Distribution Network Optimization to Reduce Process Variability and Improve Throughput for an Online Retailer

by

Michael T. Schoder
B.S.E., Princeton University (2010)

Submitted to the MIT Department of Mechanical Engineering and
MIT Sloan School of Management
in partial fulfillment of the requirements for the degrees of
Master of Science in Mechanical Engineering and
Master of Business Administration
in conjunction with the Leaders for Global Operations program
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Michael T. Schoder
MIT Department of Mechanical Engineering
MIT Sloan School of Management
May 14, 2021

Certified by .....
Hermano Igo Krebs, Thesis Supervisor
Principal Research Scientist, Department of Mechanical Engineering

Certified by .....
Stephen C. Graves, Thesis Supervisor
Abraham J. Siegel Professor of Management, MIT Sloan School of Management

Accepted by .....
Nicolas Hadjiconstantinou, Chair
Mechanical Engineering Committee on Graduate Students

Accepted by .....
Maura Herson, Assistant Dean, MBA Program
MIT Sloan School of Management

Distribution Network Optimization to Reduce Process Variability and Improve Throughput for an Online Retailer

by

Michael T. Schoder

Submitted to the MIT Department of Mechanical Engineering and
MIT Sloan School of Management
on May 14, 2021, in partial fulfillment of the
requirements for the degrees of
Master of Science in Mechanical Engineering and
Master of Business Administration
in conjunction with the Leaders for Global Operations program

Abstract

In moving from standard two-day to single-day shipping, Amazon Fulfillment Centers (FC) must stock an increasing variety of product in inventory. Amazon uses a hub and spoke model where Inbound Cross-Dock (IXD) facilities split and ship large quantities of vendor product efficiently to numerous FCs, a process known as transshipment. Depending on the volume of product required for optimal inventory placement at the receiving FC, product may depart an IXD either in its original vendor corrugate case packaging or in an Amazon standard plastic yellow tote if the original case was split apart at the IXD.

Furthermore, case and tote containers may be transshipped either in traditional palletized trailers or in floor loads (trailers loaded directly with stacked cases or totes). In an effort to reduce transportation costs, Amazon is transitioning to using floor loads for an increasing proportion of its North American transshipments. Floor loaded trailers enable higher volume utilization by eliminating the gaps between pallets and the pallet material itself, and also reduce the indirect labor associated with palletizing cases and totes at the IXD, and with breaking down pallets at the receiving FC. A *hybrid* floor load is a trailer that contains both totes and cases mixed together, which provides further flexibility needed to improve product placement and optimize trailer fullness. As Amazon increases IXD throughput and requires each IXD to support an increasing number of destination FCs, hybrid loads have become the norm and continue to increase as a proportion of total transshipments, growing substantially in number between 2019 and 2020.

However, while hybrid floor loads bring several advantages, they also increase the complexity and variability within downstream processes, particularly in inbound freight processing at receiving FCs. In particular, newer Amazon Robotics FCs (2019 generation buildings and beyond) use a process known as *decant*, where product arriving in cases is immediately removed from its corrugate packaging and placed into standard plastic totes to enable uniform processing further down the line. Hybrid loads cause large and unpredictable variability in the inflows of different container types, which results in the decant line often being either over-saturated or starved for work, resulting in inefficient use of labor and lower overall throughput, as well as further compounding effects on dependent downstream processes. Lost labor costs from the decant process alone totaled more than $20M in 2020, and without intervention these costs will continue to increase as Amazon expands its number of decant-enabled sites.

The aim of this project was to investigate the root causes of variability in case flow to the decant process, and to assess different means of reducing this variability in a cost-effective and

sustainable manner. The research presented in this thesis consisted of four stages: analysis of the current processes to understand the problem, developing hypotheses for potential solutions, testing hypotheses through a combination of simulation modeling and onsite testing, and finally analyzing results to present scalable process change recommendations. This thesis presents results of analysis in three distinct but related areas: FC inbound dock processing, trailer receive scheduling, and IXD trailer loading. Results include a set of process change recommendations to improve FC inbound dock operations, as well as an optimization-based trailer scheduling program to minimize variability across multiple dimensions of inbound product flow. Analysis of the IXD loading process leads to the conclusion that changes to this set of operations would not be cost-effective at this time and should not be undertaken given current constraints. The conclusions in this thesis are specific to Amazon's network, but the models and frameworks presented may be generalized to a wide variety of networked logistics operations, including applications in warehousing, container shipping, and supply distribution.

Thesis Supervisor: Hermano Igo Krebs
Title: Principal Research Scientist, Department of Mechanical Engineering

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

Acknowledgments

I would like to thank Amazon for sponsoring this research project, and for providing the access to resources which enabled this thesis. In particular, my project supervisor, David Pistorino, deserves a tremendous debt of gratitude for his guidance and mentorship throughout. Additionally, I would like to acknowledge the continuous help and support provided by Leo Tabilin, Ravi Lote, Zachary Simon, Zack Pezzner, Connor McIntyre, and Jeremy Lieu.

This thesis would never have been possible without the continuous direction, advice, and encouragement of my thesis advisors, Professor Stephen Graves and Doctor Hermano Igo Krebs. Thank you for your tireless support.

To the Leaders for Global Operations Program, I owe my gratitude for the the tremendous learning experiences and for the opportunity to be part of such a diverse and thoughtful community. To my classmates, who have been exceptional teachers and friends, thank you.

Finally, I would like to thank my family for their steadfast support and encouragement.

Note on Proprietary Information

In the interest of protecting Amazon's competitive and proprietary information, figures presented throughout this thesis may have been disguised, are solely for the purpose of illustration, and may not represent actual Amazon data.

Contents

List of Figures

List of Tables

List of Equations

List of Acronyms

ARS	Amazon Robotics Sortable (building type)
ASIN	Amazon Standard Identification Number
AWS	Amazon Web Services
CAGR	Compound Annual Growth Rate
CPH	Cases per Hour
CPSAT	Constraint Satisfaction Problem
DMAIC	Define, Measure, Analyze, Improve, Control
EOL	End of Line
FBA	Fulfillment by Amazon
FC	Fulfillment Center
FL	Floor Load, also Fluid Load
GCU	Gross Cube Utilization
HFL	Hybrid Floor Load
ITS	Inventory Transfer Service
IXD	Inbound Cross-Dock
LTL	Less-than-Truckload
MILP	Mixed-Integer Linear Program
NST	No Stow Turnaway
OOW	Out-of-Work
PID	Parcel Identification Device
RWC4	Robotic Work Cell-4
SKU	Stock Keeping Unit
SP	Small Parcel
TCPU	Transportation Cost per Unit
TPH	Totes per Hour
UPT	Units per Tote
VCPU	Variable Cost per Unit
WIP	Work in Progress

Chapter 1

Introduction

1.1 Amazon's Business Model

Amazon.com, Inc. (Amazon) was incorporated in 1994 as an online book retailer, and has since grown to become the third largest company in the world (by market capitalization, behind Microsoft and Apple, as of 2019). Amazon is the world's largest online retailer and second only to Walmart in overall retail by net revenue. Amazon Web Services (AWS) is also the largest provider of cloud computing services. Amazon's 2019 net sales were $280.5B and net income was $11.58B. Amazon's core businesses have expanded to cover nearly every segment of both online and physical retail, grocery retail and delivery, digital media and content publication, and electronic devices. Amazon's operations are organized into three segments: North America, International, and AWS, which reflect both its management of operations and measures of profitability [3].

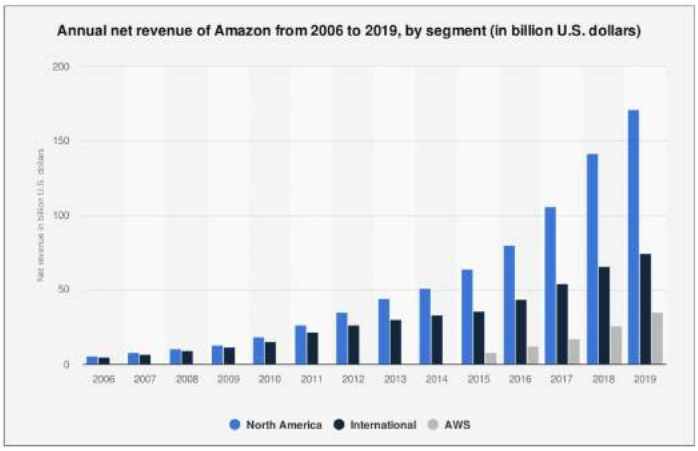


Figure 1-1: Amazon's net revenue growth by business segment, 2006-2019 [6]

In its bid to become "Earth's most customer-centric company," Amazon has focused relentlessly on growth since its inception, with a goal of making the widest selection of products and services at the lowest price possible available to consumers. Amazon's founder and CEO, Jeff Bezos, is credited with the defining the company's growth strategy through the "virtuous cycle" (Figure 1-2), where growth is the self-sustaining engine which enables better customer experience through wider product selection and lower prices. Indeed, customer loyalty is strong, with more than 150M

customers paying $119 annually for Amazon Prime membership [3]. With the acquisition of the Whole Foods grocery chain in 2017 and the inclusion of grocery delivery within Prime, Amazon has greatly expanded its core services and customer reach.

Technological innovation has been central to the company's growth as well. In addition to its leadership in the cloud computing arena, Amazon developed the Kindle, the first popular e-reader in 2007, and has remained at the top of this product category. More recently, in pioneering the Echo smart speaker and the Alexa AI assistant in 2014, Amazon created yet another billion-dollar market and secured a dominant position.

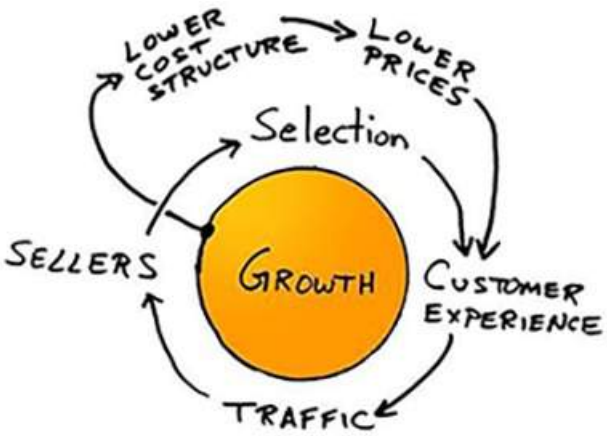


Figure 1-2: Amazon's *Virtuous Cycle* business concept, designed by CEO Jeff Bezos

To support its rapidly growing retail business, Amazon's fulfillment operations network continues to expand at an impressive clip as well. In 2019, Amazon grew its North American footprint by 33.98M square feet, and launched more than 20 new warehouses and sortation centers [2, 3]. Because the vast majority of Amazon's 840,000 employees comprise hourly associates working at these sites to process the five billion orders fulfilled each year, Amazon is investing heavily in automation and process standardization to support future automation goals. The 2012 acquisition of Kiva Systems (now Amazon Robotics, or AR) for mobile automated inventory storage and retrieval, and 2019 acquisition of Canvas Robotics for autonomous material handling are two examples of large-scale transformational investments. Beyond these major projects, however, Amazon's operational leaders continuously seek to implement technical solutions at all scales to reduce manual labor requirements and repetitive tasks.

1.2 Amazon Fulfillment Network

The Amazon fulfillment network constitutes the full set of processes and the end-to-end supply chain that enable retail orders from Amazon.com to be delivered quickly and cost-effectively to customers. Across North America, Amazon has more than 85 Fulfillment Centers (FCs), and an even larger number of smaller special-purpose buildings such as sortation centers and delivery stations. Amazon's distribution network is organized as a hub-and-spoke model, where vendor product is shipped to and aggregated in hubs, known as Inbound Cross-Docks (IXDs), and then transshipped to FCs where individual items are held in inventory until needed to fulfill an order. FCs are further classified into "AR-sortable" and "non-sortable" building types. AR-sortable FCs handle typical small and medium sized items such as books, toys, and housewares, and are able to

rely on a higher degree of automation and standardization. These building types use robotic drives to convey inventory pods across a large storage area to workers performing stow and pick roles at the area perimeter. Non-sortable FCs handle larger and bulkier items such as furniture and heavy equipment. The focus of this thesis is on the IXD and AR-Sortable FC, which together process the large majority of customer orders.

In addition to IXDs and FCs, the fulfillment network also consists of several types of specialty fulfillment centers, optimized to handle clothing, shoes, perishable groceries, and other items. Other building types consist of sortation centers, which consolidate packaged customer orders from various FCs for delivery to more specific locations, and delivery stations, where orders are routed for rapid last-mile delivery. While an important and growing part of Amazon's fulfillment network, this thesis does not explore operations in these building types.

1.3 Problem Statement

In the latest generation sortable FCs, transshipments from IXDs comprise more than 95% of inbound product flow. Because transshipped trailers may arrive in a variety of configurations with different container types, these loads pose a particular problem for the receiving FC, both in planning for and processing the trailers' freight contents while attempting to maintain continuous product flow. Transshipped trailers may be either palletized or floor-loaded, and may contain product in either corrugate cases or in Amazon's standardized yellow plastic totes, shown in Figure 1-3. Particularly problematic are hybrid floor loads, which contain arbitrary configurations of cases and totes mixed together, and must be separated so that individual containers may proceed into distinct process paths after unloading.

Figure 1-3: Amazon standard yellow totes, which can hold items up to 18" long, can be either nested or stacked, depending on orientation.

In the latest generation of sortable FCs, corrugate cases go through a process called decant, where workers remove items from each case and place them into totes, which are then sent to the follow-on stow process where product is moved into inventory. Using decant shifts the corrugate removal process further upstream from stow, and brings benefits in standardization and stow rate improvements. For these reasons, full decant of all inbound cases is the new process standard, and all current and future FC designs include provisions for this process.

Transshipments, and hybrid floor loads in particular, result in disruption to the flow of cases available to the decant line, and to the flow of product in general throughout the inbound department. These disruptions cause "out-of-work" or "dry line" instances where decant and stow workers are not fully utilized due to a lack of available work in their queues. In 2020, two factors have magnified this lost labor problem. First, the number of hybrid floor loads as a proportion of total transshipments is rising rapidly, more than doubling in 2020. Second, numerous new AR-Sortable FCs launched throughout the course of the year, all fully decant-enabled. Based on these developments, the lost labor cost from decant alone was more than $20M across its North American network in 2020, and these costs are expected to grow substantially without changes to current processes.

1.4 Project Research Approach

In examining the decant out-of-work problem, this thesis examines several distinct yet related parts of Amazon's inbound operations – dock operations, trailer arrival scheduling, and trailer loading. Each process occurs in a different physical location and has a separate owner. As such, the analysis for each process is carried out in a somewhat discrete manner, but uses a common set of metrics in decant utilization and labor hours to measure and compare results where applicable.

The general project approach follows the Six Sigma DMAIC (Define, Measure, Analyze, Improve, Control) process, and is split into five constituent phases [18]. In the Define phase, a process flow map describing each process in depth was developed from observations at multiple FC and IXD locations and interviews with managers and workers performing each component function.

In the second phase, Measure, data on all relevant components was gathered. In some instances, large amounts of historical data were readily available for analysis in Amazon's large internal data warehousing system. Other processes required defining new metrics and capturing data through observation. This was the case for estimating trailer unload rate and for the critical measure of decant out-of-work time, where new metrics and data capture methods are suggested.

In the Analyze step, statistical analysis of historical data was performed, and correlations between various process metrics were examined. Simulation models for the FC inbound dock and IXD loading operations were developed and validated against historical data, to be used for exploration of potential solutions to the variable case flow problem.

The fourth phase, Improve, consisted of testing various proposed solutions. Adjusting buffer sizing and labor allocations on the FC inbound dock, and trailer loading constraints at the IXD were all tested through simulation. A series of models were developed to level-load inbound product flows by batching and sequencing trailer arrivals from the FC yard. Solutions assessed to be viable in simulation were either tested during a brief pilot at one FC, or were handed off to a network team for incorporation into related software tools currently in development.

The final phase in DMAIC, Control, is arguably the most important and most difficult; it is the process of making change stick within an organization. While the most promising solutions identified in this project have not yet been fully implemented at the time of writing, feedback was captured from all stakeholders during pilot and testing phases to identify friction points and best practices for larger scale implementation. These findings are discussed, in particular with regards to change management for inbound dock processes.

1.5 Summary of Hypotheses

This thesis examines each of the three major components of Amazon's transshipment process, specifically with respect to hybrid floor loads, and seeks to assess root causes of variability and propose solutions specific to each domain. The initial hypotheses for each component of the transshipment process are summarized below; Chapter 3 provides the formal context for generating these proposals.

1. **FC inbound dock processes:** Improvements to the current trailer unload and decant processes can be identified and standardized to increase visibility of case flow, improve labor utilization, and reduce re-work.
2. **Trailer scheduling process:** Using an optimization-based approach, it is possible to develop a standardized trailer scheduling process for the FC inbound dock that will substantially reduce the variability in container type and item size over the course of a shift.
3. **IXD loading process:** Changes are possible to the container sortation and trailer loading processes within the IXD that result in positive savings when transportation and labor costs in both the IXD and FC are considered.

1.6 Thesis Overview

Chapter 2 begins with a detailed overview of Amazon's inbound operations process. In this section, we begin with a summary of warehouse and automation trends which parallel the rise of e-commerce, and then describe more specifically the relevant stages of the transshipment process between the IXD and AR-Sortable FC. An analysis of current state metrics is given here as well for each of the major processes.

Chapter 3 provides a closer look at the lost labor problem within the FC decant process, and describes the methodology developed to measure the problem and assess its net cost. A root cause analysis is then presented, which assesses variability associated with each component of the transshipment process and estimates its effect on labor usage. Chapter 3 concludes with a description of the hypothesis selection methodology, as well as identification of the proposed solutions that are evaluated in subsequent chapters.

Chapter 4 examines the FC inbound dock process. A discrete event simulation model for the trailer unload and decant process is introduced, and is used to evaluate the effects of altering system parameters such as the case buffer size, unload rate, and trailer load configuration. Results from a short pilot at one FC are summarized, along with physical and human implementation challenges. The chapter concludes with a discussion of further opportunities for process improvement.

Chapter 5 presents a mixed-integer optimization model to select and sequence inbound trailers such that several measures of inbound product flow variability are minimized while physical capacity processing and priority constraints are respected. Results from model back-testing are provided, along with a summary of implementation challenges that resulted in the model being handed off for incorporation into another planning tool.

Chapter 6 considers potential process changes within the IXD, and concludes that in light of current constraints and external demands on the cross-dock system, alternative trailer loading strategies are neither practical nor economical. Transportation costs are shown to be high relative

to labor costs, and current state processes which maximize flexibility and trailer density through the use of hybrid floor loads are cost-optimal.

Finally, this thesis concludes with a summary of findings and a set of recommendations, as well as discussion of related follow-on projects that might lead to further improvements within Amazon's inbound operations.

Chapter 2

Amazon Fulfillment Inbound Operations

2.1 Fulfillment Network Overview

2.1.1 Trends in E-Commerce Supply Chain Operations

The e-commerce retail industry has seen explosive growth over the past decade, and projections indicate that it will continue to grow rapidly. The global e-commerce market is estimated at $3.5 trillion in 2019, and projected to grow to more than $6.5 trillion by 2022 [14]. Even in the US, where e-commerce penetration is higher than the global average, the online retail market accounts for only 10% of total retail sales. With a CAGR of 11.8% [17], the opportunity for e-commerce expansion is still large. Indeed, with the onset of the global COVID-19 pandemic, many sellers have seen significant volume spikes in 2020 which may accelerate a permanent shift toward more online retail shopping for many consumers.

In order to support higher order volumes, more frequent purchasing, and faster delivery time expectations – all trends for which Amazon has been a primary driver – online retailers are experiencing pressure to quickly expand their warehousing footprint and optimize processes to enable greater SKU diversity and faster cycle times. One study estimates that the number of distinct SKUs in warehouses in the US is growing at 18% annually [28]. To cope with the increasing operational complexity and manage labor costs that average between 50-70% of warehouse operating costs [27], many retailers are implementing increasingly sophisticated levels of automation while also redesigning ancillary processes to accommodate the requirements and limitations of automated systems.

A second trend in e-commerce distribution networks is the proliferation of the hub-and-spoke model. For larger retailers with a multitude of warehouse locations, the hub-and-spoke model allows large central distribution hubs to take advantage of economies of scale in transportation and certain processing functions while feeding a wider array of fulfillment warehouses located closer to customers. In a just-in-time supply chain model, the hub locations are often called cross-docks, as these locations are not designed to hold inventory, but rather to facilitate splitting apart large vendor shipments and spreading them across multiple distribution nodes. While this system is somewhat more complex than a traditional supplier-distributor network and requires a larger logistical footprint initially, the hub-and-spoke model offers several benefits. Savings on transportation costs can be achieved by aggregating shipments along spokes to ensure that trucks are fully utilized. Cross-docking can

also reduce lead times by allowing vendors to aggregate product shipments to a single location, which in turn allows distributors to maintain lower inventory levels. Splitting vendor shipments across multiple destination nodes also allows for a higher diversity of product SKUs, enabling faster fulfillment and shorter final-leg shipping to customers. Finally, economies of scale for high-volume, repeatable processing tasks are possible at hub locations, whereas similar automation might not be cost-effective within individual distribution warehouses [12].

2.1.2 Amazon's Inbound Supply Chain

As Amazon continues to expand and standard customer delivery times drop from two-day Prime shipping to one-day and faster, Amazon's inbound supply chain has come under increasing strain to simultaneously grow and become more efficient in order to keep pace. The inbound supply chain must receive product from vendors in a rapid and accurate manner, while minimizing costs. The two primary drivers of change in Amazon's inbound network are shipping costs and product placement – both key to building inventory affordably and in proximity to the customer, which serves Amazon's greater mission of providing the best possible selection, service, and price to its customers.

With these criteria in mind, Amazon began transitioning to a hub-and-spoke model in 2013. Amazon's IXDs are the hubs which serve the spokes (FCs) within its inbound supply chain network. Previous work has directly assessed the effectiveness of this model for Amazon specifically – in his 2012 thesis, Olufemi Oti estimated substantial savings in costs and lead times from moving to the cross-dock model, which Amazon has since done [24]. In fact, in its current form, Amazon's massive distribution network is actually a multi-layered hub-and-spoke model, as in many cases FCs route packaged orders to Sortation Centers which then feed Delivery Stations, all designed to provide efficient and timely local delivery to customers. However, this thesis focuses on the inbound supply chain, which comprises the first two building types, the IXDs and FCs.

Amazon's inbound supply chain begins at the vendor. Product from vendors may be classified as Amazon Retail or Fulfillment by Amazon (FBA). Amazon Retail products constitute the traditional Amazon-sourced and owned inventory. This inventory may come from third party vendors, and also includes items sourced under a private label such as the Amazon Basics brand. Amazon launched FBA in 2006, a service that allows smaller merchants to list their products on the Amazon.com website. Amazon holds FBA merchandise in its fulfillment centers, and manages the customer order, fulfillment, shipping, and returns processes. Amazon charges FBA vendors an inventory fee and also takes a percentage of each sale.

Amazon also offers vendors two transportation options, one where vendors pay for and arrange shipping, and another where Amazon handles these logistics. In the second option, Amazon leverages its large internal distribution network to send vendor freight to the nearest IXD, where it can then be optimally distributed throughout the network via standard transshipment arcs at the lowest cost possible. This is the preferred option for most vendors. In the vendor-managed shipping option, vendors are able to use their preferred shipping carrier, but must bear the cost of sending freight directly to whichever destination FCs Amazon determines to be optimal for inventory placement, usually across multiple locations[1].

In most cases, Amazon prefers to receive vendor shipments at one of its Inbound Cross-Dock facilities. These hubs serve the primary purpose of breaking apart large vendor shipments and

distributing product efficiently to a broader set of fulfillment centers. IXDs each typically serve a fixed number of destination FCs along predefined transshipment arcs. IXDs may disaggregate full vendor truckloads into smaller segments, combine Small-Parcel (SP) and Less-Than-Load (LTL) truckloads, or break apart pallets or even individual cases for distribution to various receiving FCs. IXDs aggregate shipments and typically ship full trailers along each transshipment arc, enabling cost-effective transportation. This type of "proactive" transshipment makes up the majority of freight that FCs must receive.

In other cases, due either to IXD capacity limitations, vendor location, or existing product inventory, Amazon may receive vendor product directly to a fulfillment center. At less than 5% of total inbound freight, freight direct from vendors makes up a relatively small portion of inbound volume for most FCs.

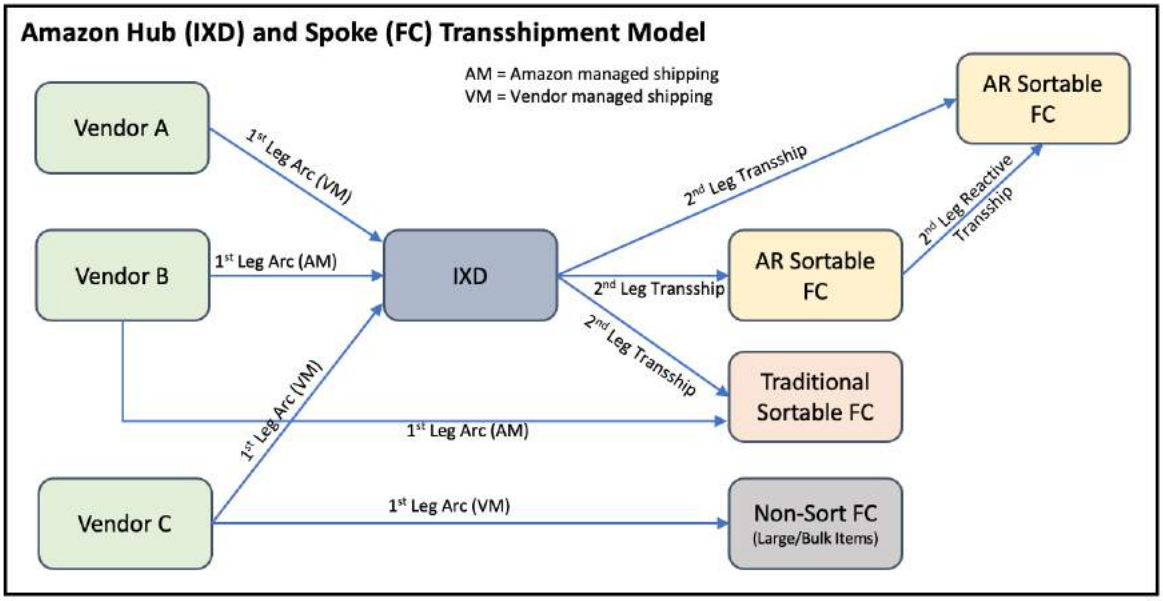


Figure 2-1: Amazon fulfillment network hub and spoke model

Product may also be transshipped "reactively," where after a customer order is placed, Amazon's fulfillment cost algorithms determine that the most cost-effective path to the customer is to fulfill an order by bringing items from one FC to another. This may occur because of FC-specific shipping costs, or in the case of multi-item orders where items are located in different FCs and Amazon can save by consolidating shipments while still meeting the delivery date promise. These reactive transshipments are prioritized upon receipt at an FC, and they arrive in plastic tote containers, as the items have already been picked from inventory at the source FC.

2.2 Inbound Cross-Dock Operations

2.2.1 Cross-Dock Overview

As of August 2020, Amazon operates more than ten IXDs within its North American fulfillment network. These IXDs are spread throughout the region, each serving roughly 30-40 FCs. The number of FCs served by each IXD is growing, however, in an effort to accommodate the

need for more widely distributed inventory. As described in the previous section, IXDs provide a highly automated means of receiving large quantities of vendor inventory and are able to maximize utilization of outbound shipping trailers to distribute inventory across a wide array of FCs at low cost, which in turn enables shorter customer delivery times. The IXD model also provides Amazon benefits in defect reduction and flow control. With individual vendors shipping to fewer locations, a product quality issue can be fixed once at the IXD level rather than at each downstream FC. IXDs enable finer control over inbound volume to each receiving FC by allocating inventory to different FCs based both on demand and on relative inventory capacity. Finally, the IXD model also benefits Amazon's vendors, who generally prefer to make fewer large shipments to a single location, rather than numerous (and perhaps less-than-truckload) shipments to multiple FCs.

While all IXDs perform similar functions, there are design, capacity, and technology differences between each that make complete process standardization difficult. However, every IXD is organized into three major process stages: inbound receive, sortation, and end-of-line, each of which is described in turn below. Freight typically spends very little time in an IXD, normally less than 12 hours. IXDs measure performance by total volume throughput and by variable unit cost, normally calculated on a per-arc, or "shipping lane" basis.

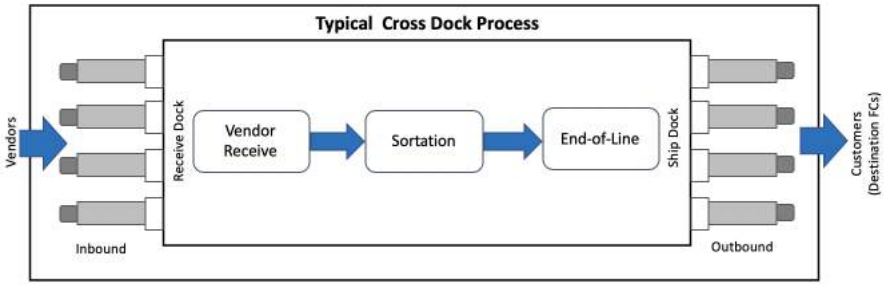


Figure 2-2: Typical inbound cross-dock process flow

2.2.2 Vendor Product Receive Process

Vendor product typically arrives at a cross-dock packaged in a corrugate *case* (i.e. a cardboard box). These cases may be palletized or stacked individually within the shipping trailer. Vendor pallets typically arrive as single-SKU pallets, where an entire pallet consists of a single, uniform product. However, some smaller vendor shipments will mix multiple SKUs on the same pallet; this is termed a mixed-case pallet. Additionally, some single-SKU pallets have no case packaging and are palletized directly in consumer-facing packaging – a typical example would be palletized bottled drinks.

Upon arrival at an IXD, each container is scanned in, which begins the receive process. The purpose of the receive process is to match incoming inventory units to a vendor purchase order. Items can be received at the pallet, case, or individual unit level, depending on how they are packaged. When a container is scanned, Amazon's product placement algorithm, known as the Inventory Transfer Service, determines the required unit quantity, destination FC, and container format. This determination relies on a combination of demand forecast, historical product sales volume, item size and volume, and current stock levels, among other factors. Depending on the source and destination container types, at this point in the process, vendor containers may go through one of the following process paths:

1. Single-SKU pallet remains intact and goes directly to an outbound ship lane.
2. Pallet (either single-SKU or mixed-case) is broken down into cases, followed by one of the below case paths.
   (a) Case remains intact and moves to sorter conveyor, and ultimately to destination lane.
   (b) Case is broken apart and individual units placed into totes, which may go to different lanes.

Figure 2-3: Left: A pallet of totes. Center: Single-SKU pallet. Right: Mixed-case pallet.

In all of the case-level handling processes listed above, IXDs are able to take advantage of costly Parcel Identification Devices (PID) that automate scanning, weighing, and measuring inbound case packages at a high rate. While such devices are generally too expensive to install in FCs, the larger receive scale of IXDs make these cost-effective for these buildings.

The last process listed, where individual units are removed from cases and transferred into totes, is known as decant in FCs and as Each Sortation in IXDs. In the IXD, Each Sortation allows the individual sellable units ("eaches") to be sorted to totes destined for separate FCs. Amazon has deployed some machines to partially automate this labor-intensive sortation process, but much of the process remains manual.

As the problem identified in this thesis centers on variability of container type in transshipments, a natural question might be to ask why IXDs do not simply send all product through the Each Sortation process and ship only totes to FCs. There are multiple reasons for the current set of processing options. IXDs are limited in capacity and sortation machines are expensive, so adding further sortation capacity would not necessarily be economical. Furthermore, because transshipped totes contain loose product of varying sizes, they have a lower item density than vendor-packaged cases, which are typically optimized with space-saving packaging to minimize shipping costs. Amazon incurs a higher per-unit shipping cost for transshipped totes than it does for vendor packaged items, so this additional cost must be balanced against the value of improving inventory placement for a particular SKU. Jeffrey Birenbaum explored these cost tradeoffs thoroughly in his 2018 thesis on optimizing the inbound network transportation cost structure, concluding that substantially increasing the proportion of product leaving the IXD in totes would not be cost effective [10].

2.2.3 IXD Sortation

Once vendor items have been received and moved into a new container if appropriate, the destination containers (cases or totes) are transferred to the ship sorter, a single overhead closed-loop conveyor which spans the majority of the building and includes off-ramps to dock doors assigned to each destination arc. The exception here is for pallets that remain intact; these are moved by forklift to the appropriate outbound buffer queue. IXD dock doors are configured physically for either floor loads or palletized loads; these configurations are fixed as different supporting equipment exists for each door type. Dock doors are typically assigned to a specific destination arc, and this assignment remains constant for at least a full shift, and often several days or more, which allows for smooth transition between consecutive trailers and minimizes changeover time as totes and cases arrive continuously from the ship sorter.

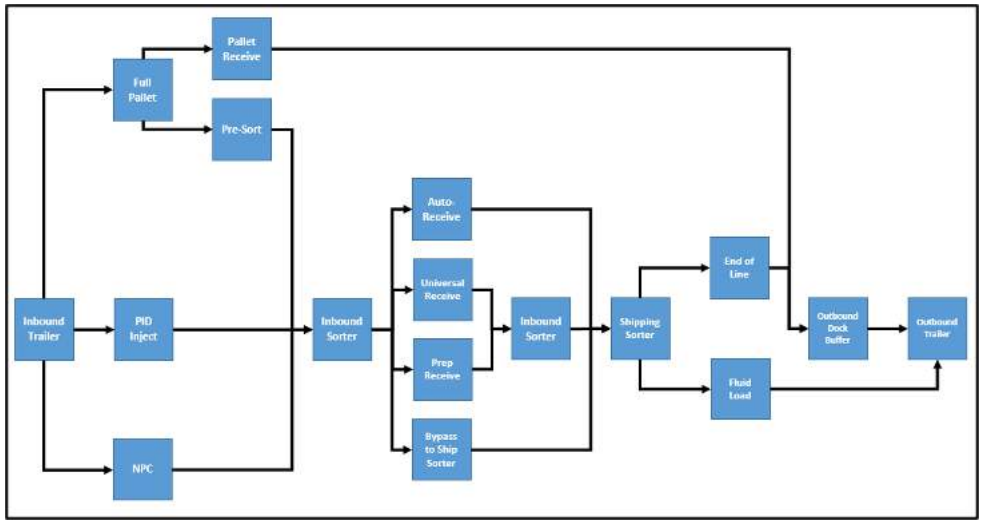


Figure 2-4: Process flow diagram for a typical IXD, depicting multiple processes paths for different container types

While the inventory placement algorithm that assigns a given SKU quantity and container type to a destination FC is by necessity centrally-controlled, lane assignments matching destination arc - container type pairs to specific dock doors are controlled at the IXD level. In general, the IXD will aim to fully utilize the capacity of floor load lanes by assigning higher volume destination arcs to those lanes, which have higher trailer volume density and require less labor than palletized lanes. However, transshipment costs vary widely across destination arcs, so the cost-optimal layout truly depends on the combined per-unit transportation and labor cost. Furthermore, many IXDs have robotic work cells used to palletize totes, so depending on the volume of totes being produced in the inbound Each Sort process, manual labor can be reduced by allocating sufficient tote volume to palletized lanes to keep these robots well-utilized. Vendor freight receive schedules are unpredictable, so projections for case and tote volumes through each transshipment arc are based off of prior week averages. In 2019, Amazon implemented a software solution to solve for the cost-optimal lane assignment at IXDs using a mixed-integer program. This program is run daily, and lane assignments are adjusted when cost savings exceed a given threshold which justifies the time and effort needed to make the transition.

2.2.4 IXD End of Line

Once containers are diverted off of the ship sorter, they arrive at the End of Line process at the outbound ship dock, where all sellable product leaves the IXD in a standard 53-foot shipping trailer. Outbound trailers may be configured in one of two ways: as palletized loads or as floor loads.

Palletized Loading

Palletized loads, as the name implies, consist of any mixture of the three pallet types described previously. A standard trailer can hold 60 pallets in a 15x2x2 configuration where pallets are stacked two-high. Typically, sturdier tote pallets will form the bottom layer and case pallets are stacked on the top layer. FCs also use pallet sleeves, sturdy plastic frames designed to hold loose cases and capable of acting as the foundation pallet layer in a trailer.

Separate process paths exist for each of the three pallet types. Single-ASIN pallets moved directly from the inbound dock to outbound dock are straightforward and require no further processing. Case pallets are constructed at End of Line via a manual process where workers retrieve cases coming down the ship sorter off-ramp, scan the case label to determine its destination, and place the case on a pallet assigned to that destination. Constructed case pallets might be of the simple case pallet or the pallet sleeve variety, depending on the need at the loading dock. This manual process is relatively labor intensive, slow, and error prone, but it is also less constrained by equipment and physical capacity, and so its use tends to vary with the total throughput demand on the IXD. The third variant, tote pallets, are built by stacking 25 totes onto a wooden pallet in a standard configuration. As mentioned in the previous section, many IXDs are equipped with robots designed specifically for sorting and palletizing totes (Figure 2-5). While the robotic tote palletizing process reduces labor requirements, it also creates an additional constraint in flow planning to ensure the robots are well-utilized.

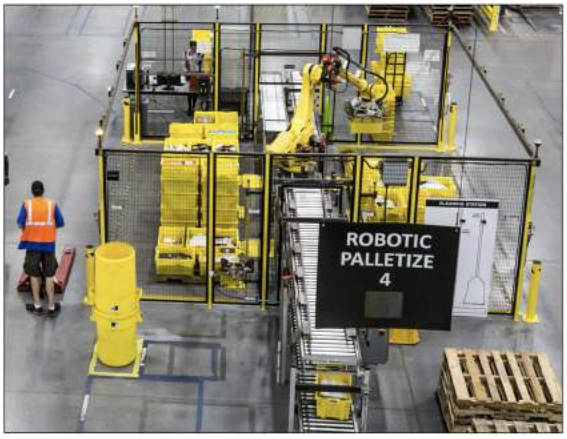


Figure 2-5: The RWC4 robot sorts totes by destination and builds pallets for up to eight distinct arcs. Workers then move the pallets to the IXD's outbound buffer queue for loading onto a trailer. Photo credit: The New York Times.

For all palletized loads, once completed pallets are built, they are manually moved via pallet jack to a staging area. At this point, a dedicated forklift driver moves the pallet to the appropriate

outbound buffer. Once these buffers are full and a trailer is available at the dock door, another forklift will load the trailer, completing the cross-dock process.

Floor Loading

Floor loads, also known as fluid loads, are trailers where individual cases and totes are loaded directly onto a trailer manually, without any pallet or additional structural support. While the actual trailer loading process for floor loads takes longer and requires more touches than loading a palletized trailer, floor loads negate the need for building pallets and the multiple associated staging moves, and so are significantly more labor-efficient as a result. Floor loaded trailers also have higher trailer volume densities, which results in fewer overall trailers required to ship a given amount of inventory. Floor loads may be further categorized into three subtypes: all-case loads, all-tote loads, and hybrid floor loads. All-case and all-tote floor loads are self-descriptive, and hybrid floor loads refer to trailers that contain a mixture of floor loaded totes and cases. While Amazon does not enforce any formal definition of these terms, for the purpose of this thesis, we consider pure tote loads and pure case loads to be trailers that contain 95% or more cases or totes, respectively.

Figure 2-6: Amazon associate building a case wall in a floor-loaded trailer

During the floor loading process, individual cases and totes arrive down a ramp extending from the ship sorter toward the outbound dock door assigned to the appropriate destination arc. Floor load dock doors are configured with an extensible conveyor that can reach the front of the trailer (Figure 2-6), allowing workers to load the trailer while minimizing distance travelled. Workers load a trailer by building a single "wall" at a time, where a wall refers to a complete floor-to-ceiling stack of cases or totes which is only one container deep. Walls may be built as either *case walls*, *tote walls*, or *hybrid walls*. In the case of hybrid walls, totes are required to be consolidated and placed on the bottom, as their neat stacking forms a sturdier base. The special design of totes allows them to be stacked without requiring a lid – the upper tote's bottom edge rests on a small lip just below the top edge of the lower tote. However, when cases are stacked on top of totes, either plastic tote lids or a large section of cardboard is required to protect the contents inside the top layer of totes. This step requires the loading worker to make a trip outside the trailer and incurs some additional labor to secure the lid or separator cardboard.

Containers arrive from the ship sorter in a random sequence, and the dock door workers have no visibility into the relative volume of totes and cases that will arrive during any given time

period. Workers loading trailers will typically segregate totes and cases as they arrive off the line. Because totes are easier to stack neatly and stacks can be slid across the floor by a single worker, totes are stacked along the trailer wall adjacent to the conveyor until enough totes have been set aside to build a section of tote wall. Depending on the relative volume of cases and totes coming down the line, the trailer loaders will make a determination as to the height of the particular tote wall. A tote wall might be anywhere between one and ten totes high, although wall heights between five and ten are most common, and full-height walls that are ten totes high are by far the most common. The maximum number of totes that will fit in single tote wall is 60 (6 across x 10 high). If the tote wall is less than ten totes high, cases will fill in the remaining space in the wall while totes are again buffered adjacent to the line in preparation for the next wall. However, the additional work required to separate totes and cases in each wall with tote lids or cardboard creates a strong incentive for workers to build the tote walls up to ten-high when possible to avoid this step.

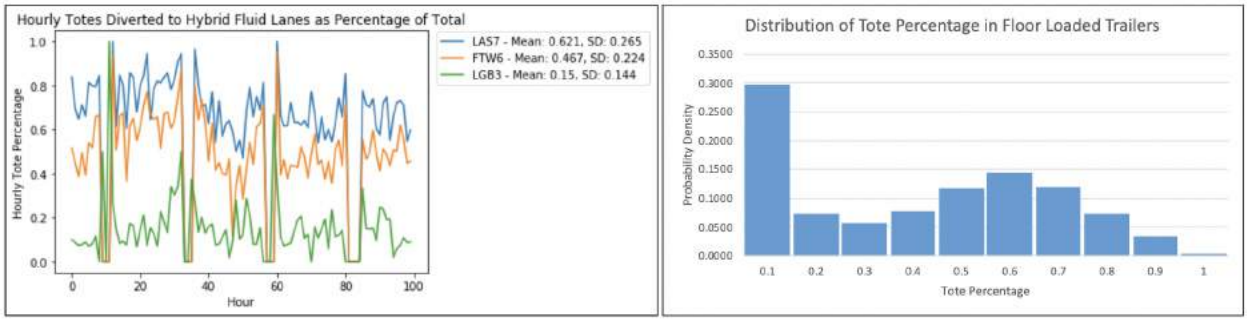


Figure 2-7: Left: Hourly distribution of totes as a percentage of total containers sorted to three separate destination lanes. Hourly variability is high, while weekly ratios are quite stable. Right: Distribution of tote percentage among floor loaded trailers, based on a five month dataset of IXD transshipments in North America

From the distribution of totes shown in Figure 2-7, nearly 30% of floor loads have historically consisted of all or nearly all-case loads. In the past, some IXDs have allocated two floor load trailer doors to particularly high volume destinations, with separate dock doors for cases and totes. This makes work easier for both employees loading and unloading trailers. However, with the more recent push for IXDs to expand the number of destination nodes served, these dual door allocations are no longer tenable, and each door must be assigned a separate destination, so the proportion of pure-case floor loads is expected to drop in the future.

It is worth mentioning that in the past, tote walls have been subject to varying height limits (first five totes, and later eight). This was intended to facilitate tote unloading and to mitigate safety concerns around unloading trailers where containers have shifted during transit. However, as IXDs continue to produce more totes to increase diversity of product placement, the flexibility to increase tote wall height for many high tote volume arcs was deemed essential to maximizing trailer density and reducing transportation costs. In 2019, a review of loading and unloading practices was conducted, and a new unloading safety practices were implemented, allowing the tote wall height restriction to be lifted entirely.

2.2.5 IXD Performance Evaluation Criteria

Within Amazon, IXDs are evaluated quarterly based on their reported individual profit and loss statements. IXDs are cost centers within Amazon and do not directly contribute to revenue.

Component profit factors include network cost savings on avoided fixed assets such as additional conveyance and scanners in FCs, as well as savings on avoided first leg transportation costs. On a daily basis, however, performance is monitored through the variable cost of operations – cost per unit transshipped, which can be broken down into component labor and transportation costs.

Labor Costs

As described previously, the IXD labor required to process and transship a unit of product depends greatly on the specific process path and container type for that unit, which in turn depends on destination allocations and several IXD capacity constraints. This variable labor cost is typically aggregated at either the IXD level or the specific shipping arc level, and is considered to be one type of variable cost-per-unit (VCPU). Labor requirements are substantially higher for the palletized transshipment process due to the additional product touches required to sort, palletize, and transport containers arriving at End of Line.

$$\text{VCPU}_{\text{IXD,labor}} = \frac{\sum \text{Labor Hours} \times \text{Hourly Labor Rate}}{\sum \text{Units Processed}} \tag{2.1}$$

Transportation Costs

Transportation costs for transshipments between IXDs and FCs, or second leg transportation costs make up the second component of the IXD's variable cost metric. Transportation cost-per-unit (TCPU) is a function of the lane cost, the all-in trailer shipping cost from the origin IXD to the destination FC, and the unit count, the number of individual sellable items contained in a given trailer. Lane cost is primarily a function of the arc length, the distance from IXD to destination FC. Amazon regularly updates lane costs, and typically the average quarterly cost of a trailer for each arc is used for computation. Unit count is easily computed for each departing trailer based on the known contents of each loaded and scanned case, tote, or pallet. Unit count depends on multiple factors, including item size, load type, and trailer fullness. Like $\text{VCPU}_{\text{labor}}$, TCPU is commonly aggregated by lane or across the entire IXD over a given time period.

$$\text{TCPU}_{\text{transship}} = \frac{\sum \text{Lane Cost}}{\sum \text{Unit Count}} \tag{2.2}$$

Item size also varies substantially between trailers, but it is an extrinsic factor outside the control of the IXD. Therefore, another metric, trailer volume utilization, is also tracked to evaluate operational performance at the IXD. As Amazon collects detailed dimensional data for all products, trailer volume utilization is simply given as the sum of individual unit item volumes in a trailer divided by the total trailer volume:

$$\text{Trailer Volume Utilization} = \frac{\sum \text{Unit Volume}}{\text{Trailer Volume}} \tag{2.3}$$

IXDs have set targets for trailer utilization, which differ by load type. Figure 2-8 provides an aggregate view of six months of transshipment data, showing how trailer utilization varies both by tote ratio and trailer load type. Totes are typically less dense than cases, and floor loading allows denser packing of containers into trailer than does palletized loading. In general, floor loaded trailers are allocated to the highest cost lanes, which takes advantage of the greatest per-unit transportation cost savings.

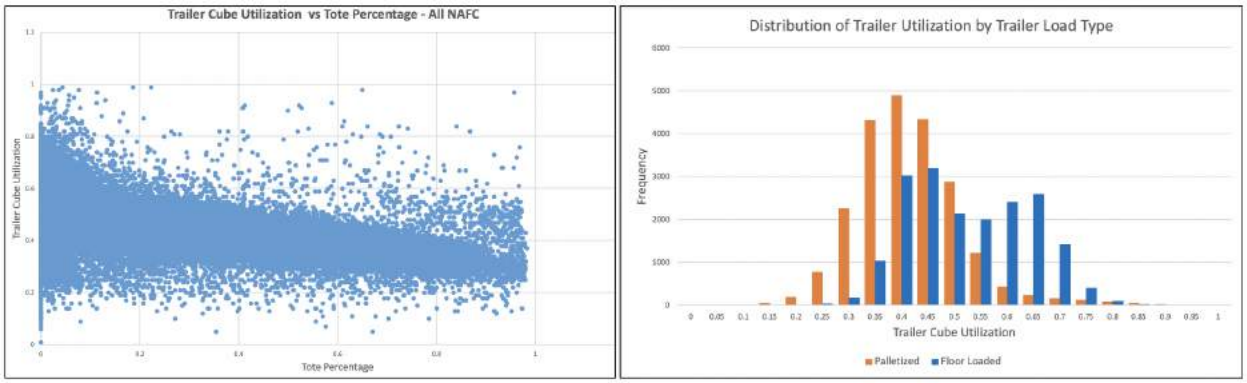


Figure 2-8: Trailer volume utilization decreases as tote ratio increases (left). Trailer volume utilization is higher for floor loaded trailers than palletized trailers (right).

2.2.6 Summary of IXD Operations

This section provided a rough overview of the major processes, decisions, and metrics required to understand Amazon's IXD operations. The process interactions and decision trade-offs are numerous and complex, but the key takeaways relevant to this thesis are listed below:

1. Floor loaded trailers are more cost effective than palletized trailers. This is true for both transportation cost and labor cost.
2. Generally, transportation costs dominate labor processing costs at the IXD and FC, unless the transshipment arc distance is very short.
3. IXDs have a limited capacity for floor loaded trailers, constrained by floor load configured dock doors and trailer loading time. Palletized load capacity is more flexible and scales easily with additional labor because the actual trailer loading time when using forklifts is much faster.
4. Hybrid floor loads are an increasing necessity as IXDs begin to serve a growing number of distinct transshipment arcs, which is driven by the need for more widespread product placement resulting from Amazon's increasingly fast delivery promise times to customers. Due to physical dock door constraints, IXDs can no longer allocate separate tote and case floor load doors for a given arc, meaning that these shipments are converted into hybrid load doors.
5. The IXD floor load dock area has very limited space with which to stack cases or totes in a buffer area. This limits the ability of the trailer loaders to make configuration changes to hybrid-loaded trailers.

2.3 Amazon Robotics Fulfillment Center Inbound Operations

Product shipments (either direct shipments from vendor or transshipments from an IXD) arriving at a fulfillment center mark the beginning of the FC inbound process. The FC inbound department is responsible for all operations starting with trailer receive and ending with product being stowed into inventory. The inbound process is the primary focus of this thesis, and specifically, we are concerned with the newer (2019 and later) Amazon Robotics Sortable type FCs. In these building types, transshipped freight from IXDs makes up more than 95% of incoming product, so we will largely ignore the separate processes associated with the other 5% of product arriving as vendor shipments.

Amazon robotic sortable buildings are characterized by their large inventory areas, where inventory is held in four-sided "pods" in an extremely dense configuration (Figure 2-9). Small autonomous robotic drive units slide under a pod and bring it to the edge of the inventory floor, to either a stow station, where received inventory is placed into a pod, or to a pick station, where inventory is removed from inventory to fill a customer order. A "floor" in this context refers to a contiguous robotic inventory area, and multiple inventory floors may exist on a single physical FC building floor level. A dedicated algorithmic system exists to orchestrate the continuous movement of pods to the appropriate stations along the floor edges.

Figure 2-9: Amazon robotics inventory floor (right), where robotic drive units (left) transport densely-packed inventory pods to stow and pick stations located along floor edges. Photo credit: The New York Times.

The physical layout of Amazon robotic sortable FCs varies, but a typical newer variant FC might be three stories tall, with inbound and outbound docks on the ground level, along with some processing and sortation functions for both inbound and outbound departments. Some FC variants have two separate inbound docks at either end of the building to accommodate the downstream stow stations, which are also located along building edges. A typical ARS FC might have five inventory floors located across three levels, typically one on the ground level and two each on the second and third levels. In addition to the docks and FCs, several miles of conveyance exists within each FC

to transport product between various processing steps.

Operations managers perform day-ahead planning to allocate labor and work for the entire inbound shift. Planning begins with a shift volume target, given as total units of product to be stowed into inventory. This target is provided to the FC from the Central Flow team, and is a function of several variables, including estimated outbound demand, estimated labor available, scheduled inbound product shipments, and the volume of not-yet-received freight dwelling in the trailer yard. From the volume target, managers allocate labor across the various inbound functions, and also schedule trailers to be received based on load type, content characteristics, and FC-specific capacity constraints. Figure 2-10 provides an overview of the primary inbound department processes.

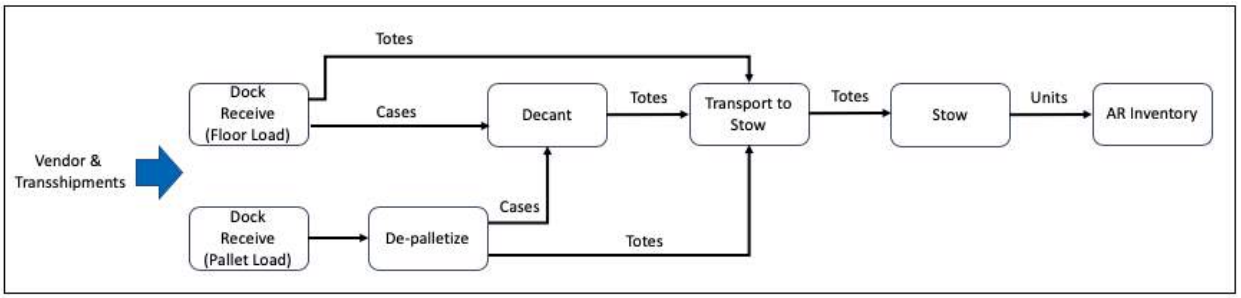


Figure 2-10: Amazon Robotics Sortable FC inbound processes overview

Within the inbound department, Amazon classifies labor functions as either '*direct*' or *indirect*. Direct labor typically refers to functions that are *in path*, or are directly required to move product into stowed inventory. These functions include decant and stow. Indirect functions include processes such as transporting pallets or carts of totes between locations, replenishing empty totes at decant, or operating the transport elevators. This distinction is somewhat arbitrary, and the main indicator of a direct function is that it is *rated*, meaning that Amazon is able to aggregate rate data to monitor the overall line throughput for particular operation against an expected target, typically measured in cases/totes per hour or units per hour.

Figure 2-11 provides a more detailed process flow map showing the specific components of the inbound department functions at a 2019-series ARS FC on which much of the analysis in this thesis is based. Flow paths for different trailer and container types are detailed up to the stow stage, where product is placed into inventory marking the end of the inbound process. The remainder of this chapter describes the specifics of each stage shown here.

2.3.1 Inbound Freight Scheduling

Transshipped trailers arrive sporadically throughout the day, and are staged initially in the trailer yard, adjacent to the fulfillment center. Inbound trailers may be classified as either *live loads* or *drop loads*. Live loads occur when the trailer is not detached from the truck cab, and the driver waits while the trailer is unloaded. These are rare occurrences among transshipments, but must be prioritized due to service-level agreements with the shipping providers. The majority of transshipments are drop loads, where a driver will deposit a trailer in the yard, and then proceed with their next scheduled load, often a pickup at one of the FC's outbound dock doors.

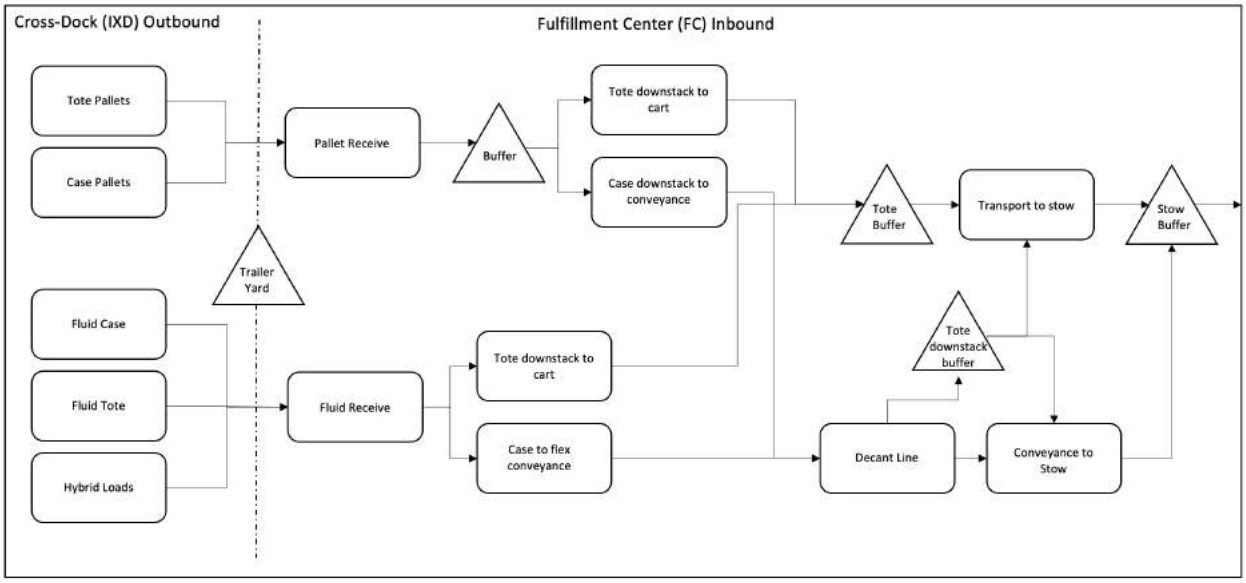


Figure 2-11: Generalized process map for an ARS fulfillment center inbound department

Inbound shift managers have a view of all trailers dwelling in the yard at any given moment, as well as trailers scheduled to arrive during the next shift. Managers can also access a variety of data about the contents of each load – the total number of units, size classification[1] mixture, and number of pallets, cases, and totes. Furthermore, Amazon assigns a standard metric known as a *priority score* to each inbound trailer, which is continuously updated every two hours until the trailer is unloaded. Priority score is a function that applies weightings to the trailer contents over a variety of attributes, including the number of units already allocated to a customer order (reactive units), trailer dwell duration, product backlog demand, product sales velocity, and special circumstance weightings for essential items such as COVID-19 protective equipment, for example.

Priority score is intended to be the primary factor that determines the order in which trailers are received, although in practice, the scheduling process is more complex. Inbound managers must consider physical building constraints such as the number of floor load and palletized dock doors, labor processing constraints, and stow demand by item size at each AR floor. Additionally, in buildings with split inbound docks, each dock feeds a subset of disconnected AR floors on each side of the building, and different constraints will apply to each side. Finally, with the implementation of full-decant in the newest buildings, an additional requirement exists to maintain a sufficiently stable inflow of cases to keep decant workers well-utilized. This is a complex resource allocation problem that is still performed manually, and the process varies substantially by manager.

In addition to the obvious difficulties inherent in performing this process without a well-defined standard procedure, trailer scheduling is also subject to a "cherry-picking" problem, where individual shift managers are incentivized to select the "best" trailers available from the yard to improve certain metrics during their shift. For example, trailers with a high percentage of smaller items will translate to more items per tote, which will make it easier to keep the stow process adequately fed. Additionally, stow rates and decant rates are higher with smaller items, and it is therefore

[1]Amazon uses a 4-size classification standard for product units. Sortable FCs process units classified as small or medium, which fit in a yellow plastic tote and are no longer than 18" and no heavier than 25lbs. Non-sortable buildings process large and heavy/bulky size classes.

easier to hit overall shift volume goals. Another example occurs frequently with single-ASIN pallets. These items often are palletized in their sellable packaging with no outer case corrugate case to be removed, so when these items arrive at decant, they require only to be scanned and placed in a tote. Single-ASIN pallets will often be held aside to be used as a means of artificially boosting decant rates if otherwise flagging, and so trailers with a high proportion of these pallets tend to be promoted to earlier in the receive schedule.

There has been some effort to standardize inbound trailer scheduling as part of a wider automated labor planning system. This type of automated planner has been in place in Amazon's FC outbound departments since 2018, and has been successful in reducing variability associated with manual decision-making. The inbound department, however, poses somewhat of a greater challenge. Many inbound labor functions are not rated, and there exists a tendency to shift labor more frequently between functions in inbound. Furthermore, inbound freight is inherently more variable along numerous dimensions: arrival times, container type, load type, and item size. Figure 2-12 shows the aggregate freight receive volumes for each shift quarter broken down by item size and container type for one representative shift. The shift quarter (roughly 2.5 hours) is the most granular labor planning increment, and the high variability between shift quarters suggests that there is opportunity to improve scheduling to level-load the volume and types of inbound work.

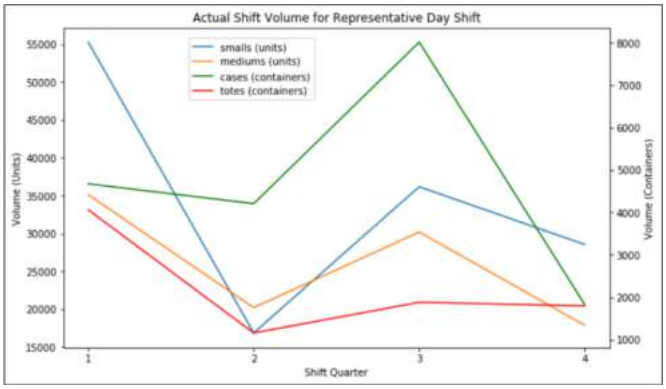


Figure 2-12: Combined transshipment receive volumes of cases, totes, small and medium units by quarter for a representative shift. The variability metric is computed as the sum of absolute deviation from the shift mean during each quarter, for each of the volume types.

There is no standard metric for quantifying the variability in product inflows to an FC over time, so Equation 2.4 is proposed as the measure to be used in this work. Variability is measured across two dimensions of inflow – container type and item size. The variability metric (denoted as V) is computed as the sum of absolute deviation from the shift mean during each shift quarter. Shift quarter is chosen as the most granular discretization of time because a shift quarter (2.25 hours) is approximately the amount of time required to unload a full hybrid floor loaded trailer. Any smaller time interval would be less meaningful as the contents distribution of specific floor loaded trailers is not known.

$$\text{Shift Variability} = V_i = \sum_{q=1}^{4} |\bar{x_i} - x_{i,q}| \tag{2.4}$$

where $i \in \{\text{cases, totes, smalls, mediums}\}$

q represents the shift quarter $\in \{1, 2, 3, 4\}$

$\bar{x_i}$ is the daily average number of units of type i

$x_{i,q}$ is the number of units processed of type i during quarter q

Amazon is in the process of developing an automated inbound planning system, to be known as Online Shift Planner. The scheduling module in the most recent development version prioritizes allocating labor first based on assumptions about worker availability, FC capacity, and shift volume goals, and then schedules trailers accordingly based on priority score alone. There are two problems with this approach. First, inbound freight is allocated based on planned staffing levels, but those levels assume an even distribution of freight type through each process path. As was shown in the previous section (Figure 2-7), the distribution of totes within trailers varies substantially. Second, this scheduling method does not account for variability in item size. The next sections will detail the relationship between item size and processing rates, as well as the variation in item size between different transshipment trailers.

2.3.2 Inbound Trailer Contents

Trailer Load Type

To assess the impact of inbound freight scheduling on all downstream processes, it is important to understand the content distribution of inbound trailers. As discussed in the previous section on IXD loading process, transshipment trailers are loaded either as palletized loads or floor loads. The distribution of load types varies slightly for each FC, but most AR Sortable FCs receive trailer mixes similar to the network breakdown shown in Table 2.1. The significant takeaway here is that floor loads are growing rapidly as a proportion of total transshipments, and that hybrid floor loads in particular are becoming the dominant type of floor loaded trailer. This change is driven in part by a conscious decision to maximize the use of floor-loaded trailers to reduce shipping costs, based on previous cost analysis noted in Section 2.2, as well as the merging of previously dedicated case and tote floor load lanes within IXDs to increase overall floor load throughput given existing lane constraints.

Trailer Type	2019	2020 (Jan-Aug)	% Change
Pallet Loads %	84.30	62.60	-25.74
Floor Loads %	15.70	37.40	138.22
Hybrid Floor Loads %	12.40	34.03	174.40
Pure Case Floor Loads %	3.30	3.37	2.09

Table 2.1: Breakdown of North America network transshipment trailers by load type between 2019 and 2020 at Amazon Robotics Sortable FCs. Floor loads are further classified as either hybrid floor loads or pure case floor loads.

Figure 2-13 shows the relative proportion of load types for one representative FCs over time.

Trailer arrivals are defined as the time when the trailer is logged into the yard. Trailer receive time is the time at which the trailer is docked and processed, which may occur anywhere from several hours to several days later. The spikes in hourly trailer receive volumes correspond to the beginning of shift segments or quarters, where new trailers are typically scanned in following a break and will be processed throughout the next time period of work.

Figure 2-13: Average number of trailers arriving and received by a representative ARS FC, shown by hour and day of the week, and categorized by trailer load type. Top row shows distribution of arrivals (when trailers arrive at FC). Bottom row shows trailers received (when trailers are processed at a dock door).

Container Type

All inbound transshipments contain inventory units packaged in either case or tote containers. Between 50-60% of the total daily inbound product volume (measured as a percentage of total sellable units) arrives in cases, which must be decanted. On average, palletized trailers contain 63% tote pallets and 37% case pallets (percentage of total sellable units, including both single-ASIN and mixed-case pallets).

Of particular interest for this thesis are the floor loads, of which the majority are hybrid floor loads, containing both cases and totes. The distributions of totes and cases in hybrid floor loads is highly variable, as shown in Figure 2-14. There are a few insights worth noting when looking at the distribution of container types in hybrid floor loads. Case volume varies more uniformly, concentrated at the 1000-1800 case range, but is also highly dependent on case size, which also varies substantially by trailer. The tote volume distribution is bi-modal, with distinct concentrations around the 120 and 1100 tote marks. This separation is a function of the trailer load type and container type decisions made at the various upstream IXDs, and allows us to classify hybrid loads

as either "tote-heavy" or "tote-light". We should also note that while approximately 9% of floor loads are pure case loads, there are virtually no pure tote floor loads. All-tote palletized loads are common, however.

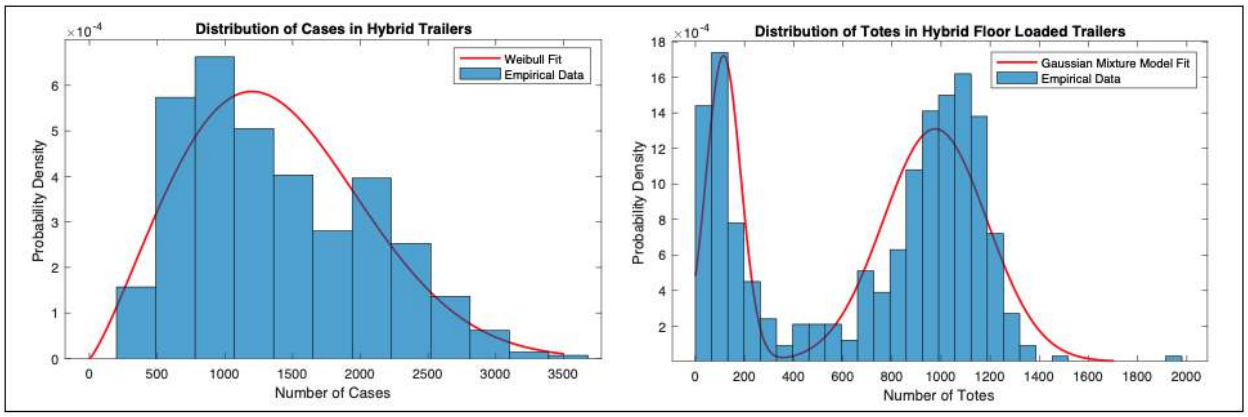


Figure 2-14: Distribution of cases and totes in hybrid floor loads arriving at a 2019 series ARS FC

Item Size

The previous section on inbound freight scheduling introduced the concept of item size classifications, where sortable type FCs are capable of processing small and medium item sizes that fit within a standard tote. ARS FCs process approximately 55% small units and 45% medium units on average. Figure 2-15 shows the distribution of small size units by day and by trailer. While the unit size composition of individual trailers varies widely, it stabilizes within a much narrower range on a daily basis. This insight is central to the trailer scheduling improvement hypothesis advanced in this thesis–when multiple trailers are received and processed simultaneously, there is opportunity through trailer selection to level-load average item size.

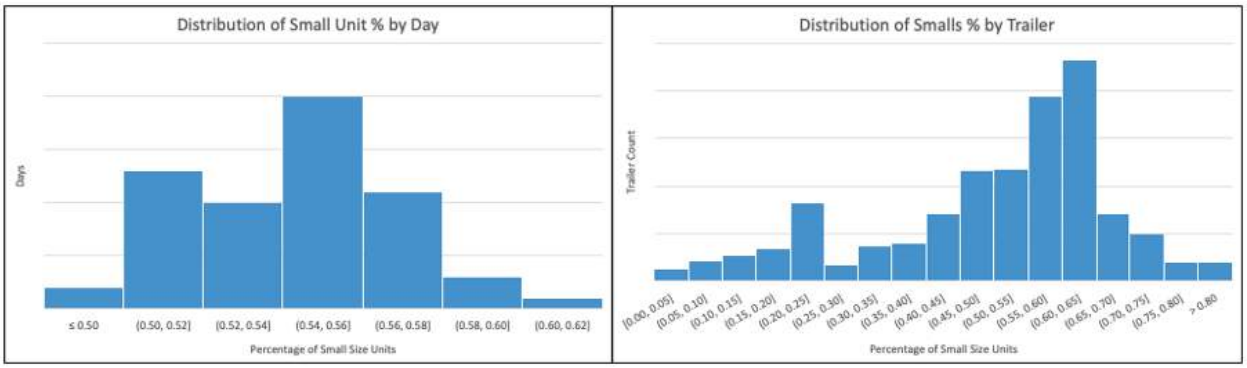


Figure 2-15: Distribution of small size units by day and by trailer received at a 2019 ARS FC. Since medium size units are the only other size category processed at ARS buildings, the medium size distributions are complimentary.

2.3.3 Trailer Unload Processes

Once a trailer arrives at an inbound dock door, the trailer is scanned, at which point its contents are virtually moved to the "receive" bucket, meaning product units are inside the building

and on the way to being stowed. Trailer unloading begins at this point. Dock doors are segmented by trailer type, as palletized trailers require heavy equipment (i.e. PIT, or forklifts), which may pose a hazard to pedestrian workers in proximity to the floor-loaded trailers. Inbound docks will also include one or more buffer areas where pallets are staged between unloading and processing (Figure 2-16). Floor-load doors are typically configured with a combination of flexible conveyance and carts onto which workers will unload cases and totes, depending on the required downstream process path and FC capabilities.

Trailer unloading is considered an indirect labor role at Amazon, and is not a rated function. Performance metrics such as unload rate are notoriously unreliable, because dock workers are constantly moving between trailers and even between adjacent roles throughout the course of a shift. Dock workers are lumped into an indirect labor pool with other supporting functions, such as cart and pallet runners (*waterspiders* in Amazon parlance) and empty tote replenishers. Overall labor utilization, or hours in relation to unit volume processed, is a key performance metric tracked within the inbound department, and translates directly to variable cost per unit. A consistent high level goal is reducing indirect labor as a proportion of total hours worked.

Palletized Unloading

Palletized loads are unloaded by forklift, where a single operator moves pallets one at a time from the trailer to the pallet buffer area. A skilled operator can unload a full trailer in less than an hour. From the pallet buffer, an associate will move the pallet to its processing area. Case pallets (either single-SKU or mixed-case) are normally brought to the side of the flexible conveyance feeding the decant line. Cases are injected into the line, either as buffer material to supplement the flow of cases from floor loads, or as the primary source of cases when the volume of palletized cases is much higher than the volume of floor-loaded cases. This process of moving case pallets from the dockside buffer to the decant line buffer is known as the buffer replenishment process, and due to the slow manual process of transporting pallets via pallet jack one at a time, buffer replenishment can be a significant bottleneck for inbound flow.

Pallets containing totes are transported directly to the stow stations, which may occur either by conveyance or by VRC (a large freight elevator). Some FCs are required to de-palletize totes before transport to stow – this happens for injection onto conveyance, or when the FC uses carts rather than pallet jacks to transport totes between the dock and stow buffers.

Floor Load Unloading

In decant-enabled FCs, floor loads are unloaded at dedicated dock doors with the help of powered flexible conveyors. Cases are unloaded onto the conveyors and transit a short distance directly to the decant line. Totes are unloaded and then stacked onto either a pallet or cart for manual transport to a stow area or to a dedicated conveyor injection point that will take the totes to a stow area, depending on the FC design. As shown in Figure 2-17, the flexible conveyance works well and allows for a continuous flow of cases out of pure case or case-heavy floor loaded trailers. However, in hybrid floor loads with higher tote volumes, removing tote walls leads to significant disruptions in the flow of cases out of these trailers to decant. To mitigate these disruptions, palletized cases are typically used as buffer material, and staged alongside the flexible conveyor feeding the decant line. This allows dedicated associates to feed the line when the flow of cases out

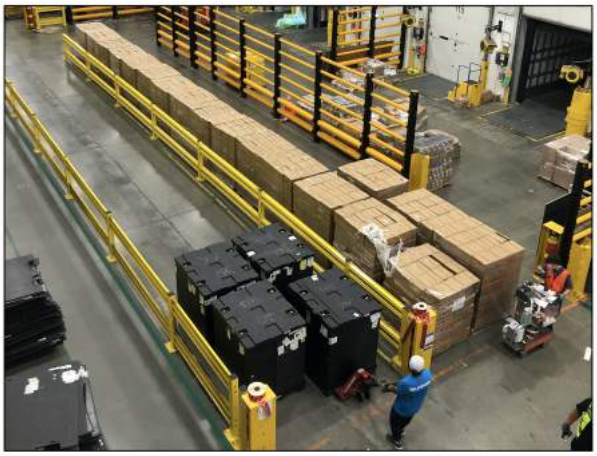


Figure 2-16: Pallet buffer, which holds palletized cases or totes between forklift unloading and manual transport to downstream processing and stow. Here, the left buffer lane contains pallet sleeves containing cases, and the right buffer lane contains standard mixed-case pallets.

of a trailer is stopped due to a tote wall. FCs vary in how much space is available to stage buffer pallets, but at the ARS FC used to model the processes in this thesis, there is space for 8-10 pallets adjacent to each line, but generally only 3-4 pallets are staged at any given time.

Figure 2-17: Various processes for unloading floor loaded trailers. Left: Unloading a section of cases using a powered flexible conveyor. Center: Unloading a hybrid floor load wall with both cases and totes. Right: Using a ladder to safely remove unstable containers that may have shifted during transit.

One complicating factor for many newer ARS FCs is that the larger footprint entails a split dock model, where the inbound dock actually exists as two separate docks on either end of the building. FCs may have different processing capabilities on each dock; at the primary site described here, only one dock is decant-enabled, whereas the second one is not. This requires case-heavy trailers to be processed on one side, and tote-heavy trailers to be processed on the other. Hybrid trailers cause additional work, where excess cases on the non-decant dock are periodically loaded back into an empty trailer and transported to the decant-enabled dock for processing. The process occurs in reverse for excess totes on the decant-enabled dock. This is not the case for every building, but more generally, split-dock designs add an extra dimension of complexity to labor planning and

trailer scheduling.

2.3.4 Decant Process

The latest generation fulfillment centers perform full decant, meaning that all cases move from the dock to the decant process, where product is removed from its corrugate packaging, scanned, and placed into a standard yellow tote for transport to stow. As outlined earlier in this chapter, decant is a means of standardizing product flow to the stow areas in the FC. The decant process brings several benefits. Stow rates are higher with decant, as removing items from a tote is faster than from a tightly packed case, which was the previous method. Empty corrugate handling is centralized at decant, so stowers do not spend time in this function, and corrugate does not have to be removed from upper-level inventory floors. Finally, standardized tote containers bring the ability to leverage transport conveyance and other advanced stow and pick automation technologies, which Amazon is currently testing. Unlike the Each Sort process at the IXD, however, decant is simpler at the FC because all product from a case can go into the same tote (provided it fits), as there is no need to split product across multiple destinations.

Figure 2-18: Left: Palletized cases are staged along the conveyor feeding the decant line as buffer material, to be injected when case flow from the trailer is disrupted. Right: Decant workers take cases from the middle (waist-height) accumulator line, remove items and place them into totes. Totes are dropped to the lower conveyor (ankle-height) for transport downstream. Empty corrugate is removed on the upper conveyor (overhead) to a compactor. Associates stand idle when there is an insufficient flow of work.

Like many other processes, the layout of decant varies between buildings. At the FC modeled in this thesis, the inbound dock has three decant lines, each with 25 stations. Cases arrive on a flexible conveyor from the dock door to an accumulator which spans the length of the decant line. Associates remove one case at a time from the accumulator and process it, first scanning its barcode, opening it, dumping the contents into one or more destination totes, and finally scanning the tote label(s) before placing the tote(s) onto the tote-takeaway conveyor (Figure 2-18). The directional line layout causes cases to accumulate at the end of the line (furthest from the dock door), such that when there is a disruption in case flow, workers at the front of the line run out of work first. Due to limited space and Amazon's encouragement of single-piece flow, decant workers are not able to build individual station buffers.

2.3.5 Tote Transportation to Stow

The output of the decant line is a steady stream of totes, which are ultimately destined for the stow area on one of the FC's robotic inventory floors. Each inventory floor has one or more staging buffers, and totes are transported from the inbound dock, either directly from a transshipped trailer or from the decant line, in one of two ways. Newer FCs have fixed conveyors that bring totes from the dock to AR floors; these conveyors include bar code scanners which enable the conveyor system to divert totes and allocate an even volume distribution to each stow area throughout the building. The capacity of these conveyors is limited however, so like older generation buildings, even the newest ARS FCs are equipped with a series of Vertical Reciprocating Conveyors (VRCs), which are large elevators capable of moving several pallets or carts of totes between floors at once. Because the decant line is connected directly to inbound conveyors at buildings which have this capability, totes leaving decant are the first to utilize this transport capacity. When running at less than full inbound capacity, it may be possible to inject totes arriving from transshipped trailers into this conveyor system as well. But at full capacity, even the decant lines alone will typically over-saturate the inbound conveyors, and excess full totes are set aside from the decant line and moved manually via VRC by dedicated cart runners. Next generation building designs increased the capacity of fixed conveyance systems, but these systems are unlikely to be expanded in current operational FCs due to physical space constraints and the relative cost of modification.

2.3.6 Stow Process

The final step in the FC inbound process is the stow function, where individual items are moved from plastic totes to inventory pods. Amazon uses a random stow process, where each item is placed in an arbitrary pod cell, and its location is recorded virtually in the inventory management database. Amazon FCs hold millions of items and hundreds of thousands of SKUs each at any given time, and distributing the inventory randomly across individual pods allows for rapid stow and retrieval with hundreds of stowers and pickers operating simultaneously. Like decant, stow is considered a direct function, and managers pay close attention to how closely each stow floor tracks its target rate, working hard to eliminate barriers to efficiency.

Figure 2-19: Amazon associate stowing inventory. The associate removes an item from one of the totes, scans the item, places it into an available cell in the inventory pod, and scans the cell location to record it virtually.

A common barrier preventing high stow rates is stowers running out of work. Stowers rely on cart runners to continuously deliver totes full of inventory and remove empty totes. With limited

space for tote buffers on the inventory floors, stowers may run out of work if an insufficient number of totes arrive from the inbound dock, or if the cart runners are working too slowly or not serving all stowers evenly. This first cause puts pressure on the inbound dock functions to ensure that stowers do not run out of work, although the lag time between flow disruptions on the dock and out-of-work instances on a stow floor can be several hours long, making it difficult to attribute causality to a specific disruption event. A robust Andon system exists to track stow out-of-work time, where associates are able to log any time periods when they are low on work directly in the computer system at their station. Stowers are incentivized to use this system when work is not available, and the aggregate data enables managers to diagnose problems both in real time and examine historical trends to make corrective process changes.

2.3.7 Summary of FC Inbound Operations

This section described the major processes and performance indicators used throughout the inbound processes within Amazon FCs. A common theme is the high variability of freight flowing through the system at multiple points, and across several dimensions, including volume over time, container type, item size, and trailer load type. Key takeaways from this section are:

1. Transshipped trailers (IXD to FC) make up the vast majority ($> 95\%$) of a FCs inbound shipments
2. Inbound trailers are received at the FC with high variability internally, both by container type and item size. Palletized trailers make up the majority of transshipments, but floor loads continue to grow as a percentage of total shipments as they reduce transportation and loading costs.
3. Floor loaded trailers tend to be either tote-light or tote-heavy; the distribution is bi-modal.
4. The current trailer scheduling process is not standardized and results in highly variable inflows between shift quarters, along both dimensions of container type and items size.
5. Cases and totes must be both split apart (from within a particular trailer) and batched (by container type from across multiple trailers) in a short time span with limited space on the inbound dock. Disruptions to this process are quickly propagated downstream.
6. Decant is a newer process being used to convert all cases into totes on arrival at the FC; it is frequently subject to out-of-work instances when the inflow of cases is disrupted.
7. Limited space for buffer material (palletized cases, typically) is available to supplement the decant line. The process of managing these buffers is not standardized.
8. Many newer FCs have dual-dock layouts, where smaller inbound docks sit on each side of the building, each receiving freight for AR stow floors on their respective sides. However, this configuration makes it difficult to balance work levels across the building, particularly when docks have different processing capabilities and downstream tote transportation capacities.

2.4 Chapter Summary

This chapter described Amazon's inbound supply chain, examining the hub-and-spoke model comprised of IXD and FC linkages. Specific attention was given to parts of the transshipment process associated with shipping trailers, namely the IXD trailer loading process, trailer receive scheduling at FCs, and the trailer unload and processing stages which occur on the FC inbound dock. With this context, Chapter 3 will examine the more specific decant out-of-work problem and diagnose its root causes through analysis of operational data and observational trends.

Chapter 3

Decant Starvation Problem Analysis

The previous chapters provided an overview of the problem of case flow disruptions between trailer unloading and decant in ARS FCs. From observation, the problem is most pronounced when unloading hybrid floor loaded trailers. The variable and unpredictable configuration of tote walls in these trailers causes gaps in the outflow of cases, which result in frequent starvation of the decant line, or out-of-work (OOW) instances. The decant OOW problem is well-established, but because decant is still a relatively new process for Amazon, cumulative OOW time has never been accurately measured or precisely quantified. We conclude that decant OOW events result in significant lost labor costs to Amazon, estimated at more than 7% of total decant labor hours incurred, at a cost of more than $20M across the North American network in 2020. This chapter will perform a current state analysis of the inbound dock processes and describe a method used to estimate the labor hours lost to decant OOW events. This is followed by an analysis of the upstream sources of variability to diagnose root causes of these events. Finally, the chapter concludes with the generation of the solution hypotheses which were introduced in Section 1.5.

A note on impact of COVID-19: The timeframe for this research project, February - August 2020, included significant disruptions to Amazon's normal operations with the onset of the global COVID-19 pandemic. Beginning at the end of March 2020, order volumes spiked across Amazon's North American distribution network as consumers shifted to online shopping while remaining at home. By August 2020, while order volumes have stabilized, they still remain significantly higher than pre-COVID levels. This has put tremendous strain on all facets of Amazon's operations. Increased volume demands at FCs and IXDs have changed the distribution of incoming freight, while processing rates have become more variable as employees learn to cope with added social distancing measures. Labor planning has become more challenging, as Amazon instituted a policy allowing employees unlimited leaves of absence during the initial months to ensure personal health and safety. Increased volume and sparse labor have placed additional strains on already heavily used equipment, resulting in broken automated equipment and temporary process changes which have been extended as suppliers struggle to meet repair and replacement orders. All of this is to say that from approximately March 20, 2020 and beyond, Amazon's operational trends have changed. The majority of baseline data in this thesis was captured before this period, and so we aim to stay consistent in the definition of "normal" or current state operations as referring to pre-COVID-19 standards. Where data is captured or additional trials are conducted after this time, it is noted appropriately.

3.1 Quantifying Decant Out-of-Work Events

3.1.1 Defining the Decant Out-of-Work Metric

To understand the magnitude of the decant out-of-work problem and appropriately scope the solution space, it is necessary to quantify the effect of decant OOW instances on the overall inbound process and on Amazon's bottom line. Unfortunately, no system currently exists to track and measure decant OOW events directly. As mentioned previously, decant is a rated function, and the decant rate, measured in cases decanted per hour (CPH) is the primary measure of performance evaluation, both for the individual associate and for the shift as a whole. Data is available at many levels of granularity, but typically, comparisons are made between average shift rates to compare shift manager performance, and between weekly FC averages to compare performance across buildings. While decant rate is certainly affected by disruptions stemming from hybrid floor loads, there are several other contributing factors as well. These factors include the number of units per case, the occurrence of tote takeaway line backups, individual unaccounted for breaks, and manager effectiveness in coaching under-performing workers, to name a few. Because of these extrinsic sources of variability, rather than use decant rate as the primary metric for determining the impact of hybrid loads on decant labor usage, it is more precise to infer the frequency and duration of decant OOW events directly. The question that must be answered is *how many hours per year are lost due to insufficient case availability at decant?*

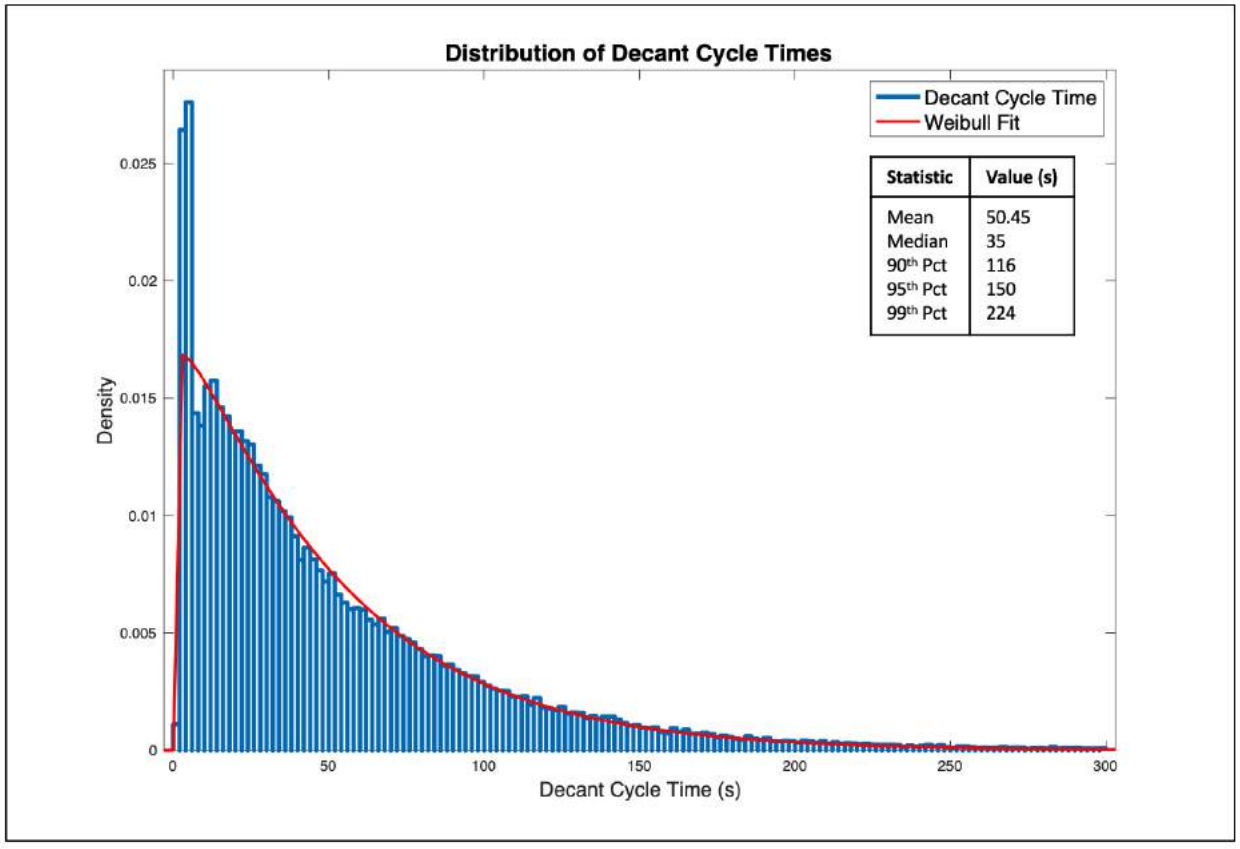


Figure 3-1: Distribution of decant cycle times at an ARS FCs using full decant. 100,000 samples over a 3-month period. 90th percentile is 116 seconds, which informs the 2 minute threshold value.

Using the most granular level of data available, the logging of individual decant employee

case scans, it is possible to estimate an *inferred OOW* metric. Decant workers perform single-piece flow, scanning the barcode on each new case retrieved from the accumulator before decanting its contents into a tote. Individual worker cycle times are measured by computing the time between successive case scans. By filtering for cycles which exceed certain threshold criteria, it is possible to estimate the frequency and duration of OOW events. To determine these criteria, we look at the distribution of individual decant scan intervals (Figure 3-1), and use the 90th percentile of cycle times, which is 116 sec. We use a threshold of two minutes for simplicity. This provides an upper limit on reasonable cycle time, and we assume that any cycle times above this threshold are attributable to some type of special cause, whether an OOW event or otherwise. Figure 3-2 shows the number of workers exceeding this threshold, sampled regularly throughout a representative day.

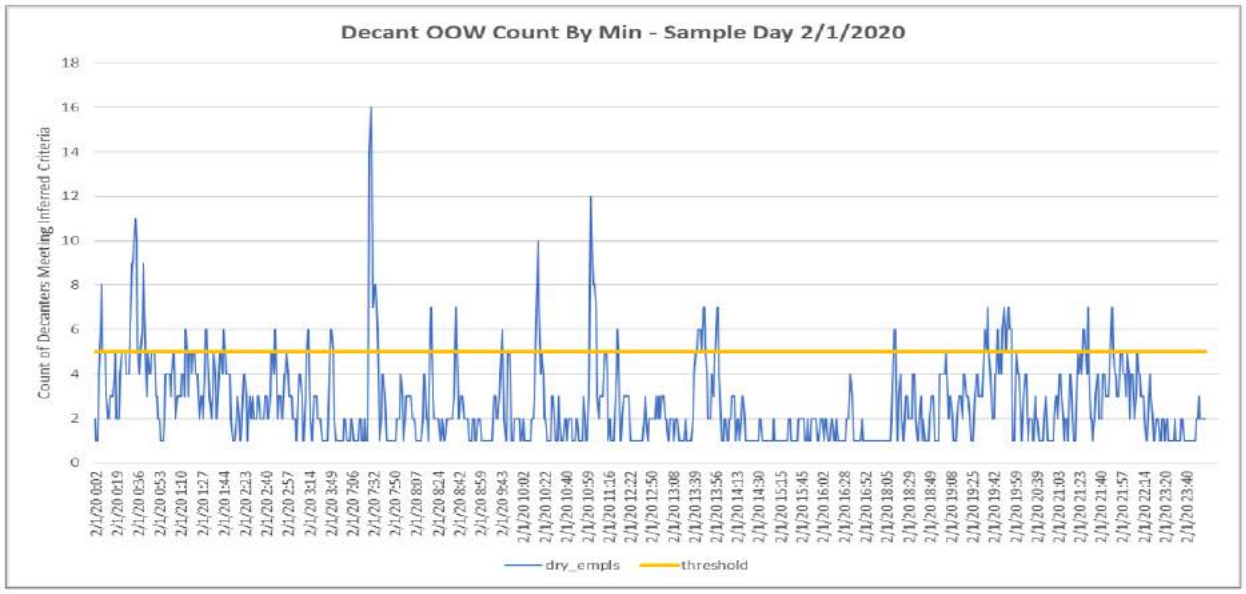


Figure 3-2: Number of decant associates recording gaps between scans longer than 2 minutes, sampled every minute for a representative day at the same FC. Horizontal line shows the five simultaneous worker threshold; areas of plot that exceed this threshold are considered OOW events.

There are numerous reasons that a worker might have an interruption to his or her decant workflow during a shift, and simply examining time between scans is insufficient as it would overestimate the the amount of OOW time, including things like unscheduled breaks, issues with product or packaging, or running out of totes to decant product into. These special cases generally occur on an individual basis, whereas during an OOW event, many workers on the affected line will experience disruptions. To filter out this first set of individual causes of disruption, we define an OOW event as a time period when multiple decant workers on a single decant line at a given FC are simultaneously experiencing cycle times greater than two minutes. We define five simultaneous workers to be a reasonable threshold after observing the decant process and interviewing managers and workers. A sensitivity analysis considering different thresholds is shown in Figure 3-3; it is clear that the choice of threshold criteria does significantly affect OOW metric values, so careful parameter selection based on FC-specific conditions is critical.

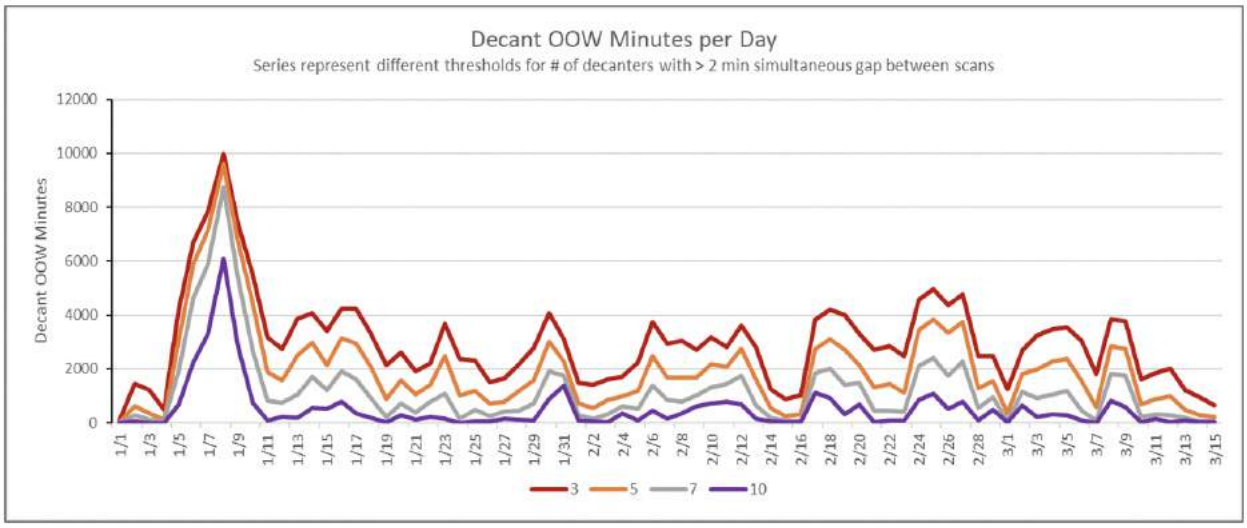


Figure 3-3: Sensitivity analysis of various thresholds for number of workers required to have simultaneous long cycle times in order to define an OOW event.

	Min # Simultaneous Cycle Times > 2 min			
OOW hrs/day	**3**	**5**	**7**	**10**
Mean	49.9	33.3	19.6	8.2
Std Dev	27.8	27.5	23.5	14.6

Table 3.1: Sensitivity analysis metrics for varying minimum thresholds for number of associates with simultaneous long cycle times. Statistics correspond to the plot in Figure 3-3.

3.1.2 Impacts of Decant Out-of-Work Events

Lost Labor Cost

With this definition for a decant OOW event, it is possible to determine the total cost to Amazon in lost decant labor. This is done by querying historical scan data and aggregating labor minutes throughout the day where the defined threshold criteria is met. Figure 3-4 shows computed daily OOW time for one FC, both in total hours and as a percentage of decant hours worked. The noticeable trend of rising OOW time beginning at the end of March 2020 is related the complicating impacts of COVID-19. Prior to March 20, the decant OOW time averages 7.7% of decant hours worked; since the onset of COVID-19, it has averaged 16.7%. Other comparable FCs show similar trends during the same period. For the analysis that follows, we will use the pre-COVID-19 estimate of 7.7% of decant hours as the OOW baseline.

It bears mentioning that the analysis so far does not yet indicate any causality from hybrid floor loaded trailers. Tying decant OOW events to hybrid floor loads will be further discussed in Section 3.2, but based on observations and discussions with floor managers, it is evident that tote wall disruptions in these trailer types are the primary contributing factor.

Additional Costs

While lost labor at decant is the most obvious effect of decant line starvation, there are a few other costs that should be considered to provide a more complete picture of the problem. The stow

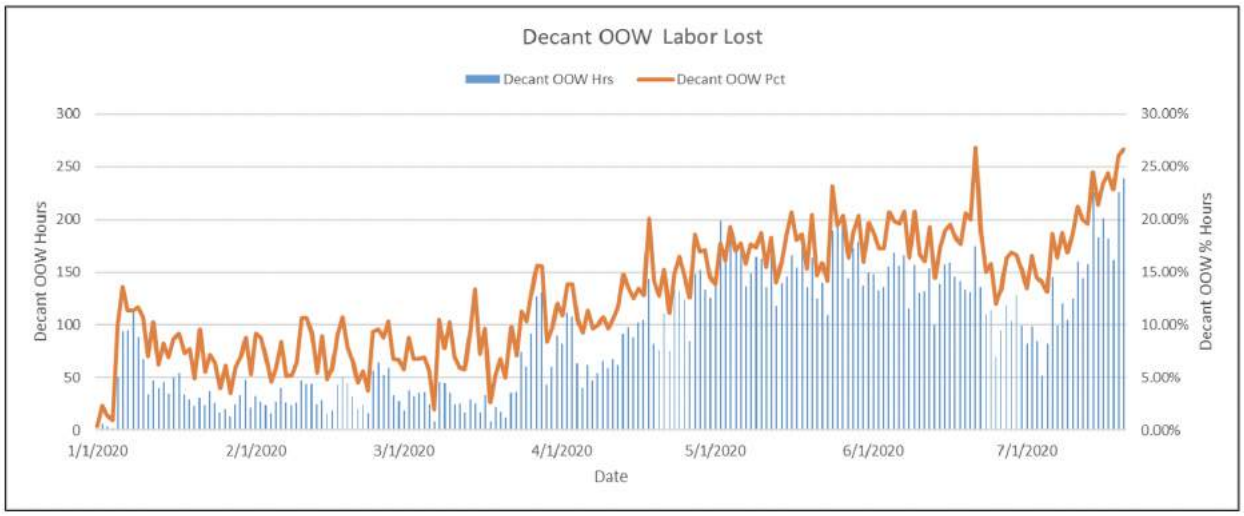


Figure 3-4: Lost decant labor to OOW events, shown from Jan 1 - Jul 31 2020 at a single ARS FC. Shown as total hours lost (blue) and percent of total decant hours lost (orange). Trend changes shape at the end of March, due to impacts on operations from COVID-19.

process, immediately downstream of decant, is the next affected process. Like decant, stow rates are closely monitored on an hourly basis and tracked against targets. The stow process is subject to even more variables than decant (tote buffer availability, unit size, units per tote, pod bin fullness, and individual performance, to name just a few), and there is a highly variable time lag between a tote leaving decant and that tote being stowed. These factors make it quite difficult to accurately estimate the effect of decant starvation on overall stow rate. However, stow stations do track OOW instances directly through the use of worker-triggered "andons", as described in Section 2.3.6. These events provide a clear signal that the stow process is starved for work. Figure 3-5 shows total daily stow OOW time plotted against decant OOW time. We see a moderate positive correlation between these two variables, although we remain mindful that this correlation is modulated by the fact that 40-50% of total daily product volume does not pass through decant on its way to stow, but instead arrives directly from the dock in totes. We do not attempt to further quantify the effect of decant starvation on stow, but simply state its presence.

In addition to stow OOW events, other aspects of the inbound process may be affected by decant line performance, but are not estimated here. Some of these might include an increase in decant quality defects as decant workers race to make up for lost time after having been previously starved. Similar quality issues may manifest at stow for the same reasons. Additionally, floor managers may respond to a visible OOW situation by shifting labor from other areas of the inbound department to supplement the inbound dock crew temporarily, but these unplanned on-the-fly moves incur switching costs and rarely result in sustained optimal flow.

Proposed Measurement System

Beyond tracking decant OOW time for the purposes of this project, visibility of the OOW metric is critical for dock area managers to understand and quantify the impacts their decisions have on maintaining continuous flow. The method outlined above provides a fair approximation but requires large dataset queries of source tables which are only updated daily, and so this metric currently cannot be exposed in real time. To increase the accuracy and accessibility of the OOW

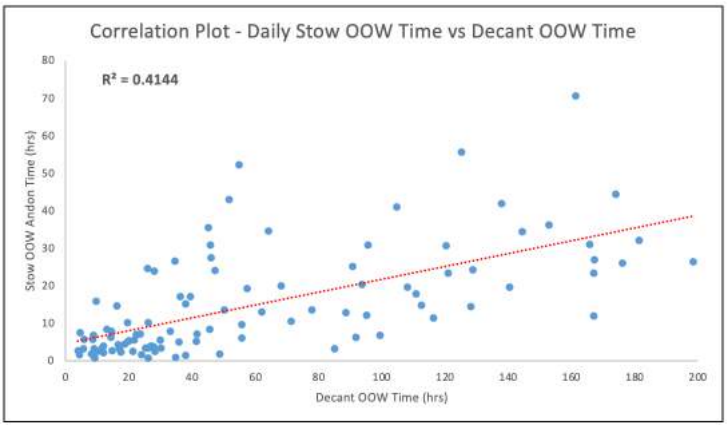


Figure 3-5: Moderate positive correlation between decant and stow out-of-work time, aggregated on a daily basis. Sample includes 5 months of data, with 2 outlier points removed due to stow data abnormalities. Stow OOW time is the daily sum of OOW minutes logged in each OOW andon, and for decant, the OOW metric described throughout this section is used.

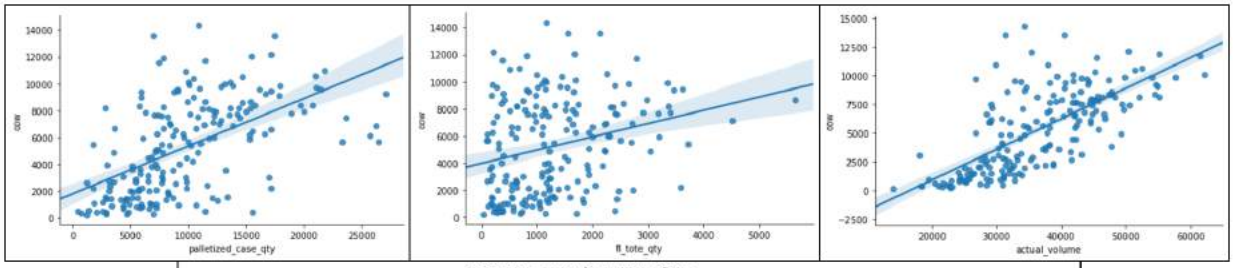
metric, a proposal for a decant OOW Andon, similar to the system already in use at AR Stow stations, has been shared with Amazon's in-house FC software development team. This project is expected to be piloted in 2021 either via a digital Andon requiring users to log OOW events, or by automatically surfacing OOW conditions using existing photo-eye sensor data to detect low case volume in the decant accumulator. If successful, such a software-based solution could be easily scaled across the Amazon network. Finally, while simply capturing and exposing a metric does not solve the decant flow problem, several studies on performance management suggest that performance often improves along the dimensions in which it is measured and made visible [25, 26], so it is reasonable to expect that exposing the decant OOW metric would likely result in some level of improvement.

3.2 Root Cause Analysis

3.2.1 Attribution to Tote Wall Blockages and Buffer Availability

With a clearer idea of the magnitude and cost of the decant starvation problem, we can begin to examine various root causes of the problem and assess the relative value of developing solutions to address each of these. We begin by using a regression-based approach to examine correlation between various potential drivers of OOW time. From six months of historical transshipment data for trailers destined to the FC modeled here, in addition to the dependence on total daily unit volume, two variables show statistically significant relationships with decant OOW time: number of totes in hybrid floor loaded trailers, and number of case pallets received (Figure 3-6). Total volume received (denoted as "actual volume") is also significant, but is used as a control variable for the expected increase in OOW with increasing total volume. The variables in this regression are aggregated at a daily level: total decant OOW minutes, number of floor-loaded totes, and total palletized cases. We note that more granular regression analysis, at the individual trailer level, for example, is not feasible due to differences between logged trailer receive times and the actual processing of trailer contents. Tote quantity in floor loads is positively correlated with decant OOW time, as a higher number of totes corresponds directly to an increase in case unload disruptions. The positive correlation between palletized case volume and OOW time is a bit more complex –

while one might expect that a higher volume of palletized cases would provide more buffer material thus reducing decant OOW time, the opposite is true. Because the FC tends to process a relatively stable unit volume through the inbound dock each day, when the proportion of palletized loads is higher, the number of floor loads will be accordingly lower. On days with highest palletized case volumes, a larger proportion of decant work must come from case pallets as opposed to floor loaded trailers. However, as described in the previous chapter, the case buffer replenishment process is slow, labor-intensive, and capacity-constrained, and so it acts as a bottleneck and throttling the flow of cases to decant in these situations.



OLS Regression Results

Dep. Variable:	oow	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.569
Method:	Least Squares	F-statistic:	86.71
Date:	Tue, 15 Dec 2020	Prob (F-statistic):	1.66e-35
Time:	08:36:03	Log-Likelihood:	-1793.0
No. Observations:	196	AIC:	3594.
Df Residuals:	192	BIC:	3607.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4452.5536	669.005	-6.655	0.000	-5772.096	-3133.011
fl_tote_qty	0.6385	0.177	3.608	0.000	0.289	0.988
palletized_case_qty	0.1775	0.036	4.937	0.000	0.107	0.248
actual_volume	0.1974	0.022	9.167	0.000	0.155	0.240

Figure 3-6: Correlation plots and regression analysis for the two statistically significant variables contributing to daily decant OOW time. As expected, decant OOW time increases with floor loaded tote volumes, as this causes more unload disruptions (left). Palletized cases, which are used as buffer material, are positively correlated with decant OOW time (center). This is likely due to reliance on the pallet buffer replenishment process, which is a bottleneck in the system. OOW time is also positively correlated with total daily processed volume (right), which also corresponds closely with total decant hours. Confounding variables which cannot be measured here, such as trailer door assignments, actual trailer processing sequence, units per case, and item size all have large influences on the decant process and OOW time as well.

While the regression model presented here is helpful in assessing historical correlations and trends, it does not present a full picture of the relevant process interactions. Indeed, the model's relatively low R^2 value suggests the presence of additional variables which cannot be measured, such as how the trailers are actually configured internally and how well labor was planned and allocated on a given day.

To take a deeper look at at the inbound dock processes that feed the decant line, we built a numerical simulation of a single decant line (presented in Chapter 4). In particular, we examine the impact on decant OOW time stemming from flow disruption caused by tote wall blockages. Figure 3-7 depicts these FC processes, beginning with trailer unloading and ending with decant. Average rates, based on an observational study at one particular FC, are used (although specific rates are redacted here). Using the median number of cases per pallet and the standard labor allocation of one worker in the buffer replenishment role for each decant line, we can calculate the cycle time for buffer replenishment in seconds per pallet, using the median number of cases per pallet. The per case

buffer replenishment time is substantially longer than cycle times for cases moving though other parts of the system, indicating a potential bottleneck here, which is consistent with the positive correlation between decant OOW time and palletized case volumes as well as observed OOW events on the decant line, where workers are unable to replenish buffers quickly enough during breaks in case flow from docked trailers.

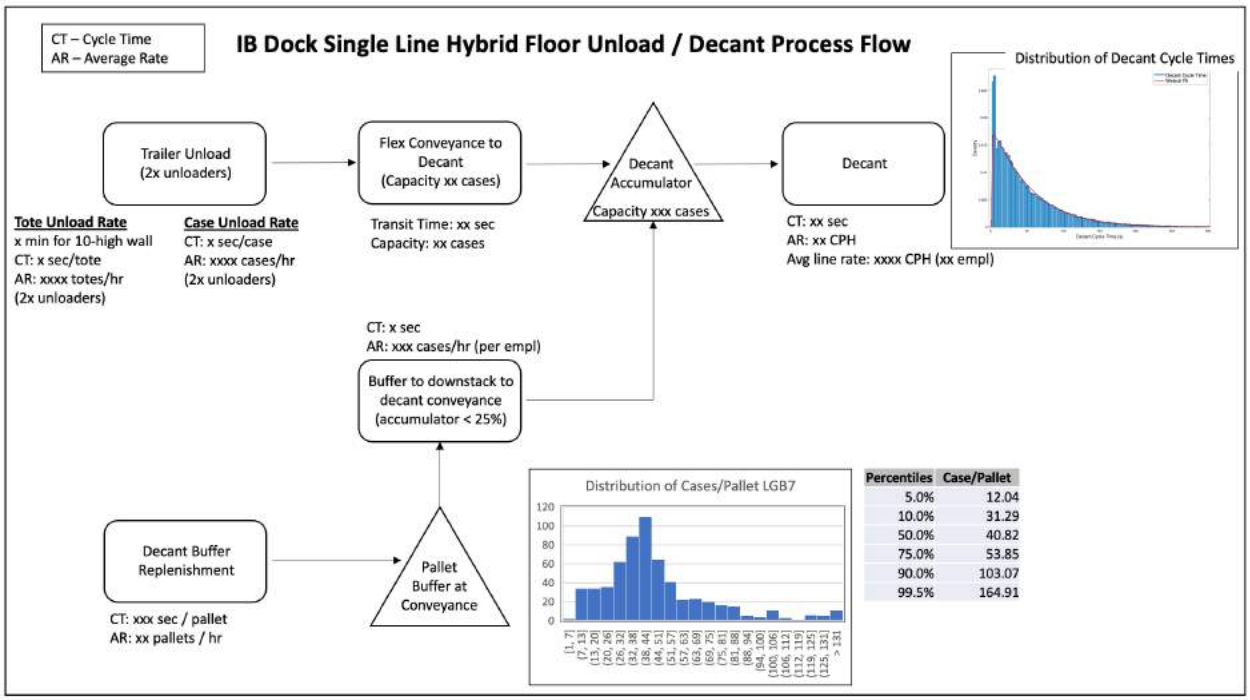


Figure 3-7: Process flow diagram depicting single-line trailer unload and decant process with associated buffer replenishment. Specific capacities and rates are redacted.

Using the rates mapped in Figure 3-7, it is possible to estimate the effect of processing a tote wall disruption given specific system conditions. Figure 3-8 details calculations comparing the duration of case buffer material with the time for a typical disruption in case flow from a tote wall blockage. Key assumptions in this comparison are the average state of three full case pallets for buffer material, and a tote wall disruption consisting of three sequential walls. The use of three case pallets along each decant line is the normal practice, although at various times observed counts ranged between zero and six. A three-wall tote blockage is considered significant but not worst-case; blockages of these types are observed to occur on the order of 10-15 times per day on each line. This calculation assumes that both the conveyor and accumulator feeding the decant line are fully saturated, a conservative estimate. It is worth noting that even a full accumulator (80 cases) provides only 2.7 minutes of buffer protection at the average decant rate.

Using these conditions as representative of a typical tote wall disruption scenario, we assume 250 cases available as buffer material when case flow from the dock is halted. At the average decant line rate, this buffer will last approximately 8.4 minutes. At the same time, using the average observed trailer unload rate for a standard two person crew, clearing three consecutive tote walls (180 total totes) will take 9.75 minutes. By analyzing even this conservative case, it is clear that the current state buffer is insufficient to fully mitigate a decant OOW event. Furthermore, once the buffer has been drawn down, the slow replenishment process causes the system to become even more susceptible to OOW events caused by even smaller tote wall disruptions.

Internal Buffer Duration		Tote Wall Disruption	
Line Case Capacity (Conveyor & Accumulator)	xxx	Totes per wall	60
Pallet Buffer Cases (3x pallets)	120	Consecutive tote walls	3
Total Buffer (cases)	250	Total blockage (totes)	180
Decant Rate (CPH)	xxx	Tote unload rate (TPH)	xxx
Decanters per Line	xxx	Unload workers	xxx
Effective Line Decant Rate (CPH)	xxx	Effective unload rate (TPH)	xxx
Buffer duration (min)	8.41	Disruption time (min)	9.75

Figure 3-8: Comparison of disruption time needed to clear a tote wall series vs. buffer supply duration in a case-saturated conveyor and accumulator. Specific rate information has been redacted.

3.2.2 Attribution to Scheduling and Time Varying Inbound Flow

The single line decant model introduced in the previous section is useful, but the fixed rates and labor staffing assumptions are based on average steady-state conditions. In reality, due to the variability in trailer types arriving at the dock, the inflow of cases and totes is also highly variable between shift quarters, as was shown in Figure 2-12. This makes labor planning difficult – component processes outlined in Figure 3-7 will often alternate between having too much and too little work, resulting in a lower overall labor utilization.

Using the shift quarter flow variability metric defined in Equation 2.4, we can assess the impact of case flow variability in particular on the decant process. Figure 3-9 shows the correlation between weekly case flow variability and decant OOW time. As we might expect, a higher degree of variability corresponds to more OOW time. Large differences in case processing volumes between shift quarters require detailed labor re-organization to avoid under-staffing or over-staffing; in practice, the pace and number of employees that each line manager is responsible for makes this type of adjustment untenable on a quarterly basis. The same positive correlation between quarterly size variability and decant OOW exists – Figure 3-9 shows the moderate positive trend for small-size units. While not depicted here, similar positive relationships exist between decant OOW and the variability metrics for quarterly tote volumes and medium-size unit volumes as well.

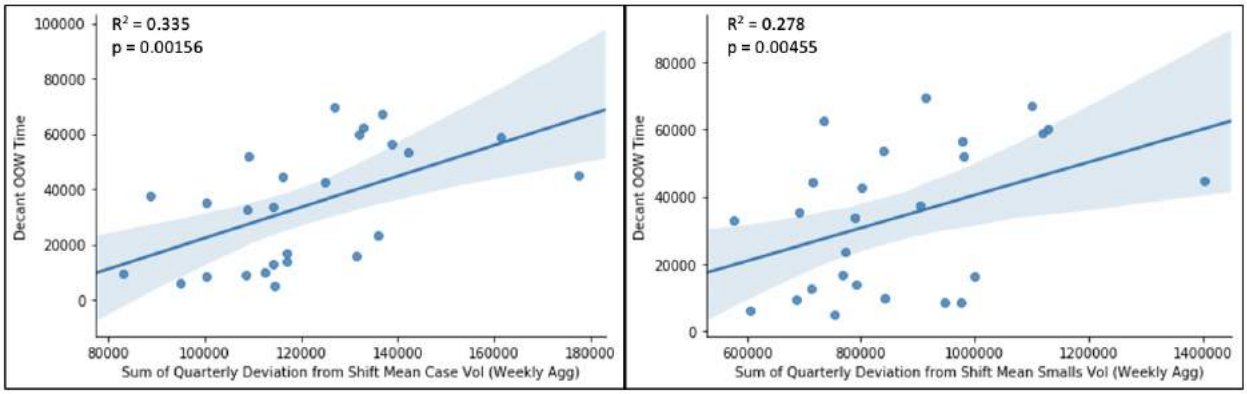


Figure 3-9: Correlation plot showing relationship between historical decant OOW time (weekly) and the sum of quarterly variation from both mean shift case volume and mean shift small-size unit volume (aggregated to weekly). Moderate relationship exists between each of the flow variables and decant OOW time. Weekly variability metrics are normalized based on average weekly unit volume to account for relationship between unit volume and decant OOW.

Dock managers attempt to mitigate some of this time-based variability by grouping loads;

for example, if a hybrid load is being processed on one unload line, it is helpful if an adjacent line has an all-case load, or if the pallet buffer is full of cases, either of which can supplement the hybrid line when unloaders encounter a tote wall disruption. Other scheduling techniques may be used to level-load by item size – dock managers are instructed to pair trailers with high proportions of small-size units with trailers heavier in medium-size units at the same time to keep the proportion of item size flowing through the building fairly consistent. These scheduling decision processes occur outside of the individual trailer level of granularity, and it is the combination of trailers being processed at a given time which affects the stability of the current flow mix of container type and item size. In practice, these considerations cannot always be satisfied, and often, when dock managers do schedule trailers with these constraints in mind, they are making trade-offs without a clear picture of the relative costs of the variability incurred. It is evident, however, that in addition to the variability introduced by individual trailers' contents and internal container configuration, the variability stemming from trailer scheduling and aggregate container volumes entering the FC over a given time period also contributes to the decant starvation problem. This source of variability will provide an additional set of improvement opportunities related to scheduling, which are discussed in Section 3.3.

3.2.3 Attribution to the IXD Loading Process

Having established tote walls as a direct cause of decant OOW events, we can look further upstream to the trailer loading process in the IXD for drivers of the specific configurations most likely to cause OOW events. Because trailers (in their current design) must be unloaded on a last-in-first-out (LIFO) basis, the way that trailers are packed at the IXD plays a large role in container-type variability during the unload process. Section 2.2.4 describes the high variability in trailer configurations, particularly for hybrid floor loads. In addition to the variable proportion of cases and totes, the heights of individual tote walls and the number of sequential tote walls all vary widely between trailers. Data for these last two attributes is not captured currently, although observations and interviews of IXD and FC dock workers confirm this high degree of variability and a lack of internal load configuration standards. The number of sequential tote walls in a trailer segment clearly affects the likelihood of causing a decant OOW event, based on calculation logic for unload disruption time presented in Table 3-8. The maximum height of totes in a particular wall also affects the chance of an OOW event. So far we assume 10-high tote walls that take up the entire height of a standard trailer; this is usually the case but not always, and in instances where totes are stacked less than 10-high, cases typically fill the remaining space within the wall. With at least some cases available in each tote wall, the severity of disruptions to the case unload process is diminished, as the wall still permits cases to be offloaded to decant, albeit at a slower rate.

The particular configuration of cases and totes in a given hybrid load depends on two factors: the physical loading process, and the order in which totes and cases arrive to the dock door from the ship sorter. The limited buffer space at the IXD outbound dock constrains the loading process to use a generally first-in-first-out (FIFO) process for containers as they arrive to End of Line. Workers loading trailers do maintain a small buffer of either cases or totes inside the trailer in order to properly sort containers and build wall sections consisting of uniform container type. The height of totes within a given tote wall depends on the ratio of cases and totes arriving off of the line, as well as the buffer space available. For example, if a loader with a half-finished tote wall is waiting for more totes, but the buffer becomes full of cases, she will have to switch to loading cases. As described previously, this entails capping all totes in the top wall layer with either plastic lids or cardboard, and so given sufficient totes, trailer loaders will generally build walls to full height.

The second factor contributing to highly variable trailer configurations is the unpredictable arrival of vendor freight at the IXD and subsequent FC allocation algorithm which assigns product to a container type and to a specific destination FC. The complexity of this system causes the arrival of cases and totes to a specific outbound ship lane to also be highly unpredictable, and effectively characterized as a random process. However, because the FC allocation system maintains a goal of sending a fixed ratio of cases to totes along each transshipment arc (the ratio is specific to each arc), this ratio stays relatively consistent between trailers on the same arc, while the specific internal configuration of trailers varies widely. Figure 2-7 provided data for a small selection of transshipment arcs illustrating this trend.

It is clear that the variability both in trailer contents and internal container configuration introduced in the IXD loading process both drive the fluctuations seen during unloading. Section 3.3 will explore solution hypotheses related to the IXD loading process.

3.2.4 Other Causes of Decant Starvation

In addition to the three root causes of decant starvation described above, several additional contributing factors were identified during the initial scoping phase of this research, but are not addressed further in this thesis. They include:

Decant labor allocation - An obvious potential cause of decant starvation is the assignment of too many decant workers during a shift. While this does occasionally happen, the shift-level staffing process is relatively straightforward, and determines a requirement based on average rate and the planned volume of case freight. Actual decant rates do vary based on item size and other factors, and the actual number of cases to be processed may change as trailers are shuffled, but in general, shift volume falls close to planned targets, and the total item size distribution for every shift remains within 10% of the annual ratio. Current labor plans are constructed by shift quarter and aim to be consistent across the shift. While it is possible to increase the granularity of decant labor planning based on the scheduled volume of cases over time, building more complex labor plans that incur higher management overhead and greater switching costs runs counter to Amazon's operational goals of simplifying processes and eliminating root causes of variability. Therefore, we focus on upstream problems such as trailer scheduling instead, which underlie the shifting demand for decant labor.

Indirect labor allocation - Indirect functions, the non-rated jobs introduced in Chapter 2, are critical to ensuring smooth flow within the inbound department, but Amazon's continued emphasis on reducing indirect hours forces department managers to make difficult labor allocation tradeoffs. Trailer unloader, cart runner, tote replenishment, and pallet buffer transport functions are all categorized into the same indirect labor bucket. These roles are currently highly utilized, and while increasing labor allocation for these roles would likely help to cope with variable process inputs, labor utilization would decrease. With the exception of examining the possibility of temporary labor shifts for the case buffer replenishment function, we do not explore the option of increasing indirect labor and instead focus on reducing process input variability as the primary means of stabilizing case flow.

FC physical layout and design - Among newer AR Sortable FCs, there exist several building design variations which affect inbound dock processes. Innovations such as automated conveyance sorting and Automated Guided Vehicles (AGVs) allow FCs to reduce headcount in indirect material transport functions and allocate that labor elsewhere. At other locations, smaller design improvements have been incorporated which significantly change the processes used. For example, some

sites use a "camelhump," a fixed conveyor which routes containers offloaded from trailers up overhead of the dock before bringing them back down to the decant line, which enables a clear path across the dock, reducing the distance that buffer replenishment workers must travel, thereby improving throughput. Another design consideration is the number of installed tote takeaway conveyors and "reverse spiral" conveyors that move totes from the dock to upper levels of the FC. Newer FC designs are capacity constrained by their reverse spiral conveyors, so much indirect labor is expended in moving totes across the FC by cart and up elevators, a requirement that reduces managers' flexibility to allocate that labor elsewhere. Each of these large physical systems are capital-intensive and highly integrated into building designs, and so they are not typically changed once a building has launched. Due to Amazon's sustained high growth rate, its dedicated FC design team prefers to focus efforts instead on incorporating process design and improvement feedback into the next generation of FCs which are several years away and will incorporate newer technologies not yet fielded. Because of this change process and extended timeline, major changes to building design are not considered in depth here.

3.3 Solution Hypotheses

3.3.1 FC Inbound Dock

The analysis of the FC inbound dock processes in the previous sections suggests that the case buffer which supplements the decant line is undersized, failing to accommodate even moderately-sized tote wall disruptions. Furthermore, the case buffer replenishment process is a bottleneck in the inbound dock system; a single worker transporting pallets cannot fill the buffer nearly as fast as it is consumed. These observations drive the hypotheses that by enlarging the case buffers and instituting a method for augmenting buffer replenishment labor, decant OOW events can be reduced. Chapter 4 will examine the impacts and feasibility both of increasing buffer size and of creating a flexible labor plan that allows the buffer replenishment process to be staffed with additional workers when needed. A process simulation model will be used to estimate optimal buffer size and labor cost savings. Chapter 4 will also consider operational impacts of newly fielded trailer unloading equipment on the decant process, and will conclude with lessons learned from a short pilot of operational process changes on the inbound dock at one FC.

3.3.2 Trailer Scheduling

Highly variable trailer contents coupled with large number of possible sequencing permutations and multi-dimensional shift performance objectives make scheduling trailers prior to each shift a significant challenge for dock managers. Poor schedules pose a barrier to consistent performance while increasing flow variability across the inbound department, resulting in decant OOW events and several other problems. As described in Chapter 2, in the current manual scheduling process, dock managers must effectively balance three priorities: processing the freight with highest priority score, allocating trailers to specific dock doors according to decant processing capacity, and maintaining the correct distribution of item size to send to stow areas. The existing manual scheduling process depends on the specific experience and priorities of the individual dock manager, and therefore invites a more robust and rigorous approach. However, any new scheduling system must also balance these same objectives.

Chapter 5 advances the hypothesis that the scheduling problem can be posed as a mixed-integer linear program (MILP) with an optimal solution which reduces variability across container

type and item size over the course of a shift while also respecting constraints on priority score and physical FC processing capacities. Trailer scheduling does not address the sub-trailer granularity of individual tote wall disruptions, but it reduces the variability in case volume over the course of the shift, which in turn enables the FC to operate with a consistent number of decanters based on the total number of cases planned for receipt. Chapter 5 describes the problem formulation, presents results of back-testing comparison against actual shift schedules, and closes with a discussion of practical implementation challenges and a proposal for further development of this solution.

3.3.3 IXD Loading Process

The analysis in the previous section emphasizes the inherent variability of case and tote arrivals to the IXD End of Line process and the lack of buffer space as causes of the unpredictable and variable floor load trailer configurations which pose a problem in unloading. We propose the hypothesis that changes to the trailer loading process are possible that will allow for greater consistency of trailer contents distribution and therefore more stable unloading inflows. Chapter 6 examines two types of solutions. First, a set of proposals to segregate totes and cases in floor loads entirely, either within the same trailer or in separate trailers, is discussed. Second, through the use of a trailer loading simulation model, we analyze the possibility of re-instituting a maximum tote height limit to ensure that in every wall, at least some cases are available to maintain flow to decant. These proposals are evaluated from both an operational feasibility standpoint and on a total cost basis, considering changes to both labor cost and transportation cost.

3.4 Chapter Summary

This chapter first described the methods for assessing the effects of flow disruptions to the decant process from hybrid floor loads, as well as the inferred out-of-work metric which can be used to estimate lost labor hours. We estimate that baseline decant OOW time is more than 7% of total decant labor hours, a substantial cost to Amazon's AR Sortable FC network. We go on to discuss three primary root causes of decant starvation: the case buffer management, trailer receive scheduling, and the IXD loading process. The following three chapters will address each of these respective causes, and will propose solutions where applicable.

Chapter 4

Inbound Dock Process Improvements

Chapter 3 introduced the hypothesis that increasing the size and replenishment rate of the case buffer upstream of the decant process would reduce the frequency and severity of decant out-of-work events. Rough calculations were introduced suggesting that the current average buffer size of three pallets (120 cases on average) is insufficient to mitigate a disruption in case flow longer than eight minutes, and depending on starting conditions of the line accumulator, potentially much less time. Meanwhile, the disruptions in case flow from unloading multiple sequential tote walls in hybrid floor-loaded trailers, and from trailer changeovers in general, are regularly observed to last ten minutes or longer, and sequential disruptions often occur with insufficient time in between for buffers to be fully replenished.

This chapter presents a more detailed model which confirms the buffer sizing hypothesis and allows various process and system parameter changes to be tested through simulation. Simulation results are then used to inform a set of suggested process changes, and also to assess newly-fielded dock equipment. The chapter concludes with a summary of test results from a pilot where a subset of these changes were trialed.

4.1 Simulation Model Introduction

In order to accurately model the trailer unload and decant process while accounting for the high degree of variability across numerous inputs, we define a system consisting of a single trailer, decant line, and buffer process, first introduced in Figure 3-7. The rates, capacities, and respective distributions outlined here define the base case system model.

In the current system, cases and totes are each unloaded at different rates, and these rates vary by worker. Net unload rates depend on the number of unloaders working in a particular trailer. Average values for observed cycle times with two trailer unloaders are used in the base case model. Additionally, the length of a disruption due to a tote wall blockage depends on the height of that particular wall. More than 90% of tote walls are observed to be of maximum height (10 totes high), so this is also used as the base case. The number of cases per pallet in the buffers, as well as buffer replenishment time, are also variable. For each buffer pallet modeled, the number of cases contained is sampled from the empirical distribution, and the average observed replenishment time is used. The case accumulator from which decanters receive work effectively provides additional

buffer capacity, and its starting volume affects how quickly work will run out – a fixed estimated capacity, based on average case size, is used here. Decant rate itself is also highly variable, and as discussed in Chapter 2, depends largely on item size. Decant rate is modeled by sampling from the bottom 90% of the actual distribution of individual cycle times (time between scans), given in Figure 3-2, which uses the same assumption made previously that the top 10% (longest) cycle times comprise the outliers corresponding to OOW events and other abnormalities. Therefore, we model separately decant processing time and OOW time resulting from tote wall disruptions, a distinction from our analysis of empirical data, where we do not have the ability to make this separation.

To account for each of these input variables and estimate the effect of various process changes on decant OOW time, a discrete event simulation is used to model the system shown in Figure 3-7. A discrete event model, as opposed to a continuous time model using rate averages, was chosen in order to describe the system at high granularity and to allow sampling from the various discrete input distributions. As discussed in [19], discrete event modeling offers the advantage of allowing sampling of the various input and rate distributions relevant to the system, and of using real discrete event data as model inputs. Furthermore, the discrete event model handles more intuitively the concepts of nested trailer, pallet, and case containers, each which have different conditions for moving through their respective process paths. The actual model used for this analysis is built using MATLAB's Simulink and SimEvents packages, chosen for its flexible and modular component architecture. A visualization of the model structure is given in Figure 4-1.

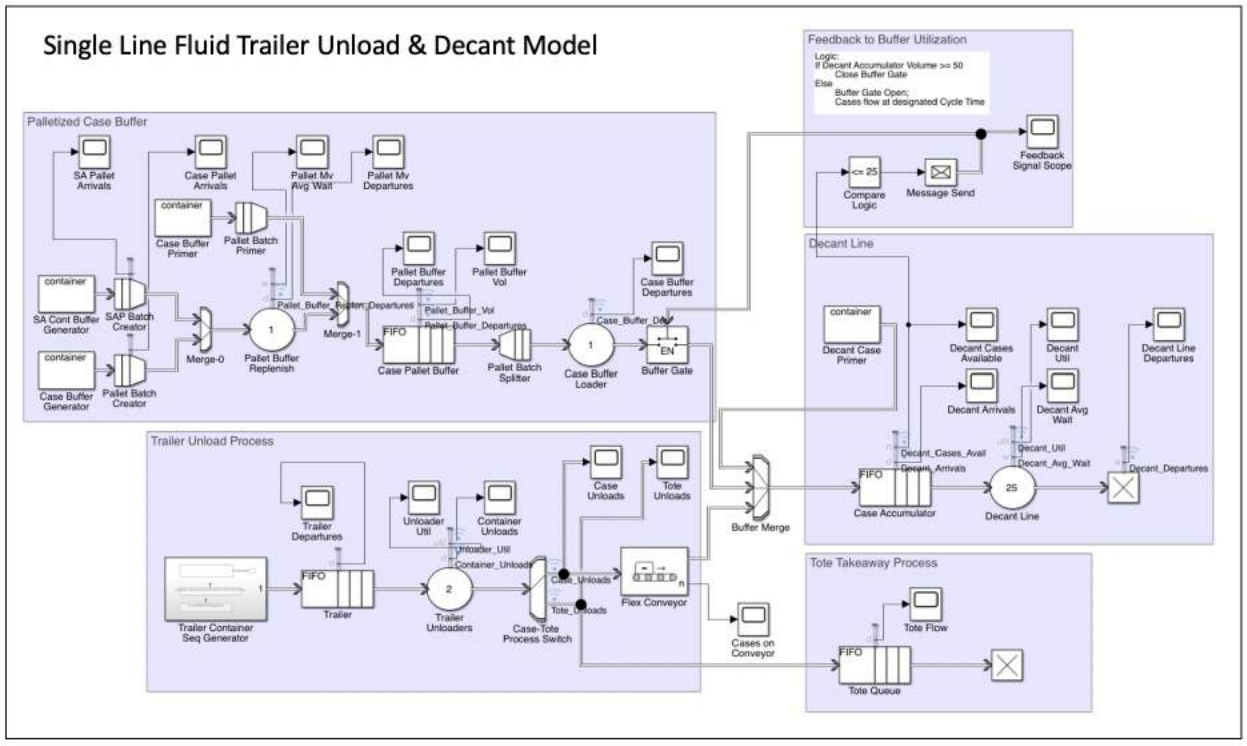


Figure 4-1: Discrete event simulation model of a single fulfillment center inbound dock floor-load trailer unload and decant processes, with inclusion of a palletized case buffer. Model is created using MATLAB Simulink and SimEvents packages.

The primary model input is a two-month historical dataset of trailer freight contents. From this dataset, floor-loaded trailer arrivals are extracted, where the total case and tote quantities in each trailer are known, but the specific configuration of containers within the trailer is not. A

virtual load configuration is constructed for each trailer, where tote walls are enumerated based on the total number of totes manifested, and their location within each trailer is randomized, to correspond with the random arrival loading process at the IXD.

Each simulation iteration consists of four trailers randomly selected from the historical set without replacement, which are then unloaded in sequence. Four floor-loaded trailers is representative of a typical shift's work for a single decant line. Trailers are sequentially unloaded in last-in-first-out (LIFO) order, one container at a time. Cases continue on to decant, and totes are removed from the system (to the stow process). The flow of cases to decant is supplemented by manual injection of palletized cases from the case buffer. Palletized cases are injected using a simple feedback rule, which is to only inject when the decant accumulator is less than 25% full. Additionally, cases cannot flow into the system if the decant accumulator is full, and any resulting backups in the system are propagated all the way back to the trailer unloading and buffer replenishment stages, halting these processes as well. The case pallet buffer is held at a fixed size, and is replenished on a one-for-one basis. Replenishment takes a fixed amount of time, which starts once there is at least one empty pallet space. For initial conditions, the decant accumulator and all buffers start of at full capacity, a conservative estimate that approximates normal conditions at the start of shift. The simulation stops once all four trailers have been processed; depending on exact contents and configuration, the overall simulation time may differ somewhat from a typical shift length.

Base Case Simulation Inputs		
Parameter	**Value**	**Notes**
# Simulations per scenario	100	
Trailers in sequence	4	Typical number of floor-loaded trailers processed during a typical shift on one line
Trailer Changeover Time Gap (min)	10	Estimated based on observations of dock
Tote Unload Cycle Time (s)	XX	Avg from obs time study, rate is per unloader
Case Unload Cycle Time (s)	XX	Avg from obs time study, rate is per unloader
Tote Wall Height	10	~90% of current loads have 10-high walls
Tote upper-bound threshold	800	Cutoff separating "Low-Tote" and "High-Tote" loads. Only Low-Tote loads are sampled
Buffer capacity (pallets)	3	Typical current state
Cases per pallet (cases)	40 (mean)	Model samples from historical distribution
Unloaders per trailer	2	Typical labor allocation
Employees managing buffer	1	Typical labor allocation
Buffer inject Cycle Time (s)	XX	Avg from obs time study, including SA case prep time, rate per worker
Buffer Replenishment Cycle Time (s)	XX	Avg from obs time study
Decant accumulator capacity (cases)	80	Measured; approximation based on average case size
Decant cycle time (s)	XX (mean)	Model samples from distribution of historical cycle times, given in Fig 3-1
Base Case Simulation Results		
Decant Utilization	0.9128	Slightly lower than 0.927 utilization time estimated from inferred out-of-work time metric

Table 4.1: Single-line decant simulation model assumptions for base case scenario. Values corresponding to labor rates (XX) are redacted.

The full set of base case model assumptions is listed in Table 4.1. It is also important to note that because of the split-dock nature of the fulfillment center modeled here, as described in Chapter 2, trailers with the highest proportions of totes will actually be processed on the smaller non-decant

dock, which in turn reduces the average number of totes and the number of tote wall disruptions to the decant-enabled dock. For this simulation, only "low-tote" trailers, or trailers with less than 800 totes, are considered, which generally approximates the heuristic used by dock managers to decide which dock should process a hybrid trailer.

The primary output metric for this model is decant utilization, or the percentage of time that a worker is actually engaged in her task. This is computed by averaging total working time for all 25 simulated workers and dividing by total simulation time. In this simulated environment, the only factor that can cause decant utilization to drop below 100% is a lack of available work in the case accumulator, so the simulated utilization metric maps closely to real decant out-of-work time. Results from 100 simulations of the base case model are given at the bottom of Table 4.1. The 91.3% utilization result is relatively close to the 92.7% utilization estimate based on the inferred out-of-work time metric presented in Chapter 3. Differences between the two values are likely due to tighter restrictions on the inferred out-of-work metric – in the inferred metric, gaps of less than two minutes are excluded to reduce noise, while the simulation model penalizes utilization for all time periods when a case is not available for even a single decant worker.

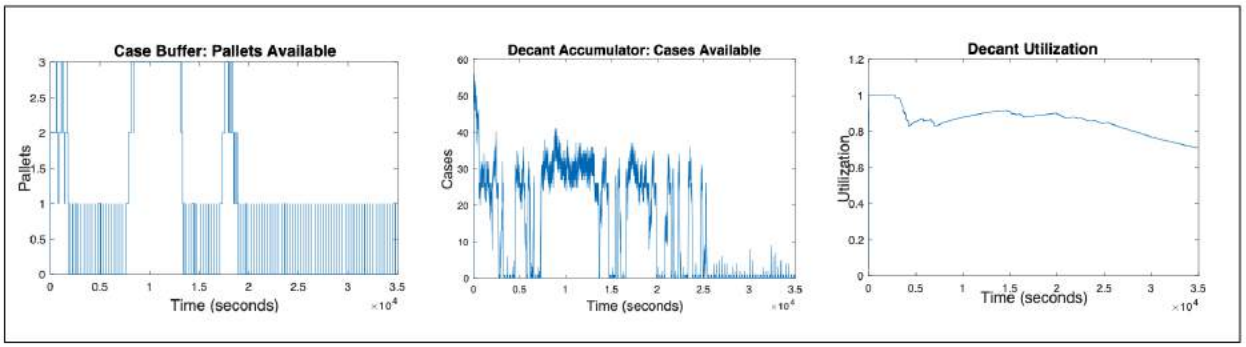


Figure 4-2: Results from a single representative 4-trailer sequence of simulated unload and decant processes. Plots show number of case pallets available over time in the case buffer (left), the number of cases available to decant workers (center), and the decant utilization rate (right).

In this section, a discrete-event simulation model of the combined trailer unload, decant, and case buffer process was presented, with parameters established for a base case that describes the current state process. Results of the base case simulation compare closely with the previously established out-of-work decant metric, and we conclude that the model provides a valid characterization of the system. Subsequent sections will use the same model to analyze the effect of making changes to the system.

4.2 Buffer Sizing

From the base case simulation introduced in the previous section, it is clear that the palletized case buffer used to augment the case flow from fluid trailer unloading is insufficient to prevent the decant line from running dry. The plots in Figure 4-2 are typical, showing the number of case pallets available in the buffer, as well as the number of cases directly available to decanters in the accumulator, over time for a single four-trailer iteration of the base case simulation. At several points during the simulation window, the number of available cases falls to zero, and supplementary cases are fed from the buffer as fast as they can be replenished. The high frequency at which the buffer runs dry makes it clear that it is undersized.

In practice, increasing the size of the case buffer is possible, but due to space limitations and the risk of letting high priority product sit for too long without being stowed into inventory, it is desirable to size buffers no larger than required. To determine the optimal buffer size, a series of simulations is run while varying buffer capacity. Buffer size remains discretized at the pallet level, as pallets are the primary means used to transport case material to the line injection point. In each simulation, the shift begins with a full buffer.

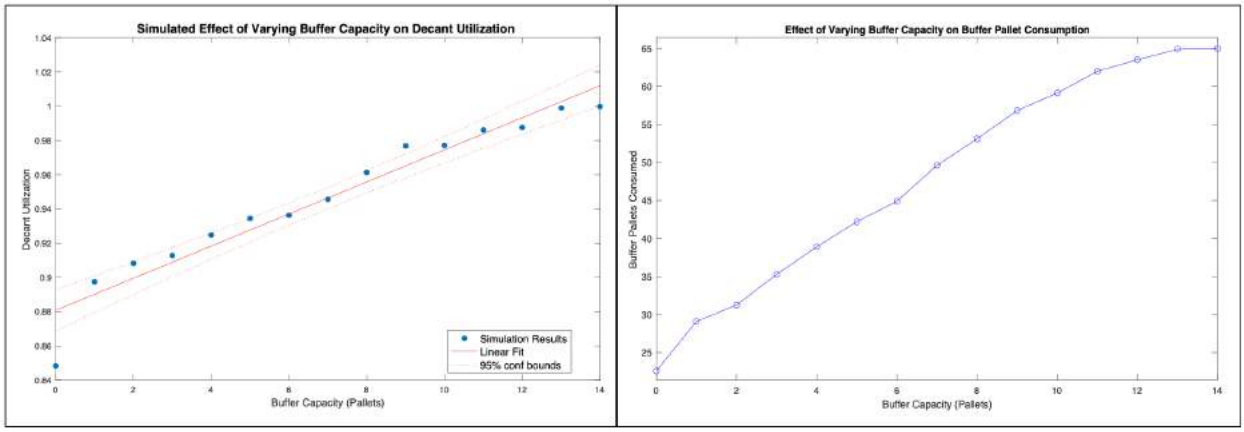


Figure 4-3: Simulation results of increasing case buffer size. Both decant utilization (left) and total buffer consumption (right) increase linearly with buffer capacity.

Figure 4-3 summarizes the results of these simulations. A generally linear increase in decant utilization can be observed as buffer size is increased. 14 pallets of cases is the required buffer size to effectively guarantee that the decant line will never run dry (99.9% utilization) given current system parameters. Furthermore, we see that the consumption of palletized cases from the buffer follows a similar linear trend, increasing with buffer size as more buffer cases are used to fill in gaps in case flow from the docked trailer. Due to space constraints, however, the inbound dock area cannot accommodate 14 pallets of buffer material ahead of each decant line; there is space for a maximum of between eight and ten pallets, depending on the line. The subsequent section on dock process changes addresses these physical layout considerations in more detail.

It is also important to note that the simulation scenarios in this section analyze the effect of increasing buffer size in isolation without considering use of additional labor to replenish the expanded buffer. Replenishment labor utilization increases as the buffer size grows and the capacity constraint preventing continued transport of pallets to a full buffer is relaxed. The following section examines the effects of adjusting labor allocation to the replenishment process. We conclude that increasing the buffer size as much as possible will be beneficial, but insufficient to solve the case flow problem alone, and therefore additional solutions must be considered as well.

4.3 Buffer Replenishment

Because the physical constraints prohibit expanding the decant case buffer beyond a certain limit, another lever to improve case flow availability to decant must be considered. One possibility is to increase the effective unload rate by adding unload workers to each trailer. However, with more than two workers in a trailer, movement becomes congested and individual rates fall rapidly, so this solution is not considered further. Instead, to achieve a greater availability of case buffer

material, the buffer replenishment rate can be increased. Given the practical constraints of the current layout and the manual pallet transportation process, boosting replenishment rate requires additional indirect labor allocation. Using the same process simulation model introduced in Section 4.1, different methods of increasing available labor to the replenishment process can be evaluated.

Table 4.2 summarizes the results of two scenarios where replenishment labor increases are simulated. In the first simulation, an additional worker is added to the buffer replenishment process for each individual line, effectively doubling the base case replenishment capacity. Replenishment cycle time remains constant, but with two workers available, the delays in replenishment are mitigated when there are more one empty pallet slot. This process modification increases decant utilization to 94.0%, a moderate 2.7% increase from the base case. The decant utilization increase is relatively limited because the buffer size is still constrained, which limits the ability of replenishment labor to be fully utilized; indeed, a decrease in utilization for the replenishment task is observed in this scenario. The overall gain at decant from this additional labor is multiplied across 25 decanters, or 67.5% of a single decanter's previous output, and therefore the increased labor expense is not considered worthwhile. Similarly, if a decanter is shifted into the replenishment role instead, a minimum increase in decant utilization of 4.2% (1/24) would be required to justify the lost decant production, which is not achieved.

Buffer Replenishment Labor Simulation Results							
Scenario	**Buffer Size (Pallets)**	**Buffer Replenisher Workers**	**Buffer Replenishment Utilization**	**Decant Utilization**	**Δ Decant Utilization from Base Case**	**Decant Productivity Gained***	**Net Labor Capacity Recovered***
Base Case	3	1	0.8336	0.9128	-	-	-
Additional Buffer Replenishment Labor	3	2	0.6877	0.9402	0.0207	0.6850	-0.3150
Buffer Size Increase Only	8	1	0.8975	0.9613	0.0485	1.2125	1.2125
Increase Buffer Size & Replenishment Labor	8	2	0.7623	0.9911	0.0783	1.9575	0.9575
Increase Buffer Size & Use Flexible Decant Labor	8	1 + 2x flex**	0.8920	0.9864	0.0736	1.8185**	1.8185
*Decant productivity gained and net labor capacity recovered are measured in terms of percentage of labor capacity of one decant worker							
**The last scenario with 2 flex laborers is modeled using a feedback system where 2 decanters shift to augment buffer replenishment only when the buffer contents falls below 2 pallets. Over 100 simulations, the flex laborers are away from their primary roles 14.6% of the time on average. Therefore, decant productivity gained is calculated using the average decanter staffing (24.708).							

Table 4.2: Simulation results for buffer replenishment labor changes.

If both buffer size and replenishment labor are increased simultaneously, we see a more substantial increase, to 99.11% if the buffer is expanded to hold eight pallets. While this higher decant utilization is desirable, the lower buffer replenishment utilization results in a scenario where additional net labor capacity recovered is still less than in the case where only buffer size is increased, and the addition of a second worker is still not justified.

Instead, a more flexible solution is proposed, where decant labor is shifted temporarily, and only when the line risks running dry. This is simulated through the addition of a simple feedback control loop in the model, which reallocates two decant workers to the buffer replenishment role anytime the buffer volume falls below three pallets. We find that, using this rule, the flexible labor is required to work in the replenishment role 14.6% of the time, on average. Taking into account the lost decant productivity during this time, this solution still provides the best gain in net recovered labor capacity, summarized in the last row of Table 4.2. Specific implementation details for a pilot test are discussed in the subsequent section on proposed process changes.

4.4 Trailer Unloading Equipment

In 2020, several Amazon sites have begun fielding a new type of powered-tilt conveyor, pictured in Figure 4-4. This equipment brings advantages in individual unload rate, efficiency, and particularly in employee safety, as it mitigates ergonomic hazards associated with unloading heavy containers from the top of a wall, reducing the distance each container must be moved. However, this equipment comes with some inherent tradeoffs. Due to the way the extension arm moves side-to-side, the machine is intended for use by a single operator only. The single operator's unload rate is significantly higher with the machine than without, but still not as high as two unloaders using conventional methods. This results in an overall net reduction in unload rate and therefore an increased reliance on buffer material from palletized trailers or elsewhere.

Figure 4-4: Powered-tilt conveyor for trailer unloading

Furthermore, the powered-tilt conveyor takes up the majority of the trailer's width, and therefore renders infeasible the previous method of quickly removing totes by sliding out stacks of totes along the trailer wall. Instead, both cases and totes must exit the trailer through conveyance when using this machine. Since only cases continue on to decant, totes must be individually removed from the conveyance at the trailer exit and stacked onto carts for transport to stow. This additional step incurs a higher labor requirement than the previous method of moving stacks directly from the trailer onto carts. While improving this process does not directly affect the flow of cases to decant, reducing the overall labor requirement would, indirectly, free up labor that could be used to manage the increased requirement for case buffer material.

For newer buildings with built-in conveyance to transport totes directly to the stow areas, an ideal process would be for totes to be touched only once during the unload process, and then diverted automatically onto another conveyor at the trailer exit, where the tote would continue directly to one of the stow areas without further human intervention. Such a system could be built at relatively low cost, using a combination of existing technologies already deployed within Amazon FCs. Figure 4-5 provides a sketch design for such a system.

Such a process would also enable a fulfillment center to log each tote's movement individually by bar code, which would provide the capability to more accurately track the volume of work in

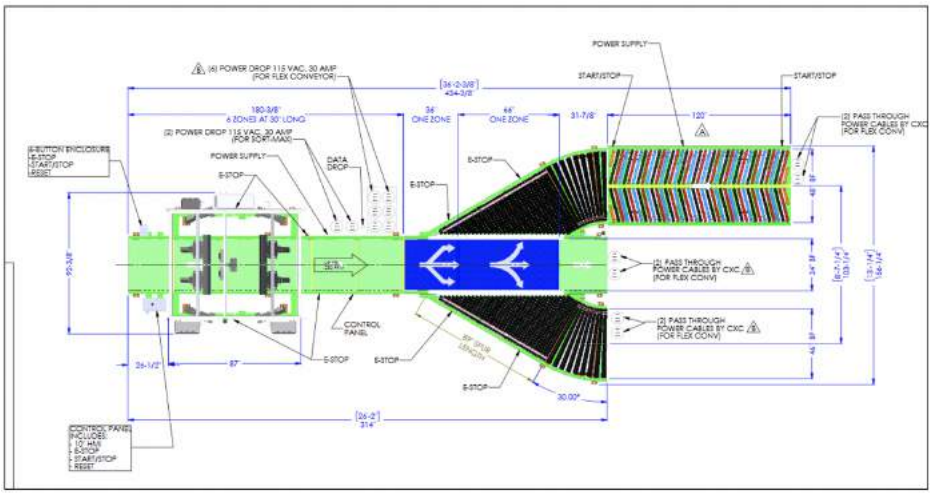


Figure 4-5: Design schematic for an automatic tote diverter, as proposed by a third party supplier. A camera or optical sensor system distinguishes between cases and totes, and a directional roller system diverts the container to an appropriate routing line.

progress (WIP) throughout the inbound department, another ongoing initiative to support stable flow improvements. Accurate WIP counts would further allow the FC to sort and divert specific totes to specific stow areas, improving the ability to balance inventory areas by fullness, item size, and SKU assortment, all metrics closely tracked in the stow and inventory management processes.

The reduction in required labor for automated tote divert and conveyor system is substantial. The tote diverter alleviates the need for one worker per unload line performing the repetitive task of removing totes from the line and placing them on a cart. Depending on the specific building configuration and density of totes in the trailers for which the system is used, follow-on conveyance moving totes would reduce the need for tote transportation by between two and four workers. The overall reduction in labor suggests an annual savings of 21,600 labor hours equivalent for one FC, assuming two 10-hour shifts per day and a conservative reduction of three total workers per shift. This estimate assumes that the FC has sufficient conveyance capacity to move all inbound totes to the stow areas; currently this is not the case, so such a system could not be tested, but newer building designs are planned to have sufficiently increased capacity allocations.

4.5 Proposed Process Changes

From the analysis in the previous sections, we conclude that in the near term, the inbound dock at this particular FC should (a) increase the size of its buffers, and (b) implement a flexible labor plan to temporarily augment the buffer replenishment role by shifting decant labor ahead of disruptions that are likely to cause the decant OOW events. In general, all ARS FCs suffering from the decant OOW problem should follow similar steps to determine appropriate buffer size and required staffing at bottlenecked processes, which may vary with physical layout between buildings. As the newer powered-tilt unload equipment is phased in and unload rate becomes further constrained, these modifications will become increasingly important. This section outlines the specific implementation details of these process changes for one FC.

4.5.1 Piloted Changes

The buffer sizing analysis and flexible replenishment plan derived from the simulation model provide a reasonable starting point for increasing the pallet buffer sizing and using flexible decant labor to augment the buffer replenishment process. However, a few practical considerations demand some additional modifications. First, the buffer replenishment process described in the simulation model is most representative of the line closest to the palletized unloading area, where forklifts deposit pallets in a separate buffer before they are transported manually to the decant buffer (the replenishment process). This is pictured in the current state diagram in Figure 4-6. To supply any decant line further down the dock, palletized cases must be transported via pallet jack all the way around the decant stations, which takes approximately three times longer than replenishing the near side. As this is impractical, the current process instead involves one worker injecting cases onto the near line, while one or two additional employees, standing between the lines, remove cases from the near line and inject them onto the far line. This is a non-value added task that serves to balance the flow of cases, but does not directly add to the overall throughput rate.

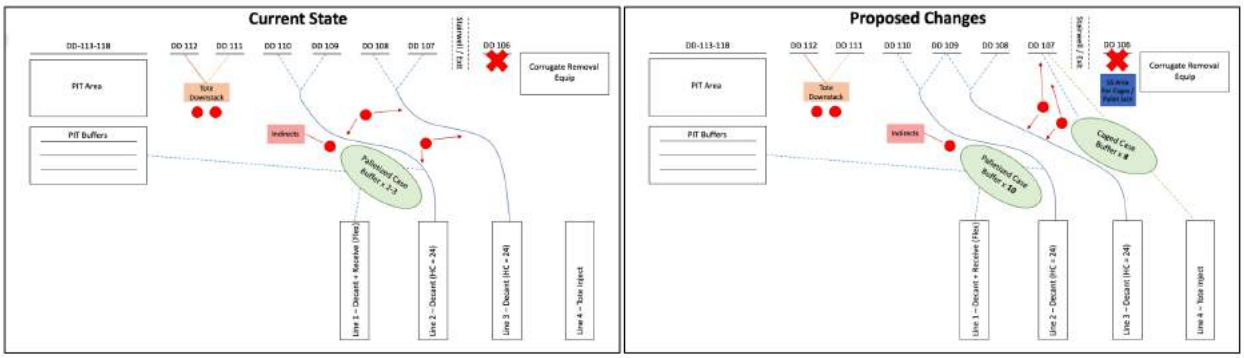


Figure 4-6: Proposed dock layout changes. Previous state (left) and proposed state (right).

An alternative proposal involves re-purposing the labor between decant feeder lines to operate a second buffer, using case material from an additional docked trailer. As shown on the right side of Figure 4-6, an additional floor-loaded trailer can be brought to the farthest dock door and unloaded slowly using carts to transport and hold cases in a buffer on the far side of the dock. Carts hold roughly the same number of cases as a pallet, and are faster to move across the floor than a pallet jack. This system effectively doubles the number of cases available as buffer material, and has the added advantage of increasing the overall unload rate by re-purposing the labor managing the buffers. Figure 4-7 shows the use of carts to supplement the conveyance lines feeding decant. In summary, we move from a single buffer maintained at a level of between two and three case pallets to a system with two separate buffers, one with 10 case pallets and a second with eight carts of cases, each equivalent to a pallet.

The second proposed change to the inbound dock process is to implement a flexible labor plan to allow for shifting decant labor to augment the dock crew and ensure adequate buffer replenishment and case injection when flow disruptions occur. To do this, the first two workers on each decant line (closest to the dock and with best visibility of the buffers) are designated as *flex workers*, and understand that they may be instructed to fill buffer replenishment roles at any time throughout the shift.

While this process change seems relatively simple and benign, in practice, these adjustments are more complex. In the simulation model, labor shifts are executed immediately, with no switching

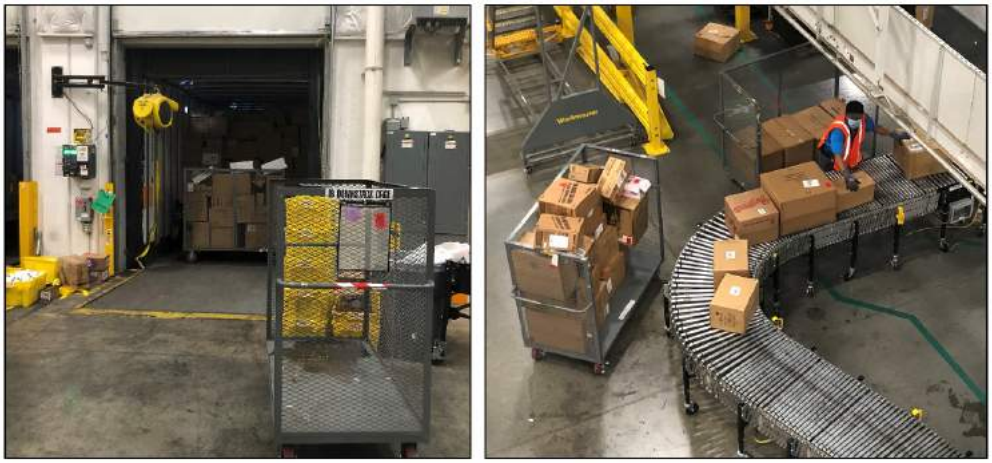


Figure 4-7: Workers using cages to unload an additional floor loaded trailer for use as case buffer material to supplement the flow of cases to the decant line (left). During an interruption in case outflow from a trailer, a worker uses cages with cases to keep the decant line fed (right). These are easier to move than pallets, and this configuration cuts down on transit time based on the FC-specific dock layout.

costs. In reality, labor shifting can be expensive, particularly if not well-executed, and doing this effectively involves numerous human considerations. To deal with the inevitable delays in shifting labor and to reduce the number of shifts overall, the conditions for shifting were modified slightly. Flex workers are moved from decant to the buffer replenishment role whenever the pallet buffer drops below three pallets, and they remain in place until the buffer volume returns to eight pallets. Additionally, the dock managers are given latitude to shift labor proactively, either just before a trailer changeover, or when trailer unloaders come across a large tote blockage that is several walls deep.

The ability to execute these labor shifts effectively relies on real-time notification from the dock floor manger to the decant manager, which means both of these individuals must be in place and paying close attention to the current situation. Furthermore, the decant manager must keep track of the labor sent to the dock and ensure they return to their stations once no longer needed. It was decided not to allow the individual decant employees to make these moves autonomously to ensure accountability, which introduced substantial lag in the labor shifting process. An additional minor cost incurred here is that managers must code labor in the tracking system to account for decant employees' time away from their stations. Finally, this plan required the up-front training of decant workers to perform the buffer management processes on the dock, as well as informing all workers of the associated change to the existing decant rotation plan.

4.5.2 Pilot Results

The proposals outlined in the previous sections were accepted for testing at one fulfillment center, and the two process changes were put in place as part of standard dock work after training and walk-throughs with floor managers and hourly employees. In general, the changes were relatively easy to adopt, and both processes have remained part of the new training standards. To measure the effectiveness of these process changes, we observe the impact on the inferred decant out-of-work metric introduced in Chapter 3 for the lines on which the changes were implemented. Figure 4-8 shows the out-of-work metric, beginning two weeks before the changes were implemented.

Unfortunately, the pilot was terminated early due to a breakdown of the FC's flexible conveyors, which caused the site to transition to a completely manual process of unloading all cases onto carts for transport to decant. The short duration of the pilot makes quantifying improvements difficult. A drop from roughly 20% to 15% OOW time is observed for the period in question, although this improvement occurred amidst already elevated OOW levels that have been consistent since the beginning of COVID-19-related changes. Furthermore, the new powered-tilt unload equipment was introduced only days prior to the pilot, and variations in case flow as dock workers adjusted to this change add another uncontrolled variable.

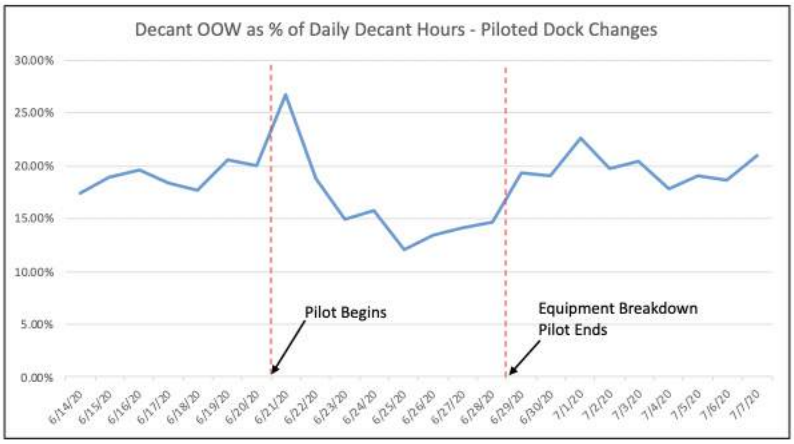


Figure 4-8: Effect of inbound process changes on decant OOW metric.

From a more qualitative perspective, the enlarged buffers did appear to significantly improve the ability of workers to maintain consistent flow to the decant line. However, the new processes also introduced some new friction points. Allocating one of the existing dock doors as a buffer trailer reduces the number of remaining doors available to process other trailers. Without two dedicated doors per floor-load line, trailer changeovers becomes a more complicated process and changeover times can become longer if not carefully planned ahead. Additionally, the available freight in the yard does not always support allocating an extra trailer for case buffer material, although it usually does. Finally, the labor shifting process was clearly effective in providing a temporary boost in feeding cases to the decant line when needed, but managing these labor shifts has proved to be a challenge for managers, who are tasked with monitoring many functions at once and may not be available to execute labor shifts immediately when needed. For this reason, it is suggested that managers consider delegating the responsibility of making temporary role shifts to the assigned individual decant workers, as they are in the best position to see when such shifts are needed.

4.6 Conclusions

While the process changes described in this chapter can be generalized to all Amazon Robotics FCs that use decant, the implementation details are highly specific to a single FC with a particular physical design. However, all decant-enabled sites struggle with the same problem of efficiently filling in gaps in case flow during trailer unloading, and so the methods presented here for determining buffer size and replenishment labor requirements should be widely applicable to any FC or warehouse facing similar problems. At FC inbound docks across Amazon, there is a need to increase buffer capacity and enable labor flexibility as the proportion of hybrid floor-loaded trailers continues to grow across the network and trailer unload rates are constrained by new equipment

types. Permanent solutions may range from the introduction of new equipment, as proposed in Section 4.4, to the simpler changes to increase buffer size and reallocate labor that were briefly piloted during this project.

Chapter 5

Trailer Receive Scheduling Optimization

5.1 Trailer Scheduling Problem Motivation

Inbound trailer scheduling is another significant factor affecting the steady flow of cases to decant, and as a result, a second order cause of decant OOW events. Due to the inherent variability in trailer contents and load configurations, the number of cases and totes, as well as the number of small and medium size units, all vary widely over the course of a typical shift, as shown in Figure 2-12. This variability results in large fluctuations in labor requirements throughout the shift which are difficult to plan for. Section 3.3 introduced the hypothesis that reducing variability of both container type and item size over the course of a shift will result in improved flow throughout the inbound department, positively impacting not only decant but also stow and indirect labor utilization as well.

The current shift planning process relies on priority score as a decision guide for selecting trailers to receive, but in practice, dock managers must take several other factors into account to ensure a feasible schedule that respects building capacity constraints while minimizing labor adjustments. The automatic schedule generated in the current scheduling system, which uses a "greedy," or first-available heuristic, is generally used as a starting point, but a highly manual process ensues to make schedule adjustments with the parallel objectives of (1) ensuring high-priority and reactive freight is processed in a timely manner, (2) allocating trailers with high case percentages to doors with decant lines to minimize manual transit and rework, and (3) level-loading the volume of small and medium size units across the shift, or even allocating rough size-based volume percentages to different stow floors based on available space. While attempting to meet these objectives heuristically, managers must also account for constraints in dock unloading capacity by trailer type (palletized vs floor-loaded doors) and decant line capacity. A final complication is that in newer FC designs with a split-dock layout, each dock serves only a subset of AR floors with minimal ability to transfer product between the two sides of the building. Therefore, allocating volume to each dock adds an additional layer of complexity. Without constraints, the trailer receive schedule for a typical shift has more than 60 trillion possible trailer-door-timeslot permutations, so it is perhaps not surprising that manually created schedules tend to be sub-optimal.

5.2 Previous Work

5.2.1 External Research

The general problem of selecting from available trailers dwelling in the yard, and assigning them to specific dock doors in a particular sequence is a well-studied problem in operations research. There are several common variants which are formulated as optimization problems [11]:

1. Job Shop Scheduling Problem - Multiple jobs need to be scheduled on multiple machines, with constraints on which machines can process each job and on the sequence in which jobs must be performed.
2. Assignment Problem - Traditionally formulated as assigning employees to tasks while minimizing total cost.
3. Parallel Machines Scheduling Problem - Multiple machines which can perform a variety of tasks are the scarce resource, and jobs must be scheduled with labor constraints to minimize overall time or cost.
4. Multi-Period Assignment Problem - A fixed time horizon is incorporated, and assignment occurs over several sequential time periods while minimizing an overall cost function.

In addition to the standard problem definitions listed above, several studies have advanced more specific solutions for a variety of trailer scheduling problems at warehouses and distribution centers. Berghman et al. introduce a time-indexed variant of the parallel machine assignment problem to solve for an optimal dock assignment trailer schedule that minimizes cumulative waiting time and delays [9]. Jain et al. advance another variant of the parallel machine problem using due date constraints, and provide formulations both as a Mixed-Integer Linear Program (MILP) and Constraint Satisfaction Problem (CP-SAT) [22]. Lim et al. contribute a cost-based optimization specifically for trailer scheduling at cross-docks to minimize the total cost of delays, unfulfilled orders, and transportation costs. Both Integer Programming (IP) and Genetic Algorithm (GA) approaches are introduced in this work [5].

Research on optimization-based scheduling specific to e-commerce inbound scheduling has also been contributed through past MIT theses. Russel Forthuber (2017) provides a model for inbound container queuing and selection from a distribution center yard, which optimizes for customer demand satisfaction and DC throughput [20]. David Jackson (2005) advances a longer term scheduling model for assigning freight to different FCs at Amazon based on priority scoring.

5.2.2 Internal Research

There is also current work underway within Amazon to improve inbound trailer scheduling as part of a larger project to automate inbound planning and flow process decisions. The Amazon Fulfillment Technologies software team has an improvement to the current scheduling system which uses an optimization-based approach similar to the job shop scheduling problem. Trailers are slotted with time constraints based on estimated unload rates and manifested container counts, with the objective of maximizing priority score and minimizing total processing time. However, unload rate estimates have proven to be inaccurate, as rates actually depend on the labor allocated to unloading by the dock manager. Furthermore, the latest version of the model under development takes neither item size nor split-dock constraints into account. Item size in particular plays an important role from an operational perspective for several reasons:

1. Item size affects rated functions. Decant and stow both process small items faster, so shift managers tend to want to increase their small items percentage and will try to select trailers with higher percentages of small units.
2. A higher small item percentage also results in higher units per tote, or tote density. This requires relatively fewer totes to be transported to the stow floors to keep stowers fed for a given duration, which translates into less indirect labor required. Similar to above, shift managers will try to select trailers with more small units.
3. Item size distribution stays relatively constant on a daily basis (55% smalls and 45% mediums at the FC modeled here), but varies widely by trailer (Figure 2-15). Because of this, it is possible to use item size via trailer selection to affect shift performance. While individual shift managers want to select trailers with high smalls percentages, senior operations managers responsible for continuity and performance across all shifts will prefer to keep a stable item size ratio.
4. A diversity of item sizes is desirable at the stow process. With multiple size options, stow associates have more opportunities for item placement in each pod. While this is not captured directly in data separate from stow rate, a reasonable hypothesis would be to capture the benefits of item size diversity in terms of reduced changeover time between pods, more units stowed per pod face, and fewer no-stow turnaways (NST), all metrics that are tracked. For these reasons, managers continuously monitor Gross Cube Utilization (GCU), a measure of pod inventory distribution by item size template, and try to minimize deviations across AR floors.

For all of the above reasons, level-loading volume inflows to the FC by item size in addition to container type is considered critical to developing a useful scheduling tool, and this is one of the main contributions of the model introduced here.

Additionally, while many variants of the scheduling problem have been studied extensively, there exists very little research considering how to schedule freight with the objective of reducing downstream variability and improving process flow. Amazon is somewhat unique in its use of two distinct container types (cases and totes) for transshipments, but the attributes of this problem make it applicable to any scheduling application where large batches comprising distinct classes of work must be sequenced to support downstream processing.

5.3 Model Formulation

The inbound trailer scheduling problem borrows elements from several of the problem formulations defined in the previous section. Individual trailers can be viewed as tasks to be processed, while available dock doors and decant lines can be viewed as machines, the scarce resource. As a cost function, the variability metric defined in 2.4 is used, and we seek to minimize the maximum variation from a steady flow volume target for each of the four flow variables (totes, cases, smalls, mediums).

To simplify the problem somewhat, we use a fixed planning horizon, which can be defined as one or more shifts. Shifts are divided into four quarters, the existing labor scheduling discretization used at Amazon. Shift quarters are also a convenient discretization for trailer scheduling because at current processing rates, a good approximation is that one full floor-loaded trailer or two palletized trailers can be unloaded during a quarter.

Binary decision variables are used to represent trailer assignments. For a given shift, there may be approximately 80 trailers available for scheduling, either in the yard or scheduled to arrive during the shift. Assignments are made across three dimensions: trailer ID, dock door number, and a time period (shift quarter):

Decision Variables:

$$x_{i,d,t} = \begin{cases} 0 & \text{if trailer is not assigned} \\ 1 & \text{if trailer is assigned} \end{cases} \tag{5.1}$$

$$\begin{aligned} \text{where} \quad i &= \text{trailer index} \\ d &= \text{dock door index} \\ t &= \text{time period index} \end{aligned}$$

We allow a further classification of dock doors into groups (simply called "docks"), as several Amazon FCs, including the one modeled here, have two distinct docks on separate sides of the building, each feeding a distinct set of non-connected stow areas. We define the following parameters and variables for use in the model and constraint formulation:

Model Inputs & Parameters:

$$\begin{aligned} h &= \text{dock identifier if building has split-dock configuration} \\ v_h &= \text{unit volume percentage allocated to dock } h\text{, if building has a split-dock configuration} \\ v_{flex} &= \text{allowable percentage overage from shift volume target} \\ n_f &= \text{number of floor loads that can be unloaded during time period } t \\ n_p &= \text{number of palletized loads that can be unloaded during time period } t \\ \text{decant}_h &= \text{decant capacity at dock } h\text{, given as a percentage of the building total} \\ \text{pri score threshold} &= \text{all trailers with priority score above this threshold must be processed during the shift} \\ p_i &= \begin{cases} 0 & \text{if trailer is floor-loaded} \\ 1 & \text{if trailer is palletized} \end{cases} \\ r_i &= \begin{cases} 0 & \text{if trailer is not reactive or a live-load} \\ 1 & \text{if trailer is reactive or a live-load} \end{cases} \\ \text{doors}_{h,d} &= \begin{cases} 0 & \text{if dock door is configured for fluid loads} \\ 1 & \text{if dock door is configured for pallet loads} \end{cases} \\ \text{D}_h &= \text{set of dock doors on dock } h \\ s_i &= \text{number of small units in trailer } i \\ m_i &= \text{number of medium units in trailer } i \\ c_i &= \text{number of cases in trailer } i \\ y_i &= \text{number of yellow totes in trailer } i \\ j_i &= \text{ETA for trailer } i\text{, given as estimated first time period when it will be available} \\ \text{priority}_i &= \text{priority score of trailer } i \\ T &= \text{total number of time periods } t \\ u_{c,i}, u_{y,i} &= \text{number of total units in cases or totes, respectively, in trailer } i \end{aligned}$$

Auxiliary Variables:
Next, we define the mean inflow volume of small and medium units, and case and tote containers, for selected trailers across all time periods:

$$\mu_k = \sum_{i,d,t} x_{i,d,t} k_i / T \tag{5.2}$$
$$\text{where } k \in \{s, m, c, y\}$$

Based on the building-specific physical and labor constraints, we can determine target flow volumes for each time period. This is straightforward for small and medium size volumes. For container type flows, we first allocate case volume according to decant capacity, and then allocate tote volumes based on remaining overall unit capacity:

$$\begin{aligned}
&\text{tgt}_{k,h} = \mu_k v_h \quad \text{for } k \in \{s, m\} \\
&\text{tgt}_{c,h} = \mu_c \text{decant}_h \\
&\text{tgt}_{y,h} = \frac{\sum_{i,d,t} [x_{i,d,t}(u_{c,i} + u_{y,i})] \frac{v_h}{T} - (\text{tgt}_{c,h} upc)}{upt} \\
&\text{where } upc = \frac{\sum_i u_{c,i}}{\sum_i c_i} \quad \text{(average units per case)} \\
&\text{and } upt = \frac{\sum_i u_{y,i}}{\sum_i y_i} \quad \text{(average units per tote)}
\end{aligned} \tag{5.3}$$

We define the auxiliary variables $\epsilon^+_{t,h,k}, \epsilon^-_{t,h,k}$ to linearize the desired absolute deviation from the mean volume per time period in the objective, using the method defined in [8]:

$$\begin{aligned}
&\epsilon^+_{t,h,k} \geq \text{tgt}_{k,h} - \sum_{i,d \in \text{D}_h} x_{i,d,t} k_i \quad \forall\ t, k, h \\
&\epsilon^-_{t,h,k} \geq \sum_{i,d \in \text{D}_h} x_{i,d,t} k_i - \text{tgt}_{k,h} \quad \forall\ t, k, h \\
&\text{where } k \in \{s, m, c, y\} \\
&\text{and } \epsilon^+_{t,h,k}, \epsilon^-_{t,h,k} \in [0, \infty)
\end{aligned} \tag{5.4}$$

Finally, we create an additional set of auxiliary variables $z_{k,h}$ which encode the maximum value of the error term $\epsilon^+_{t,h,k}$ and $\epsilon^-_{t,h,k}$ across all time periods:

$$\begin{aligned}
&z_{k,h} \geq \epsilon^-_{t,h,k} \quad \forall\ t, k, h \\
&z_{k,h} \geq \epsilon^+_{t,h,k} \quad \forall\ t, k, h
\end{aligned} \tag{5.5}$$

Objective Function: Our objective is to minimize the maximum deviation from our level target time period volume goals for each of our k flow metrics. We use $\lambda \in [0, 1)$ as a weighting parameter

signaling the relative value of level-loading flow by item size versus container type:

$$\min_{z} \quad \sum_{h} \lambda \left[z_{s,h} + z_{m,h} \right] + (1 - \lambda) \left[z_{c,h} + z_{y,h} \right] \tag{5.6}$$

Define Constraints:

1. Only floor-loaded trailers can go to floor-enabled dock doors, and only pallet-loaded trailers can go to pallet-enabled doors:

$$x_{idt} = 0 \quad \forall\ i \text{ where } p_i \neq \text{doors}_d \tag{5.7}$$

2. Trailer cannot be received before first available time period j_i:

$$x_{idt} = 0 \quad \forall\ i, d, t \text{ where } t + 1 \leq j_i \tag{5.8}$$

3. Trailer must be received during first available time period j_i if load is "reactive" or "live":

$$\sum_{d} x_{idt} \geq r_i \quad \forall\ i, t \text{ where } t = j_i \tag{5.9}$$

4. Every trailer is assigned to no more than one door-time period combination:

$$\sum_{d} \sum_{t} x_{idt} \leq 1 \quad \forall\ i \tag{5.10}$$

5. Total volume must be within the defined shift volume goal range:

$$\begin{aligned} \sum_{i,d,t} x_{i,d,t}(u_{c,i} + u_{y,i}) &\geq \text{shift vol goal} \\ \sum_{i,d,t} x_{i,d,t}(u_{c,i} + u_{y,i}) &\leq \text{shift vol goal} \times (1 + v_{flex}) \end{aligned} \tag{5.11}$$

6. Trailer must be processed during shift if priority score is above a given threshold:

$$\sum_{d,t} x_{i,d,t} = 1 \quad \forall\ i \text{ where priority}_i \geq \text{pri score threshold} \tag{5.12}$$

7. Set the number of floor and pallet trailers that can go to a given door type during each time period:

$$\begin{aligned} \sum_{i} x_{idt} p_i &\leq n_p \quad \forall\ d, t \\ \sum_{i} x_{idt}(1 - p_i) &\leq n_f \quad \forall\ d, t \end{aligned} \tag{5.13}$$

5.4 Model Validation

In order to validate the trailer scheduling model, start-of-shift trailer data was captured over a two week period at one FC, providing a snapshot of available trailers for the particular shift as well as the relevant trailer properties, such as priority score and contents count by container type, item size, and total units. This data was captured one hour prior to the start of both the day and night shifts, at the time when shift planning and scheduling for the shift is normally performed. In addition to this data, actual trailer receive times are logged in a central system which allows for comparison of the actual trailer receive schedules to the "optimal" schedules produced by the scheduling model. This enabled multiple iterations of the model to be developed and improved upon, ultimately producing the model as formulated in Section 5.3.

		Sum of Quarterly Variations from Shift Mean											
		Projected				Actuals (AFT receive time)				Percent Change			
Date	Shift	Cases	Totes	Smalls	Mediums	Cases	Totes	Smalls	Mediums	Cases	Totes	Smalls	Mediums
5/11/2020	D	863.29	1555.13	36145.34	16412.00	12480.75	10905.17	96638.84	93773.67	-93.08%	-85.74%	-62.60%	-82.50%
5/11/2020	N	1770.79	1318.22	46808.73	31609.79	19822.96	11950.55	129392.75	54093.32	-91.07%	-88.97%	-63.82%	-41.56%
5/12/2020	D	3118.52	1628.07	43426.96	32597.61	3903.80	7395.53	193375.07	67513.58	-20.12%	-77.99%	-77.54%	-51.72%
5/12/2020	N	6001.13	1557.18	30711.75	14548.46	12602.65	8399.30	61987.97	63978.47	-52.38%	-81.46%	-50.46%	-77.26%
5/13/2020	D	862.73	765.70	37059.17	30887.57	13105.16	3336.85	81143.61	68638.81	-93.42%	-77.05%	-54.33%	-55.00%
5/13/2020	N	5366.60	1565.10	46592.91	18738.53	13804.31	4350.18	127878.23	82265.06	-61.12%	-64.02%	-63.56%	-77.22%
5/14/2020	D	7566.65	2177.75	42252.97	24726.56	20295.32	5764.22	144769.73	100042.59	-62.72%	-62.22%	-70.81%	-75.28%
5/14/2020	N	1709.74	1935.38	30186.30	25367.36	13123.56	6587.44	134685.40	84632.46	-86.97%	-70.62%	-77.59%	-70.03%
5/15/2020	D	5050.50	2542.07	28040.38	14483.99	15114.90	5130.73	96820.06	71475.05	-66.59%	-50.45%	-71.04%	-79.74%
5/15/2020	N	4722.27	1246.47	40331.39	24753.65	19758.37	14130.16	111381.69	60474.51	-76.10%	-91.18%	-63.79%	-59.07%
5/16/2020	D	3337.73	1776.30	64889.50	25033.80	13775.94	8105.54	147233.15	46760.91	-75.77%	-78.09%	-55.93%	-46.46%
5/16/2020	N	4687.18	1611.16	36500.99	26510.57	11278.21	4227.25	132506.94	88681.09	-58.44%	-61.89%	-72.45%	-70.11%
5/17/2020	D	3265.81	2173.72	15203.18	26402.16	15583.37	7094.39	162795.11	49928.08	-79.04%	-69.36%	-90.66%	-47.12%
5/17/2020	N	1159.50	1431.01	46551.95	9050.85	20163.46	8439.24	86504.74	56818.94	-94.25%	-83.04%	-46.19%	-84.07%
5/18/2020	D	2879.26	2531.47	44144.75	18278.61	8998.94	3183.04	123122.85	83441.13	-68.00%	-20.47%	-64.15%	-78.09%
5/18/2020	N	4668.52	2202.51	35377.82	14024.87	11452.06	10745.61	87624.89	104157.49	-59.23%	-79.50%	-59.63%	-86.53%
5/19/2020	D	3095.47	2051.56	45811.11	30254.36	14256.45	5544.95	168357.13	70472.14	-78.29%	-63.00%	-72.79%	-57.07%
5/19/2020	N	766.59	1817.84	26945.10	17582.08	11578.60	8774.48	73058.30	93981.34	-93.38%	-79.28%	-63.12%	-81.29%
5/20/2020	D	3706.97	980.30	25880.06	29130.82	12961.43	5194.11	108636.97	44830.56	-71.40%	-81.13%	-76.18%	-35.02%
5/20/2020	N	2157.99	1724.93	35355.58	22332.85	20294.13	4872.61	139031.47	85996.38	-89.37%	-64.60%	-74.57%	-74.03%
5/21/2020	D	3149.66	2693.34	42631.46	31420.93	20134.82	5339.74	73457.27	85559.33	-84.36%	-49.56%	-41.96%	-63.28%
5/21/2020	N	5397.59	1378.28	34566.41	12127.99	7160.31	6764.68	198868.81	84191.71	-24.62%	-79.63%	-82.62%	-85.59%
5/22/2020	D	949.77	1878.04	38497.72	15032.65	9449.00	7096.00	153552.97	46954.54	-89.95%	-73.53%	-74.93%	-67.98%
5/22/2020	N	858.10	1896.79	33832.29	25275.41	7954.59	5237.49	133002.15	49563.35	-89.21%	-63.78%	-74.56%	-49.00%
5/23/2020	D	5555.85	2113.72	51212.12	32346.95	14246.84	6052.77	103706.79	70550.39	-61.00%	-65.08%	-50.62%	-54.15%
5/23/2020	N	890.23	2564.05	65745.18	19452.47	14096.87	6347.15	66162.79	74918.00	-93.68%	-59.60%	-0.63%	-74.03%
5/24/2020	D	1348.68	1565.09	56541.21	29149.19	17479.91	8224.38	115356.42	69616.76	-92.28%	-80.97%	-50.99%	-58.13%
5/24/2020	N	3716.16	1672.11	39764.13	37250.71	8286.34	8307.35	120700.86	46498.75	-55.15%	-79.87%	-67.06%	-19.89%
Mean		**3165.1**	**1798.3**	**40035.9**	**23385.1**	**13684.4**	**7053.6**	**120419.7**	**71421.7**	**-73.61%**	**-70.79%**	**-63.38%**	**-64.33%**
Standard Dev		**1874.7**	**463.6**	**10934.9**	**7283.9**	**4332.3**	**2545.1**	**35456.7**	**17333.8**	**19.61%**	**14.40%**	**16.57%**	**16.66%**

Table 5.1: Results of a two-week back-testing period for the trailer scheduling optimization program, comparing projected quarterly variability metrics prior to each shift with actual variability measurement after running standard planning processes.

Table 5.1 provides results from the two week model back-testing period. Using the variability metric described in Equation 2.4, the sum of quarterly variability for cases, totes, smalls, and mediums was computed for both the optimization-based schedule prior to the start of shift and the actual receive plan, based on trailer receive times. Table 5.1 shows the substantial reduction in all variability metrics, with an average reduction in quarterly case variability of 73.6% across each shift. This represents a substantial improvement in stabilizing the time-varying flow of work to various inbound processes, and in particular to decant with a more stable flow of cases and an even

distribution of item size. Figures 5-1 and 5-2 provide visualizations of the improvement in stable quarterly flow typical for a given shift, comparing the results of optimized scheduling with current practices.

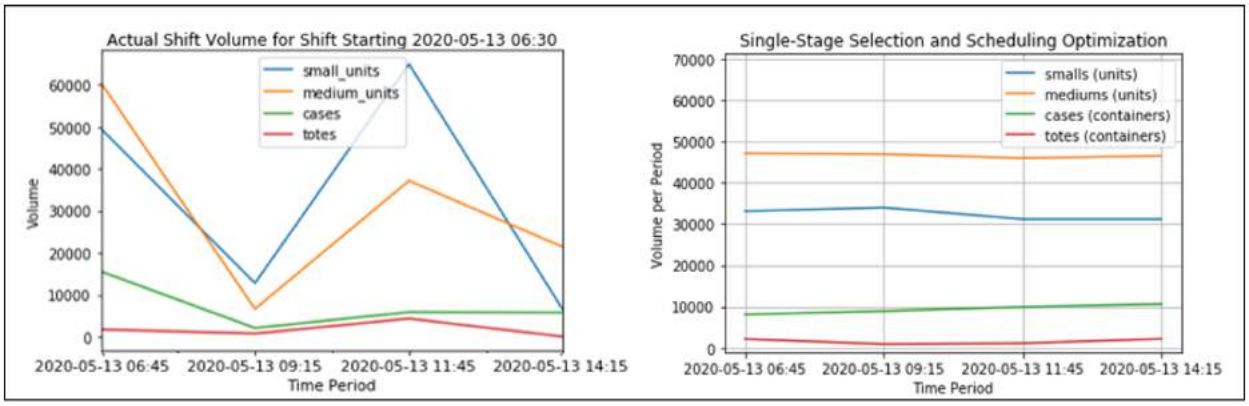


Figure 5-1: Comparison of actual quarterly transshipment receive volume (left) and optimal level-loaded schedule (right) for a representative shift during the back-testing period.

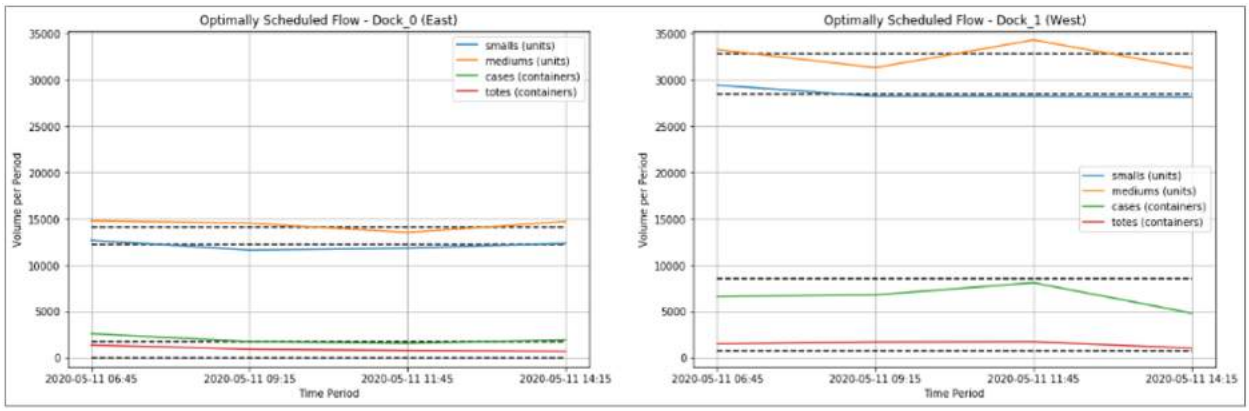


Figure 5-2: Split-dock flow output from optimal scheduling model, where a percentage of volume is allocated to each dock based on processing capacity. Dotted lines show the quarterly target volume for each flow category. This FC's East dock (left) has a quarterly target of zero cases because it does not have decant stations on that side of the building.

We note that the specific outputs of this model somewhat depend on parameter selection. For FCs with split-dock configurations, the v_h parameter which allocates a percentage of volume on each dock depends not only on processing capacity at each dock (as this is somewhat flexible dependent on labor staffing) but also on inventory pod space availability on respective AR floors, as well as the overall container mix scheduled. This is a parameter which must be set at the start of each shift; for the schedules produced here, an average of the fixed relative dock capacity and the total tote-case ratio in the yard was used. Additionally, while the back-testing results show reductions in all flow variability metrics, the relative reductions between measured flows (case and tote vs. small and medium) depend on the specific λ parameter weight in the model objective. The results here reflect a $\lambda = 0.2$, which weights case and tote quarterly variations at four times that of small and medium size quarterly variations, to account for higher relative number of individual units as compared to containers. With an average of roughly nine units per inbound container, this weighting has net effect of prioritizing level-loading by container type at twice the overall weight of level-loading by item size.

5.5 Model Implementation

5.5.1 Technical Implementation

The development version of the scheduling optimization model presented in Section 5.3 is written in Python 3.7, and uses the commercially-licensed FICO Xpress Solver as the numerical optimization engine. The scheduling program runs a series of scripts to first pull and parse latest data from various internal sources, including assigned FC processing volume targets, trailer manifests, scheduled arrival times, yard logs, and priority scores. Because this data is subject to change over time, being able to retrieve the most up-to-date data prior to running the model is imperative.

The size of the optimization problem itself depends primarily on the desired shift volume as well as the number of trailers available in the yard prior to the given shift, which might range anywhere from 30 to 80 trailers, depending on throughput, backlog, and time of year. In the example result given in Figure 5-3, a scheduling problem selecting from 40 trailers available for two separate docks and five total processing lines results in an mixed integer optimization problem containing 840 decision variables and 904 constraints, and a total of 52,516 non-zero elements in the linear program. In general, the number of decision variables and the number of constraints both scale linearly with the number of available trailers. Optimal solution times vary widely for the branch-and-cut solver, but typically range between 30 seconds and 4 minutes. In practice, however, we terminate the solver at 60 seconds and use the best solution available, which improves usability of the tool and the ability to iterate by testing different model parameter selections without degrading solution quality substantially. The below problem resulted in 21 feasible solutions after 60 seconds, the best of which is given here.

Finally, the program includes an interface to display the model output as a human-readable schedule which is usable for dock managers, shown in Figure 5-3.

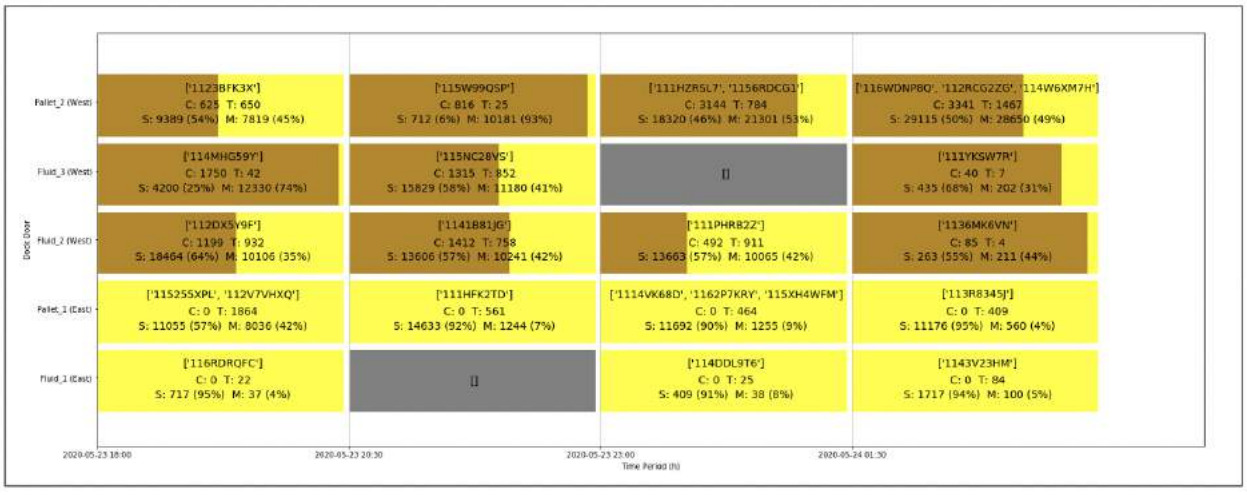


Figure 5-3: Sample optimization-based trailer scheduling model output, showing trailer assignments by trailer ID (in square brackets) and quarterly volumes allocated to each unload process line. Yellow and brown coloring represents relative volume of totes and cases over the shift quarter for a specific unload line. C=case, T=tote, S=small, M=medium. Gray shading indicates no scheduled activity for that period.

5.5.2 Implementation Challenges

Having validated the model on historical data, a natural next step was to pilot the scheduling program for live use in an FC. This was done over a period of several days, with mixed results. The model was able to produce effective schedules at the start of shift, but changes to underlying trailer routings, arrival schedules, and freight priority throughout the course of the shift often required substantial manual deviations from the original plan by the end of shift. In particular, the following specific barriers prevented full adherence to optimized schedules:

1. Data integrity issues and mismatches between the four primary trailer data source systems
2. Trailer routings and manifest updates occurring frequently throughout the shift, requiring the scheduler to be re-run periodically to fetch updates
3. Limited accessibility of scheduling program for dock managers - development program runs locally and requires command-line coding proficiency to manage data inputs and model parameters
4. Uneven inventory pod bin fullness between AR floors, requiring specific trailer routings for re-balancing, which is not accounted for in the level-loading model

5.5.3 Future Improvements

To overcome these implementation issues, several improvements are suggested. First, a permanent software solution that is scalable across Amazon's FC network must be built using a robust cloud-based service architecture, as all software services are at Amazon. This would allow more reliable and real-time connections to source data systems and existing services which provide updates to time-varying data such as trailer transit schedules and priority scores.

Ideally, the scheduling program would run with a fixed-planning horizon default of two full shifts. In testing, this time horizon was identified as providing the most benefit in reducing both inter- and intra-shift flow variations, while also using input data that is somewhat stable. Trailer schedules more than 24 hours out are otherwise too variable and tend to be missing yet-to-be-manifested routings. To deal with the frequent updates, the scheduling algorithm should be re-run automatically at set intervals, likely at the beginning of each shift quarter, and re-plan taking into account freight processed since the last update, as well as updates to the trailer input data sets. This type of dynamic re-planning is much more feasible with a web-service based architecture, rather than a program running locally on a single laptop.

To make the program more user-friendly for dock mangers, an improved user interface should be added, allowing simple adjustments to building-specific configuration parameters, such as volume allocations to separate docks, objective function weightings, and priority score thresholds.

Finally, several opportunities exist for such a scheduling program to be integrated to support other ongoing initiatives. The issue of uneven inventory bin fullness on different AR floors was briefly mentioned as a common cause of manual schedule adjustments. This is another operations process for which no clear standard procedures exist, and practices followed by different FCs are highly manager-specific. As part of a separate ongoing project, the Amazon Fulfillment Technologies team is testing different methods for tracking and allocating in-process inventory by bin size to the appropriate inventory area through a variety of physical sortation technologies. With a more granular, real-time demand signal for specific item size properties, this data could be integrated

into a future version of the scheduler program to support decisions at the upstream trailer receive decision point.

For all of the above-stated reasons and requirements identified, further FC testing of the scheduling model was put on hold, and the model has been passed off to the AFT team working on the Inbound Online Shift Planner, introduced in Section 2.3.1. The team has agreed to incorporate this model into their labor planning and scheduling logic, to be released in 2021. This decision allocates proper resources to developing the prototype model into more robust software that is tightly integrated with Amazon's existing planning and information systems.

5.6 Conclusions

This chapter introduced the detailed motivations behind the trailer scheduling assignment problem, and reviewed existing research in scheduling algorithms for operations and specifically for trailer scheduling and assignment in the distribution center context. A formal optimization model formulation was presented, followed by an analysis of model performance improvement through back-testing against historical data along the dimension of quarterly flow volume variability for container types and item sizes entering the FC. Estimating or attributing performance improvements to specific processes, such as decant OOW hours, is difficult without conducting a well-controlled experiment using the model for actual trailer scheduling over an extended period of time, which was not possible in this case. However, the analysis presented in Chapter 3 makes it clear that highly variable flows along container type and unit size dimensions have adverse effects on multiple inbound processes, so an automated scheduling solution that imposes negligible operational cost but demonstrates potential for large variability reductions is of great interest to Amazon. At a minimum, we would expect to see improvements within the following processes as a result of implementing an optimal scheduling tool:

1. Reduction in decant OOW time as a result of more steady inflow of cases
2. Reduction in idle time for trailer unloading as a result of fewer conveyance line backups
3. Reduction in stow OOW time as a result of the previous two effects
4. Improvements in stow rates as a result of a more even distribution of item sizes over time
5. Reduction in labor switching costs, particularly for buffer replenishment and tote transport tasks on the inbound dock, as a result of more consistent product inflows

With the adoption and further development of this model by the AFT team, it will be possible to evaluate the real effects of using an optimization-based scheduling system to level-load inbound product flows in the near future.

Chapter 6

Inbound Cross-Dock Process Changes

To complete the life cycle analysis of the hybrid floor loaded trailer, the IXD loading process, which is the fundamental source of variability in the configuration of containers within the trailer, is examined. Chapter 3 lays out the reasoning attributing variability in trailer load contents configuration to decant OOW events. The specific layout of case and tote containers in hybrid floor loaded trailers depends on the frequency, volume, and variability of containers arriving at fluid lane diverts off of the main IXD ship sorter, as well as on the actual loading method. The product sortation and allocation algorithm, briefly described in Chapter 2, is an entirely separate and complex component of Amazon's supply chain optimization framework, and as such, changes to this system are out of scope for this project. This chapter will focus on the details of the End of Line trailer loading process, specifically for floor loaded trailers, and will assess trailer loading strategies that might reduce container type variability in the unloading process.

In total, three possible changes to the trailer loading process are identified and evaluated, each with the potential to improve the unload process and consistency of case flow to decant at the receiving FC:

1. Separating fluid trailers into case-only and tote-only trailers
2. Longitudinal or lateral separation of totes and cases in trailers
3. Tote wall height limit, to ensure some number of cases is available in each wall during unloading

6.1 Previous Work

Jeffrey Birenbaum's 2018 thesis examined network transportation and process cost reduction strategies for Amazon's transshipment network. Specifically, this research considered the labor and transportation cost differences between floor loaded and palletized trailers, concluding that the dominance of transportation costs make the more densely-packed floor loaded trailers generally more cost-effective [10]. Labor savings from reducing manual palletization in the IXD was roughly offset by additional labor incurred from unloading floor loads in the FC. However, this study was undertaken prior to decant being implemented as a primary processing mechanism in FCs, so additional costs due to decant OOW events are not accounted for in this work.

Additionally, Birenbaum's research considers a hypothetical transition to all-tote floor loaded transshipments. Even while ignoring physical IXD processing limitations, he concludes that the

lower unit volume utilization of totes as compared to cases renders this method impractical, as the increased transportation costs from all-tote transshipments vastly outweigh labor savings in the FC.

In the intervening two years, Amazon's transshipment strategy has remained largely aligned with the conclusions and recommendations presented by Birenbaum. While remaining consistent with current product destination and container-type allocation systems, the following analysis considers more granular changes to existing loading processes which seek to reduce container-type variability during trailer unloading.

6.2 Separation of Cases and Totes

6.2.1 Separation of Cases and Totes into Different Trailers

Instead of processing hybrid loads, if floor loaded trailers arrived to FCs as separate pure case and pure tote loads, trailers could easily be allocated to different doors with different processing capabilities, greatly alleviating the problem of disruptions in case flow to the decant line. From the FC perspective, this is an obvious optimal solution, but one that poses significant challenges at the IXD outbound dock.

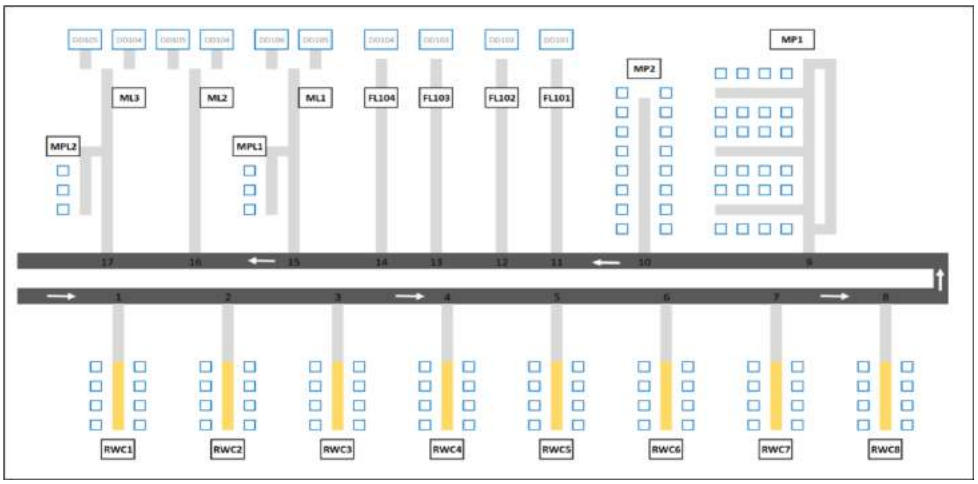


Figure 6-1: Typical IXD Outbound Lane Layout. Multi-load lanes, where a single sorter divert feeds two dock doors, are shown in the upper left. Sites are actively converting these into single fluid lanes to increase the number of distinct FC destinations served.

Dividing a stream of containers destined for one particular transshipment arc into separate pure case and pure tote floor loads is possible in one of two ways. Some IXDs have "Multi-Load" lane configurations, where a single sorter divert feeds two separate dock doors (Figure 6-1). Generally, these lanes already do allocate one door to each container type, as long as the ratio of cases and totes arriving allows both trailers to be filled in the allotted time. This process is generally effective, producing pure or near-pure case and tote floor loads. However, for this process to be expanded, more lanes would need to have two dock doors allocated. Instead, IXDs have a mandate to serve a larger number of distinct arcs to meet network product placement and transship efficiency targets, and so multi-load lanes are already being converted into single fluid-load lanes in many cases. Dock doors are a constraining factor here, and so any dock door allocated as a second door for a given arc would need to demonstrate value above the network efficiency and product placement gains expected from the IXD serving an additional destination.

Because of the variable arrival stream of cases and totes diverted off of the sorter, without a second dock door, containers would need to be sorted by type. This would require a buffer capable of holding up to a full trailer's worth of either cases or totes. Given that current outbound dock space supports a buffer of roughly 150 cases or totes, more than ten times the current available space would be required to support this. Additionally, such a strategy would incur substantial additional loading labor costs, as buffered containers would each incur two additional process touches. This additional labor outweighs any gains at the receiving FC, so this strategy is not considered further.

6.2.2 Separation of Cases and Totes within Same Trailer

Longitudinal Separation

From the FC perspective, if hybrid floor loads are required, an ideal load configuration would be to split the trailer into distinct tote and case sections. In a longitudinal separation configuration, cases and totes would be split between the front and rear sections of the trailer, and the FC dock team could effectively treat the load as two distinct trailers, using the most appropriate unload and downstream process for each section. Cases would flow to decant uninterrupted, and tote unloading would occur at a different time and use a separate set of designated equipment and labor.

In the IXD, however, while this strategy is not as costly as the separation by trailer policy described above, facilitating this type of loading process is still cost-prohibitive and physically infeasible. Similar dock space constraints exist that prohibit the large buffer area required to offload either cases or totes, while the other container type is loaded into the trailer. In the worst case, for a lane with an even 50-50% split of cases and totes, a longitudinal separation strategy would require a buffer with capacity for approximately 750 totes or 1300 cases (half of a trailer's volume). In a best-case scenario, a trailer with 80% cases and 20% totes, more than 300 totes would have to be held in a buffer while cases are loaded (Figure 6-2). With a current per-lane buffer capacity of approximately 150 containers, physical space is a limiting factor.

Lane Tote Percentage	20%	50%
Buffer Space Req'd (totes)	312	780
Addl Labor / Trailer (h)	0.867	2.167

Figure 6-2: Buffer space and labor requirements for splitting hybrid floor loaded trailers longitudinally by container type

In addition to the space constraint, operating a larger buffer incurs a substantial labor cost. Each case or tote that would be offloaded from the line and moved to a buffer, and later retrieved from the buffer and loaded into the trailer, requires additional touches and labor hours. Using an estimated five second transport time per container, occurring twice per container (insertion and removal from the buffer), the additional labor requirements are given in Figure 6-2. As compared to the estimated average savings of 3.8 decant labor hours per trailer from entirely eliminating interruptions in case flow, a longitudinal separation strategy in trailer loading may be cost effective in some cases, but a substantial portion of the savings entitlement is eaten by additional IXD labor. Even in the 80-20% container mix case, the marginal gain would have to be weighed against the opportunity cost of creating more buffer space in the building, and as IXDs are already seeking to expand the number of fluid load doors available, it is unlikely that this would have a positive net benefit.

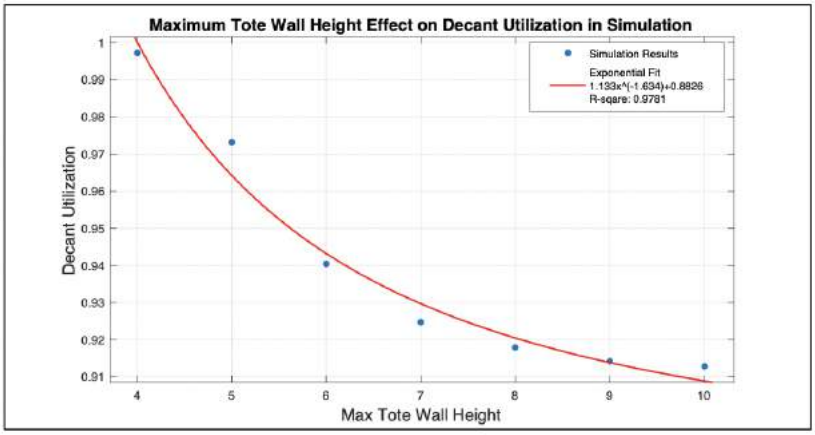
Lateral Separation

Lateral separation of cases and totes in a trailer is a similar scheme, but one that alleviates the need for as large of a buffer area external to the trailer. In this method, cases and totes are each loaded as they arrive, on a different side of the trailer. This requires an up-front decision about the case-tote ratio in a particular trailer; based on the substantial hourly variability in container ratios introduced in Figure 2-7, this ratio would be difficult to predict. In addition, the variability in container arrivals means that one side of the trailer risks being built significantly faster than the other. Sending an employee deeper into a trailer to add totes to a wall, for example, when the adjacent walls of cases are already stacked to the trailer ceiling, presents high risk of walls toppling and causing severe injury. Furthermore, lateral separation decreases the overall load stability of each wall, and increases the likelihood of containers shifting during transit, creating further hazards in unloading. Because of these inherent safety considerations, lateral separation of cases and totes is not considered further.

6.3 Tote Wall Height Limit

6.3.1 Effects on Trailer Unloading and Decant

If hybrid trailers cannot be separated into distinct case and tote sections, another option is to implement a maximum tote wall height policy, effectively limiting the allowable height of totes within each container wall, while providing additional space in each wall for cases. Such a policy would reduce the impact of tote wall disruptions on the FC inbound flow process by ensuring that at least some number of cases are available in each wall during the unloading process.



Tote Wall Height Limit	Decant Util
10	0.9128
9	0.9142
8	0.9179
7	0.9247
6	0.9404
5	0.9731
4	0.9972

Figure 6-3: Results of running trailer unload and decant simulation model for various tote wall height limits. Decant utilization increases at the FC as tote wall height decreases, because more cases are available in each wall, so tote walls do not completely block the outflow of cases from each trailer.

Using the same trailer unload and decant simulation model presented in Chapter 4, the base case simulation is altered by sequentially implementing different tote wall height limits of between five and ten totes. A tote wall height limit simulates a restriction on trailer loading, prohibiting walls with totes stacked higher than the given limit. To simulate this, the same historical trailer manifest data is used from the base case model, but virtual trailers are constructed respecting the given tote limit. In the event that a particular trailer manifest does not have enough cases to fill the

space above totes in each wall, the wall is terminated early (i.e. empty space). For each tote wall height limit, 100 simulations of four randomly selected historical trailers each are run, and average decant utilization across all of these simulations is summarized in Figure 6-3.

As expected, limiting the height of tote walls enables a more steady flow of cases out of each trailer, which in turn reduces decant OOW event and improves utilization. However, adding constraints on the trailer loading process increases the likelihood that trailers will be under-utilized. Specifically, for high tote percentage arcs, lower wall height limits will inevitably result in trailers being built with less-than-full walls, increasing the number of trailers required and thus increasing overall transportation costs. Therefore, to accurately assess the impact of a tote wall height limit policy, we must compare the gains in downstream FC processes with the additional transportation and IXD labor costs.

6.3.2 Loading Process and Trailer Volume Utilization

Loading Simulation Model

To estimate the impact of a tote wall height limit on trailer volume utilization, we must consider the distribution of cases and totes at each IXD floor load lane, and then estimate the impact of adjusting the tote height limit. The resulting trailer configuration depends on the specific loading logic, the distribution and variability of the incoming container stream, and the tote height limit policy, as well as specific decisions made by the trailer loading crew. The loading process logic with multiple container types becomes fairly complex, so a discrete-event model is used to simulate loading trailers based on a given container arrival stream. Figure 6-4 details the process logic modeled for loading a hybrid floor load trailer.

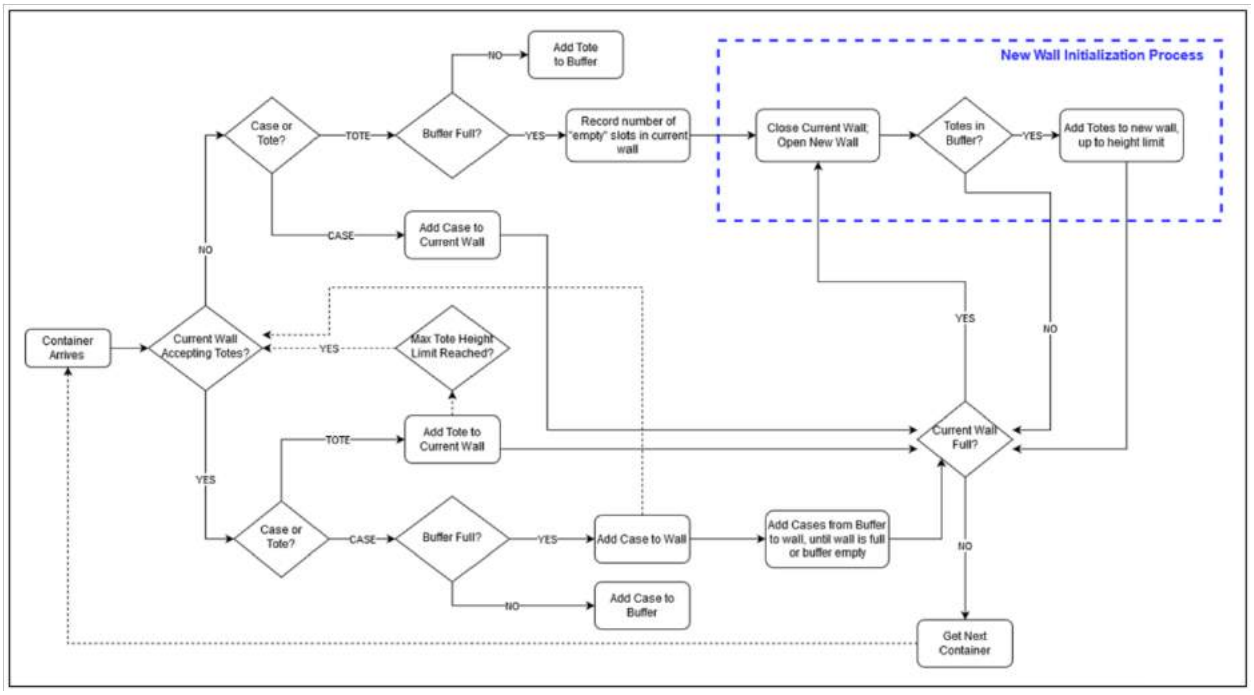


Figure 6-4: IXD floor load process logic for trailer loading, based on observation of actual trailer loading process for multiple lane configurations.

The arrival of cases and totes diverted from the IXD sorter to specific lanes can be modeled as a Bernoulli process, where cases and totes arrive in a continuous stream with complementary probabilities. Probabilities for each container type are considered to be fixed at the set ratio for a specific shipping arc, as these tend to remain stable on a weekly basis. The simulation model ingests the stream of input containers, one at a time, and following the loading process logic (Figure 6-4), determines whether to place the next container into the current wall, the small 50 container buffer, or start a new wall. Using 100,000 containers in each simulation, numerous virtual trailers are constructed, and we estimate average trailer volume utilization.

We acknowledge the significant hourly deviation in container ratios from the weekly mean (Figure 2-7) as a limitation of this model, although results from simulating trailer loading using real minute-level container arrival data are within 2% of simulation results that take random Bernoulli-generated container streams as input when aggregated at the weekly level.

This simulation model required some specialization to deal effectively with the logic of creating virtual "wall" and "trailer" groupings and to keep track of empty spaces within each trailer. The limitations of existing commercial simulation software made this challenging, and instead, a purpose-built simulator was written in Python 3.7, and all results presented in this section make use of that simulation system.

Model Results

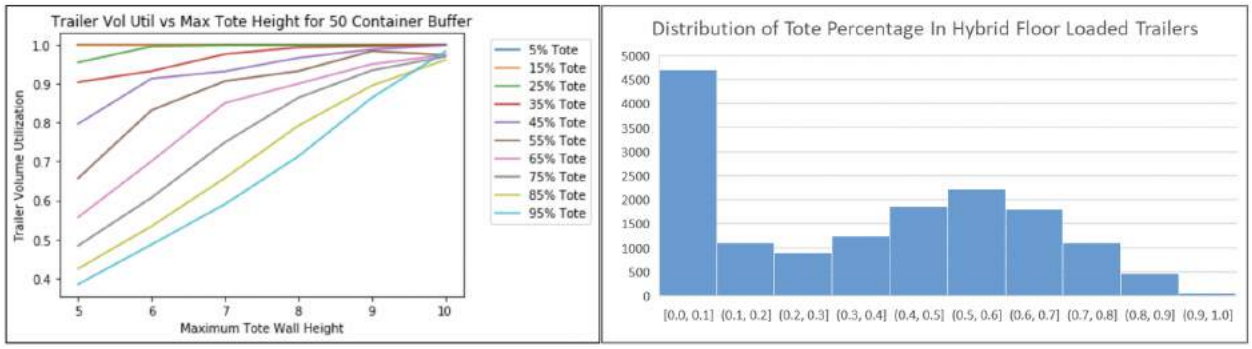


Figure 6-5: Left: Simulated trailer volume utilization vs maximum tote heights for various transshipment arcs with different case-tote distributions. Loading 100,000 randomly ordered containers at the specified case-tote ratio and tote wall height limit is simulated for each iteration. Right: Distribution of tote percentage in network transshipment hybrid trailers. Data is sampled from all North American network transshipments during the first five months of 2020.

We find that as the tote wall height limit decreases (i.e. becomes more restrictive to tote usage), trailer utilization drops, and this effect is more pronounced for trailers with higher tote percentages. Within tote-heavy lanes, it is increasingly likely that the small loading buffer will already be full of totes once the tote limit for a given wall is reached. In this case, trailer loaders have no choice but to begin the next wall, leaving empty space at the top of the previous wall. This has the effect of reducing overall trailer volume utilization and therefore increasing the number of trailers needed to transship the required freight along a given arc. Figure 6-5 gives results of simulations for all combinations of tote height limits and tote percentages, using 100,000 containers per simulation. To estimate the effect of implementing a tote wall height limit across the network, we use the historical distribution of all hybrid floor loaded trailers by tote percentage across the North American FC network to compute a weighted average trailer volume utilization for each possible

tote height limit. This utilization value can then be used to determine the additional trailer capacity required, and based on the average network hybrid floor load trailer cost, the additional cost per trailer incurred can be calculated as well. These calculations are shown in Figure 6-6, with results given as labor hour equivalents to protect proprietary cost information.

Simulated Data - Weighted Average Trailer Utilization Estimate								
Trailer Samples	15472		Wall Height					
Tote % Bucket	N (trailers)	% of Total Trailers	5	6	7	8	9	10
0.1	4602	0.2974	1	1	1	1	1	1
0.2	1137	0.0735	1	1	1	1	1	1
0.3	895	0.0578	0.9555	0.9965	1	1	1	1
0.4	1205	0.0779	0.9045	0.9331	0.9768	0.9943	0.9977	1
0.5	1815	0.1173	0.7972	0.9137	0.9325	0.9672	0.9889	1
0.6	2249	0.1454	0.6569	0.8324	0.9073	0.9331	0.9846	0.9744
0.7	1851	0.1196	0.5565	0.7011	0.8511	0.9008	0.9509	0.9748
0.8	1139	0.0736	0.4834	0.6072	0.7497	0.8646	0.935	0.9704
0.9	506	0.0327	0.4243	0.5334	0.6578	0.7925	0.8957	0.962
1	73	0.0047	0.3839	0.4864	0.5903	0.7137	0.8643	0.984
Wtd Avg Cube Utilization			0.8035	0.8777	0.9274	0.9560	0.9816	0.9898
Avg Trailer Cost*		66.46						
Cost per Trailer (Lost Utilization)*			13.06	8.13	4.82	2.92	1.22	0.68
Addl Cost per Trailer (above current state)*			12.38	7.45	4.14	2.24	0.54	0.00
* All cost data normalized to labor hours equivalent to protect proprietary transportation and labor cost data.								

Figure 6-6: Simulation results used to compute weighted average trailer volume utilizations based on number of network transshipment trailers at each volume threshold. This is done for each tote wall height policy. Utilization is then used to compute an incremental trailer cost based on tote wall height threshold.

6.3.3 Cost Comparison

With average per-trailer cost estimates of the effect of tote wall height limit policies on FC inbound processes, IXD loading processes, and transportation requirements, it is possible to assess the overall benefit of these policies on Amazon's operations. Table 6-7 provides a comparison of these variables. Included here is a relatively minor increase in IXD labor cost – this reflects the added labor cost incurred above the current state by requiring dock workers to add tote lids to the top layer of totes in each tote wall, when cases will be stacked above. For a given transshipment arc tote ratio, as the tote wall height limit is lowered, the number of tote walls increases, and so the lid requirement increases as well.

	Incremental Costs (Savings) per Trailer			
Tote Wall Height	Addl Trailer Cost	IXD Labor Cost	FC Labor (Savings)	Net Change
10	0.00	0.00	0.00	0.00
9	0.54	0.10	-0.06	0.59
8	2.24	0.12	-0.22	2.14
7	4.14	0.13	-0.51	3.76
6	7.45	0.15	-1.19	6.41
5	12.38	0.19	-2.61	9.96
* All cost data normalized to labor hours equivalent to protect proprietary transportation and labor cost data.				

Figure 6-7: Comparison of projected costs and savings for tote wall height limit implementation

Overall, it is clear that the increase in transportation cost due to lower trailer utilization is nearly an order of magnitude greater than the savings expected from continuous flow improvements on FC labor. Because the various models that underpin this analysis rely on numerous strong assumptions, a sensitivity analysis is included in Table 6-8. Three scenarios are considered where system parameters are adjusted boldly in the direction required to favor a tote wall height limit.

However, even with these assumptions, implementing a tote wall height limit does not produce a cost savings (negative cost number) in any scenario. We therefore conclude that a tote wall height limit is not a cost-effective means of reducing FC unloading variability.

Sensitivity Analysis - Net Cost Change for Scenarios A, B, C				
Tote Wall Height	**Base Case**	**A**	**B**	**C**
10	0.00	0.00	0.00	0.00
9	0.59	0.56	0.42	0.31
8	2.14	2.05	1.47	1.02
7	3.76	3.56	2.52	1.69
6	6.41	5.94	4.17	2.68
5	9.96	8.95	6.24	3.77
Scenario Descriptions				
A - Increase Decant Hours Savings by 50%				
B - Decrease Avg Trailer Cost by 30%				
C - Trailer loading leaves 50% less empty space				
* All cost data normalized to labor hours equivalent to protect proprietary transportation and labor cost information.				

Figure 6-8: Sensitivity analysis for three additional scenarios described in relation to the tote wall height analysis. Positive tabulated values reflect additional cost per hybrid FL trailer. All values are positive, indicating that none of these scenarios provides any net cost savings when considering FC, IXD, and transportation costs.

6.4 Conclusions

This chapter considered several options for modifying the current IXD floor load trailer loading process in such a way as to reduce container type variability during the unloading process. Both physical separation of totes and cases, as well as a limit on the height of tote walls within trailers were considered. In summary, all proposals were either infeasible due to safety or floor space constraints, or they were not cost effective. The first two proposals, physical separation of cases and totes, require more buffer space to perform the separation than is currently available, and would require additional doors allocated to each shipping arc, which conflicts with the mandate for IXDs to free up doors to increase the number of destination FCs served. Implementing a tote wall height limit suggests substantial benefits from the FC perspective, but simulation of the trailer loading process shows a reduction in trailer volume utilization which incurs additional transportation costs that outpace FC labor savings. This is consistent with previous analysis which triggered the removal of a tote height limit in 2019. Moreover, implementing recommendations described elsewhere in this thesis will reduce the adverse effect of hybrid loads on the decant process within the FC, further reducing the benefit of any of the IXD-related changes examined in this chapter.

Chapter 7

Conclusions and Future Work

7.1 Summary of Recommendations

Hybrid floor loads will continue to grow as a proportion of transshipments as Amazon's IXD network continues to expand and increase the number of FC nodes served per hub, and FCs will bear the brunt of the inherent container type variability in unloading. Moreover, the decant out-of-work problem will continue to grow as projected 2021 new ARS building launches will more than double the number of fully decant-enabled FCs. Several improvements to mitigate these adverse effects are possible however, and this thesis proposes the following recommendations:

1. **Inbound Dock Processes**
   - FCs should track decant OOW time and report it as a key performance metric. The cost of lost labor to the network is high, but currently only decant rate is tracked, which encompasses several extrinsic factors. A system to track decant OOW events in real time will be required to do this.
   - FCs can perform site-specific analysis to determine appropriate case buffer size for their decant lines, based on the frequency and duration of disruptions, as well as available floor space. Additional trailers may be used for buffer material when floor space is a limitation.
   - FCs may consider using flexible labor assignments to manage variable-demand labor processes such as decant buffer replenishment. Depending on the specific material transport process used, buffer replenishment is likely to be a bottleneck process with adverse downstream effects on continuous work flow, but one that is not immediately obvious to floor managers due to the time delay between replenishment and buffer depletion.
   - In newer buildings with more robust fixed conveyor capacity in the inbound department, Amazon should consider supplementing the newly-fielded powered unload machinery with an automated divert system to reduce repetitive container touches and manual transport.
2. **Trailer Scheduling** - Optimization-based trailer scheduling to level load time-varying freight inflows has the potential to significantly improve several metrics across the FC inbound department, including decant OOW times. The Inbound Online Shift Planner scheduling module, currently under development by the AFT software team, should ultimately consider item size

and include constraints for split-dock FCs, in addition to the existing process path and building specific capacity constraints. AFT is currently incorporating these features from the model presented in this thesis into a more widely scoped tool.

3. **IXD Trailer Loading** - No changes to the current IXD processes for allocating destination lanes and loading hybrid trailers are recommended, as the options for modifying loading processes explored in this thesis do not yield expected cost reductions.

7.2 Future Work

This thesis primarily examined improvements to the IXD - FC transshipment process through the lens of addressing disruptions to the continuous flow of work in the FC inbound department, particularly to the newer decant process. The broad problem scope and limited timeframe required narrowing the solution space to practical changes that can be made in the near term. Furthermore, as this project was carried out in partnership with Amazon's Process Engineering Team, the problem scope was constrained to existing processes and systems, and therefore did not consider some larger-scale changes to FC building designs or network systems, initiatives for which other teams within Amazon are responsible. However, throughout the course of this research, several additional opportunities for potential process and system design improvements across Amazon's transshipment operations were exposed, some of which may deserve closer inspection as part of future research:

1. **Increasing the number of dock doors and decant lines** This option is not considered in the context of this thesis as it would require a change to new building design. Increasing the number of trailers being simultaneously unloaded, but with a lower standard staffing for each line, would potentially enable easier labor shifting across decant lines to absorb flow variability, while also providing FCs with a higher peak capacity for use during the holiday season. A more comprehensive model of costs and expected utilization would be required to value such a project.
2. **Optimal Capacity Sizing of Inbound Conveyance** At several points throughout this thesis, we remark on product routing decisions and indirect labor staffing as being dependent on conveyor capacity which carries totes between the inbound dock and a FC's various stow floors. Newer buildings are increasing the capacity of these conveyors, but still leaving much of the expected transport volume to be done manually. Modeling the full costs and savings from inbound conveyance, including labor reductions in upstream activities such as trailer unloading and container sortation, as well as line jams resulting from insufficient capacity, could be done to estimate the optimal investment for this type of automation.
3. **Use of AGVs for Material Handling** While some Amazon FCs are already using Automated Ground Vehicles for repetitive transportation tasks, most are not. Assessing the savings impacts, solving the implementation challenges in specific building layouts, and designing safe processes for human interaction are all projects which carry large potential benefits from a flexible technology that can be quickly re-purposed as requirements change.
4. **Automated Trailer Unloading Technologies** Since 2019, several companies have unveiled prototype robotic systems for unloading trailers. Assessing the efficiency of these systems and developing potential integration solutions with surrounding dock processes could dramatically change the way that inbound inventory is processed.
5. **Methods to Increase Tote Volume Utilization** One of the primary reasons for using hybrid case and tote trailers for transshipments in the first place is that totes have a lower volume utilization than densely-packed vendor cases, and so are only transshipped when breaking apart cases is required. The current automated product sortation system simply drops

items into a tote, but advances in robotics and computer vision might allow for a more tightly-configured packing of items into totes that could change this dynamic entirely. This would help Amazon to further standardize downstream FC processes and accelerate the implementation of further automation.

6. **Supplier Packaging** Amazon already works with major suppliers to reduce excess packaging to work towards sustainability and transportation cost reduction goals. Further opportunities may exist to standardize and further improve or eliminate packaging to reduce the need for decant in some cases. Possibilities for providing standardized reusable containers (totes or similar) to suppliers also exist, and further study could explore the tradeoffs between increased transportation costs and reduced warehouse processing labor.

Bibliography

[1] Amazon. *Getting Started with Fulfillment by Amazon (FBA)*. URL: https://sellercentral.amazon.com/gp/help/external/ (visited on 02/10/2021).

[2] *Amazon.com 2018 Annual Report*. 2018. URL: https://ir.aboutamazon.com/annual-reports-proxies-and-shareholder-letters/default.aspx (visited on 04/16/2020).

[3] *Amazon.com 2019 Annual Report*. 2019. URL: https://ir.aboutamazon.com/annual-reports-proxies-and-shareholder-letters/default.aspx (visited on 04/16/2020).

[4] *Amazon.Com, Inc. Form 10-K*. Dec. 31, 2018. URL: https://www.sec.gov/Archives/edgar/data/1018724/000101872419000004/amzn-20181231x10k.htm (visited on 03/30/2020).

[5] Andrew Lim, Hong Ma, and Zhaowei Miao. "Truck Dock Assignment Problem with Operational Time Constraint Within Crossdocks". In: *IEA LNAI* 4301 (2006), pp. 262–271. URL: https://link.springer.com/chapter/10.1007/11779568_30.

[6] *Annual Net Revenue of Amazon from 2006 to 2019, by Segment*. Statista, June 2020. URL: https://www.statista.com/statistics/266289/net-revenue-of-amazon-by-region/.

[7] Youssef Aroub. "Container Removal and Replacement Automation: Design and Development of Robotic Work Cells for Warehouse Automation". Cambridge, MA: Massachusetts Institute of Technology, June 2019. URL: https://dspace.mit.edu/handle/1721.1/124574.

[8] Benjamin Granger, Marta Yu, and Kathleen Zhou. *Optimization with Absolute Values*. URL: https://optimization.mccormick.northwestern.edu/index.php/Optimization_with_absolute_values (visited on 04/11/2020).

[9] Lotte Berghman, Roel Leus, and Frits C. R. Spieksma. "Optimal Solutions for a Dock Assignment Problem with Trailer Transportation". In: *Annals of Operations Research* 213.1 (Feb. 2014), pp. 3–25. ISSN: 0254-5330, 1572-9338. DOI: 10.1007/s10479-011-0971-7. URL: http://link.springer.com/10.1007/s10479-011-0971-7 (visited on 08/21/2020).

[10] Jeffrey Birenbaum. "Inbound Supply Chain Optimization with Ship-Mode Variation in a Fixed-Capacity Fulfillment Center". Cambridge, MA: Massachusetts Institute of Technology, June 2018. 74 pp. URL: https://dspace.mit.edu/handle/1721.1/117981.

[11] Peter Brucker. *Scheduling Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. ISBN: 978-3-662-12944-9 978-3-540-24804-0.

[12] Paul Buijs, Hans W. Danhof, and J.Hans C. Wortmann. "Just-in-Time Retail Distribution: A Systems Perspective on Cross-Docking". In: *Journal of Business Logistics* 37.3 (Sept. 2016), pp. 213–230. ISSN: 07353766. DOI: 10.1111/jbl.12135. URL: http://doi.wiley.com/10.1111/jbl.12135 (visited on 03/30/2020).

[13] Rafał Burdzik, Maria Cieśla, and Aleksander Sładkowski. "Cargo Loading and Unloading Efficiency Analysis in Multimodal Transport". In: *PROMET - Traffic&Transportation* 26.4 (Aug. 7, 2014), pp. 323–331. ISSN: 1848-4069, 0353-5320. DOI: 10.7307/ptt.v26i4.1356. URL: https://traffic.fpz.hr/index.php/PROMTT/article/view/1356 (visited on 03/30/2020).

[14] J. Clement. *Emarketer.Com Survey*. June 2019. URL: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

[15] D Collis. "Walmart Inc. Takes on Amazon.Com". In: Harvard Business School Publication (May 2018), p. 30. URL: https://hbsp.harvard.edu/product/718481-PDF-ENG.

[16] David O. Jackson. "Managing and Scheduling Inbound Material Receiving at a Distribution Center". Thesis. Cambridge, MA: Massachusetts Institute of Technology, June 2005. URL: https://dspace.mit.edu/bitstream/handle/1721.1/34853/63199983-MIT.pdf?sequence=2&isAllowed=y.

[17] *E-Commerce in the United States*. Statista, May 2020. URL: https://www.statista.com/study/28028/e-commerce-in-the-united-states-statista-dossier/.

[18] George Eckes. *The Six Sigma Revolution: How General Electric and Others Turned Process into Profits*. New York: John Wiley, 2001. 274 pp. ISBN: 978-0-471-38822-7.

[19] George S. Fishman. *Discrete-Event Simulation: Modeling, Programming, and Analysis*. Springer Series in Operations Research. New York: Springer, 2001. 537 pp. ISBN: 978-0-387-95160-7.

[20] Russel G. Forthuber. "Inbound Container Queuing Optimization Model for Distribution Centers". Cambridge, MA: Massachusetts Institute of Technology, May 2017. URL: https://dspace.mit.edu/handle/1721.1/111265 (visited on 04/06/2020).

[21] Jean-François Houde, Peter Newberry, and Katja Seim. *Economies of Density in E-Commerce: A Study of Amazon's Fulfillment Center Network*. w23361. Cambridge, MA: National Bureau of Economic Research, Apr. 2017, w23361. DOI: 10.3386/w23361. URL: http://www.nber.org/papers/w23361.pdf (visited on 08/11/2020).

[22] Vipul Jain and Ignacio E. Grossmann. "Algorithms for Hybrid MILP/CP Models for a Class of Optimization Problems". In: *INFORMS Journal on Computing* 13.4 (Nov. 2001), pp. 258–276. ISSN: 1091-9856, 1526-5528. DOI: 10.1287/ijoc.13.4.258.9733. URL: http://pubsonline.informs.org/doi/abs/10.1287/ijoc.13.4.258.9733 (visited on 08/21/2020).

[23] Jeffrey Bezos. *Amazon.Com 2019 Letter to Shareholders*. Apr. 16, 2020. URL: https://ir.aboutamazon.com/annual-reports-proxies-and-shareholder-letters/default.aspx.

[24] Olufemi Oti. "Hub and Spoke Network Design for the Inbound Supply Chain". Cambridge, MA: Massachusetts Institute of Technology, May 2013. URL: https://dspace.mit.edu/handle/1721.1/81007?show=full (visited on 03/30/2020).

[25] Andrey Pavlov and Mike Bourne. "Explaining the Effects of Performance Measurement on Performance: An Organizational Routines Perspective". In: *International Journal of Operations & Production Management* 31.1 (Jan. 11, 2011), pp. 101–122. ISSN: 0144-3577. DOI: 10.1108/01443571111098762. URL: https://www.emerald.com/insight/content/doi/10.1108/01443571111098762/full/html (visited on 05/12/2020).

[26] Mary Bryna Sanger. "Does Measuring Performance Lead to Better Performance?: Professional Practice". In: *Journal of Policy Analysis and Management* 32.1 (Jan. 2013), pp. 185–203. ISSN: 02768739. DOI: 10.1002/pam.21657. URL: http://doi.wiley.com/10.1002/pam.21657 (visited on 05/12/2020).

[27] Dan Spitzer. *E-Commerce & Online Aucations in the US*. 45411a. IBIS World, May 2020. URL: https://my.ibisworld.com/us/en/industry/45411a/.

[28] *The Trend towards Warehouse Automation*. White paper. Westernacher Consulting, Dec. 2017. URL: https://westernacher.com/white-paper-the-trends-towards-warehouse-automation/ (visited on 05/12/2020).