Amazon UK Transshipment Stow Process Analysis through Discrete Event Simulation
by
Eric Young

B.S. Engineering, Harvey Mudd College, 2009

Submitted to the Department of Electrical Engineering and Computer Science and the MIT Sloan School of Management in partial Fulfillment of the Requirements for the Degrees of

Master of Science in Electrical Engineering and Computer Science
and
Master of Business Administration

In conjunction with the Leaders for Global Operations Program at the
Massachusetts Institute of Technology
June 2016

©2016 Eric Young. All rights reserved.
The author hereby grants MIT permission to reproduce and to distribute publicly copies of this thesis document in whole or in part in any medium now know or hereafter created.

Signature of Author ____________________________ Signature redacted

Department of Electrical Engineering and Computer Science
MIT Sloan School of Management
May 20, 2016

Certified by ____________________________ Signature redacted

Dr. Luca Daniel, Thesis Supervisor
Professor of Electrical Engineering and Computer Science

Certified by ____________________________ Signature redacted

Dr. Stephen C. Graves, Thesis Supervisor
Abraham J. Siegel Professor of Management Science

Accepted by ____________________________ Signature redacted

Leslie Kolođziejski, Chair of the Committee on Graduate Students
Department of Electrical Engineering and Computer Science

Accepted by ____________________________ Signature redacted

Maud Hef,ón, Director of MIT Sloan MBA Program
MIT Sloan School of Management
This page has been intentionally left blank
Amazon UK Transshipment Stow Process Analysis through Discrete Event Simulation

by

Eric Young

Submitted to the Department of Electrical Engineering and Computer Science and the MIT Sloan School of Management on May 20, 2016 in partial Fulfillment of the Requirements for the Degrees of Master of Science in Electrical Engineering and Computer Science and Master of Business Administration

Abstract
This thesis investigated Amazon’s implementation of transshipment, the act of moving inventory between fulfillment centers to facilitate order fulfillment, inventory rebalancing, and operational cost reduction. The role of transshipment in Amazon’s fulfillment center network was explored and a proposed transshipment stow process is modelled to examine potential reductions in transshipment cycle time.

As Amazon’s fulfillment network has grown in size, the function of individual fulfillment center has evolved from shipping customer orders to providing a network-level, low-cost fulfillment option through transshipment. The complexity of this process, from the source fulfillment center to transportation and finally to the destination fulfillment center is examined in-depth to provide insight into transshipment processes that are typically not considered or modelled in research.

One particular process—transshipment stow—was determined to have a large impact on transshipment cycle time. A proposed transshipment tote stow process was envisioned and modeled using discrete event simulation in order to determine cycle time savings versus volume efficiency tradeoffs. While the proposed transshipment tote stow process simulation was able to reduce transshipment cycle time by more than 80%, the tote stow process required an additional 7.5x amount of space compared to the current stow process. Based on initial simulations, the benefits of transshipment tote stow cycle time reduction does not outweigh the costs in space utilization. Transshipment-specific item stowing areas with periodic inventory consolidation may be the best tradeoff between cycle time, additional processes, and space.

Thesis Supervisor: Luca Daniel
Title: Professor of Electrical Engineering and Computer Science, Department of Electrical Engineering and Computer Science, MIT

Thesis Supervisor: Stephen Graves
Title: Abraham J. Siegel Professor of Management Science, MIT Sloan School of Management
This page has been intentionally left blank
Acknowledgements

The completion of this internship and thesis would not have been possible without the support of countless individuals along the way. I would like to thank Amazon UK and the LGO program office for their help making this internship possible and providing an unforgettable six month experience.

From MIT, I would like to thank my thesis advisors Stephen Graves and Luca Daniel for their invaluable guidance in understanding the complex systems I encountered at Amazon as well as in writing the thesis.

From Amazon, I would like to extend my sincerest gratitude to my manager Henry Low and my sponsor John Tagawa for their unequivocal support throughout my entire internship. In addition, I would like to thank Amazonians Tattiana Villalobos, Tobias Fusser, Christopher Smith, Brian Stout, Shreeshastha Kumar, Peter Kansteiner, Neil Humphryes, and Edwin Lei for their patience in bringing me up to speed throughout the six months.

Finally, I would like to thank fellow LGO and Amazon intern Elizabeth Ohrt for providing support throughout the internship.
# Table of Contents

Abstract ............................................................................................................................... 3

Acknowledgements ................................................................................................................ 5

Glossary ............................................................................................................................... 10

1 Introduction ....................................................................................................................... 11
   1.1 Project Motivation ..................................................................................................... 11
   1.2 Problem Statement ................................................................................................... 11
   1.3 Project Goals ............................................................................................................ 11
   1.4 Thesis Overview ...................................................................................................... 12

2 Operations at Amazon UK .......................................................................................... 13
   2.1 Amazon Fulfillment Background ............................................................................. 13
   2.2 Amazon UK Network ............................................................................................... 13
   2.3 Amazon Fulfillment Center Operations (UK Perspective) ....................................... 15
   2.4 What is Amazon Transshipment ............................................................................ 15
   2.5 What are Transshipment’s Advantages and Disadvantages? ............................... 16

3 Literature Review ........................................................................................................... 20
   3.1 Transshipment ........................................................................................................ 20
   3.2 Discrete Event Simulation ....................................................................................... 23

4 Current Transshipment State: Fulfillment Center (FC) ........................................ 25
   4.1 Transshipment Overview ......................................................................................... 25
   4.2 Inventory ................................................................................................................ 25
   4.3 Transshipment Pick (Source FC) ............................................................................. 27
   4.4 Transshipment Merge/Palletize (Source FC) .......................................................... 29
   4.5 Transshipment Outbound Dock (Source FC) ............................................................ 30
   4.6 Transportation (Third Party) ................................................................................... 30
   4.7 Transshipment Inbound Dock (Destination FC) ...................................................... 31
   4.8 Transshipment Receive (Destination FC) ................................................................. 32
   4.9 Transshipment Stow (Destination FC) ..................................................................... 32
4.10 Transshipment Quality ................................................................. 34
4.11 Transshipment Timing ................................................................. 34
4.12 Transshipment Improvements .................................................... 39

5 Transshipment Stow Process Discrete Event Simulation ................. 40
  5.1 Current Transshipment Stow Process ............................................ 40
  5.2 Proposed Transshipment Stow Process ......................................... 41
  5.3 Data ............................................................................................ 43
  5.4 Initial Feasibility Study ................................................................. 45
  5.5 Current Transshipment Stow Model Formulation & Inputs .......... 47
  5.6 Results ......................................................................................... 52
  5.7 Discussion .................................................................................. 58

6 Conclusions, Feasibility of Implementation, and Future Work ............ 60

References ........................................................................................ 62
List of Figures

Figure 1: Map of the UK with 10 Amazon Fulfillment Centers .............................................. 14
Figure 2: Standard Order, Reactive Transship and Transfer Differences .................................. 16
Figure 3: No Transshipment versus Transshipment Comparison ........................................... 17
Figure 4: Basic Discrete Event Simulation Walkthrough .......................................................... 23
Figure 5: Discrete Event Simulation with 2 Pick Processes ....................................................... 24
Figure 6: Amazon Transshipment Processes ........................................................................... 25
Figure 7: Standard FC Processes in a Single FC with Inventory States .................................... 26
Figure 8: Expanded Pick Process .............................................................................................. 27
Figure 9: Expanded Merge/Palletize Process ............................................................................ 29
Figure 10: Expanded Outbound Dock Process .......................................................................... 30
Figure 11: Expanded Inbound Dock process ............................................................................ 31
Figure 12: Expanded Receive Process ....................................................................................... 32
Figure 13: Expanded Stow Process ........................................................................................... 32
Figure 14: Standard Order Timing and Deadlines .................................................................... 35
Figure 15: Transshipment Order Timings and Deadlines .......................................................... 36
Figure 16: Timing of an Example Shipment ............................................................................. 38
Figure 17: Transshipment Item Pick Time after Successful NTRBD Stow .................................... 45
Figure 18: Distribution of Number of Items in Each Tote ............................................................ 47
Figure 19: Distribution of Transshipment Item Volumes .............................................................. 48
Figure 20: DES Structure of a Single Shipment in the Current Transshipment Stow Process ... 49
Figure 21: Item Pick Delay Probability ...................................................................................... 50
Figure 22: DES Structure of a Single Shipment in the Proposed Transshipment Stow Process .... 51
Figure 23: Current Stow Process Volume Use Over 5 Days ...................................................... 53
Figure 24: Current Stow Process Volumes Over 20 Days .......................................................... 54
Figure 25: Current Transshipment Stow Process with Consolidation after x Days .................. 55
Figure 26: Single Shipment Proposed Process Volume Use Over 5 Days .................................. 56
Figure 27: Single Shipment Proposed Process Volume Use Over 20 days ................................. 57
Figure 28: Proposed Stow Process Volumes Over 20 Days ...................................................... 57
Figure 29: Current Transshipment Stow Process with Consolidation after x Days .................. 58
**Glossary**

**ASIN:** Amazon Standard Identification Number, a code that identifies items in Amazon’s system.

**CPT:** Critical Pull Time. Time when an item needs to be picked, palletized, and placed on a truck for transshipment.

**Customer Order Affinity Transshipment:** This type of transshipment is initiated because Amazon believes that more orders for this particular ASIN will be requested in the DFC region and thus has proactively transshipped inventory to the DFC. E.g. Amazon has fulfilled 100 units of Harry Potter Book 1 from FC A and had to further transship 20 units from FC B to fulfill demand (Customer Order Reactive). Amazon can initiate an affinity transshipment of 10 units of Harry Potter Book 1 proactively to meet future demand (Customer Order Affinity).

**Customer Order Reactive Transshipment:** This type of transshipment is initiated because a customer has placed an order for this particular item and it is most efficient to transship to the DFC for fulfillment.

**DFC:** Destination Fulfillment Center; FC receiving the transshipment.

**Domestic Cross-Dock Transfer:** This type of transfer is initiated at specific cross-dock capable sites to facilitate efficient inventory receiving for the network. A vendor would send all purchased inventory for the network to the cross-dock facilities and the cross-dock facility would separate them out and transship to their final destination. This is in contrast to the vendor shipping each FC their allocated inventory.

**EU:** European Union; represents all European FCs.

**Fulfilled by Amazon (FBA):** Inventory that Amazon holds for its merchants—instead of Amazon owning the inventory, merchants contract Amazon to provide warehousing and logistics services.

**Network Inventory Rebalancing Transfer:** This type of transfer is initiated if Amazon determines there is an inventory imbalance in the system: E.g. all 1,000 units of ASIN A in the network are at a single FC. Amazon will attempt to rebalance this inventory by transshipping units across the network.

**NTRBD:** Need to Receive By Date. This deadline is a bit misleading as it does not represent the actual need to receive by date, but rather the time when the item needs to be available in a prime location for picking. Stow deadline for transship purposes.

**Prime Location/Prime:** A bin in the FC that can be picked for orders.

**Promised Delivery Date (PDD, Promise):** Delivery date promise that Amazon has made to the customer.

**SFC:** Source Fulfillment Center; FC sending the transshipment.

**SLA:** Service Level Agreement.
1 Introduction

1.1 Project Motivation

Amazon.com, started by Jeff Bezos in 1995, has grown tremendously since its inception, expanding from a garage-based shipping operation into a global fulfillment center network. As Amazon’s network continues to grow, new and creative ways of using the fulfillment center and the fulfillment network emerge, allowing for improvements to Amazon’s fulfillment performance as a whole. One of the largest changes to the fulfillment operations is a process known as transshipment—the movement of inventory between fulfillment centers instead of shipping directly to the customer. Transshipment has become a core component of Amazon Fulfillment Operations and is the focus for improvement in this project.

1.2 Problem Statement

Transshipment, the movement of goods between fulfillment centers (FCs), facilitates inventory rebalancing, bulk inventory receiving, and customer order fulfillment and is a vital and growing part of Amazon FC processes. This process is a contributor to improved customer satisfaction due to fewer shipped boxes, cost-reduction from faster, cheaper inventory transfers, and savings on purchasing and inventory holding costs from efficient inventory placement, among other benefits. As transshipment becomes more important to fulfillment operations, Amazon would like to further improve transshipment cost, quality, productivity, and reliability.

1.3 Project Goals

The project goals were to produce functional and actionable results for Amazon to improve transshipment at the individual FC process levels. Multiple recommendations were proposed to Amazon, including network-wide inventory classification changes, inbound and outbound
process changes, as well as metrics changes. However, during the full analysis of Amazon’s transshipment structure, an area of improvement was theorized to dramatically decrease transshipment process times as well as increase item availability. This thesis provides a simulation of the current process, a simulation of a proposed process and an analysis of the performance-critical characteristics of the new process that would impact Amazon operations.

1.4 Thesis Overview

This thesis will provide an extensive overview of the actual execution of transshipment within Amazon and initiate research into a major change in the transshipment process. While the project resulted in multiple conclusions and recommendations, only a specific portion of the work will be documented in this thesis.

Chapter 2 provides a brief overview of the overall operations of the Amazon fulfillment network in the UK as well as a high-level overview of transshipment. Chapter 3 contains an overview of the academic literature on transshipment in as well as the discrete event simulation modeling technique used in this thesis. Chapter 4 overviews the individual transshipment process in the transshipment system as well as overarching quality and timing concerns. Chapter 5 highlights the current state of the current transshipment stow process, proposes a new stow process, examines the underlying assumptions of the new process and models both processes to examine overall tradeoffs. Finally, Chapter 6 summarizes the overall findings in this thesis.
2 Operations at Amazon UK

2.1 Amazon Fulfillment Background

Fulfillment at Amazon is concentrated in fulfillment centers (FCs) which house both Amazon inventory as well as Fulfilled by Amazon (FBA) inventory for Amazon Marketplace. These individual fulfillment centers are located in geographically advantageous locations to fulfill certain regions of the country and are managed on a FC by FC basis. Best practices are shared throughout the FCs when applicable, although there can be some divergence of processes in each FC. However, the overall processes performed in each FC in order to bring in or ship out inventory are very similar.

Individual FCs were originally built to ship to the customer—the FC brings in inventory from various vendors and ships out customer orders. As the FC network became more dense, FC transshipment—the process of shipping inventory between FCs rather than just to the customer—was established to improve supply chain flexibility and reduce overall fulfillment cost. This transshipment process has now grown to be a large part of Amazon FC processes and results in large cost reductions throughout the Amazon system.

2.2 Amazon UK Network

The Amazon UK network at time of writing consists of 10 FCs located throughout the UK, all of which are mapped in Figure 1. These fulfillment centers specialize in certain types of goods in order to facilitate efficient processes throughout the network. There are two primary characteristics that a FC is classified into, sortable versus non-sortable and crossdock-enabled or not crossdock-enabled.
A sortable distinction for a FC represents an FC that only carries items that can be sorted into a specific size of box. This allows for use of totes or trays in conjunction with automated conveyance to transport items throughout the FC instead of manual carts or pallet jacks. Non-sortable FCs hold inventory that are not sortable and generally consist of the larger items that Amazon sells on its website or items that are sold as full cases. While an FC can hold both distinctions, holding both sortable and non-sortable inventory, this is not common in the network. Overall the FC network is heavily biased towards sortable systems due to the larger volume of sortable goods sold. This paper will focus on sortable processes—particularly sortable transshipment processes—which dominate the shipping and transshipping bulk of the FC network.

A second distinction, crossdock, is a process in which an FC receives inventory, but need not store the inventory into the receiving FC; rather, the receiving FC can ship (i.e. crossdock)
some or all of the inventory out to other FCs\(^1\). The inventory that is crossdocked is primarily used for inventory replenishment at the destination FCs and is done in large quantities. There is only one FC that is crossdock-enabled in the UK network—this particular FC performs both customer orders fulfillment as well as crossdock processes. The thesis primarily focuses on non-crossdock enabled FCs, though the process changes would apply to this particular crossdock FC.

2.3 **Amazon Fulfillment Center Operations (UK Perspective)**

The FC-specific operations in Amazon UK are split into two major categories— inbound and outbound. Inbound encompasses the process that brings inventory into the FC and is generally split into three major steps: Inbound Dock, Receive, and Stow. Outbound encompasses the processes that send inventory out to the customer (orders) or other fulfilment centers (transshipment) and are split into Pick, Pack and Outbound Dock for orders and Pick, Palletize/Merge, and Outbound Dock for transshipment. A brief overview of standard order processes can be found in Section 4.1 while an in-depth discussion of transshipment FC operations can be found throughout Chapter 4.

2.4 **What is Amazon Transshipment**

Amazon transshipment, broadly speaking, is the movement of inventory between two separate FCs for any reason. This is in contrast to the standard operating processes of an FC—bringing inventory into the FC for storage and sending inventory out of the FC to the customer.

An Amazon transshipment can be triggered in three different ways, known as requesting sources: Customer Order, Network Inventory Rebalancing, and Domestic Cross-dock. Additional information on the definition of these sources can be found in the Glossary. These three sources can be grouped into two categories for execution purposes—reactive transshipment (Customer

---

\(^1\) Note that crossdock is a part of transshipment by definition as inventory is shipped between FCs.
Order) and proactive transfers (Network Inventory Rebalancing, Domestic Cross-dock). Each transshipment has three defining characteristics: the transshipment container type, the transship item composition, and the transship miss impact, which are summarized in Figure 2. Reactive transshipments are a ‘reactive’ response to a customer order—if a customer order can be fulfilled more efficiently through transshipment and the customer shipping experience is not impacted, Amazon will move inventory between FCs to fulfill the order. Proactive transfers are not directly tied to the fulfillment of a customer order, but are performed for inventory replenishment and to rebalance inventory across the network. Reactive transshipments address a near-term need while proactive transfers address a long-term need.

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Standard Order</th>
<th>Reactive Transship</th>
<th>Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Container Type</td>
<td>Customer Order</td>
<td>Customer Order</td>
<td>Source Inventory Balancing, Bulk Receiving</td>
</tr>
<tr>
<td>Container Unique ASINs</td>
<td>Lower</td>
<td>Higher</td>
<td>Black Tote, Cases, Pallets</td>
</tr>
<tr>
<td>Container Quantity/ASIN</td>
<td>Lower</td>
<td>Lower</td>
<td>Lower</td>
</tr>
<tr>
<td>Failure Impact</td>
<td>Promise miss/Increased shipping cost</td>
<td>Promise miss/Increased shipping cost</td>
<td>Difficult to quantify</td>
</tr>
</tbody>
</table>

**Figure 2. Standard Order, Reactive Transship and Transfer Differences**

### 2.5 What are Transshipment’s Advantages and Disadvantages?

Transshipment allows for more efficient use of network resources and reduces costs of fulfillment and replenishment throughout the entire FC network.

Reactive transshipment responds to customer orders and moves inventory to advantageous locations for order fulfillment. Shipping costs are a primary cost driver for packages and is highly dependent on the number of packages shipped. As such, reactive transshipments aim to reduce the ‘splits’ per order, or number of boxes that are shipped per order. Additionally, reactive transshipment allows Amazon to use time that otherwise might be underutilized. Last mile delivery services such as UPS, FedEx, USPS in the United States and Royal Mail in the United Kingdom have specified package pick up times—there is little need to rush package fulfillment when the pick
up time has passed as the package would not get shipped until the next day. These pickup times are scheduled generally in the afternoon. As Amazon FCs are 24-hr or near 24-hr operations, FC to FC reactive transshipment allows Amazon to transship ordered inventory in bulk during less busy periods to reduce the overall cost of order fulfillment and still make order shipping deadlines. A visual example is provided in Figure 3, where a customer orders an item after the shipping deadline (4 PM in the example) and the item is only present in FC A which is far away from the customer. Without transshipment, this order would not be picked until the next day, a few hours prior to 4 PM—the order cannot leave the warehouse until 4 PM anyways. However with transshipment, Amazon can spend the time shipping the item to a closer FC B and have it arrive at the same time as the no-transshipment package. If the transshipment cost plus the outbound shipping cost from FC B is less than the outbound shipping cost from FC A, Amazon has saved money and has not delayed delivery to the customer.

![Diagram](image)

**Figure 3: No Transshipment versus Transshipment Comparison**
(No delivery shipments can be made past 4 PM)
In contrast, transfers are used for cost savings on replenishment to fulfillment centers. Traditionally, cost savings are realized by having vendors bulk ship to a crossdock FC instead of shipping small amounts of inventory to individual FCs. Inventory replenishment orders from multiple vendors are consolidated at the crossdock FCs and then bulk shipped to individual FCs. In addition, inventory can also be rebalanced based on projected demand, essentially ‘staging’ inventory in locations where it will most likely be needed. Overall, transfers represent the traditional advantages of transshipment—supply chain flexibility.

However, transshipment’s benefits come with costs—increased complexity as well as additional labor. First, unlike standard order fulfillment which originates and ends at a single FC, both reactive transshipments and transfers involve three separate parties in the overall process—the source FC (sending FC), the transportation carrier, and the receiving FC. Secondly, processes have to be developed to address the process flow of FC-FC inventory. The addition of other parties into the standard shipping process requires additional labor—instead of a single flow from the order fulfillment FC, inventory needs to be handled at the source FC and by transportation, increasing labor costs. Thirdly, timing also becomes extremely important and more complex in the transshipment process—a single delay at any step could cascade into an order miss or an inventory replenishment delay.

Of particular interest for this thesis are the customer order reactive transshipments. Customer order reactive transshipments are highly time constrained as these transshipments are used for order fulfillment. A delay or error in a reactive transshipment can result in Amazon not being able to fulfill a delivery time commitment to a customer, resulting in a ‘customer promise miss’. Thus, while proactive transfers will reduce the overall cost of fulfillment, reactive transshipment can directly
impact the customer experience and create a ‘customer promise miss,’ an important metric in the Amazon system.
3 Literature Review

3.1 Transshipment

Transshipment as a whole has been investigated by researchers for many years. In particular, research relevant to Amazon transshipment is termed lateral transshipment in literature and represents the movement of goods within the same echelon of the supply chain. However, transshipment literature is generally focused on the high-level optimization of what to transship, where, and efficiency. A comprehensive overview of lateral transshipment models can be found in Paterson’s 2011 paper which summarizes the papers and work performed in the academic community. In particular, transshipment has been examined from the late 1950’s to now, focusing on varying number of items, warehouse echelons, warehouses, ordering policies, and cost analysis. In addition, both proactive and reactive transshipments have been extensively analyzed by academics with a variety of techniques. A brief summary of the findings from Paterson’s paper for both proactive and reactive transshipment is briefly examined in the next two paragraphs.

Proactive transshipments have been examined from the simplest model assuming a single-period, single-transshipment with no network inventory replenishment model, to a very complex model assuming multi-period, multi-transshipments with network inventory replenishment. The basic models focus on the timing and the impact of the transshipment itself while assuming negligible transportation time for the transshipment. The primary investigation area for these models is the characteristics of the transshipment—should transshipment be performed and when in the period should transshipment be performed. More complex models focus on the transshipment network and look at multi-period, multi-transshipment models that allow for network inventory replenishment. These studies attempt to establish comprehensive inventory
policies that dictate when items should be transshipped and when inventory should be replenished, similar to standard inventory replenishment policies such as the newsvendor model. Overall, the models highlight an expected result of transshipment assuming all goes to plan—safety stock in the network is reduced, especially in cases with a large network, a high service level, or a long replenishment lead time.

Reactive transshipment research can be split between periodic versus continuous review and further into single-echelon, multi-echelon, and decentralized systems. Single-echelon systems represent a single level in the supply chain, multi-echelon represent a warehouse (distribution) and retailer (sales) network, and decentralized systems introduce game theory with each location attempts to maximize their own benefit. Of particular interest to this thesis are the single-echelon transshipment systems as Amazon FCs at this level all perform similar functionality—store and ship to customers. These models primarily look into the benefits of partial pooling of inventory (holding inventory in reserve) versus complete pooling of inventory (fulfilling all demand). However, these models can come to different conclusions based on the simplifying assumptions that are made—for periodic review transshipment complete pooling was found to be the most efficient if there are negligible transshipment times in a two warehouse system while partial pooling became more efficient for a multi-location system. As a whole however, the models are extremely restrictive and generally assume a two warehouse system, single item transshipment, full knowledge of demand, or negligible transshipment times.

More recent models further increase the complexity but primarily in a focused area. In Ramakrishna, Sharafali, et al. 2013 paper, a two-item, two-warehouse periodic-review inventory model with transshipments and ‘rush’ orders system is examined and costs analyzed. Noham and Tzur 2014 improved on previous methods of modeling single and multi-item transshipment
problems by implementing a fixed transshipment cost component—previous research assumed
that there were minimal fixed costs to execute a transshipment and assumed a per unit cost.
Torabi, Hassani, et al. 2014 addresses exactly what Amazon performs—transshipment in e-
tailing. In their paper, a Mixed Integer Linear Programming (MILP) model is created and
logistics cost minimized. However, shipment and transshipment lead times are ignored. Finally,
Xie and Ouyang discovered that a non-cyclic hexagon shape results in an optimal transshipment
facility layout, assuming an infinite homogenous plane.

This document’s primary focus is not on the allocation and planning of transshipment, but
the actual execution. In this area, the research is less comprehensive and moves away from the
standard newsvendor modelling and optimization. In terms of execution, Amazon’s
transshipment planning mirrors single-echelon, partial-pooling, multi-location models with an
extra layer of complexity: Amazon calculates all possible plans for order fulfillment upon order
receipt and chooses the cheapest option that meets the customer delivery promise. While most of
literature is focused on the cheapest option for transshipment, Amazon’s highest priority is the
customer delivery promise. As Amazon is in control and fully understands the timing of
transshipment processes, transshipment cycle time is fully calculated into the possibilities of
order fulfillment. Note that this method ‘brute forces’ all possible fulfillment possibilities
throughout the entire network and may become intractable as Amazon’s transshipment network
gets larger or fully globalizes (current transshipment is primarily regional).

Overall however, academic transshipment papers are focused on high-level network supply
chain optimizations and changes and on the actual execution of the transshipment. As such,
while the research papers are interesting, they are not directly applicable to transshipment
processes improvements. However, a full understanding of the transshipment network and execution is necessary in order to provide meaningful process improvements.

3.2 Discrete Event Simulation

In order to simulate the transshipment system, discrete event simulation (DES) was chosen to thoroughly analyze the timing and effects of a transshipment process. DES models a system as a discrete sequence of events in time and allows for instantaneous time skips. Each item under simulation is given a particular timestamp where a process needs to be executed and the execution duration. The simulation keeps track of these timestamps and executes them in time ascending order. On execution, the state change of the system can be recorded and the system analyzed. The primary advantage of DES is that time is not incremented based on a timestep—instead the simulation relies on the sorted array of timestamp actions (state changes) and simply advances time to the next action in the array.

Thus a simple system example of item A picked at time $t = 5$ and B picked at $t = 8$, a DES would not need to increment through $t = 0$ through $t = 8$ and could instead immediately jump to $t = 5$ and then $t = 8$, realizing a state change occurs within the system at those times. For the case of cascaded process (say, picking the item and shipping the item), the simulation simply reinserts the item back into the array, with the tag that the item needs to perform the second action. Using the same example above, if the pick process takes 2 timesteps before the item completes pick and is ready to ship, the simulation would step through as follows.

<table>
<thead>
<tr>
<th>System Time = 0</th>
<th>System Time = 5+</th>
<th>System Time = 7+</th>
<th>System Time = 8+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>Item/Action</td>
<td>$T$</td>
<td>Item/Action</td>
</tr>
<tr>
<td>5</td>
<td>Item A/Pick</td>
<td>7</td>
<td>Item A/Ship</td>
</tr>
<tr>
<td>8</td>
<td>Item B/Pick</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 4: Basic Discrete Event Simulation Walkthrough
DES also has the advantage of allowing for parallel processes to be implemented and constrained. For example, if three items need to be picked at time $t = 5$, but only two “pickers” are available, only 2 of the items would be advanced forward to the next step—the remaining item would be delayed until the pickers are free ($T = 7$).

<table>
<thead>
<tr>
<th>System Time = 0</th>
<th>System Time = 5$^+$</th>
<th>System Time = 7$^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>Item/Action</td>
<td>T</td>
</tr>
<tr>
<td>5</td>
<td>Item A/Pick</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>Item B/Pick</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>Item C/Pick</td>
<td>7</td>
</tr>
</tbody>
</table>

**Figure 5: Discrete Event Simulation with 2 Pick Processes**

DES fits well with Amazon processes as a whole—individual processes are broken out into independent steps only executed after previous processes are completed and flow through the overall system in a single direction with no loops. While DES could be used in a more complex manner with state probabilities similar to a Markov chain, this functionality is not needed for Amazon processes—probability of a loopback in the Amazon system is nearly nonexistent.

DES is not a new technique used to investigate individual processes. In fact, three previous Amazon LGO theses (Kasenga, 2013; Faranca, 2004; Price, 2004) use DES to model Amazon’s sorting process, picker to conveyor process, and outbound process flow respectively. In particular, DES was used to analyze floor volume use, reduce container use in processes, identify bottlenecks, and improve productivity.
4 Current Transshipment State: Fulfillment Center (FC)

4.1 Transshipment Overview

This section will provide a comprehensive overview of current transshipment processes within Amazon. An initial explanation of how inventory is perceived within the Amazon system is provided in order to provide full understanding of the transshipment cycle. All seven major steps of transshipment are then examined—a brief diagram of the steps can be seen in Figure 6. Additionally, two major metrics of transshipment, transshipment quality and timing, are briefly reviewed to highlight additional complexities of transshipment. Finally, potential areas for improvement in the transshipment process are explored.

![Figure 6: Amazon Transshipment Processes](image)

4.2 Inventory

While inventory is not technically a process in the transshipment process, it is the start and the end of the transshipment process as a whole. Thus the state of inventory is extremely important to understanding the success and failure of transshipment. A brief diagram of the standard order processes in a single FC can be found in Figure 7.
Figure 7: Standard FC Processes in a Single FC with Inventory States

Inventory at Amazon is generally stored in the inventory tower—multi-story shelving that contains a variety of items. Inventory is categorized as ‘Prime’ or ‘non-Prime’, where Prime inventory is inventory available for customer fulfillment (and physically stored in the tower) while non-Prime inventory is not available for customer fulfillment and is enroute to the tower or stored in long term (reserve) storage.

When vendor inventory arrives into the fulfillment center as new inventory for Amazon, it is scanned in as inventory at receive stations. At this receive scan point, the vendor inventory has now been acknowledged as Amazon inventory and is available to sell to Amazon customers. However, the inventory is not in a Prime location at this time as it is not physically present in the inventory tower. It is not until the inventory is physically placed into the inventory tower through the stow process that the inventory changes into Prime inventory and can be physically picked for Amazon customers.

For customer order fulfillment, items are picked from the Prime locations, lose their Prime inventory designation and are immediately assigned to a customer order—at this point in time, identical items are non-fungible; an item assigned to a customer cannot be used to fulfill
another customer’s identical order. This is due to Amazon’s ‘batching’ system that they use for orders with multiple items\(^2\). The item(s) is then sent to pack and shipped out to the customer.

The following sections will fully describe the transshipment processes and some of the difference between standard FC and transshipment processes.

### 4.3 Transshipment Pick (Source FC)

![Figure 8: Expanded Pick Process](image)

Pick is the first step in the transshipment process for reactive transshipments. Picks are initiated when an order arrives, requires transshipment, and can be transshipped out the same day. Transshipment picks do not occur for transshipments planned a week from now—if a transshipment pick is initiated, it should be shipped on the next transshipment truck. Note that the order itself may be planned to ship days after transshipment arrival—the system will naturally prioritize the transshipment of orders in the near future, but if an order scheduled to be shipped a week later can be transshipped now, the system will attempt to do so.

Process-wise, a transshipment pick and standard order pick are nearly identical—an associate is instructed to scan a tote, scan the bin (Prime location), scan the item, pick the item in the bin, place it into a tote (yellow tote for orders, black tote for transshipment), and once the tote is full, place the tote on a conveyor. The only difference in transshipment pick is inventory

\(^2\) Additional information about the Amazon batching system can be found Luna 2015.
assignment—standard order inventory assignment is done at pick and is non-fungible and while transshipment inventory is also assigned to an order at pick, it is fungible and non-binding. However, even if it is non-binding, upon transshipment pick, a Need to Receive By Datetime (NTRBD) is generated for the item/order pair. The NTRBD represents the time that the item needs to be stowed (not received!) at the source FC so it can be used to fulfill the assigned order. Stowing the item before NTRBD at the destination FC is the definition of success for transshipment. For example, at source FC pick, an item is assigned an NTRBD of 5 PM the next day. If the item is stowed at 4 PM the next day at the destination FC, the transshipment has succeeded, if it is stowed at 5:01 PM the next day at the destination FC, the transshipment has failed.

If an item is not picked in time (exact timing details provided in Section 4.11) the transshipment is cancelled and the customer order is replanned to another fulfillment option. If the item is picked successfully at the source FC, the Amazon system assumes that the item will be available to fulfill orders at the NTRBD at the destination FC. While the inventory is not located in a Prime location (or even in the right FC) at this point in time, the assumption is that the transshipment process will succeed; that is, in x hours, defined by the shipping route/transshipment processes, the item will be in a Prime location at the destination FC before the NTRBD and be available for order fulfillment.

Note that Amazon’s transshipment system attempts to integrate processes across the fulfillment centers into a seamless whole. In particular, extraneous actions in the process are minimized by moving inventory across FCs in Amazon’s picking totes. Items travelling from FC A to FC B are picked into black totes during pick at FC A (identical to standard picking totes except for color; yellow for orders, black for transshipment) and the entire tote is shipped to FC
B. There is no need for inventory repackaging for the transshipment—the picking tote itself is directly transshipped to the destination FC.

4.4 Transshipment Merge/Palletize (Source FC)

Once the item is picked, placed into a tote, and sent forward on the conveyor, it will head to the transshipment merge and palletize area. This step is distinctly different from standard orders—normally totes head to the pack area and individual items removed to prep for shipment to customers. However, as transshipment inventory is sent to other Amazon sites, the totes themselves are shipped out to other Amazon FCs—essentially the tote becomes the shipped package instead of repackaging the inventory. If tote space is not used efficiently, a ‘merge’ process is performed and the contents of multiple totes are combined into fewer totes. Note that inventory is not checked at this process step—the only time when transshipment inventory is counted in the source FC is during the picking process.

Transshipment totes are aggregated by destination and palletized in pallets of twelve totes or sixteen totes. The totes are designed to interlock vertically and are very rugged; assuming the totes are picked correctly, palletizing is simple as stacking totes like LEGOs. Once a pallet is full, it is wrapped up, a pallet label with a list of totes and item quantity is applied, and the pallet
is sent to outbound dock. The pallet label contains the earliest NTRBD of all the items on the pallet so that the destination FC can prioritize which pallets to receive first.

4.5 Transshipment Outbound Dock (Source FC)

![Figure 10: Expanded Outbound Dock Process](image)

Outbound dock aggregates all pallets bound for a certain destination and loads them onto necessary trucks when they arrive. This step is particularly important as all the totes on each pallet were scanned onto a specific pallet and the pallet is “departed” virtually from the source FC. If an item is picked incorrectly, a tote is not scanned in the palletize process, pallet packed incorrectly, pallet put on the wrong truck, or one of many other possible fault scenarios occur, then the physical inventory and the scanned inventory differ and cause transshipment process issues as well as inventory accuracy issues. The time that the truck is scheduled and needs to ship out is known as the critical pull time (CPT)—all items need to be on the truck and ready to ship by this time.

4.6 Transportation (Third Party)

Trucks arranged by Amazon arrive to pick up transshipment pallets and drive to the destination FC. These pick up times are pre-negotiated by Amazon and are consistent day to day—an FC A to FC B route is scheduled to ship from FC A and arrive at FC B at the same time.
every day. Route timings are closely monitored and expected travel times are based on historical average drive time with some padding for traffic.

4.7 Transshipment Inbound Dock (Destination FC)

The destination FC inbound dock offloads the transshipment and either places the pallets in a holding area if receive is full or immediately takes the pallets to the receive area for receipt. The truck is noted as having arrived in the destination FC and all the items in the shipment are acknowledged as to have arrived in the destination FC. While the destination FC has a web portal that indicates how many pallets, how many totes, and which items are present in the shipment, only the offloaded number of pallets is checked at this time.
4.8 Transshipment Receive (Destination FC)

Once the transshipment is ready for receive, two totes from a pallet are scanned and the entire shipment is loaded into the destination FC. When this scan is complete, all details about the shipment are available for lookup within the destination FC. The totes are then sent to the inventory tower for stow. There is still no count or check of items within the shipment.

If an unexpected tote is loaded into the system (due to a mistake at the source FC), the tote will not scan and will be kicked out for troubleshooting.

4.9 Transshipment Stow (Destination FC)

This is the final step of transshipment and is when the transshipment items are moved from non-Prime to Prime inventory. Each tote is moved to a staging area, loaded onto carts, and then...
stowed. Process-wise, transshipment stow is identical to standard inventory stow: an associate
scans a tote, and then for each item in the tote, the associate first scans the item, scans a location
to stow and then physically stows the item. Note that there are stow rules, the most important for
this investigation being that Prime locations can only stow a certain number of distinct items
maximum. This causes extra movements for transshipments as the characteristics of the tote are
dramatically different from standard inventory stow—transshipment totes contain a high mix of
different items and generally have lower quantity. Thus transshipment stow generally has a
longer per unit stow time average as the associate needs to spend more time moving to open stow
space (non-value add) instead of executing the stow.

The time of the item stow must be earlier than the NTRBD for transshipment to be
successful. While the impact of missing NTRBD may not be immediate—the Amazon order
planning system may not ask for the item immediately—this is the earliest potential time that the
order planning system can ask for the item. If the item is not stowed into Prime prior to NTRBD,
then the order may be fulfilled elsewhere, rendering the transshipment useless.

It is important to note that even though the item is “bound” to an order at transshipment
pick, it is not necessarily bound at the end of transshipment. Once the inventory is placed into the
inventory tower and becomes Prime inventory, all inventory becomes fungible. The inventory
system will choose any item in Prime locations based on order timing need, resulting in the
bound item being used for other orders. For example, suppose we have two units of an item, with
unit A bound to order 1 for a NTRBD of 2 PM and unit B bound to order 2 for a NTRBD of 3
PM. Both units are transshipped and unit B is stowed at 1:30 PM—order 1 can request item B for
its order (as it needs it earlier) and force order 2 to use unit A once it is stowed.
4.10 Transshipment Quality

Transshipment quality is measured as the difference between what the source FC indicates was picked and what the destination FC says was actually received and stowed—essentially, what items did the source FC say it sent and what was actually received at the destination FC. There are no standard check steps in between the two processes in order to reduce transshipment cycle time. If an error is found at the destination FC, the inventory at the destination FC is adjusted to match what was stowed. However, inventory at the source FC is not changed. Due to the lack of quality checks within the transshipment process, it is extraordinarily difficult to determine whether an error was caused by the source FC or destination FC as quality errors can occur throughout the entire transshipment process—items are picked incorrectly, shipped incorrectly, or stowed incorrectly. Thus, while the numbers are adjusted at the destination FC to reflect inventory reality, the source FC depends on random cycle count inventory audits to catch transshipment errors. Inventory quality is extremely important to Amazon as poor inventory accuracy results in incorrectly planned customer orders as well as additional clean-up costs from inventory audits.

4.11 Transshipment Timing

The time from source FC pick to destination FC stow on each transshipment route is carefully planned out and monitored. Timing in the Amazon system can be split into two major components—process cycle time and timing checkpoints. Process cycle time represents the planned process cycle time that Amazon expects the process to take given standard labor assumptions while timing checkpoints represent actual times of the day that a process is expected (or needs) to complete. Amazon’s most important goal is the timing checkpoint known as customer promise or Promise Delivery Date (PDD)—the date that Amazon promises the
customer that the package will arrive at their house. For non-transshipped order, timing is shown below working backwards from the PDD:

![Diagram showing package delivery process]

**Figure 14: Standard Order Timing and Deadlines**

Individual timings are described below—ship method, dock time, pack time, and pick times are cycle times while PDD, OCPT and Pick Deadlines are checkpoints.

**Standard Order Cycle Times:**

- **Ship Method** = Time needed for last mile delivery to customer. This cycle time is owned primarily by the third-party, last-mile shipper.

- **Order Outbound Cycle Time (OOCT)** = Amount of time allowed for Amazon order outbound process. This cycle time is 
  \[ OOCT = Pick\ Time + Pack\ Time + Dock\ Time \]

With these cycle times, Amazon establishes timing checkpoints that individual FCs work towards; the FCs are primarily judged on meeting the checkpoints, not the individual cycle times.

**Standard Order Checkpoints:**

- **Customer Promise/Promised Delivery Date (PDD)** = Date Amazon promises delivery. This is a hard checkpoint—if this is missed, there are no possible recovery options.

- **Order Critical Pull Time (OCPT)** = Time when package needs to depart Amazon. This checkpoint is set based off the ship method cycle time and primarily controlled by the third-party, last-mile shipper. Making this checkpoint is crucial—if a package misses this
checkpoint, either PDD is missed or a large upgrade in shipping method is needed (e.g. standard shipping upgraded to next day).

*Order Critical Pull Time (OCPT) = PDD − Ship Method*

**Pick Deadline** = Last possible time when pick can be initiated and still meet OCPT. While there are some recovery options for this deadline (individually walking items to certain parts of the process), for the standard process, this is a hard checkpoint.

*Pick Deadline = PDD − Ship Method − OOCT*

Thus, in order for Amazon to fulfill the Customer Promise an item needs to be available in Prime prior to the pick deadline; e.g., if Amazon promises a customer a May 4th delivery and Amazon ships next day air (Ship Method = 1 day), and Amazon processes take 1 day (Order Outbound Cycle Time = 1 day), the inventory needs to be available on May 2nd for picking (Pick Deadline).

In the case of transshipment the *Pick Deadline* becomes the Need To Receive By Date (NTRBD) of the transshipment. Transshipment timing, working backwards from the PDD can be seen in Figure 15:

*Figure 15: Transshipment Order Timings and Deadlines*

Where Ship Method, OOCT, SLA, Transport Time, and TSOCT are cycle times and PDD, OCPT, NTRBD, TAT, TCPT, and Pick Deadlines are timing checkpoints. Ship Method, OOCT, PDD and OCPT are unchanged from the standard order.

**Transshipment Order Cycle Times:**
Service Level Agreement (SLA): This is the transshipment inbound cycle time at the destination FC and covers dock, receive, and stow cycle times. This cycle time is targeted to x hours at Amazon can be adjusted by assigning more/less associates to the process.

\[ SLA = Dock Time + Receive Time + Stow Time \]

Transport time = Transshipment transportation time

\[ Transport Time = Average Transshipment Arc Transportation Time + Traffic Buffer \]

Transshipment Outbound Cycle Time (TSOCT) = Cycle time of transshipment outbound

\[ TSOCT = Pick Time + Palletize Time + Dock Time \]

Transshipment Order Checkpoints:

Need to Receive By Date (NTRBD) = Time an item in a transshipment must be stowed in order to proceed towards standard order fulfillment. This is the same as Pick Deadline from a standard order shipment. Stow prior to NTRBD represents a successful transshipment.

Transshipment Arrival Time (TAT) = Transshipment arrival time at the destination FC. This is a soft checkpoint—transportation is expected to drop the shipment off at this time, but could drop it earlier or later. If it is dropped off later than expected (late TAT), NTRBD is still expected to be met, meaning a reduction in the standard SLA is required.

\[ TCPT = NTRBD - SLA \]

Transshipment Critical Pull Time (TCPT) = Time transshipment needs to leave the source FC. This checkpoint can be hard or soft depending on the failure—if the item is on the truck but is shipped late (late TCPT), then the item may still be able to make NTRBD. However, if the item does not make the truck that is leaving at this TCPT, it will almost certainly miss the NTRBD and result in a cancelled transshipment.

\[ TCPT = NTRBD - SLA - Transport Time \]

Pick Deadline = Time an item needs to be picked to make it onto the transshipment truck. This is a hard deadline—if this is missed, none of the other checkpoints can be made and the transshipment request will be cancelled.

\[ Pick Deadline = NTRBD - SLA - Transport Time - TSOCT \]

Thus, in order to meet the customer promise (Figure 15 for visualization), a transshipment order must meet its Pick Deadline, NTRBD, and OCPT. Unfortunately, OCPT is a fixed checkpoint, because the last-mile, third-party trucks are scheduled at certain times and Amazon FC processes must work to that time. Similarly, the Ship Method cycle time is constrained by the
last-mile delivery service. While TCPT is also based on a third-party transportation truck, Amazon directly contracts these trucks and is able to set pickup times based on their need. However, the cycle time of TCPT is dictated by the physical distance between the fulfillment centers and the transportation mode. Thus, the ‘easily’ improved times are the SLA, which is driven by the destination FC inbound processes and affects the NTRBD, and the TS Outbound Cycle Time, which is driven by the source FC outbound processes and affects the TCPT and the Pick Deadline.

Transshipment timings are important to the transshipment process as it determines which orders are considered for transshipment. For example, when Amazon promises a customer that a package will arrive May 4th at 8 PM at their house, the timing checkpoints with sample cycle times are as follows (example only):

<table>
<thead>
<tr>
<th>Checkpoint</th>
<th>Cycle Time</th>
<th>Duration</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order PDD</td>
<td>SM</td>
<td>1 Day</td>
<td>May 4, 8 PM</td>
</tr>
<tr>
<td>OCPT</td>
<td>OOCT</td>
<td>5 hours</td>
<td>May 3, 5 PM (Scheduled)</td>
</tr>
<tr>
<td>NTRBD</td>
<td>SLA</td>
<td>5 hours</td>
<td>May 3, 12 PM</td>
</tr>
<tr>
<td>TAT</td>
<td>Transport Time</td>
<td>5 hours</td>
<td>May 3, 7 AM</td>
</tr>
<tr>
<td>TCPT</td>
<td>TSOCT</td>
<td>5 hours</td>
<td>May 3, 2 AM (Scheduled)</td>
</tr>
<tr>
<td>Pick Deadline</td>
<td></td>
<td></td>
<td>May 2, 9 PM</td>
</tr>
</tbody>
</table>

Figure 16: Timing of an Example Shipment

Note that OCPT is scheduled at 5 PM due to the inflexibility of the shipper while TCPT is scheduled right when Amazon requires it, as Amazon contracts the third-party trucking companies. Amazon’s planning system in this case would only be able to transship items that were ready to pick at the source FC prior to May 2, 9 PM. Amazon would prefer to catch as
many orders as possible for transshipment, thus any reduction in cycle time would help increase the value of transshipment.

4.12 Transshipment Improvements

Transshipment transparency, quality, and cycle time were all areas of investigation. In particular, new metrics, new quality controls, and a network-wide item classification change was proposed in order to improve transshipment throughout the Amazon network. However, these changes highlighted incremental changes to the network, not a profound process change.

Amazon is particularly concerned about the cycle time of transshipment. The cycle time of a particular transshipment route is important as it limits the number of items that can be transshipped. In literature, many models assume transshipment is instantaneous in its models; unfortunately, the Amazon transshipment process takes a non-trivial amount of time. Any reduction in transshipment cycle time would result in an increase in the number of orders that can benefit from transshipment.

Based on the transshipment process investigation, one particular area of transshipment seems capable of being shortened: the stow process at the source FC. In particular, spending time stowing a reactive transshipment item when we know that the item is destined to be picked for an order in the near future does not seem efficient. Thus the feasibility of an abbreviated transshipment stow is examined and modelled to provide an idea of the tradeoffs between a shorter cycle time and new process complications.
5 Transshipment Stow Process Discrete Event Simulation

5.1 Current Transshipment Stow Process

The current inbound stow process involves three major steps: Gathering transshipment totes onto a push cart, moving the push cart to open areas in the inventory tower, and then stowing individual items from the totes into available locations in the inventory tower.

The process of gathering transshipment totes onto a push cart is primarily performed by 'cart builders' and is staffed based on the rate of tote arrival from the conveyor. Cart builders pull totes off the conveyor line, build carts of 'like’ totes (yellow for receiving inventory, black for transshipment), and then place the built carts into a staging area in quasi-FIFO order. Transshipment carts are prioritized and always placed at the start of FIFO in order to meet the transshipment cycle time goals of Amazon. The cart building process is not seen as a bottleneck in the transshipment stow process.

The stower picks up the FIFO marked cart from the staging area and then proceeds into the inventory tower to stow items. The stower scans all six totes on the cart and then activates a tote to begin stowing. Multiple stowers work in parallel to stow, a single transshipment could have been stowed by 10+ stowers overall.

Finally, each individual stower moves to an open Prime location in the inventory tower, scans the item to be stowed, scans the Prime location, and stows the item into the location. There are stow etiquette limits for each Prime location—only a certain number of unique items are allowed in a location, only a certain total number of items are allowed, and only a certain total expected volume is allowed.

For the purposes of the discussion and model, the following rates and timings will be used (not actual Amazon values, selected for ease of calculation):
- Transshipment shipment composition: 500 totes with about 5,000 items in total
- Required stow to Prime time: 5 hours
- Cart building time: Instantaneous (buffer of transshipment carts exist)
- Individual stow rate: 100 items per hour

Thus, in order to stow all the items in a single transshipment shipment within a 5 hour time frame, a minimum of 10 stowers would be needed.

5.2 Proposed Transshipment Stow Process

The proposed transshipment process is to completely remove the individual item stow step. Instead, an area dedicated to transshipment would be created and totes would immediately be stowed. This has one primary advantage—items would become nearly immediately available for picking, allowing for a shorter transshipment cycle time. Essentially, instead of loading totes onto carts, moving to an open location, stowing one to two items, and then finding another location to stow the tote, the associate would simply find an empty tote ‘cubby’ and slot the tote into the cubby.

This change would come with a few costs however. First, the turnover of transshipment inventory is not well defined; while associates mention that a majority of transshipment is picked very quickly, there are no concrete numbers behind this assumed knowledge. Second, tote volume usage is thought to be less efficient than standard inventory stow. Third, the amount of extra processes needed to consolidate and clear out empty or low efficiency cubbies and totes are currently unknown.

Inventory volume efficiency in Amazon FCs is extremely important. As the size of the Amazon FC is fixed, the more inventory that the Prime locations can hold, the more items Amazon can offer to its customers. It is assumed that inventory volume efficiency in totes are
less than standard stow inventory efficiency, resulting in a drop in the amount of inventory in the same amount of space. However, if there is a pronounced increased in stow speed as well as high amount of churn in the area, it may be worth it to use the less efficient cubby space. Currently ‘stow’ procedure is done by zones—first one zone is packed as full as possible and then the available stow area is moved to a different zone. Transshipment is believed to have a high turnover; stowed transshipment items are quickly picked, resulting in suboptimal use of the stowing zones in a traditional stow process. Thus by concentrating less volume efficient tote stows in a single area and turning over the area quickly could be better than the current stow process.

The proposed transshipment stow process is a double-edged sword. By performing a tote stow, inventory becomes available in Prime nearly immediately (place tote in a cubby) and the overall number of stow processes decrease as each item no longer needs to be individually stowed. Stow labor is reduced as well, as the number of stow actions decreases dramatically. However there will most likely be increased processes to remove empty totes, consolidate totes with little inventory, and other clean-up activity. The trade-off is essentially faster access to inventory to fulfill customer orders, but work is required afterwards to clean up the transshipment tote stow areas. Currently, the ratio of transshipment cycle time savings versus tote cleanup is unknown. Additionally, while tote stow labor decreases, there may be an impact to picking rates from totes—totes have a higher mix of items and it may be difficult to locate the correct item at pick. While cycle time improvements are very desirable, if the new transshipment stow process requires constant cleanup or has a large impact on picking times, the tradeoff may be prohibitive.
Thus, this improvement boils down to three major process attributes of interest—how much time is saved in the transshipment process, how often is tote consolidation needed, and how is inventory space utilization impacted. Note that this paper will only provide guidelines on the tradeoff between transshipment cycle time savings, increased consolidation efforts, and decreased volume efficiency. Tote stow item pick rates are not examined as the rate impact was not trialed prior to the end of the internship. Whether or not such tradeoffs are desirable in Amazon’s current state will be left to Amazon to pursue.

For the purposes of the discussion and model, the following rates and timings will be used (not actual Amazon values):

- Tote stow rate: 60 totes per hour

5.3 Data

In order to model the current and proposed transshipment processes, a good understanding of the current processes through investigation of Amazon data is needed. In particular, we need to compile data on transshipment quantities and timings. Ideally, actual transshipment data would be used in the transshipment model to provide the closest match to stow processes.

Amazon collects as much data as possible on all of its inbound and outbound processes. Each item state move (e.g. an item is picked from a Prime location to a tote) is tracked within the Amazon system, theoretically allowing for real time identification of any particular item or container anywhere in the Amazon network. In practice however, individual FC real-time data is only accessible through the individual FC’s data portal to minimize the amount of queries to production databases. Once a day, all the production data is scraped into Amazon’s Data Warehouse for long term storage and to facilitate analysis. These FC data tables are further refined into specific analysis tables containing data relevant to analysis.
For this analysis, multiple data tables spanning multiple processes and FC operations were combined into relevant datasets for analysis. A brief list of the primary data fields and field explanations are provided below.

- **Source Fulfillment Center**
- **Destination Fulfillment Center**
- **Shipment ID:** Transshipment shipment ID, specific to all items shipped on a transshipment truck
- **Tote ID:** unique ID to each tote in the system
- **ASIN:** Amazon Standard Identification Number, a code that identifies items in Amazon’s system
- **Item Volume:** Physical volume is calculated by standard Length x Width x Height (LxWxH).
- **Stow datetime:** Date and time that a transshipment item is stowed into a Prime location
- **Pick datetime:** Date and time that a transshipment item is picked (identical item as stow)

Note that this dataset quality is limited by two primary fields, item volume and pick datetime. Item volume data is fragmented throughout multiple databases within Amazon. This dataset contains data from the most complete database and covers ~95% of the examined items. All transshipment items without an item volume are excluded from the analyzed dataset. Pick datetimes were only gathered for 5 days after the stow datetime to reduce query complexity and lookup time. Thus data for item picks after stow are limited to 5 days. Simplifications of pick datetimes will be discussed in the model formulation section.
The initial raw data set contains both reactive and transfer transshipment data, resulting in a dataset of 450,000 individual item shipments.

5.4 Initial Feasibility Study

A feasibility study was performed to determine the turnover of transshipment goods within the system. The contents of eight shipments, five reactive transshipments and three transfers, were tracked throughout the Amazon processes, from arrival, to stow, and finally to pick for customer. This chain of movement was then aggregated and plotted over time. Figure 17 shows the percentage picked from shipments from the NTRBD day and four days after. The pick behavior is clearly delineated between the two types of shipments: reactive transshipments where ~50% of the shipment is picked the same day, and transfers which have a much lower pick rate. For reactive transshipments greater than 50% of the inventory is picked within the day, increasing over the next few days to ~85-90%, indicating high turnover of the items.

Figure 17: Transshipment Item Pick Time after Successful NTRBD Stow
In addition, item volume efficiency in storage was examined in order to determine
differences in volume efficiency between an item stowed in the inventory tower and an item in
transshipment totes. Inventory tower volume efficiency is calculated as

\[
\text{Inventory Tower Volume Efficiency (ITVE)} = \frac{\text{Volume of Items in Pick Tower}}{\text{Usable Space}}
\]

Note that ITVE varies dramatically based on the time period—as Amazon reaches the holiday
season, volume efficiency rises. Usable space is an Amazon defined volume based on the design
specifications of the inventory tower. Transshipment tote volume efficiency is examined at
arrival of the tote at the destination FC and is calculated similarly as

\[
\text{Tote Volume Efficiency (TVE)} = \frac{\text{Volume of Items in Tote}}{\text{Total Occupied Tote Volume}}
\]

There are two caveats with the TVE calculation. First, total occupied tote volume is calculated as
the entire space the tote occupies, e.g. a simple LxWxH calculation based on the footprint of the
tote. This is because with tote stows, the entire occupied volume of the tote becomes reserved.
Second, tote volume efficiency can be greater than 1. Item volume is calculated as LxWxH but
the item in question may be hollow. Thus if multiple small, stackable baskets were transshipped
in the same tote, the volume of items in the tote could be greater than the total tote volume.
While this can also occur in inventory tower volume efficiency, because the inventory tower data
is aggregated over a much larger inventory pool, the probability of ITVE > 1 is lower. Overall
however, ITVE was more efficient compared to TVE; that is, inventory tower space usage was
more efficient than tote space usage. Based on these findings, the following assumptions are used
in the transshipment model (not actual Amazon values)

- Only reactive transshipments are modelled: ~298,000 items in 27,500 totes (down from
  450,000 items)
Inventory tower volume efficiency (ITVE): 66% or an item volume multiplier of 1.5 (1.5 = 1/0.66). This is to adjust item volumes to the volume the item would take in the inventory tower (model numbers only).

5.5 Current Transshipment Stow Model Formulation & Inputs

A transshipment stow model was created performed in SimPy based on a single transshipment route between two fulfillment centers. Essentially, a destination FC expects a single transshipment from a source FC per day. While a destination FC would normally receive transshipments from multiple source FCs per day, the model has been simplified to look at the base case of a daily recurring transshipment from a single source fulfillment center.

A single shipment has been defined for this model as 500 totes per shipment, with the quantity of items in each tote determined by randomly sampling the item compositions of 27,500 totes in the examined dataset. The distribution of the quantity of items in each tote can be seen in Figure 18; note that the y-axis is log_{10} to display the rareness of very high item totes. This model assumes an average of 10 items per tote.

![Distribution of Quantity of Items in Tote](image)

**Figure 18: Distribution of Number of Items in Each Tote**

(vertical axis is Log_{10})
Similarly, the volume of the items in the tote is generated by sampling volumes from the 298,000 available datapoints from Amazon databases. Figure 19 displays the distribution of transshipment item volumes. Note that this technique does not check for a Tote Volume Efficiency of greater than 1; this simplification was done as TVE can exceed 1 in the actual system. A sample of 10 model shipments and their TVEs was validated against the actual TVE from the sample data, resulting in similar TVEs.

![Distribution of Transshipment Item Volumes](image)

**Figure 19: Distribution of Transshipment Item Volumes**

(vertical axis is \(\log_{10}\))

The modelling of the current transshipment process is depicted in Figure 20. The overall model is running in minutes—all rates and timestamps are converted to minutes for time analysis.
The first step in the model is cart building. As discussed in Section 5.1, cart building is not a bottleneck of the stow process thus viewing a single shipment independently, cart building is performed instantaneously (e.g. all items in a single shipment are available to stow at time \( t = 0 \)). If multiple shipments are chained together, the cart building of the second shipment is offset by 24 hours or 1440 minutes. Thus all of shipment 0 can start stow at time \( t = 0 \), shipment 1 at time \( t = 1440 \), shipment 2 at \( t = 2880 \), etc. A simplification to the model needed to be made at this step. Instead of a stower taking a cart with 6 totes and stowing all items in the 6 totes consecutively, the model allows a stower to stow any item in the shipment. Thus, instead of carts with totes being stowed in parallel, the entire shipment is stowed in parallel. This limitation was due to difficulty in manipulating the container object in SimPy.

The second step in the model is the physical stow process. Based on current transshipment process assumptions, 10 stowers are needed to complete transshipment stow prior to the 5 hour stow deadline. Thus the model limits the number of stowers to 10. The rate used for the stower is
an average of 100 items per hour (0.6 items per minute) with a standard deviation of 5% to provide some process variation. Once the item is stowed, the item’s volume is added to the overall system volume.

The third step of the process is the item pick delay where Amazon data is used to inform the overall distribution (not actual Amazon data). The item pick delay distribution is generated based on Figure 17. For this model, an item has a 57% chance of being picked uniformly on the same day it was stowed, 18% 1 day after, 6% 2 days after, 4% 3 days after, and 3% 4 days after. The remaining 12% is uniformly distributed between the next 15 days. Amazon data only extends to 4 days after the stow; the remaining 15 day distribution assumption was made to allow the model to reach equilibrium. The item pick delay distribution for the model is shown in Figure 21.

Finally the item pick is performed. The item pick is assumed to be instantaneous—once the pick action is initiated, the model ends. Additionally, the number of pickers is assumed to be infinite. As the item pick delay is based on the pick completion time, the pick time constraint is
already calculated into the item pick delay. Once the item has been picked the item’s volume is subtracted from the system volume. The total system volume is determined each minute and is then multiplied by 1.5 (corresponding to ITVE = 66%) to find the modeled actual required volume in the inventory tower.

The proposed transshipment stow process model is schematically the same. However, in order to deal with tote stowing, two DES models are run back to back. Additionally, volume calculations are now dependent on the state of the tote, not the item itself.

Figure 22: DES Structure of a Single Shipment in the Proposed Transshipment Stow Process

In the tote DES, step 1, cart building, remains unchanged; all totes are available for immediate stowing. Totes from a different shipment again have their start times offset by 24 hours or 1440 minutes. Step 2, the physical stowing is again split between 10 stowers but stowers stow at a rate of 60 totes per hour or 1 tote per minute. This is equivalent to an item stow average of 600 items per hour based on an average of 10 items in a tote; a dramatic reduction in stow time is expected as the standard item stow process assumes 100 items per hour. A 5% stow time
standard deviation is again applied to provide process variation. Tote volume is immediately added to the system volume upon stow.

In the item DES, the items are available for picking once the tote completes stow. Items are then individually picked based on the item pick delay and the system ends when the final item is picked. Tote volume is only removed from the system when all items in a tote have been picked. There is no volume efficiency multiplier applied to tote volume as the volume a tote occupies is static.

Note that all randomly generated values—tote composition, item volume, stow time, pick delay—were generated using a seeded pseudorandom number generator in Python and stored in a ‘Shipment Data’ file prior to input into the model for repeatability. The DES model takes values fed to the system from a ‘Shipment Data’ file and does not generate any random numbers itself; all randomness is generated and stored in ‘Shipment Data’.

5.6 Results

All plots shown have the same vertical axis increments, unless noted, to allow for like-for-like data comparisons between the two models.

The baseline model was examined by looking at simulation runs for a single shipment. Results for five randomly generated shipments are shown in Figure 23. The current transshipment model has a consistent volume and curvature even with some sampling and randomness added in. Overall, these simulation results highlights the baseline assumptions of the model—the stow volume is relatively consistent even with random sampling, the stow is complete in 5 hours, and the picks should be occurring based on the prescribed probabilities per day.
Figure 23: Current Stow Process Volume Use Over 5 Days

More complex analysis was performed by looking at chaining shipments and examining the volumes the current system reaches equilibrium under the current assumptions. If Amazon dedicates a specific area to transshipment and allows the volume to attrit naturally, how much inventory space would be required? This result for two shipments is shown in Figure 24. Based on the model, volume use reaches steady state after 15 shipments and at maximum, uses ~3.2x a single transshipment's volume.
A further analysis was performed by examining how item consolidation efforts could help reduce the amount of space needed for the current transshipment stow process. This simulation assumes a different pick time model; instead of allowing the item to be picked at its natural pick time as determined by the item pick distribution (and thus leaving the system volume at the pick time), stowed items were consolidated to another location after $x$ days in storage. In the FC this would be represented by slow-moving transshipment inventory; if the inventory has not been picked within $x$ days from the transshipment stow area, a consolidator would be asked to consolidate the item out of the area, freeing up space for higher turnover items. Figure 25 shows the results of the same random-seed simulation under each of the various conditions. The greatest volume gains can be found by consolidating after 5 days when total volume usage drops to approximately 2x shipment volume. However, $\sim12\%$ of a shipment’s items would need to be consolidated per day. If consolidations were performed daily on $\sim45\%$ of the shipment’s items, volume use can be further reduced to 1x the shipment volume.
Figure 25: Current Transshipment Stow Process with Consolidation after x Days

Similar analyses can be done for the proposed transshipment model. The basic analysis was performed by running the same five random seeds through the proposed transshipment model and plotted in Figure 26 (a current process run is plotted as reference). While the proposed process allows for a stow time reduction from 5 hours to less than 1 hour, the new process uses over 2.3x more space compared to the current process. Additionally, an exponential decrease in volume usage is not observed; instead, volume usage decreases linearly and remains high even after five days.
Figure 26: Single Shipment Proposed Process Volume Use Over 5 Days
(a single current stow process in blue is provided for reference)

Extending the simulation timescale to capture the entire stow and pick process in Figure 27 reveals that volumes used for the proposed stow process reach 10x compared to the current stow process for a majority of the 20 day period. Volume decrease in the proposed process is roughly linear throughout the 20 day period.
A 20 day, 20 shipment analysis was performed for the proposed process and is shown in Figure 28. The tote stow process uses more space than the current transshipment process, reaching steady state at \(~7.5x\) the volume of the current process and nearly \(24x\) the volume of a single shipment.
Consolidation was also modelled in the proposed process as seen in Figure 29. The largest volume use improvements can be obtained by doing a consolidation after 5 days—each further day of consolidation decreases the volume required by ~50% of the required shipment volume. Additionally, while consolidating daily would allow the volume use to remain at the proposed processes single shipment volume of 2.3x the current process volume of a shipment, 90% of the totes would need to be consolidated, reducing the value of the tote stow process.

5.7 Discussion

Based on the DES findings the proposed transshipment stow model would almost certainly not be worth implementing in Amazon FCs. While there is greater than 80% efficiency in stow speed and a similar reduction in stow cost, the space requirements from a tote stow dramatically overshadows the gains from cycle time.
Additionally, consolidation times were not directly calculated for the two processes—in the current system no consolidation is required—the item can sit in the inventory tower like standard inventory. If transshipment items were stowed in a transshipment only area, consolidation can be used to reduce the volume footprint needed to maintain that area. However in the proposed process, consolidation labor is required as part of the process in order open up locations for stow. Once tote stow is finished picking an empty tote remains in the cubby and requires consolidation to free up space.

Finally, this model and simulation does not touch on an important aspect of the transshipment—transshipment quality. In the current system, transshipment quality is checked during the item stow. Under the proposed system, transshipment quality would not be checked until either the pick or at a consolidation step. Even a daily tote consolidation would result in a near 24 hour delay to transshipment quality checks.

Overall, while transshipment tote stow would result in shorter transshipment stow cycle times, based on initial modelling, the transshipment tote stow benefits do not appear to outweigh the drawbacks. However, an area dedicated to transshipment stow seems to be much more amenable on the consolidation and volume tradeoffs; while the stow rate would not increase dramatically, the volume required for transshipment could be reduced with consistent consolidation. While some FCs do have an area dedicated to transshipment stow, consolidation efforts are usually done on a threshold basis—consolidate when the area is xx% full—rather than on a consistent schedule. Based on the model results, a consistent consolidation after inventory has been sitting for 5 days would only require a total 2x shipment volume space at steady state for the transshipment stow area or as low as 1x shipment volume required when consolidating daily.
6 Conclusions, Feasibility of Implementation, and Future Work

This thesis provides a comprehensive overview of how transshipment is actually executed in a real-life fulfillment center and provides insight into some of the complexity required in a functioning transshipment system. Through this investigation, a new process to reduce transshipment cycle time and allow for more transshipments was proposed—transshipment tote stow. Transshipment tote stow depends on high inventory turnover—instead of stowing items into the inventory tower to be picked in the near future, the tote itself is stowed and the items removed from the tote. Once the tote is empty, a single consolidation to remove empty totes could be performed. While the proposed process was not physically tested, a model was created to simulate some of the advantages and tradeoffs of the new process within the Amazon FC system. In particular, the time and volume tradeoffs were examined to determine if the proposed process was feasible within current operations.

A discrete event simulator model was generated and run for both the current and the proposed transshipment stow using dummy rates mirroring actual Amazon operations. Overall, while the proposed stow process would reduce stow times by greater than 80%, the increase in maximum required inventory volume was too high at 7.5x the maximum volume of the current stow process. Additionally, the proposed stow process would also delay quality checks in the system—a key part of how Amazon inventory and guarantees customer satisfaction. The volume and quality trade-off seems too steep for an 80% reduction in transshipment stow cycle time, rendering this process highly unlikely to be further pursued.

While additional modeling could be done to make the models more accurate, such as matching the pick distribution more closely to actual picks, the proposed transshipment tote stow process in its proposed form does not seem to be a viable method of reducing transshipment
cycle time. Future work should focus on stowing transshipment in a dedicated area, improving the quality of the transshipment and look for cycle time improvements, or labor reductions in other areas of the transshipment process.
References


