supply of totes at the cross-dock.

$$Tote\ Recycle\ Cost = \frac{Trailer\ Cost - Pallet\ Savings}{\#\ Totes\ Per\ Trailer}$$

Equation 6: Tote Recycle Cost

3. *Sortation Challenges:* Sortation from cases (arriving from vendors) into totes is currently expensive, slow, and manual. As a result, the increase in labor and facility requirements to sort or receive all product into totes would be prohibitively expensive.

Several other improvements to the network processes that would benefit the fulfillment center were theorized, but overall, similar constraints were seen. As a result, a focus on internal FC process selection and capital improvements were theorized, as seen in the following section.

4. Inbound Process Improvements

The following section will outline the investigation that was undertaken to determine the magnitude of improvements possible within the Amazon Robotics fulfillment center. A linear optimization model was created in an effort to analyze all of the various processing options within the fulfillment center. In particular, the focus of this thesis is on the utilization of the following (1) decant (see sections 2.6.2 and 4.4) as a process improvement and (2) the implementation of a simple inbound conveyance system to move product from the dock to the stow stations. This section highlights the first objective of utilizing a container standardization process (decant).

4.1 Model Introduction

Linear programming has been utilized widely within industry since the Simplex method was invented by George B. Dantzig in 1947, though linear programming was originally applied in finite resource allocation problems by Leonid Kantorovich and Tjalling Koopmans in 1939 [10]. Linear programming, in this thesis, will be applied to calculate the minimum number of hours to stow a given amount of product, summing the indirect and direct hours expended within this effort.

4.2 Model Formulation

The linear optimization model was formulated by completing a variety of time studies for all of the major processes in the inbound department. These time studies were utilized to create various scaling coefficients, some of which are listed in Appendix A. These scaling coefficients (delineated by α) connect the stow hours with indirect hours, and are critical for calculating a total cost (in hours) to move a product from the trailer to an inventory location. Certain constraints were employed to keep the model realistic with the current state process of the inbound department in the fulfillment center. The formulation is outlined in this section and a simplified flow chart of the model can be seen in Appendix B.

Objective: Minimize total labor hours spent across the inbound department to stow all required product into the Robotic Storage Platform (RSP) over a given period.

Variables and Definitions: In this case, four RSP floors with ten zones (see Appendix D for single-floor outline with zones delineated) as per the standard AR building architecture discussed in section 2.6.

$i \in [1,4]$: floor number
$j \in [1,10]$: zone number
$k \in [1,6]$: freight type
$l \in [1,2]$: processing architecture

Equation 7: Variable Definition, Number of Floors, Zones, Freight Types, Process Selection

Definitions for k:

1: Palletized Tote Units	3: Palletized Single-ASIN Units	5: Fluid-Loaded Case Units
2: Palletized Case Units	4: Fluid-Loaded Tote Units	6: Vendor Receive Units

Definitions for l:

1: No Decant
2: Decant

Note for l: Decant only applies to freight types 2 and 5. Types 1,3,4, and 6 will never be decanted in this model. For instance, $c_{1,1,3,2}$ (defined below) will always be 0.

$c_{i,j,k,l}$ = labor hours allocated to stow by floor, zone, and freight type / process architecture
$d_{i,j,k,l}$ = labor hours allocated to indirect functions in support of stow for that floor, zone, and freight type / process architecture
$v_{i,j,k,l}$ = volume of product allocated by floor, zone, and freight type / process architecture

Objective Function: Minimize the total number of hours spent in the various paths within inbound. As discussed previously, the total hours spent in stow (for a given amount of volume) are the indirect hours and direct hours, summed over a given period.

$$MIN \sum_{i=1}^{4} \sum_{j=1}^{10} \sum_{k=1}^{6} \sum_{l=1}^{2} (c_{i,j,k,l} + d_{i,j,k,l})$$

Equation 8: Objective Function, Minimize Total Hours Required

Direct Inputs

$I_k \in [1,6]$: Number of units in each category of freight type. In the case of this analysis, this is the product received during the period being analyzed. The only difference between this value and the v values is that v is an output (on a per zone, floor, type, process basis) while I is an input describing the unit types to be handled during the selected period being analyzed

1: Palletized Tote Units	3: Palletized Single-ASIN Units	5: Fluid-Loaded Case Units
2: Palletized Case Units	4: Fluid-Loaded Tote Units	6: Vendor Receive Units

$SR_C \in [1,3]$: Expected average stow rates per container type based on employees in each learning curve. For instance, if half of the associates stowing on a shift are in level 1 of the learning curve and half of the associates stowing on a shift have completed the learning curve, the expected average stow rate would be 86% of mature stow rates (per Table 4 on page 37). Without knowledge of each rate out of each container, a summary rate expected for the day can be chosen and assigned to all three.

1: Stow Rate out of Cart	2: Stow Rate out of Tote	3: Stow Rate out of Case

$B_i \in [1,4]$: Desired floor balance, in percentage of stow hours placed on each floor. Generally, the specified floor balance will be 25% of hours utilized on each of the four floors to keep the floors balanced in terms of inventory levels and station utilization. However, there are times when floors become unbalanced in terms of inventory levels or space availability, and thus, one could use this input to change the stow percentage per floor. For instance, if the 4th floor had more space available, B_4 could be set to 30% to achieve a higher inflow of inventory into that area. Overall, $\sum B$ will equal 100%.

1: % of hours, floor 1	2: % of hours, floor 2	3: % of hours, floor 3	4: % of hours, floor 4

h: Period length, in hours (length of shift, day, or week i.e. processing horizon). This determines the amount of time that all of the input units are stowed in. For instance, if h equals 18 hours, the analysis is being conducted for two continuous shifts. This input is critical, as it determines how much product must be stowed per hour, and therefore, how many employees are needed to stow that product in a given hour.

$Z_{i,j} \in [0,1]$: Availability of zone (not used for count, pick functions). For instance, if all of the stations in a zone are being used for count, or are undergoing maintenance, this binary available allows the user to disable the availability of those stations to ensure that the solver does not allocate those stations to stowers.

$S_{i,j}$: Number of stations in zone not used for count, pick functions. This varies based on the size of the zone and based on the pre-determined amount of stations available for stow. This is useful if a single station is offline for instance, so that a zone is not allocated a stower that cannot be placed at a specific station.

$U_{i,j}$: Percentage of time during shift stowers actually spend stowing, tracked as a metric called fast starts within Amazon (removes time at standups, walking to and from station during shift). For instance, a stower may be clocked in for a shift that is ten hours but may only spend 9.5 at the station due to the start-of-shift meeting with their manager or their transportation from the breakroom to the station, which results in a percentage of time utilized in stow at 95%. Depending on the input metrics being used, this value may either not be utilized (and thus set to a value of 1) or utilized with a decimal value between 0 and 1.

Decision Variables

Allocated hours of stowing to a specific zone in the robotic storage platform (RSP)

$$c_{i,j,k,l} \geq 0, i \in [1,4], j \in [1,10], k \in [1,6], k \in [1,2]$$

Equation 9: Decision Variable. Allocation of Hours Per Zone

Note: $d_{i,j,k,l}$ (indirect hours) and $v_{i,j,k,l}$ (stow volume) are directly calculated from $c_{i,j,k,l}$ (using the scaling coefficient, α) as each path to a floor and zone has an implied indirect "cost" in hours associated with the touches prior to the product being stowed. There is also an implied volume based on the number of stow hours and the stow rate.

This is a critical assumption for calculating indirect cost. By using a combination of time studies and travel distance measurements, I calculated the average cost (in hours, converted from seconds) of each product touch and placed those costs in a matrix. For each path from the dock to a stow zone on a per-container basis, I summed each of these costs to calculate an indirect zone cost. For instance, a tote has a lower cost than a case, because a case must be cut open by an indirect associate, which is another touch that has a cost of 10 seconds. This indirect cost calculation is depicted in Figure 22:

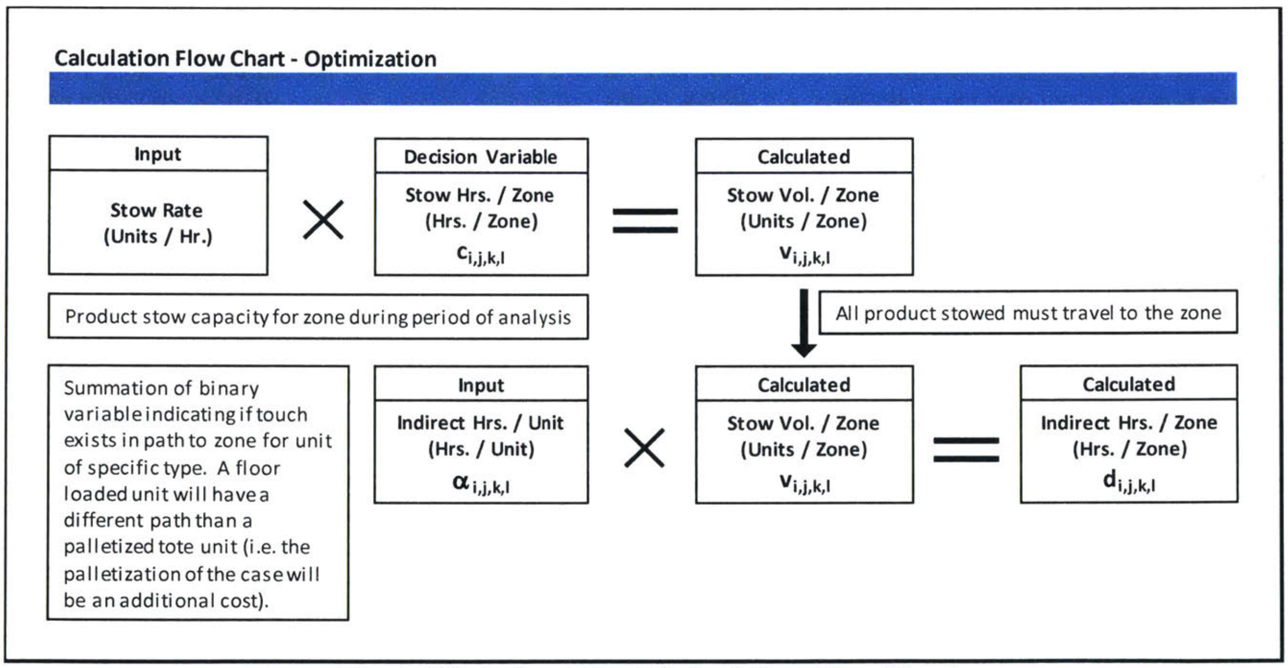


Figure 22: Indirect Cost Calculation Flow Chart

The output of this analysis were scaling coefficients α corresponding to each floor, zone, freight type, processing type combination. For instance, $\alpha_{4,10,5,2}$ (fluid-loaded cases decanted and stowed on the 4th floor in zone 10) will be a higher relative value because of the number of touches required to process that product, including the slower trailer unload process, the decant hours required, and the distance from the dock to floor four and zone 10 (which are the farther from the dock). On the other hand, $\alpha_{1,4,1,1}$ (palletized totes stowed on the 1st floor in zone 4) will have a very small value because the processing path for this product is simple. Thus, the optimization algorithm will favor these paths generally to minimize d.

An example of the α calculation, with normalized task times, is shown in the following table, with the example being for a unit within a floor-loaded case being stowed in zone six on the 2nd floor (coefficient representation is $\alpha_{2,6,5,1}$):

Action	Time / Unit	Data Source
Place case from trailer onto conveyor	6 seconds / case = 0.6 seconds	Time Studies
Place pallet near conveyor for floor load pallet build	20 seconds / pallet = 0.2 seconds	Time Studies
Place case from conveyor onto pallet	3 seconds / case = 0.3 seconds	Time Studies
Wrap pallet with plastic wrap	30 seconds / pallet = 0.3 seconds	Time Studies
Move pallet from conveyor to buffer	20 seconds / pallet = 0.2 seconds	Time Studies
Move pallet from buffer to VRC buffer	30 seconds / pallet = 0.3 seconds	Time Studies
Open VRC (2x)	20 seconds / four pallets = 0.05 seconds	Time Studies
Load / unload pallet into / from VRC (2x)	20 seconds / pallet = 0.2 seconds	Time Studies
Close VRC (2x)	20 seconds / four pallets = 0.05 seconds	Time Studies
Move pallet from VRC to buffer	10 seconds / pallet = 0.1 seconds	Time Studies
Move pallet from VRC buffer to zone buffer	90 seconds / pallet = 0.9 seconds	Time Studies
Move pallet from zone buffer to stow station	30 seconds / pallet = 0.3 seconds	Time Studies
Remove plastic wrap from pallet	10 seconds / pallet = 0.1 seconds	Time Studies
Load case onto stow station sled	5 seconds / case = 0.5 seconds	Time Studies
Cut open case to expose product	5 seconds / case = 0.5 seconds	Time Studies
Quality Support	x	Avg. Calculated Labor Tracking Rate
Process Assistant	y	Avg. Calculated Labor Tracking Rate
Other Support (Corrugate management, pallet management)	2.0 seconds	Estimated Rates and Time Studies
Total Indirect Labor	**6.6 + x + y seconds**	
Stow Labor	z	Avg. Calculated Labor Tracking Rate
Total Direct Labor	**z seconds**	
α (scaling coefficient)	**6.6 + x + y seconds**	

Table 6: Dock-to-Stow Path Calculation Example

As shown above, depending on the path the product takes to a stow station, the cost, in seconds, to stow that product could vary based on ship-mode or container. For instance, a tote doesn't require an associate to open it, which would remove the "Cut Open Case to Expose Product" step and a palletized load wouldn't require the first five steps. Similarly, the location of the stow station would impact the moving cost of a pallet and would add (or subtract) incremental time from the Total Indirect Labor bucket.

Note on Container Properties: In this case, 10 units per case and 100 units per pallet were used. In each calculation using the model, these properties are calculated based on the freight data inputs.

Note on Variability: on any given shift, the costs of each product touch vary based on the associate completing the task. In the case of this analysis, since there are thousands of product touches in any given period analyzed, I assume that the average cost seen during a period conforms to the average cost I calculated during my time studies. On some shifts, based on many factors, there could be significant variability seen here which could mean that additional associates or fewer associates are required to complete the same amount of indirect work.

Primary Constraints

The primary constraints of this model focus on ensuring that the freight inputs are completed subject to actual building processing steps (i.e. floor-loaded totes are processed as floor-loaded totes).

1. Stow all of the input product in the given period, and ensure that product distribution stowed matches the distribution input, by freight type. This ensures that all product is stowed, as the optimization would minimize cost by stowing no product in a period without this constraint.

$$\sum_{i=1}^{4}\sum_{j=1}^{10}\sum_{l=1}^{2} v_{i,j,k,l} \geq I_k$$

Equation 10: Stow Completion Constraint

2. Hours of stowing a specific type of product (by container and process type) is at least the number of hours required to stow all of the volume of that type of product. This ensures that associate efficiency and container stow rates are connected appropriately to volume by container type.

$$\sum_{i=1}^{4}\sum_{j=1}^{10}\sum_{l=1}^{2} c_{i,j,k,l} \geq \frac{\sum_{i=1}^{4}\sum_{j=1}^{10}\sum_{l=1}^{2} v_{i,j,k,l}}{U_{i,j} \times SR_C} \; (with\, SR_C\ corresponding\ to\ the\ container\ type)$$

Equation 11: Stow by Container Completion Constraint

3. Hours of stowing in a zone does not exceed available capacity of that zone.

$$\sum_{k=1}^{8} c_{i,j,k,l} \leq Z_{i,j} \times S_{i,j} \times h$$

Equation 12: Floor Stow Hour Capacity Constraint

Secondary Constraints

The secondary constraints allow the user of the model to set conditions that imitate building behaviors based on what is actually seen. In some sense, these constraints are soft constraints; they only keep the model honest with actual performance and are not technically required to satisfy the optimization. However, they are essentially required in order to ensure the outputs of the model are consistent with fulfillment center operations.

4. Hours of stowing on a floor are distributed according to desired floor balance.

$$\sum_{j=1}^{10}\sum_{k=1}^{8}\sum_{l=1}^{2} c_{i,j,k,l} = B_i C$$

Equation 13: Stow Hour by Floor Allocation Constraint

5. Containers are allocated evenly across the floors in terms of stow hours committed (soft constraint) and as related to the floor balance. The purpose here is to ensure that one floor does not receive all totes, as the size and type of product in totes and cases is different, and this could cause a floor mismatch in terms of inventory.

$$\sum_{j=1}^{10}\sum_{l=1}^{2} c_{i,j,k,l} = 0.25 \sum_{i=1}^{4}\sum_{j=1}^{10}\sum_{l=1}^{2} c_{i,j,k,l}$$

Equation 14: Container Allocation Constraint

4.3 Programming

An Excel-based solver, named OpenSolver, is employed by the base optimization model that was built for this study. OpenSolver utilizes an open-source linear and mixed-integer programming solver, named COIN-OR Branch-and-Cut (CBC), which includes a linearity check to ensure that a global minimum is achieved in each subsequent trial. This solver was selected because it is globally recognized across multiple platforms, including JuMP and PuLP (a Python-based integration of CBC) [11]. This is important in the event that the base model was transitioned to another platform such as Python. It is also not considered to be a proprietary platform, so the utilization of the solver would not infringe upon any intellectual property rights upon writing this thesis.

4.4 Utilization of Model for Analyses

This model allows the user to analyze different scenarios that the inbound department will face, including daily scenarios involving slightly different freight and container mix, or boundary conditions, such as a daylong period with only floor-loaded case arrivals or palletized-tote arrivals. The model also allows the user to change product touches and the path of product to stow stations by selecting zones that are fed by inbound conveyance (which removes palletization of floor loads and pallet movements). Furthermore, it allows the user to implement the decant process, and see how indirect cost and station utilization change as the result of moving all case product to totes and carts.

Much of the analysis of this thesis is completed utilizing this model. For instance, to aid in calculating the savings for inbound conveyance, the model was run in its base mode (without inbound conveyance) and then compared to a model run where inbound conveyance was supporting several stow zones. With inbound conveyance, the indirect costs (in labor hours) for moving product to the stow stations fell, and the reduction calculated helped inform savings possibilities. The model allows for this ability by recalculating specific α scaling coefficients associated with the improvement project completed. By having a lower α value, $c_{i,j,k,l}$ will be lower for given floor, zone, freight type, and processing type paths. When running the model with these lower α values and comparing that to the original model calculations, the optimal solution will be different based on routing changes. In some sense, this is a classical allocation model, minimizing a value by routing product as efficiently as possible.

An example of the output is shown below in Table 7. For this model run, the processing quantity for the weeklong period was 3.5 million units with a representative distribution of the different freight types, representative stow rates, and no decanting.

Output Variables	Current-State – No Inbound Conveyance	Future-State – Inbound Conveyance	Difference
Output Hours	25,641	24,856	- 785
Indirect Hours	7,973	7,125	- 848
Stow Support Hours	1,752	1,533	- 219
Dock Support Hours	1,249	596	- 643
Direct Hours	17,668	17,730	+ 62
Stowers	121	121	~ 0
Stow Support Associates	19.2	17.9	- 1.3
Dock Associates	7.7	3.3	- 4.4

Table 7: Model Output Example

This example above, in the current state, shows a 68% direct allocation and 32% indirect allocation, with 59% of the indirect hours utilized in the larger "Stow Support" bucket, when comparing Table 7 to Figure 16. These allocations are within a margin of error that is appropriate for this study, given that Figure 16 is a snapshot at a single fulfillment center over a short period that had more totes and pallets than floor loads (which contributed to the positive direct-to-indirect ratio. A validation attempt of the model outputs is discussed in the following section.

4.5 Model Validation

In order to determine if this model is an accurate estimator of inbound performance in the current state, a validation study was completed. The validation involved aggregating various data sources that constituted the inputs and outputs of the model, and then running the inputs through the optimization model to determine if the theoretical outputs matched the actual outputs in comparison to past performance. The inputs were aggregated from several sources; the PPR (Process-Path-Rollup) database provided overall volume, rate, and hours-worked information by employee function. A trailer manifest created to a SQL query provided trailer arrivals and data by container and ship-mode type. RoboScout, a database that collects data on the Amazon Robotics inventory management system, provided rough information on floor balance and station utilization. Two performance metrics were explored in comparing the model output to actual performance: Inbound Throughput-Per-Labor-Hour (TPH) and Indirect Labor Hours – both of which are provided directly by PPR and computed directly by my optimization.

Both Inbound TPH and Indirect Labor Hours were selected because both are being estimated by the optimization model, and because both data points are easily available in various time increments within Amazon's labor tracking database. The verification study was completed for two months of data (60 days) on a daily basis for which data was easily available and for which I was able to witness any abnormalities in performance and attribute them directly to observable causes. As such, I ran the model 60 separate times to review each day separately.

The following histograms show the 60 data points and the percentage difference in the outcome in comparison to the data recorded within Amazon's labor tracking systems. The percentage differences were calculated using a simple percentage change equation:

$$Percentage\ Difference = \frac{Modeled\ Output - Actual\ Output}{Actual\ Output} \times 100$$

Equation 15: Percentage Difference Calculations

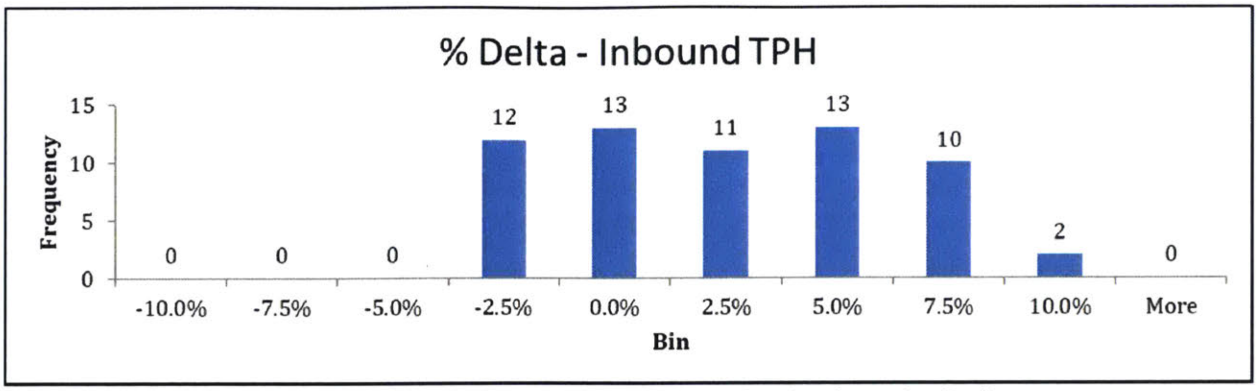


Figure 23: Inbound TPH Verification Study, Actuals, Daily

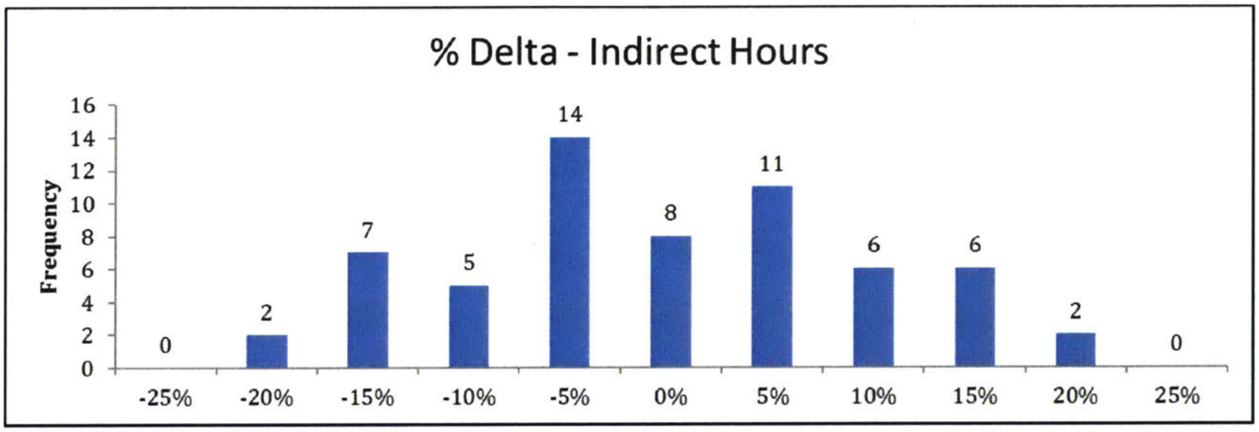


Figure 24: Indirect Hours Verification Study, Actuals, Daily

Translating these figures into confidence intervals, the following confidence intervals for model estimation apply:

95% Confidence Interval of TPH: + 3.27% ± 1.22% | 95% Confidence Interval of Indirect Hours: - 8.94% ± 3.12%

Equation 16: Model Performance Confidence Intervals

Translating this TPH data to weekly data, the following chart shows that TPH is slightly overestimated, as shown by the confidence intervals above.

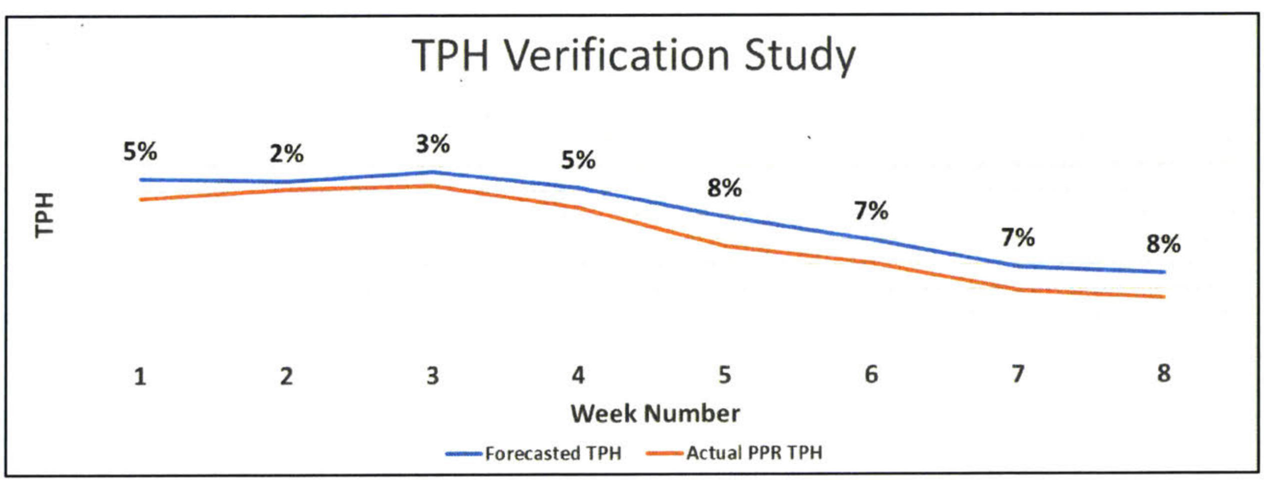


Figure 25: TPH Verification Study Time Series, Actuals, Weekly

In general, overestimation in the case of utilizing an optimization model to predict process performance is expected. Management is constantly trying to optimize the performance of their process, and this model removes (or doesn't account for) any inefficiencies that management sees on a daily basis. The inefficiencies that management sees include the following:

- *Labor moves*: causing under-utilized labor hours.
- *Re-merching and de-merching*: the names for the processes utilizing labor hours to rebuild and empty queues from zones or stations.
- *Fast start inefficiencies*: a degradation in the transition efficiency of associates from break periods to active periods of work.
- *End-of-shift attrition*: employees utilizing personal time to leave work early causes inefficient utilization of remaining indirect employees and labor shifts.

The conclusion from this model verification study is that utilizing the model to examine tradeoffs and predict overall performance should be satisfactory, as long as the inputs are validated and the outputs are taken to be estimations of actual performance, within a certain error interval as described by the confidence intervals. This is with the caveat that recognizing the overestimation of performance predicted by the model should be considered if the model is being used for prediction purposes.

4.6 Inbound Process Selection

Currently, a minority of the fulfillment centers utilizes the decant process to reduce waste flow to the robotic storage platform and to improve stow rates (and thus reducing station requirements). As discussed in section 2.6.2, the decant process employs associates to move product from corrugate cases to totes or carts; the waste produced from cases is easily discarded at this point by corrugate takeaways and products are, at this point, placed in a stower-friendly container. There is a cost of quality to this process as well; containers with higher item densities (such as carts) require additional counts to ensure that inventory accuracy is maintained. Essentially, decant improves stow rates and reduces indirect hours spent removing corrugate waste, but increases the number of hours spent cleaning quality errors and costs hours itself by having associates move product from one container to another.

Stow rate estimation by container has not been finalized to date, with several internal studies indicating a varying degree of improvement by changing the stow container from cases to totes. Similarly, the decant rate varies. As a result, decant has two unknowns (the improvement in stow rates and the decant rate) that can be varied at the same time to produce a sensitivity analysis of different possible results.

Utilizing the linear optimization model, I modeled the impact of decant on station utilization, which can be seen in Appendix C. I completed this analysis by forcing the model to make certain decisions. By forcing the model to decant certain amounts of volume and then comparing the outputs from those model runs to the output of the unconstrained, optimal model runs, I could determine the difference in variable cost, station allocations, and staffing requirements. I was able to force the model to decant certain volumes by setting artificially high costs on non-decant paths, such that all volume going through those paths would be transitioned to decant volume.

As shown, the projected increase in stow rates based on decanting into totes or cases is between 6-25% based on several studies (completed prior to my research) which attempted to answer the difference in stow rates by stowing from cases versus a standard container (totes or carts). Given the uncertainty in stow rate improvements from stowing from a standard container over cases, and given the achievable decant rate, the analysis recommended not decanting.

If stowing capacity is not a binding constraint, then the cost of decanting outweighs any benefits from decanting. However, if decanting allows for higher stow capacity due to stow rate increase, there could be a sufficient benefit from decanting that outweighs the cost of decanting. The stowing capacity constraint idea and importance are discussed extensively in a previous thesis written by Aaron Small [12]. An illustration of this is shown below; decanting 50% of the daily inbound volume can result in seven additional stations being available if stow rates improve by 22%, allowing for additional inbound volume to be stowed. In the following example, the number of stow stations being utilized in approximately 200, though that number varies from 190 stations to 210 stations as stow rate is varied.

Stow Station Requirements Analysis

Stow Rate \ Decant Vol.		0.0%	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%	87.5%	100.0%
Case Stow	Cart Stow	N/A	Floor Loaded Cases				Palletized Cases			SA
Baseline	0%	0	1	2	3	5	6	7	8	9
Baseline	3%	0	1	1	2	3	3	4	5	5
Baseline	6%	0	0	0	1	1	1	1	1	2
Baseline	10%	0	0	0	-1	-1	-1	-1	-2	-2
Baseline	13%	0	-1	-1	-2	-3	-3	-4	-4	-5
Baseline	16%	0	-1	-2	-3	-4	-5	-6	-7	-8
Baseline	19%	0	-1	-3	-4	-6	-7	-8	-10	-11
Baseline	22%	0	-2	-3	-5	-7	-9	-10	-12	-14
Baseline	25%	0	-2	-4	-6	-8	-10	-12	-14	-17
Baseline	29%	0	-2	-5	-7	-10	-12	-14	-17	-19
Baseline	32%	0	-3	-5	-8	-11	-13	-16	-19	-21

Break-Even Point to Reduce Station

Expected Station Requirements Reduction if Floor Loaded Cases Decanted

*Includes required count stations for extra counts generated by decant as a result of cost of quality

Figure 26: Stow Station Requirements Analysis

As shown, at a stow rate increase of 6% or greater, the stow station requirements begin to decrease. As a result, if a building is station constrained and can achieve a stow rate improvement of 6% or greater from decanting, it *might* be of value to decant in this situation. This depends on how critical it is to release additional stations, or if other levers cannot be pulled to improve overall throughput. It also depends on the decant rate, and if decant proves too costly to utilize despite the fact that decanting can release a small amount of additional capacity in the RSP.

Based on this analysis, it doesn't currently make sense to decant cases within the Amazon Robotics fulfillment centers. However, if decant rates could be increased, this investigation would have different results. Possibilities for this are discussed in chapter 6.

4.7 Future Improvements to the Optimization Model

One of the difficulties of this project was finding a stakeholder who had immediate use for the model such that developing the model further with a software development team would have a material payback period. Throughout the project, the model was socialized with teams from the Inbound Supply Chain Team (to explain differences in cost structures of different ship-modes and containers) to Amazon Robotics (to be utilized in as a part of future and development) to the ACES (Amazon continuous improvement) team. A stakeholder within Amazon Robotics was eventually identified, and with that, several improvements can be acknowledged:

1. Transition the model from a zone optimization approach to an individual station optimization approach.
2. Transition the model from an Excel-based OpenSolver optimization program to a Python-based optimization program to allow for more dynamic sensitivity analyses and easier changes to the inputs, outputs, and constraints.
3. Build in distributions into each of the input rates, allowing for dynamic simulation to determine a range of performance estimates, rather than in its current discrete form.
4. Build in functionality to transition the optimization model easily between different building designs (which is much easier with Python).
5. Build a web application that allows the utilization of this optimization model by many different teams within dynamic inputs and outputs.

4.8 Comments on Organizational Structure

The primary difficulty of finding a team to utilize the optimization model in perpetuity was that there isn't a single team that has ownership over the entire inbound space; many of these teams are disconnected and focused on individual pieces of the value chain (such as the fulfillment center or inbound transportation). As a result, it becomes difficult to align stakeholders for projects that affect two or more areas within the supply chain. This is part of the reason that I focused my primary efforts on the inbound conveyance analysis, while keeping in mind that there were several projects that could be more lucrative if I had more time to align stakeholders in different parts of the organization.

This difficulty is partly the result of the functional organizational structure that has evolved at Amazon [13]. There are several weaknesses for this organizational structure including customer unfriendliness, longer lead times, and difficulty in interdepartmental communications [14]. Amazon does an excellent job at combatting customer unfriendliness and longer lead times by making its organizational goals about optimizing those two qualities. However, their current focus on scaling to meet customer demand doesn't include the ability to smooth communication lines between groups and doesn't allow for groups to be completely aware of what other groups within Amazon are doing. This was especially obvious during this research project; often times, there were teams working on the same project that weren't aware of each other's efforts. Sometimes this was purposeful; allowing two (or more) individuals or groups to attempt to address the same problem sometimes yields completely different answers that can be weighed against each other to get an optimal solution. However, there are also downsides to this approach; namely, individuals with different perspectives don't have a global view or perspective when they are approaching the problem. In the future, Amazon could benefit from having a unifying team that helps to address this deficiency and it looks like Amazon is trying to address that with the increasing size of the Amazon continuous improvement team.

5. Impact of Inbound Conveyance and Other Potential Automation Improvements

This section covers the utilization of inbound conveyance in reducing the amount of indirect labor hours required within the inbound value stream. Inbound conveyance is equipment that is used to move product (at the container level) from the dock to the sleds (queues) that are placed adjacent to stow stations, thereby reducing processing steps in between the dock and stow. Inbound conveyance can span over a mile within the fulfillment center to connect disparate locations.

Currently, inbound conveyance is not used with Amazon Robotics Sortable fulfillment centers, and all movements of product are completed using pallet-jacks and carts. Within traditional sortable fulfillment centers, inbound conveyance allows operators to route certain product to certain areas to be stowed. This conveyance has some intelligence; photoelectric sensors allow managers and associates to see containers (and their contents) at various locations in the facility and ensures that containers on the conveyance do not collide. In summary, inbound conveyance is not a new technology for Amazon and has been used in a passive sense, to simply move product from one location to another and manage flow. Likewise, conveyance is utilized heavily in outbound, where flow is managed from central control room and conveyance flow is monitored very closely.

In the past, the issue with implementing inbound conveyance has focused around the cost / benefit analysis and the desire by Amazon senior management to reduce capital expenditure within new fulfillment centers in any way possible. This has resulted in analyses on inbound conveyance in Amazon Robotics FCs that are skewed in the direction of not implementing conveyance. For instance, a prior analysis of inbound conveyance focused solely on the value proposition of conveying totes removed from pallets to stations, which has the least value of all of the ship-modes I have discussed.

The value proposition, that I propose, of inbound conveyance is two-fold: (1) to reduce extra processing resulting from the increasing volume of floor loads arriving that require palletization, and (2) to reduce supply chain deficiencies seen as a result of fulfillment of product from a sub-optimal FC as a result of labor shortages.

5.1 Savings Calculations

In order to calculate the savings for a capital project in a warehouse environment, a framework was developed to determine the areas of savings that would be seen in a fulfillment center from the implementation of capital. The principle innovation of this framework is that capital improvements impact the cost structure of the fulfillment center from a labor cost perspective, and in some cases also influence the cost structure of the supply chain.

In section 2.6.2, this concept was introduced during the discussion regarding load-balancing cost-per-unit. The theory introduced through this framework is that load-balancing CPU can be reduced if labor gaps are reduced, and labor gaps are reduced when capital reduces the labor required. Specifically, labor hours that were previously utilized in non-value-added tasks within the fulfillment center (container palletization or pallet moving) can now be utilized in direct, value-added functions. This framework, called the Integrated Supply Chain Capital Allocation Framework, is outlined in Figure 26.

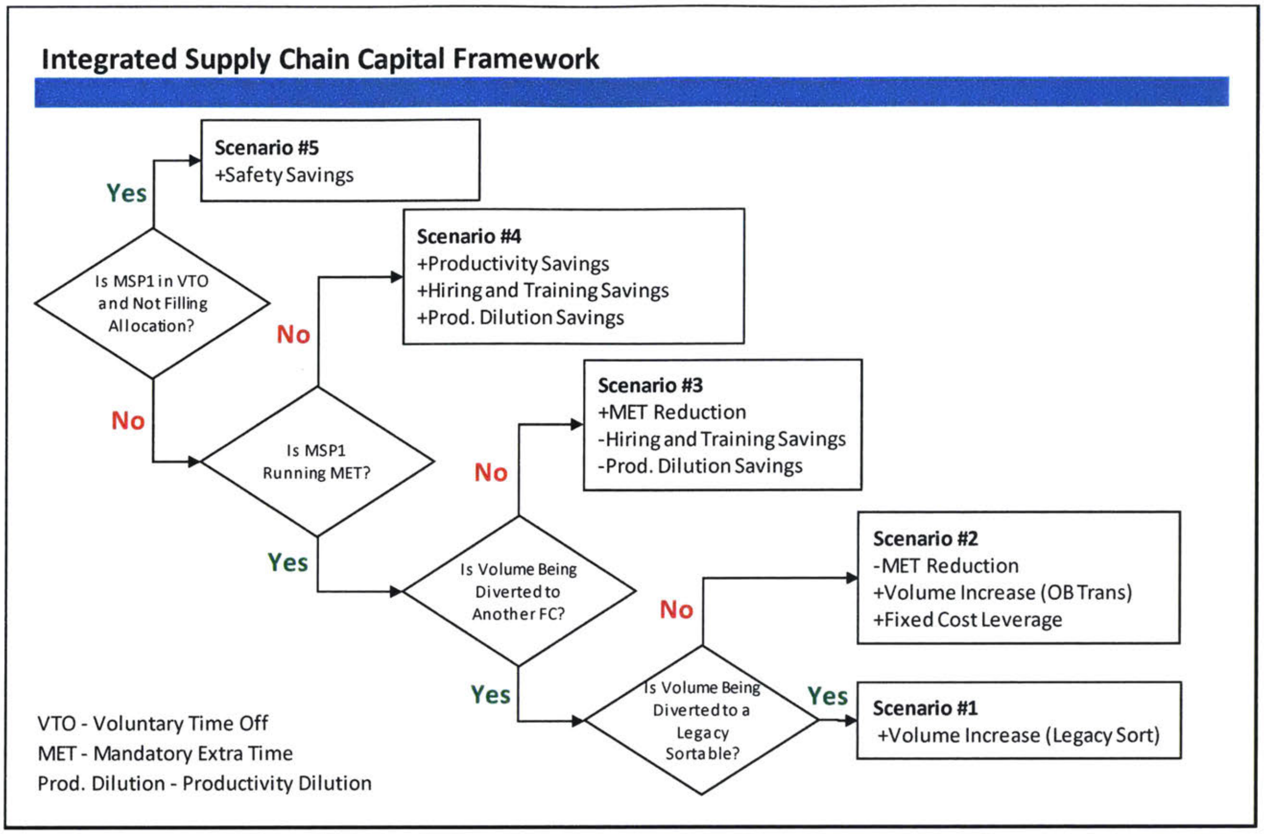


Figure 27: Integrated Supply Chain Capital Framework

This model outlines that viable savings are based on the "scenario" that the fulfillment center is facing. There are five primary scenarios that the fulfillment center can be facing, listed in order of labor shortage severity.

1. **Volume Diversion to Traditional Sortable Building:** Product optimally fulfilled from the fulfillment center in question is re-routed and fulfilled from a sub-optimal, less efficient fulfillment center.
2. **Volume Diversion to Robotics Sortable Building:** Product optimally fulfilled from the fulfillment center in question is re-routed and fulfilled from a sub-optimal, equally efficient fulfillment center.
3. **Overtime Called:** Product is fulfilled from the (optimal) fulfillment center, but overtime is required to ensure that fulfillment occurs.
4. **Standard Operation:** Product is fulfilled by the (optimal) fulfillment center.
5. **Overstaffing Situation:** Product is fulfilled by the (optimal) fulfillment center, but labor supply overages cause incremental costs as a result of a lower labor utilization rate.

The effect of scenario #1 is depicted in the following chart:

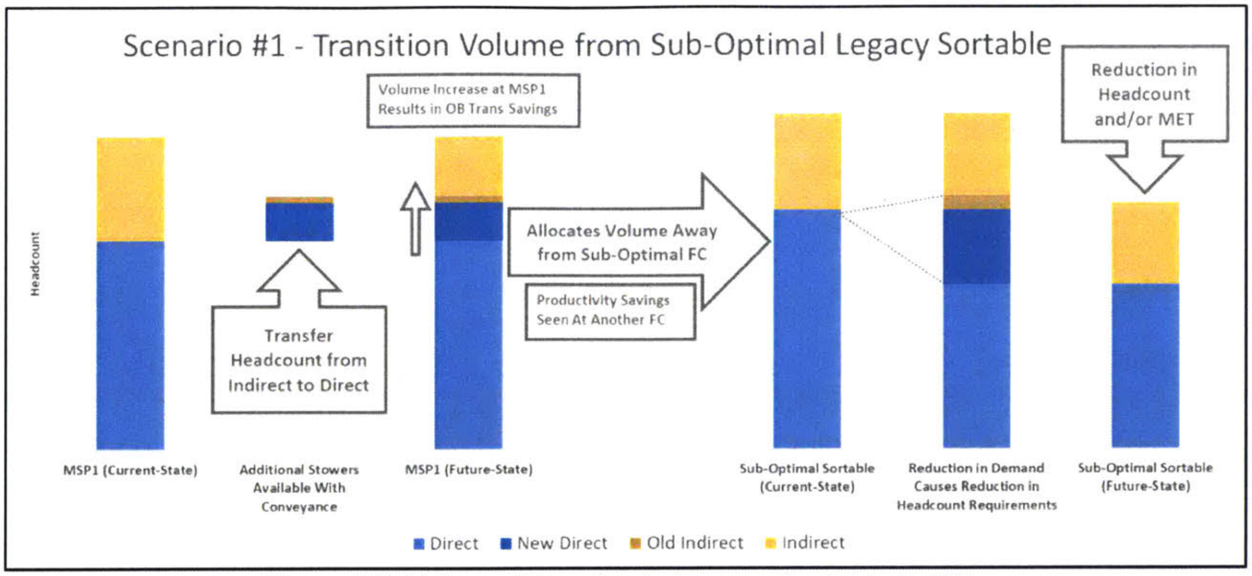


Figure 28: Volume Diversion to Legacy Sortable Building

As depicted above, if the fulfillment center in question has a labor deficiency that is causing volume to be re-routed to a legacy fulfillment center, labor savings is seen at another fulfillment center, while supply chain savings are achieved as the result of a more-optimal allocation.

Each savings category indicated in Figure 27 is calculated separately. Calculating estimated savings is difficult; estimating when labor shortages will occur (and how drastic they will be) is difficult to forecast. In the case of the project undertaken here, past data was utilized for the first six months of savings calculations, and full-year projections were utilized for the last six months of savings calculations. The following page shows the savings calculations per category:

1. Safety Savings

Safety savings encompasses the reduction in injury recordables that occur as a result of eliminating the palletization process for floor loads, which is the process introduced in section 2.6.1. A six-month period was analyzed to calculate the occurrence rate of injury recordables as a function of floor load volume. Savings is calculated by multiplying the expected reduction in palletization volume by the cost of an injury recordable as a function of the floor load volume.

2. Productivity Savings

Productivity savings encompasses the reduction in the number of labors hours. Savings is calculated by multiplying the number of labor hours reduced per floor load volume by the amount of floor load volume directed through the inbound conveyance system. The number of labor hours reduced is then multiplied by the effective hourly labor cost rate.

3. Hiring and Training Savings

Hiring and training savings encompasses the reduction in hiring costs and training costs associated with onboarding an associate. The costs are reduced when the fulfillment center reduces its overall headcount, as peak and off-peak hiring is reduced. Hiring costs are calculated utilizing a flat rate and multiplying that rate by the expected reduction in yearly hires. Similarly, training costs are calculated by multiplying the expected number of training hours expended per new hire by the expected reduction in yearly hires.

4. Mandatory Overtime Reduction Savings

Mandatory overtime reduction savings encompasses the reduction in overtime hours as a result of not having to call overtime at slightly higher volumes with the same number of associates. These reductions are difficult to calculate, since it is not immediately clear exactly what volumes constitute a situation where a marginal reduction in non-value-added work resulted in the avoidance of overtime. Nonetheless, in this case, overtime reduction savings are calculated by multiplying the expected reduction in overtime hours by the effective hourly labor cost rate. It is expected that mandatory overtime reductions will occur before and after periods of substantial increases in load balancing costs, indicating a period in which volumes are increasing and overtime is about to be called.

5. Productivity Dilution Savings

Productivity dilution savings encompasses the reduction in learning curve hours, where learning curve hours are defined as the period when associates were less productive as a result of fewer hours spent in a direct role. The learning curve within a fulfillment center in the direct functions is very well defined. As a result, the productivity dilution savings is calculated by multiplying the reduction in hires by the number of lost hours per hire, at the effective hourly labor cost rate. The expected number of lost hours per hire, if they experience the average Amazon learning cure, is 32 hours.

6. Volume Increase – Outbound Transportation Savings

Outbound transportation savings constitutes the transportation savings realized as a result of shipping product from the optimal fulfillment center, as opposed to from a sub-optimal fulfillment center. The savings here is easily calculated by multiplying the load balancing CPU by the increase in capacity during the period being analyzed.

7. Volume Increase – Legacy Sortable Opportunity Cost

Legacy sortable opportunity cost savings is the savings realized by the reduction in volume fulfilled by a less optimal fulfillment center. To calculate legacy sortable opportunity cost savings, the difference of the respective rates of the more efficient robotics fulfillment center and the legacy fulfillment center are taken and multiplied by the number of units transitioned to the more efficient building. Given that it is incredibly difficult to predict when this occurs, these savings were not included in the overall analysis, though there are certainly times during peak periods when this savings is valid.

8. Fixed Cost Leverage

Fixed cost leverage savings encompass the increase in outbound volume seen from transitioning non-value added labor hours to outbound functions, thereby increasing outbound volume. The savings is calculated by multiplying the increase in outbound volume by the fixed cost-per-unit. This savings is seen only if there is a material decrease in the amount of fixed cost allocated as a result of this project, such as a reduction in the number of fulfillment centers required. At the time of this project, the finance team recommended proceeding with the assumption that this savings is valid, though there were still questions in my mind as to the likelihood that there would be a reduction in the number of fulfillment centers built or that the fixed cost leverage savings was definitely valid.

5.2 Scenarios Analyzed

This analysis focused on two improvement scenarios utilizing inbound conveyance; a system that supports two floors and system that supports four floors of the robotic storage platform. A 2-floor system provides direct inbound conveyance to a total of 48 stations located on two floors, with a total capacity of approximately 2-3 floor loaded trailers per shift depending on stow rates and units per floor load. A 4-floor conveyance system provides direct inbound conveyance to a total of 96 stations located on four floors, with a total capacity of approximately 4-6 floor loads per shift depending on stow rates and units per floor load. The inbound conveyance line utilizes spirals that can convey totes and cases directly from the trailer at the dock to a few feet away from each stow station, at which point associates unload containers from the conveyance line and place those containers onto sleds like the current process. As such, this eliminates pallet building, pallet movement, and VRC usage for the floor loads that are processed through the conveyance system.

In summary, automation utilizing the recommended 2-floor system eliminates approximately $1.55 million in costs per year related to indirect labor and outbound transportation cost (through the reduction of labor constraints). Approximately 3% of labor hours are removed from the current-state inbound department by the implementation of this project. The implementation of the 2-floor system would cost approximately $1.43 million, for installation and materials associated with the conveyance system. These estimations are based on 2017 Q1 and Q2 actuals and 2017 Q3 and Q4 forecasts based on the 2017 full-year plan (FYP). Forecasting the volume of floor loads and load balancing requirements were incredibly difficult, and are the least rigorous inputs to the savings calculations. One way that I ensured my results were conservative was to utilize the regression results for estimating floor loads in Figure 18 without including trending increase in overall floor loads forecasted for the future (when the conveyance system would be installed). Furthermore, average load balancing CPU was utilized and several colleagues in Seattle were consulted with to improve the reliability of the labor shortage forecast that was included in my model.

5.3 Savings Calculations

Savings were calculated for a 2-floor and 4-floor inbound conveyance system. These two systems were chosen for a few reasons:

- Both systems had the same design, utilizing a single conveyor and divert unload system that then allocated product to two separate floors. In the case of the 4-floor system, two of the modular designs were implemented with one system supplying the 1st and 2nd floors and the second system supplying the 3rd and 4th floors.
- A 1-floor system doesn't provide enough capacity and the truck unload economics don't make the system viable.
- A 3-floor system could be analyzed in the case that it wasn't clear if a 2-floor or 4-floor system was better, but a 3-floor system doesn't involve the modular design.

I completed the analysis at the FC that I was at, and then extrapolated the analysis to other buildings, augmenting certain properties to account for differences, like load balancing CPU and building layout. Based on an analysis involving the four labor-constrained buildings, the following payback period analysis for a 2-floor system was completed.

Payback Period Analysis

Site	FC1	FC2	FC3	FC4
Classification	**Non-Decant**	**Non-Decant**	**Non-Decant**	**Decant**
Building Type	**1-MOD**	**2-MOD**	**1-MOD**	**1-MOD**
Labor Constrained	**Yes**	**Yes**	**Yes**	**Yes**
Load Balancing CPU	**$0.10**	**$0.17**	**$0.22**	**$0.23**

Site	FC1	FC2	FC3	FC4
Floor Loads Per Period	1021	783	753	862
Increase in Outbound Capacity	1,562,000	1,562,000	987,000	1,562,000
Savings	$ 1,475,000	$ 1,590,000	$ 1,550,000	$ 1,680,000
Cost	$ 1,425,000	$ 2,850,000	$ 1,425,000	$ 1,425,000
Payback Period	**0.97**	**1.79**	**0.92**	**0.85**

Table 8: Payback Period Analysis: Inbound Conveyance Systems at Labor Constrained Sites

This analysis was completed using the optimization model discussed throughout this thesis; simulating scenarios with and without conveyance supporting different zones, with outputs mirroring those shown in Table 7. In addition, Figure 30, on the following page, shows a buildup graph that visualizes the savings for a 2-floor system at FC3. In Figure 30, the area covered by each colored area represents a different savings category (i.e. the light blue area represents productivity savings).

The Y-axis represents the savings that can be attributed to the inbound conveyance system each week. The X-axis relates the week to the condition that is seen within the fulfillment center.

For instance, during week 8, the fulfillment center was providing voluntary time off (VTO) to employees. During this period, the savings within the fulfillment center for this project is zero, because the facility is not reducing labor hours by utilizing inbound conveyance, and because additional volume is not being re-allocated to other fulfillment centers.

Similarly, during week 17, the fulfillment center was capacity constrained from a labor availability standpoint, and as a result, customer orders were being fulfilled from a sub-optimal location. In this situation, with inbound conveyance, the savings is according to Figure 29 (on page 57), where safety, productivity, outbound transportation, and fixed cost savings (if valid), could be seen.

Volumes of freight at 100,000-unit intervals were used for ease of calculation, and conditions for the fulfillment center were applied based on projected labor plans (and shortages) and historical data. Average load-balancing costs (for outbound transportation) were utilized, as was the average effective labor cost and full-year plan (FYP) for Q3 and Q4 of 2017.

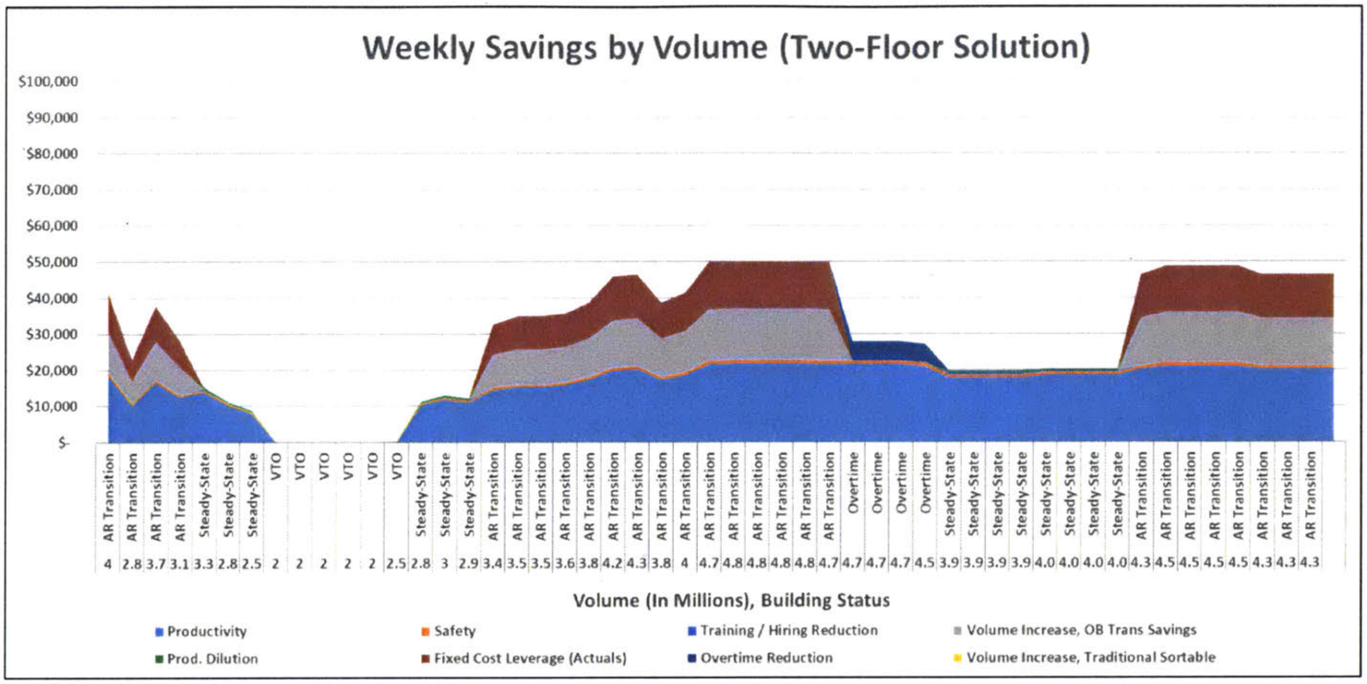


Figure 29: Buildup Chart of Savings for 2-Floor System, FC3

The 4-floor system was not recommended because the volume of floor loads received at each of these sites was not high enough (except during peak periods) to require the entire system's capacity. With a 2-floor system, this never occurs as the capacity of a 2-floor system serves as a base processing system, for which any non-served floor loaded capacity can be palletized as in the current process. Finally, the capacity increase is based on an estimated labor profile for each of the sites, with FC1, FC2, and FC4 having the same estimated labor profile. The labor profiles are based on actuals for the first half of 2017 and predictions for the second half of 2017.

5.4 Comment on the Value of Options

This project contains options within it: the amount of floors of conveyance added to the fulfillment center as a part of this project can be varied and floors can be added over time. For instance, and 2-floor system could be implemented to start and the implementation of this system would not disallow (barring spatial constraints) the addition of one or two more floors of conveyance-enabled floors later on. Based on the way the system is engineered where one divert supports two floors, it makes sense to implement the conveyance in two floor increments to avoid re-engineering the divert system later one.

The value of the option is inherent in the incremental value that it could provide to the business and also from the fact that the option is very inexpensive. In the current state, the project cost would increase by 100% but only 25% incremental value (in the form of peak productivity improvements and a further reduction in outbound transportation cost from labor constraints) would be delivered from implementing a 4-floor inbound conveyance system. Given the cost of the option is very small (2-5% of project cost, all related to project management costs), the value of reserving the option instead of completing the entire project immediately is very high.

5.5 Overall Effect on Operations

Inbound conveyance has several non-quantitative impacts on the fulfillment center and its employees, which were studied through the 2-month pilot of the system. This 2-month pilot was an experiment completed at FC1 where the continuous improvement team and I consolidated the fulfillment center's inventory of flexible conveyance and assembled it to imitate a single-floor inbound conveyance system in two zones of the inbound department. This flexible conveyance line fed totes and cases from the dock to stow stations, such that associates then removed product from the conveyance and moved that product to queues a few feet away at each station. These impacts are listed below:

- Product "cherry-picking" can occur in a non-regulated conveyance system. Within the inbound conveyance system, indirect associates are asked to load the sleds of 8-12 stations. Because indirect support associates prefer to lift lighter items and totes (which have ergonomic handles), stowers at the front of the conveyance line generally stow smaller items and stow more items from totes, leading to higher stow rates due to favorable item and container mix. This phenomenon can be seen below in Figure 28 and pictured in Appendix E. Achieving stow rate parity for stowers is critical for two reasons, (1) to ensure that rates can be compared appropriately between stowers for performance management and (2) to ensure that stowers don't perceive unfair treatment.

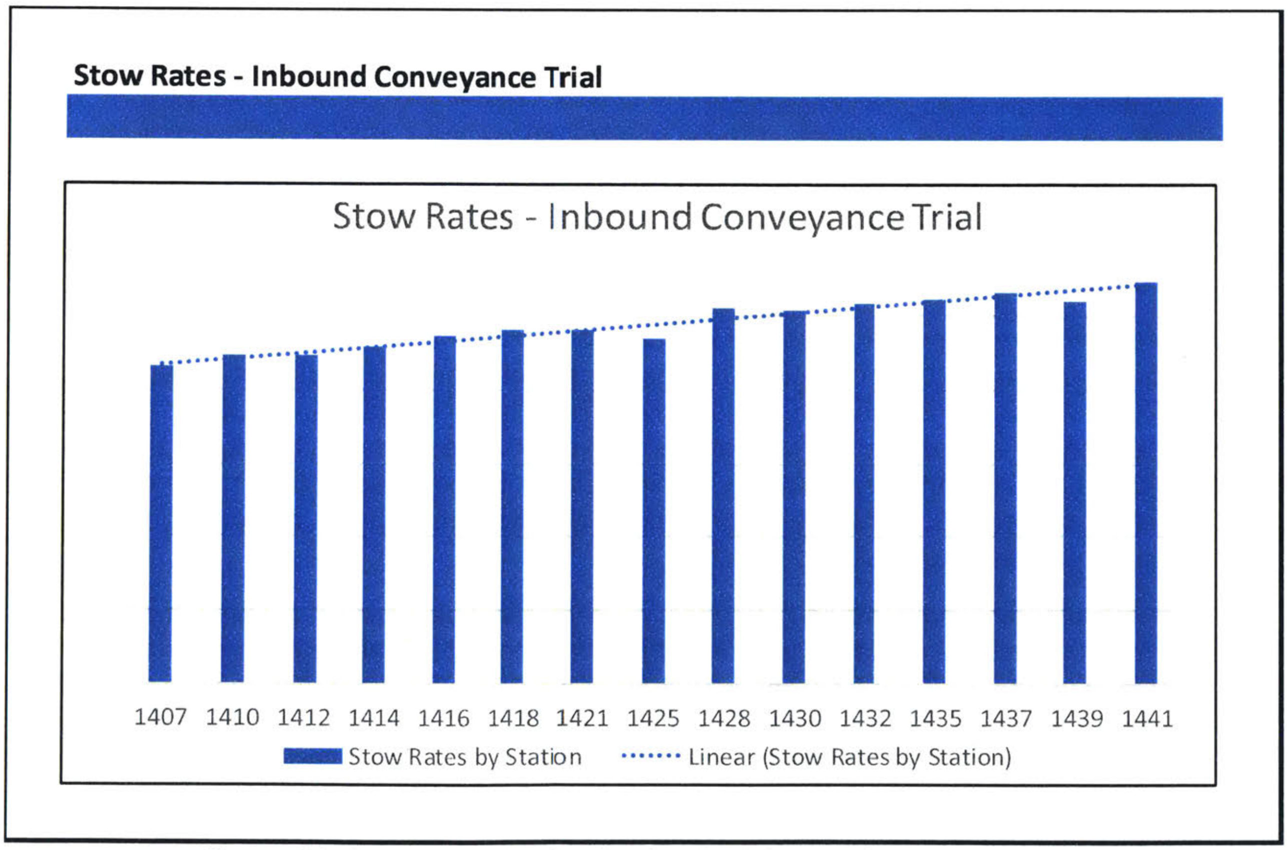


Figure 30: Stow Rates by Station for the Inbound Conveyance Trial

- There was a small, but non-zero increase in the requirements of maintenance personnel responding to issues with the conveyance system. Part of this can be attributed to the fact that the conveyance utilized for the trial wasn't meant for continuous and heavy usage, but having additional conveyance in the fulfillment center likely means additional maintenance requirements.

6. Conclusions and Further Areas of Improvement

This thesis argues that the utilization of a capital model involving internal improvements (i.e. within the fulfillment center) and external improvements (i.e. within the inbound supply chain) must be considered in efforts to select capital improvement projects optimally. This thesis provides an example of a project (inbound conveyance) where utilizing this model would lead to a decision to implement a partial system at some sites, versus the current decision to implement the system at no sites.

Furthermore, this thesis concludes that changes to upstream processes would adversely affect network-wide cost. As a result, it is recommended that no changes be made in the current state; though future state analyses will undoubtedly uncover longer-term changes that can be made to reduce overall cost in the inbound supply chain.

As a result of this thesis, several areas of potential improvement were identified. Two primary areas include the following:

- Advanced Containerization Concepts – there exists the possibility to continue to improve the utilization of trailers and increase unit count by moving away from corrugate containers and Amazon yellow totes.
- Advanced Decant Concepts – there exists the possibility to provide stowers with uniform, stower-friendly containers in a cost effective manner by reducing the labor requirements for the decant process.

6.1 Advanced Containerization Concepts

A major issue with implementing an all-tote or unified container supply chain concept for Amazon is an increase in transportation CPU as a result of a lower utilization of trailers. Figures 14 and 17 show this: an increase in totes clearly reduces trailer utilization. However, this doesn't mean that a standard container can't produce the same utilization results of supplier-provided cases due. At Amazon's scale, containerization concepts could be a major enabler of both fulfillment center automation and improved trailer utilization, if the container and surrounding processes are designed together.

There are two obvious and possible options for container standardization: (1) implement a standard set of recyclable containers that freight is transitioned into at the point of receiving product from a vendor or (2) implement a standard set of disposable containers at the vendor site that can be transitioned easily and cost-effectively throughout the supply chain. For the first idea listed above, there is several criteria that would need to be considered, including the following:

Metric	Success Criteria
Trailer Utilization	New container concept reduces trailer requirements and improves trailer utilization and trailer unit count
Automation Potential	New container provides opportunities for increased automation within the Amazon inbound value chain
Container Cost	Container cost to Amazon does not insurmountably increase
Sortation and Decant	Increase in sortation or decant labor does not insurmountably increase
Container Recycling	Container recycling cost does not insurmountably increase
Container Durability and Handling	Container must be easily handled by employees and must also be durable enough to be transported between sites

Table 9: Containerization Concept Criteria

Essentially, in order for a new containerization concept to improve upon the current state, trailer utilization and automation improvements must significantly outweigh the cost of recycling the container and transitioning all product into that container.

6.2 Advanced Decant Concepts

A major issue with utilizing the decant process within the fulfillment center is that the labor requirements and associated rate of actually performing the decant (moving product from case to standard container) task make the process cost-prohibitive. If you eliminate the labor requirements (by automating the process) and the cost of quality of the process, there is significant savings that can be captured, depending on the improvement in stow rates. Essentially, this analysis considers decant as the process chosen for all incoming product located in cases, but removes the cost of quality and improves the decant rate to reduce the cost of decanting a product to approximately nothing, as if the process was fully automated and the quality was perfect.

The initial savings results from indirect improvements in corrugate removal and container management. The subsequent increases in savings are seen from the improvement in stow rates. This savings analysis varies based on the percentage of case volume that is decanted (the higher the case volume, the larger the savings).

The calculations were made by running the optimization model in its current state (without decant as a process) and with varying case and cart stow rates, and then running identical scenarios with a decant rate of 10,000 units per hour (essentially, a labor-free decant rate) and with a $0 cost of quality (and everything else held constant). By setting a high decant rate, when the decant process is enabled within the future-state simulation, all case volume flows through the decant process and is consequently stowed from carts. Because cost of quality has been set to $0 and decant labor is minimal, VCPU drops from the improvement in corrugate management and container management, as well as increasing stow rates.

Case Stow Rate	Cart Stow Rate	Estimated Baseline VCPU - Normalized	Estimated New VCPU w/ Auto-Decant - Normalized	Potential Savings
189	189	$0.1000	$0.0936	$12,312,537
189	195	$0.0999	$0.0899	$19,118,670
189	201	$0.0998	$0.0876	$23,497,981
189	207	$0.0998	$0.0856	$27,195,616
189	213	$0.0997	$0.0837	$30,684,883
189	219	$0.0996	$0.0819	$33,982,973
189	225	$0.0996	$0.0802	$37,105,243
189	231	$0.0995	$0.0786	$40,065,448
189	237	$0.0995	$0.0771	$42,877,541

Table 10: Savings Calculations for Automated Decant

There are obvious difficulties associated with automating this process; the cases arriving are all in various shapes, sizes and conditions and would need to be carefully opened to avoid damaging the product. In addition, the product would need to be transferred quickly and accurately to a new, unified container; there are obvious amnesty issues that would likely occur in this process. A future Leaders for Global Operations thesis will evaluate the possibilities of this in the Fall and Summer of 2018.

7. Further Areas of Study

This thesis has focused on areas that are of immense importance within the current industrial economy, capital investment and automation.

As it relates to capital investment, Amazon utilizes a conservative payback period when determining if capital investments for continuous improvement projects should be made. This policy is in accordance with one of Amazon's primary leadership tenants of "Frugality".

Amazon, with its usage of automation, provides an interesting environment to consider the value in human-powered processes versus the value in increasing automation capabilities, as a result of their impressive Amazon Robotics fulfillment center technology.

7.1 Capital Investment in a Rapidly-Changing Fulfillment Environment

Amazon utilizes a discounted payback period analysis to determine the viability of capital investment projects, from small projects such as the continuous improvement projects discussed here to large-scale projects like building new fulfillment centers. Each expected future cash flow is discounted back to its present value utilizing the firms weighted average cost of capital. This represents an improvement over the payback period metric utilized in this thesis, which is considered to be theoretically deficient due to its lack of incorporating risk and the time value money [15].

The form that is completed, called an NACF CAR (Capital Appropriation Request), allows entries for cash flows on a quarterly or monthly basis, and is utilized for capital investments as large as new fulfillment center openings, and as small as fulfillment center improvements. It is expected that projects not deemed essential from a safety or system functionality perspective will have a short payback period as noted in section 7.

One issue with a conservative payback period is that there is little justification for the selection of the specified "cutoff" payback period. It is widely acknowledged that these determined cutoffs generally have little scientific backing, but in some cases, a payback period of less than half of the economic life of the investment is chosen to reduce risk of the investment [16]. In this case, the economic life of an investment Amazon makes to a newly built fulfillment center is hard to estimate, but it can be assumed that a capital investment such as conveyance would be useful to the facility in question for at least five years. As a result, the shorter payback period justification cannot be based solely on this justification, but more so on the fact that Amazon expects to change rapidly its technological portfolio within the fulfillment center environment over the next decade. This change has an assumed impact on payback period selection – that this change will cause capital investments to become obsolete quicker than the economic life of that investment. This discussion begs the question: is Amazon not completing valuable projects with low-risk returns as a result of forecasted technological change that could render these projects obsolete sooner than their economic life?

A future study could be undertaken to determine if a less conservative payback period makes sense, or if a different discount rate (higher than the weighted average cost of capital) could be used to "price in" risk of a project being obsolete in the near future due to technological advancement. In the case of Amazon, "Frugality" might be allowing Amazon to focus on the future of fulfillment center technology, or it might be getting in the way of fulfilling product at a lower cost to customers; the outcome of this policy remains to be seen.

7.2 The Value of Human-Powered and Automated Processes

Industry 4.0, broadly defined as the incorporation of smart factory technologies to allow for the "electronic flow of production processes", is typically applied to the factory environment, but is equally applicable in Amazon's global supply chain [17]. In these environments, it is expected that the future growth in talent requirements will be the management of these smart factory technologies will the requirements of individuals to perform simple, more manual tasks within these environments will shrink [18]. This can already be seen within Amazon's fulfillment centers with the deployment of its Floor Health team, which manages the Amazon Robotics system's operating areas to ensure that obstructions and other issues don't impede the efficiency of the Amazon Robotics system. In some sense, the individuals on the Floor Health team have replaced those that would be picking or stowing product in a less efficient inventory system. In the project discussed here, some of the remaining manual, unskilled tasks within the fulfillment center would be replaced with the usage of conveyance.

Outside of the implications on factory employment and the types of jobs that are available, there is some other criteria to consider when selecting automation projects, including the following:

- Does automation drastically increase system efficiency or performance during all operating periods? Is the implementation of automation improve peak performance substantially, or is it considered an improvement during normal production periods?
    - *Example:* The Amazon Robotics system has increased stow rates by over 20% in comparison to traditional Amazon fulfillment centers.
- Does automation increase the safety performance of the facility or department in question? Does it reduce the amount of hours utilized in areas that have high safety incident rates?

    - *Example:* The reduction in lifting of heavy cases results in less back injuries to employees lifting those cases.
- Does automation reduce flexibility? Without automation, a facility could implement as "scrappy" solution if there is a system failure. Does automation impede the implementation of backup solutions due to physical barriers of the new solution?
    - *Example:* The Amazon Robotics drives within the inventory management system in an Amazon Robotics fulfillment center are controlled via an internet network and algorithm. As a result, if the internet network goes down within a building, orders cannot be picked and the entire fulfillment center shuts down.

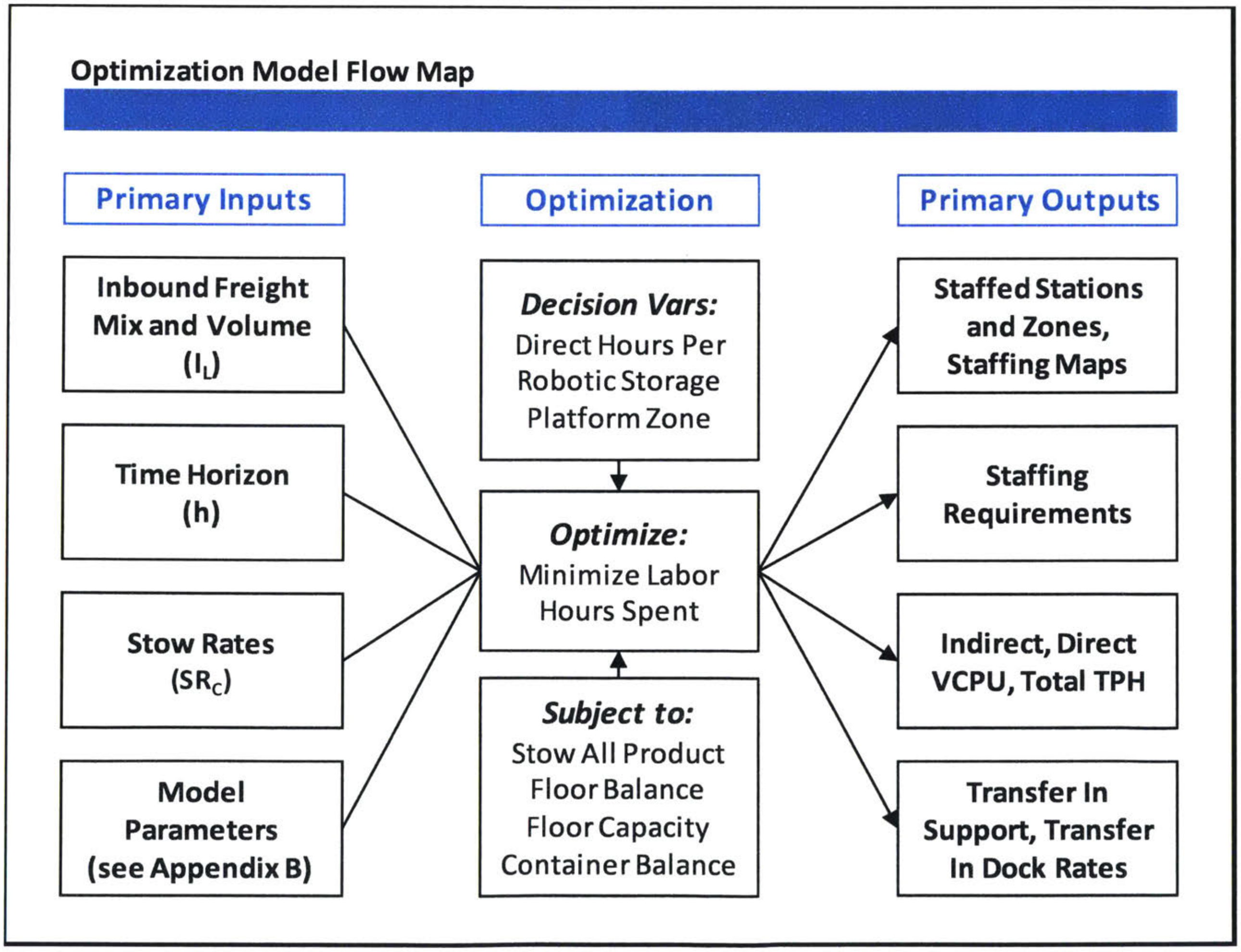
As Amazon continues to analyze what areas of its fulfillment network to automate, the final decisions will be driven by these questions just as much as by the labor constraints issue that was outlined in section 2.6.2. However, one thing is for certain; there will be value in both automated and human-powered processes in Amazon's future fulfillment network.

8. Appendices

A: Model Parameters Abbreviated List

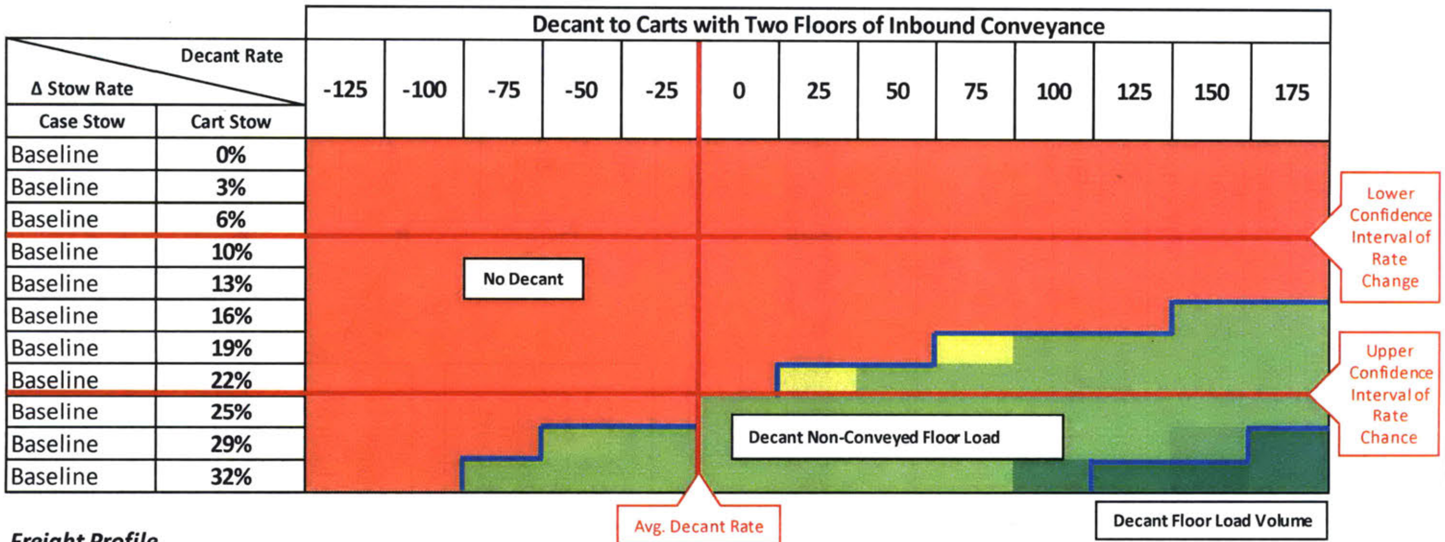
Stations Per Zone	Tote Palletization Rate	Southeast WIP -> Zone 8
Cases / Gaylord	Case Palletization Rate	Tote Sled Load Rate
Totes / Recycled Pallet	Pallet Preparation Rate	Case Sled Load Rate
Totes / U-Boat	Dock WIP -> North VRC WIP	Process Guide Rate
Cases / Fluid Load Pallet	VRC Load Rate	Process Assistant Rate
Pallet Unload Rate	Pallet Management Rate	Receive Rate
Floor Load Unload Rate	Conveyor Setup Rate	Build Gaylord

B: Optimization Model Flow Chart



C: Decant Analysis: Stow Rate vs. Decant Rate

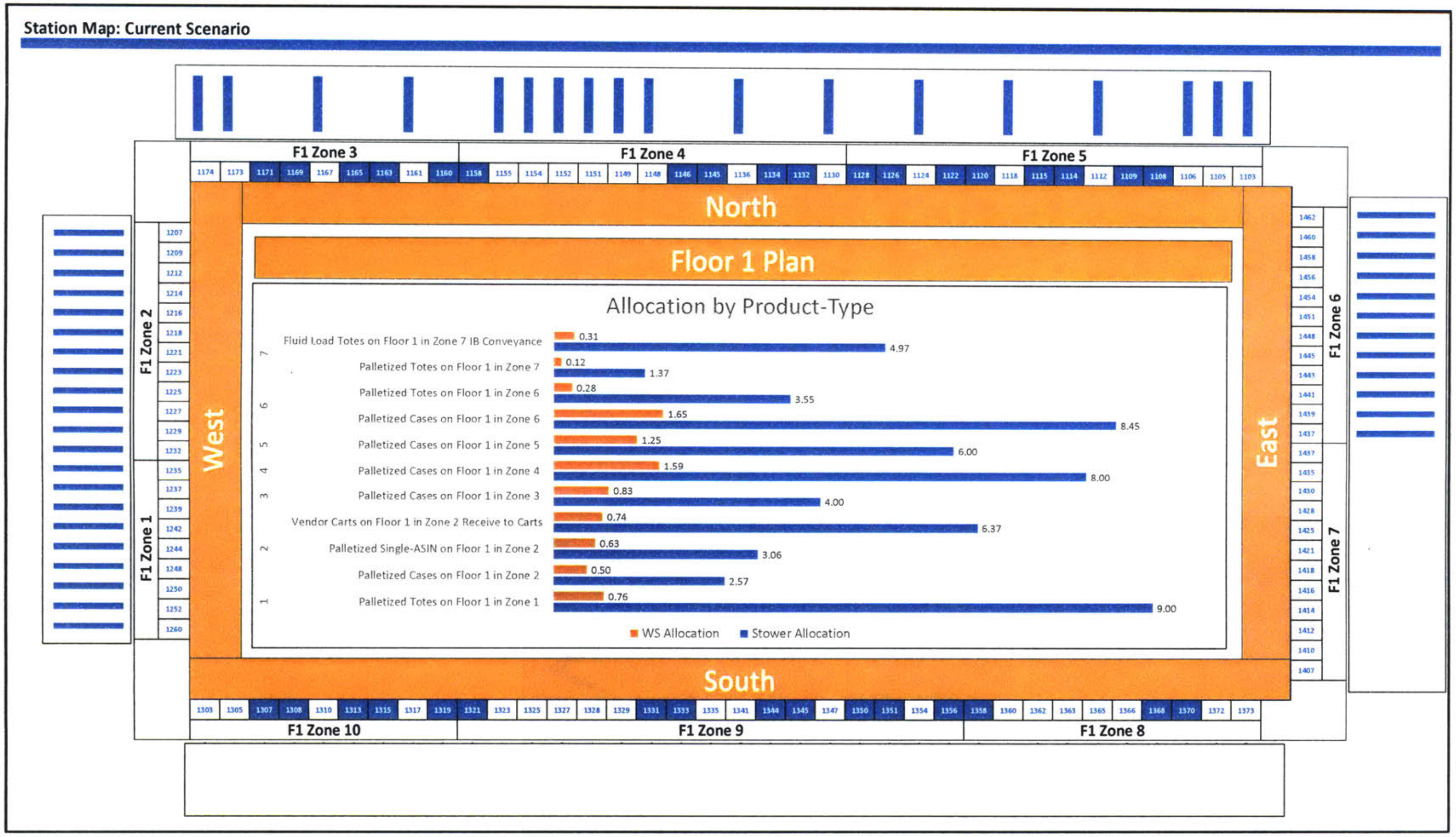
Decant Break-Even Analysis



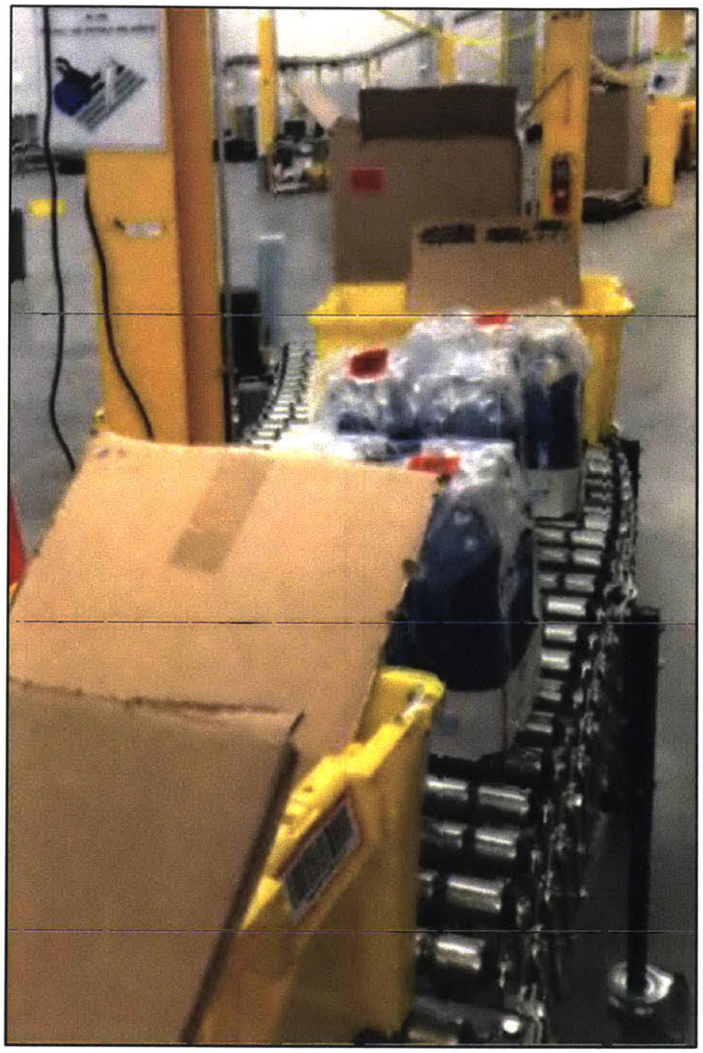
Freight Profile

Palletized Totes	30%	**Fluid Load Totes**	5%
Palletized Cases	25%	**Fluid Load Cases**	30%
Palletized Single-ASIN	5%	**Vendor Units**	5%

D: Station Staffing Map from Optimization Model



E: Cherry-picking Inbound Conveyance Trial Photo



Bibliography

[1] "Amazon.com, Inc. Form 10-K," *United States Securities and Exchange Commission*, 31-Dec-2016. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1018724/000101872417000011/amzn-20161231x10k.htm. [Accessed: 28-Oct-2017].

[2] G. Minculete and P. Olar, "New Approaches to Supply Chain Management Concept. Logistics Integration of 'Hub and Spoke' Model," *Valahian J. Econ. Stud.*, vol. 5, no. 2, pp. 21–32, Apr. 2014.

[3] J. B. Skipper, W. A. Cunningham, C. A. Boone, and R. R. Hill, "Managing Hub and Spoke Networks: A Military Case Comparing Time and Cost," *J. Glob. Bus. Technol.*, vol. 12, no. 1, pp. 33–47, Spring 2016.

[4] F. G. (Fernando G. Dobrusky, "Optimal location of cross-docking centers for a distribution network in Argentina," Thesis, Massachusetts Institute of Technology, 2003.

[5] P. Buijs, H. W. Danhof, and J. (Hans. C. Wortmann, "Just-in-Time Retail Distribution: A Systems Perspective on Cross-Docking," *J. Bus. Logist.*, vol. 37, no. 3, pp. 213–230, Sep. 2016.

[6] B. Golden, S. Raghavan, and E. Wasil, Eds., *The Vehicle Routing Problem: Latest Advances and New Challenges*, vol. 43. Boston, MA: Springer US, 2008.

[7] O. (Olufemi O. Oti, "Hub and spoke network design for the inbound supply chain," Thesis, Massachusetts Institute of Technology, 2013.

[8] "10xLogisticsExperts, LLC - Leading Revolutionary Change: Continuing on Cube Utilization - Secondary Cube." [Online]. Available: http://10xlogistics.blogspot.com/2012/09/continuing-on-cube-utilization.html. [Accessed: 28-Oct-2017].

[9] "Theory of Constraints: A Literature Review - ScienceDirect." [Online]. Available: http://www.sciencedirect.com.libproxy.mit.edu/science/article/pii/S1877042814051532? [Accessed: 28-Oct-2017].

[10] "Linear Programming." [Online]. Available: https://cs.nyu.edu/overton/g22_lp/encyc/article_web.html. [Accessed: 25-Nov-2017].

[11] "Guide to Solvers," *OpenSolver for Excel*, 15-Jun-2015. [Online]. Available: https://opensolver.org/guide-to-solvers/. [Accessed: 24-Dec-2017].

[12] A. A. Small, "Gamification as a means of improving performance in human operator processes," Thesis, Massachusetts Institute of Technology, 2017.

[13] P. Meyer, "Amazon.com Inc.'s Organizational Structure Characteristics (An Analysis)," *Panmore Institute*, 03-Jul-2016. .

[14] J. Peters, "On structures," *Manag. Decis. Lond.*, vol. 31, no. 6, p. 60, 1993.

[15] C. M. Boardman, W. J. Reinhart, and S. E. Celec, "The Role of the Payback Period in the Theory and Application of Duration to Capital Budgeting," *J. Bus. Finance Account.*, vol. 9, no. 4, pp. 511–522, Winter 1982.

[16] S. Yard, "Developments of the payback method," *Int. J. Prod. Econ.*, vol. 67, no. 2, pp. 155–167, Sep. 2000.

[17] N. Szozda, "Industry 4.0 and Its Impact on the Functioning of Supply Chains," *VIERTE Ind. Revolut. IHR EINFLUSS AUF Funkt. Mod. LIEFERKETTEN*, vol. 13, no. 4, pp. 401–414, Oct. 2017.

[18] "Technology Transforms the Industrial Workforce," *Trends Mag.*, no. 151, pp. 16–21, Nov. 2015.