Standardization of Workflow in a Large Distribution Center

By

Stephen Michael Greenlee

B.S. Mechanical Engineering, Texas A&M University, 2013

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT AND THE MECHANICAL ENGINEERING DEPARTMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREES OF

MASTER OF BUSINESS ADMINISTRATION
AND
MASTER OF SCIENCE IN MECHANICAL ENGINEERING

IN CONJUNCTION WITH THE LEADERS FOR GLOBAL OPERATIONS PROGRAM AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUNE 2019

Signature of Author: Signature redacted
Mechanical Engineering, MIT Sloan School of Management

Certified By: Signature redacted
Dr. Stephen Graves, Thesis Supervisor
Professor of Management

Certified By: Signature redacted
Dr. Maria Yang, Thesis Supervisor
Professor of Mechanical Engineering

Accepted By: Signature redacted
Nicholas Hadjiconstantinou
Chair, Mechanical Engineering Committee on Graduate Students

Accepted By: Signature redacted
Maura Herson
Assistant Dean, MBA Program, MIT Sloan School of Management

Standardization of Workflow in a Large Distribution Center

By
Stephen Michael Greenlee

Submitted to MIT Sloan School of Management
on May 10, 2019 in partial fulfillment of the requirements
for the Degree of Master of Business Administration
and Master of Mechanical Engineering

Abstract

The retail industry is shifting to enable companies to respond faster to consumer demand and expectations. For any retail company, this requires speed in their supply chain from new product generation to final order delivery. Companies that store product in centralized distribution centers must shorten the time it takes to ship a product from when an order is placed.

This thesis describes the detailed operations within a large distribution center and uses it as a basis for improving delivery time of a product or order within the four walls of the building. The current system is subjected to increased variability in workflows from work planning to work completion, causing delays within sequential work functions and a longer overall delivery time. These effects are magnified by the inherent tradeoffs in the work process format and the work behaviors of the employees.

A new system of work was developed to standardize the workflow at a large distribution center and decrease observed order delivery times. This solution was a work scheduling system that established clear expectations for work completion as well as tools needed to reduce the variability in the system. Under this new system, the average order delivery time is expected to decrease to a third of its current cycle time.

This research was conducted in partnership with Nike Inc.

Thesis Supervisor: Dr. Stephen Graves
Title: Abraham J. Siegel Professor of Management

Thesis Supervisor: Dr. Maria Yang
Title: Professor of Mechanical Engineering

Acknowledgements

The author would like to thank the many individuals for their support and guidance through this project.

Firstly, I would like to thank my fiancé Catie for her endless support of my education and our lives together. From the decision to attend this program, through the frequent location changes, down to the final career decision, I will always be grateful for her understanding, flexibility and encouragement. I would not like to thank our cats, Barley and Hops.

I would like to thank my academic advisors, Stephen Graves and Maria Yang for their advice and thoughtful questioning throughout the internship. Their weekly replies to my rambling emails helped me gain clarity in the project and made the final deliverable much more thorough and complete than it would have been otherwise. Their faith in the exploration process helped calm the nerves of a student anxious to define a project.

Thank you to Angharad Porteous for being the most supportive, understanding, practical and energetic supervisor I have had in my career. There was never a moment where I felt that my efforts were forgotten or that I couldn't tap into her extensive network of lifelines for my project. Her faith in her team's abilities and her consistent drive to learn together is something I hope to model in my own career.

Dzmitry Velikan, my supervisor in the Memphis facility, always made time to hash out new concepts and talk through execution of solutions. Many of the ideas discussed in this thesis were formed during early conversations with Dzmitry in a team room or across the ping-pong table. His help with getting engrained in the operations team were critical to the success of this project.

Lastly, I would like to thank the many people who accepted me into the Nike culture and allowed me to contribute freely as part of the team. This includes Vanessa Frumau, Joe Williams and the SRS night crew, Mario Waijiya, Mark LeBoeuf, Gareth Olds, Liz Walker, Ankit Goswami and Michael Burton.

Notes on Nike Proprietary Information

To protect information that is proprietary to Nike, Inc., the data presented throughout this thesis has been modified and does not represent actual values. Data labels have been altered, converted or removed to protect competitive information, while still conveying the findings of this project.

Contents

1. Introduction

1.1 Thesis Objectives

This thesis aims to lay out a system for enabling large and complex distribution centers (DCs) to send products in a fast, efficient and efficacious manner. Large retail companies face challenges within their distribution network as their infrastructure can be built for speed, volume or for optimizing certain order profiles. Commonly these priorities can compete with one another during day-to-day operation depending on the scale of the operation. Issues can arise when companies must fulfill orders from multiple customer types within the same facility, each with their own sets of needs.

This thesis takes a granular view of the inner workings of any large distribution center, which is relevant for the operations of any large-scale retailer that fulfills wholesale customers. The facility faces the challenge of adapting the current infrastructure to new customer trends in the marketplace. This research was done in cooperation between Nike Inc. and the Leaders for Global Operations Program at MIT.

The documentation for this study will begin with an overview of the company, including the current business strategy and market trends, focusing mostly on the North American market. The next chapter will describe the inner workings of a distribution center and provide context for the thesis. Chapter 3 will discuss academic perspectives concerning the common issues and tradeoffs observed in a distribution center. These concepts will be tied back to practical application within distributions centers and provide insight to specific decisions in work completion processes. Chapter 4 identifies the specific items within the workflow of the DC that can be changed to create better outcomes, and chapter 5 outlines the new workflow system for implementation and the expected outcomes. A summary of findings as well as a condensed set of recommendations will be found in chapter 6.

Confidential values concerning the research supply chain and distribution network, such as time and volume, have been obscured for the purposes of this thesis. Values expressed in this thesis should not be taken as accurate representations. These values will, however, maintain their relative values to provide justification for recommendations.

1.2 Nike Overview and Strategy

1.2.1 Nike Overview

Nike is a large retail company that specializes in the design, fabrication and distribution of athletic equipment. Nike Inc. consists of multiple brands, including Hurley, Converse, Air Jordan as well as many more. Together they cover over a third of the athletic footwear market and generate over 30% more revenue than their closest competitor, Adidas (Reference for Business, 2019). Worldwide, Nike sells in over 160 countries and distributes over 1 billion units of clothing, equipment and footwear a year.

Product is sold through multiple channels within Nike's distribution network including wholesale vendors, Nike stores and online sales. Wholesale vendors may include familiar stores like Foot Locker, Dick's Sporting Goods (DSG) and Kohl's and represent a significant amount of the throughput of product for Nike. Wholesale vendors can also include smaller boutique shops that can represent strategic doors that Nike must serve in order to gain influence in strategic markets. These important customers do not produce a significant volume, but they represent a customer that needs high levels of service for product delivery[1].

Nike brands have direct sales stores that are treated similarly to any other Nike owned store. Combined, Nike operates almost 1,200 direct channel stores globally and almost 400 in the US alone (Statista, 2018). These stores are similar to the wholesale customer group, as they demand high volume and require fast turnarounds to maintain relevant product in stores. Lastly, Nike also sells product directly to individual consumers through online channels such as Nike.com, the SNKRS app and the Nike app. All the consumers who purchase product through Nike owned stores and through online purchases are considered by Nike to be 'Direct' sales.

1.2.2 Nike Strategy

In 2017, the company's CEO Mark Parker introduced a new strategy for Nike called the 'Triple Double'. The core of this strategy was laid out in three main pillars, "2X Innovation, 2X Speed and 2X Direct connections with consumers" (Nike, 2017). In the release, the 2X innovation references

[1] For reference, Nike refers to wholesale companies as customers, while individuals who purchase a final product are called consumers

the acceleration of development for new product lines and technology research while the 2X speed refers to speeding up the development pipeline to shorten the time period between idea inception to sales.

The 2x direct pillar referred to the expansion of direct consumer interaction and sales. This referred mostly to sales through Nike operated stores and, even more importantly, sales and interaction on Nike's internet and mobile platforms. This focus on direct-to-consumer sales allows Nike to learn more from consumers than through sales through traditional wholesale channels. Additionally, Nike can retain a higher sales margin through direct sales without working through wholesale entities.

The 'Triple Double' represents a reaction to the changing marketplace for retail companies such as Nike. There are many opportunities to take advantage of in the new retail marketplace, but challenges must be overcome. Wholesale customers require more flexibility and speed in the supply chain to ensure they better meet consumer trends that can shift weekly. Consumers are now accustomed to shopping online with speedy turnaround periods that can put a strain on the retailer's supply chain. The more effective Nike is at responding to these demands, the better equipped they will be at solidifying their position as the market leader.

1.3 Nike North America

In 2015, Nike opened the North American Logistics Campus (NALC) in Memphis, Tennessee. At the time of construction, this was the largest DC in the world, with over 2.8 million square feet of space. This facility was meant to house and distribute all the product to be sold in the US across the three primary product channels; footwear, apparel and equipment. This facility was built to optimize delivery based on the demand profile of the time, which was centered around wholesale customer fulfillment. Since its construction, Nike has grown their online presence and heavily increased their direct-to-consumer sales. Today, NALC has been fitted to fulfill online orders through the facility.

1.4 Project Context and Motivation

Nike's distribution network is a global operation that fulfills demand across the world. Across this supply chain, Nike operates its own distribution centers. While minor changes to the

infrastructure of their DCs have allowed Nike to keep up with demand, the company, and any large retailer, stands to gain significant rewards by proactively upgrading their supply chain within this single node.

Customers across wholesale sales require faster turnaround times for smaller order quantities. Previously, a customer may have placed bi-weekly orders for 10,000 units; currently, they may be moving to weekly orders for 5,000 units. Individual consumers are becoming accustomed to the Amazon Prime model of 2-day shipping as a standard service, requiring retailers to respond immediately to orders as they occur to get the orders to contract delivery services as quickly as possible.

This means any large retailer DC must reduce the cycle time of an order within the facility while maintaining throughput volume and service quality. When we refer to the cycle time of an order, we refer to the time taken from when an order is released to the floor to when it is completed and shipped. For the purposes of this thesis, the current state is fulfilling orders with a typical cycle time of 12 hours with a 99% service quality (i.e. 99% of product shipped in full and on time) and a shift volume of 100,000 units. The target for this operation is to maintain or increase service quality and volume but to do so with an average cycle time of eight hours. The following thesis will explore and evaluate options for how this operation can adapt to fulfill these goals.

2. Operational Background and Key Metrics

This chapter will give a high-level overview of the operation within a typical large DC, including the infrastructure, workflow norms, and tracked metrics in the facility. This information was gathered using interviews and participation in ongoing operations. This section is meant to show the current state of the facility, while the next chapters will explain root causes and developed solutions.

2.1 Operations Overview

Unless otherwise noted, the role of automation in this facility is quite low within the listed work centers. All tasks described in the following section are done by human laborers. The largest role of automation is the conveyor system throughout the facility that ensures products are delivered to the appropriate areas. After delivery, human interface is needed to ensure products receive the necessary labor.

2.1.1 Work Centers and Infrastructure

The DC is divided into four primary work centers and six support work centers all connected by a series of conveyors, belt sorters and information channels. A typical DC will have a much more complex interaction between work centers, but this general framework will be suitable for considering options in any general DC. Figure 1 shows a high-level overview of process flow.

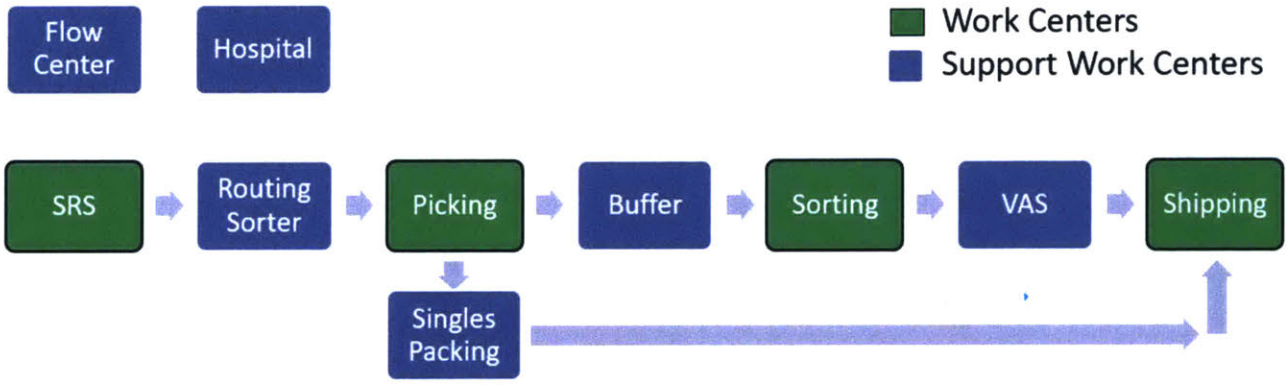

Figure 1. High level overview of process flow in a DC

Storage and Retrieval Services (SRS)

Primary responsibilities:

- Receive product inbound to the DC for storage
- Store product to fulfill future orders
- Pick Cartons as needed to fulfill orders

Storage:

All product within SRS is contained in cartons. A single carton holds multiple units of items, typically six shoes per carton (or 12 per carton for children's sizes).

There are two primary areas within SRS, Pallet Reserve and Case Reserve. High quantity SKUs that can fill up a single pallet are placed in pallet reserve. Pallets never contain more than one SKU per pallet. Lower quantity SKUs are stored in Case Reserve, where individual cartons are placed on racks for later retrieval.

SRS is organized into zones, aisles, rows and bays. Each zone can contain multiple aisles (side by side). Each aisle has multiple rows (stacked to the ceiling). Each row holds dozens of bays. Each bay can contain only one pallet at a time in pallet reserve.

Workflow:

Cartons are received as they arrive at the inbound dock bays. If the cartons are bound for Pallet Reserve, they are palletized and collected by a forklift for storage[2]. If the cartons are bound for Case Reserve, they are collected by a single forklift driver and individually placed on the racks. Product is stored on the racks in Pallet and Case Reserve until a customer order is placed.

When SRS receives a request for a product, a driver in a forklift will drive to the product location and retrieve a carton[3]. After that driver collects a set number of cartons, he or she will drive to the conveyor belt and deposit the cartons (called the 'throw line'). Here, another worker will collect the cartons and feed them onto the conveyors, all cartons placed onto conveyors in this work area will be sent to the routing sorter.

[2] No product is received already palletized

[3] A carton is a box of multiple units. One carton typically contains six pairs of shoes

Capacity Constraints:

- Limited number of forklifts for worker use
- Due to safety concerns, only one worker allowed in aisle at a time. Too much traffic in SRS can cause major wait times in aisles with fast moving product
- Routing sorter has mechanically constrained throughput value. If SRS exceeds routing sorter capacity, conveyors will halt.
- Work density effects (see Chapter 4. Root Cause Analysis)

Picking

Primary Responsibilities:

- Receive cartons from SRS, unbox them, and store units in picking slots
- Hold single unit inventory for use in later orders
- Pick units as needed to fulfill orders

Storage:

All product within Picking is stored as individual units. One unit represents one pair of shoes, a package of socks, etc.

There are three main areas of Picking: Unit Pick, Cart Pick and the Pick Module. Low volume product is contained in unit pick, and the pick slots can contain maximum two cartons of product (for a typical product size). Medium volume product is stored in Cart Pick, where the picking slots are larger, able to contain ~4 cartons depending on the product. The Pick Module contains the highest volume product, and stores units on rolling racks that can accommodate a large amount of storage.

Like SRS, Picking has zones, aisles, rows and bays. These work in the same way as SRS.

Workflow:

When Picking receives cartons from SRS, a worker will collect the cartons, unbox the units, and place them in a designated picking slot.

When Unit Pick or Cart Pick receives a request for a unit, a worker will pick the unit from a slot and put this unit in a tote. Totes can contain ~10 units depending on size. After collecting a pre-determined number of units, the worker will place their totes on a conveyor to send it on to a secondary sorter.

The Pick Module works differently than the other picking areas. Here, a tote will move around the Pick Module to specific locations based on product needs. Workers at these stations will collect these totes and fill each tote as needed. They will then place the tote back on the conveyor where it will move to a new location in the Pick Module, or down to a secondary sorter.

It is reasonable to claim that ~50% of the product moved through picking runs through the Pick Module, with ~35% through Cart Pick and ~15% through Unit Pick.

Capacity Constraints:

- Limited number of carts in Unit Pick or Cart Pick. Limited number of stations in Pick Module
- Work density effects (see Chapter 4. Root Cause Analysis)

Buffer

Primary Responsibilities:

- Accept totes from picking and cartons from SRS and hold order batches together
- Hold orders until they are complete enough to send to the sorters

Workflow:

The buffer is in place to ensure that when products are sent to the sorters, all the units for an order arrive near the same time. The sorters have limited capacity of outbound cartons. Ensuring that all units assigned to an outbound carton arrive at the same time allows the chutes to be cycled quickly, reducing the number of chutes needed in steady-state operation.

Capacity Constraints:

- Physical capacity (i.e. storage) limited by size of buffer and number of lanes

- Each group of orders requires own section, limits maximum number of order groups outstanding for the DC

Sorting

Primary Responsibilities:

- Sort and organize units into outbound carton selection and quantities as per specific order requirements
- Pack units into outbound shipping cartons[4]

Layout:

Sorting has three primary sorters; the footwear sorter, apparel sorter and the mixed sorter. The footwear sorter can only handle product that fits within a standard shoebox. The apparel sorter can only handle product in soft packaging, such as socks. The mixed sorter is equipped to handle either type of product and is used on orders that contain a wide variety of product types.

Workflow:

When sorting receives product from SRS or Picking, it inducts single units onto a circular conveyor. This conveyor will carry the product to a chute and deposit the unit in the chute. The sorter will deposit units in the order into this chute until the units on this order are ready to be packed and completed. Once this happens, a light will turn on and a worker will open the chute and pack the product into the outbound carton. This carton will then be sent to shipping or Value-Added-Service (VAS).

Capacity Constraints:

- Mechanical throughput limit on sorter, however this limit usually exceeds demand placed on sorter
- Limited number of induction conveyors, capping maximum product input

[4] Outbound and inbound product are contained in cartons, which are packages of multiple units. Inbound cartons contain only one type of sku, while outbound cartons can hold many sku varieties

- Limited number of chutes on sorters, limiting number of outbound cartons available for sorting

Singles Processing Area

Primary Responsibilities:

- Receive totes from Picking and pack individual units into outbound cartons
- Send cartons to Shipping for shipment

Workflow:

Most direct online orders are for single units. These units do not need to travel through sorting, so they are sent to a dedicated packing station that is equipped for efficient single orders. Here a worker will receive a tote from Picking and pack the units in each tote into outbound cartons. These workers will also pack a receipt, apply a shipping label and send the package to Shipping.

Capacity Constraints:

- Limited number of stations for packing product

Value-Added-Service (VAS)

Primary Responsibilities:

- Complete services on product requested by the customer (attach tags, security devices, etc.)
- Apply shipping labels to outbound cartons

Workflow:

Cartons will arrive into VAS from either SRS or sorting. Most product will only receive a shipping label and be sent to shipping via an automated system. A small amount of product will be routed to workers who will apply sales tags, security tags or provide other value-added service for the product. For footwear and equipment, the demands of this additional service are low, and can be estimated that ~5% of the actual daily volume needs this service.

Capacity Constraints:

- Limited number of stations for performing service on products

Shipping

Primary Responsibilities:

- Receive products from VAS and SRS and organize into outbound shipments
- Load outbound trailers

Workflow:

As orders are completed throughout the facility, product will be routed to shipping. Shipping has dozens of dock doors that allow it to be flexible with staging trailers and completing orders. For wholesale customers, shipping receives an arrival schedule for the trucker and works to pack the appropriate orders in a known time window.

Capacity Constraints:

- Number of dock doors limits number of trailers available for packing

Hospital

Primary Responsibilities:

- Monitor completion of orders and resolve problems with missing cartons/units
- Re-process units that have errors associated with the units (bad labels, picked in error, etc.)

Workflow:

The hospital is the Work Center responsible for fixing all issues that happen with task completion throughout the facility. It is centrally located and does not appear connected to the high-level process flow as it does not work with successfully completed tasks.

The hospital periodically checks the current work pool to monitor if any orders are having difficulty finding specific units. If a unit is shown to be in error, the hospital will accept the task

and search for a replacement unit. This unit will then be placed in the part of the distribution center that needs it; typically sorting or VAS.

If a unit or carton is picked in error, or if the label is damaged beyond reading, it will be routed to the hospital for reprocessing. The hospital will typically send all fixed product to Case Reserve, where it will be stored for later picking.

Capacity Constraints:

- Hospital is centrally located, so distance to job is major capacity constraint
- Work managed by single work area supervisor (single entity)
- Number of 'open' jobs capped by hospital shelf space

Flow Center

Primary Responsibilities:

- Group customer orders into batches and distribute work to the DC
- Monitor workflow throughout facility and coordinate work centers for efficiency

Workflow:

The flow center is responsible for selecting orders to put into the workflow and ensuring they are completed in a timely and efficacious manner. The flow center decides when to release new work to work centers and when to cut off a wave and route outstanding units to the hospital. All major decisions for handling the workflow are handled within the Flow Center.

2.1.2 Tasks, Waves and Work Center Interaction

Tasks

For clarity, individual workers do not interact with customer orders directly. Instead, customer orders are translated into a series of tasks that tell each worker what picking slot to travel to, and how many of each unit to pick as described in the section above. There are dozens of paths a pair of shoes can take to be shipped in the facility, but there are only five tasks that are executed and tracked. Figure 2 gives a high-level view of task types for the facility, work centers without trackable tasks have been excluded.

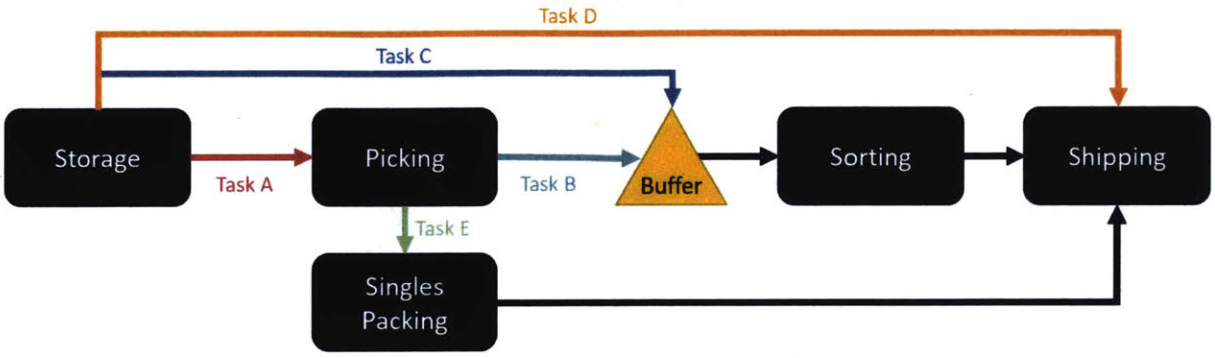

Figure 2. High level overview of task flow at a distribution center

Task A - typically called a 'replenishment task'. In this task, SRS sends a full carton of units to Picking where the carton is unpacked, and a picking slot is replenished. When timing this task, the task begins when it is released to SRS and is marked as 'complete' when the picking slot is fully replenished. Task As are grouped for workers to optimize paths within storage picking areas. Only one sixth of product output from SRS are under Task As.

Task B – a standard picking task. Picking will receive instructions to pick units from a slot to a tote and send the tote to the buffer. The task begins when the task is released to Picking and ends when the requested unit is released from the buffer[5]. Task Bs are grouped for workers to optimize paths within active picking areas. Task Bs represent most of the output of Picking, as digital singles volume is much smaller than wholesale volume.

Task C – a special task built for efficiency in orders needing sorting. When a batch of orders requires at least a full-case amount of a single SKU, SRS will send a carton of the product directly to the buffer to bypass the picking work area. This task begins when it is released to SRS and ends when the carton is released from the buffer. Task Cs are grouped for workers to optimize paths within storage picking areas. About one third of tasks output from SRS are Task Cs.

[5] This makes timing this task difficult, as the release from the buffer is not dependent on time spent on the task by workers but by other factors like the shipping schedule of the facility

Task D – the most direct task for shipping product. When one customer order requires at least one full case amount of a single SKU, SRS will send a full carton[6] directly to shipping and bypass Picking and Sorting. The task begins when it is released to SRS and ends when the task receives an outbound shipping label. Task Ds are grouped for workers to optimize paths within storage picking areas. About half of the output of SRS consists of these tasks.

Task E – a task made specifically for singles orders. Picking will collect units as in Task B and send the tote to the Singles Processing Area. Like Task Bs, the task begins when it is released to Picking, and completed when the tote is sent to the Singles Processing Area. Task Es are grouped for workers to optimize paths within active picking areas. Task Es represent only a small fraction (<5%) of volume output from Picking.

We note that all the tasks that are administered and tracked are contained within SRS and Picking. In every other work center, tasks are completed as they arrive. Once an order, unit or parcel arrives at any other work area, the Flow Center does not manage the task. The task will then travel through the proper work centers to the final shipping dock without restriction. More detail on this will be described in Chapter 3.1.3.

Waves

Unlike an e-commerce distribution center, large brand retailers typically send most of their volume to wholesale customers. These customers require the DC to send product via numerous shipping lines and carrier types. These carriers are not interchangeable, and it is imperative that a customer order all leaves on the same shipping truck. To coordinate the fulfillment, it is easier to track the progress of large batches of orders instead of single orders in the facility. The Flow Center is responsible for generating these batches, which they call 'waves'. Waves are the typical resolution that the flow center works in, where all orders are tracked via their parent wave. The buffer stores orders according to their wave type and wave number.

There are multiple types of waves that are generated in the DC. These are differentiated based on what type of orders they contain.

[6] A box of units is referred to as a carton, but a standard pack quantity of a single sku is called a 'Full Case' quantity. In this report, the term case and carton are interchangeable.

Normal Wave – Contains wholesale orders or any order that is not direct to consumers. Construction methods for these waves are listed in the section below. These waves represent most of the volume that travels through the DC. For the purposes of this thesis, we can say that 12 normal waves are produced a day (or 6 per shift).

Digital Wave – Contains orders shipped direct to consumers and contain more than one unit in the order. These waves are priority for most retailers, as a late shipment is felt immediately by the consumer and can damage the image of the company.

Digital Single Wave – Contains orders shipped direct to consumers and contain only one unit in each order. These are the only waves that produce Task Es in the workflow

Digital NDA Wave – Contains orders where the consumer has paid extra to have the product shipped via Next Day Air (NDA). These orders are handled by a single team that gathers the required orders by hand and do not use the typical flow paths in the facility. At this time, the volume of product included on Digital NDA Waves is small enough for this team to accommodate, but in the future this method of wave completion will not be practical given the expected volume.

Digital and digital single waves are produced semi-hourly and at the same time. To produce these waves, planners will accept all existing digital orders and package them together into the wave. For the purposes of this thesis, we can say there are 12 of each type produced per day, making 24 separate waves (or 12 per shift). Only two digital NDA waves are produced per day (or 1 per shift).

Chase Wave – This is a wave automatically generated every half hour by the overarching warehouse control system (WCS). As workers complete tasks, they may encounter problems with missing inventory or damaged products. These exceptions are bundled together into a chase wave and sent to the hospital, where workers will resolve the tasks quickly to ensure the wave is completed in-full and in a timely manner.

Normal Wave Construction

Wave planners generate waves on request based on the operational needs. Planners make waves immediately before release to ensure that unexpected rush orders can be included as needed.

To create a normal wave, a wave planner will pull a list of every available order. These orders will contain important information for the wave planner to consider:

- Customer type and delivery date, determining order priority
- Sorter that the order is assigned to based on product type (Footwear, Apparel or Mixed)
- Number of chutes required from each sorter to fulfill the order (i.e. number of outbound parcels traveling through the sorter)

The planner will select orders for the wave based on priority until the number of chutes required by the wave match the number of actual chutes on each sorter. Once the wave planner submits the wave to the WCS, the system generates the task list for distribution to the individual work centers. The wave planners will not know the exact number of tasks generated by their wave until after the wave is submitted to the WCS.

Task Generation from Orders in Waves

It is important to understand how tasks are generated as this is the only method for completing work in the DC. Any change to the work processes must first consider the algorithm by which tasks are generated from individual product orders. Before outlining the algorithm, we will define a critical aspect of customer orders.

- **Full Case (FC) Shipping available –** Many wholesale customers allow the DC to deliver them shoes in full case cartons of the same SKU (Task D). Others, however, prefer they repack (RP) these cartons into size runs for each case. If a customer does not allow FC shipping, all units will be sent across the sorters

To generate tasks for the wave, the WCS will first send as many units via Full Case shipping as possible if allowed. These FC shipments directly generate Task Ds, which require the least amount of effort from the DC (as it bypasses both the picking and sorting work centers).

Next, the WCS will group all the SKUs scheduled to arrive on each sorter. If the quantity of a SKU is larger than the standard quantity in a full carton of units (i.e. FC quantity, usually 6 shoes per carton), then the WCS will generate Task Cs for that SKU. This is seen as more efficient, because those shoes will bypass Picking work centers.

The WCS will then generate the remaining demand on the sorters from Picking, generating Task Bs for each SKU and each sorter. If there is not enough inventory in the Picking work area, the WCS will generate Task As, or replenishment tasks, to fill that inventory gap.

Work Balance and Complexity

The allocation of work through the WMS is difficult to predict due to the variety of degrees of freedom within the system such as FC shipment preferences, inventory availability and sorter assignment of orders. A given wave with ten units of a single SKU may have different work allocations and balances. In one case, units could be fulfilled via one full case shipment and the remainder through Picking (one Task B and one Task D). The units could all be sent via Picking if the orders are split across multiple sorters (two Task As, two task Bs). When looking at large waves consisting of thousands of units, estimation of final work content for a single wave is difficult to predict.

It is critical to note that, while the wave planners control the orders in a wave, they do not have direct control over the work content of a wave. Planners can easily see the amount of FC shipments in the wave (i.e. Task Ds), but they currently have no ability to predict and influence the work balance between Picking and SRS. That is, they do not have the ability to predict or create Task Bs versus Task Cs.

Task Closure and Wave Completion

Each wave contains a package of tasks for each work center. While tasks are largely independent from one another in the system, DCs have aspirational rules for managing task and wave completion. Below is reasonable list of rules that may be employed at any DC.

- Task As must be closed before Task Bs can be released for the wave
- Task As, Task Bs and Task Cs must be closed prior to releasing a wave from the buffer to the sorters
- Shipping must be prepared to receive the wave prior to the wave's release from the buffer

- Task Ds are released to be worked in SRS when the wave is ready to be received by shipping[7]
- A wave can be closed only after all Tasks have been closed[8]

While the work is completed by the work centers, the Flow Center is monitoring progress and managing the operation by the minute. There are multiple decision points in the life of a wave that the flow center will control to ensure the DC operates in the best way possible. These decisions are listed below:

Wave Creation – Waves can only contain orders that have been cleared for fulfillment. Orders tend to arrive throughout the day, and many orders must be fulfilled within 24 hours. As a result, planners try to push back wave creations to allow for any higher priority orders to arrive before pushing work into the system.

Task Release Time – After the wave is created, the Flow Center's goal is to have all the units for the wave arrive at shipping near the same time to ease the coordination of trucks. While the rules listed above guide decisions of tasks within waves, there are no set guidelines for releasing new waves to the floor. Currently, tasks are released to work centers when they are running low on open tasks and employees need new work content.

Task Priority – The priority of the task is the primary method of influencing the order in which tasks are completed on the floor. Workers will automatically accept the highest priority task available in their work area.

Wave Completion Time – Waves are not typically completed to 100% of tasks fulfilled. Usually, the final tasks are left outstanding due to system error, or employees have passed over the task in favor of working in more productive zones. To allow the wave to be released from the buffer, the Flow Center must force close all outstanding tasks. To close the wave, all outstanding tasks must be forced closed. This is the only way that the orders contained in the wave can be cleared

[7] Shipping must be ready to receive Task Ds from shipping as there is no buffer for these outbound cartons. If cartons arrive in shipping prior to this, they will cause congestion and could shut down the conveyors

[8] The flow center will not close all of the wave at the same time. Rather, it will close different types of tasks to move the wave forward in the process flow

for shipment, and the truck can leave the DC. This decision regarding when to close outstanding tasks will be discussed at length throughout the rest of the thesis.

2.1.3 Metrics in a DC

To guide the priorities of workers within the operation, DCs work to satisfy specific metrics. These metrics are meant to encourage high performance and many of them closely align with the overarching business strategy described in the introduction of this thesis. This section will list and explain the most important metrics tracked and their impact to the operation.

Primary Metrics

Wave Cycle Time – Waves are batches of orders to customers, therefore measuring the cycle time of a wave is synonymous with tracking the cycle time of a specific order. As mentioned in the introduction, retail customers are demanding shorter lead times from order to delivery to ensure the brick-and-mortar stores can quickly adapt to customer demand. Additionally, online consumers are purchasing products online with the expectation of two-day shipping. Both market trends indicate that DCs must reduce their wave cycle time to better serve customers and individual consumers.

Wave cycle time is measure as the time from when the wave is created[9] to when all tasks have been completed and the wave is closed or when the Flow Center force closes a wave. At the end of the wave cycle time, all trucks are assumed complete and are released from shipping. For this thesis, we are assuming a current average cycle time of 12 hours with a desired cycle time of four hours for normal waves. Digital waves have higher priority and an average cycle time of six hours with a desired cycle time of three hours or less.

Product Shipped In-Full on Time (SIFOT) – SIFOT is the primary quality metric for Distribution Centers. As in the title, it measures the percentage of product shipped in full on time. This shows how well the operation is fulfilling customer orders. Customers who do not receive their full

[9] As waves are created on demand, the time between wave creation and wave release is negligible for the considerations of this thesis

shipment are entitled to charge the retailer for missing product and the relationship between the companies or between the retailer and the consumer can be damaged.

To define the measurement, if SIFOT was 99%, it would mean that out of 100 orders, 99 had all the needed product on the order and these orders were completed and shipped on time. The one order left out could be missing a single unit, many units or could miss the assigned shipment date.

For any large DC, it is difficult to measure SIFOT in real time. Instead, SIFOT is measured retrospectively when considering the amount of back charges that are reported by customers and consumers. Working with the DC, it is much easier to use the amount of volume being routed to chase waves as a proxy for SIFOT. This relationship will be discussed later in the thesis.

Volume Shipped (Capacity) – The total product shipped for the day. This represents the retailer's sales to both retail customers and individual consumers. This is also the number that is reported to headquarters and displayed during the earnings call when representing total sales for the company.

This measurement is simply the amount of product that left the facility over the course of the day. As previously mentioned, this thesis will assume a daily capacity of 200,000 units, or 100,000 units per shift.

The maximum capacity is difficult to measure as it is largely determined by the target set by headquarters for the day. Each day, the flow center will push the required volume into the system, and each work center will complete the assigned work for the rest of their shift. This causes fluctuations in capacity during the day based on work availability and keeps DCs from fully understanding their limits.

Secondary Metrics

These metrics are not listed in the primary section because they have not been the direct focus to enable fulfillment of the new marketplace. They are still tracked daily and hold weight during discussion of improvement projects.

Worker Utilization – A measurement of how occupied each worker is in the facility for a given day. DCs use a mixture of temporary laborers and permanent laborers to flex the capacity of the facility as needed to meet market demand. By tracking this number, management tries to ensure that all workers are always completing productive labor, and the operation is not paying for idle hands.

This metric is calculated by monitoring how often a worker is logged into a productive task. The final number is expressed as a percentage of productive time, and DCs strive to maximize this number. Worker utilization is driven by two primary effects, worker capacity and work availability. For worker capacity, if there are too many workers in the work area, utilization will suffer. Work availability can impact utilization as well. The Flow Center controls the release of tasks to work areas. If the Flow Center delays the release of tasks or waves to the system, the lack of work starves the workforce, leading to lower utilization numbers. For worker utilization, an acceptable range can be above 80%, but work areas and the Flow Center strive to achieve as high numbers as possible as this metric is reviewed daily.

Worker Productivity – A measure of how productive a worker is when signed into a task. DCs use this metric to track employee performance and assess the impact of continuous improvement initiatives within the facility.

For a given work area, worker productivity can be influenced by many factors. Perhaps the largest influencer is work density. A full discussion of this concept can be found in Chapter 4. Root Cause Analysis.

Work areas strive to maximize this number as it is tracked daily. If workers fail to meet expectations on productivity metrics, they could also be subject to performance review. For this reason, workers will tend to stay in the areas of the DC with the highest density of work.

Inventory Accuracy – This metric is a measure of the accuracy of the inventory data within WMS. In a task assignment, the WMS will tell a worker to travel to a location to pick a SKU in a specified quantity. If the worker travels to the location to find that the SKU is wrong, or not enough of the SKU is present to fulfill the task, the worker will send an 'exception' message back to the system

to be remediated by the Hospital. Inventory Accuracy is defined as the number of exceptions for the number of tasks completed overall.

Cost per Unit Shipped (CPU) – This metric is the culmination of many other metrics for the facility. It represents the effective rate for sending a single unit through the facility, averaged over all units for that day. DCs strive to reduce this number as much as possible.

Average Open Waves – A measure of how many waves, on average, are open within the DC at any given time. This includes normal waves, all digital waves, and chase waves. DCs strive to reduce this number to incentivize the completion of waves and the further reduction of wave cycle time. For this thesis we can claim that the average open waves are ~35.

Buffer Capacity – During periods where there are many waves open and waves are not being completed effectively, inventory can start to build in the buffer. If the problem persists, the buffer can reach its limit and the order of operations in the facility begins to break down. As a DC works to reduce the average open waves it will largely mitigate the impact of buffer capacity on the operation, but the Flow Center still tries to reduce the inventory stored in the buffer as much as possible during operations.

Delivery Window – While this metric is not well tracked, it is felt by both the retailer and its customers. This metric measures the amount of time needed to fill the back of a delivery trailer and send the trailer off the dock. If waves are not being completed in full, it can be difficult to load the trailer all at once. Trailers must then rely on the hospital to fulfill the last outbound cartons; else the DC will impact the SIFOT metric.

Delivery drivers are not employed by the DC or the retailer and will charge the DC for time in excess spent loading the trailer. To manage this, the facility will prepare large orders before they are due at the dock. DCs can palletize these outbound cartons or load them in a dummy trailer. These cartons are then cross-docked to the correct trailer during the appropriate time window. A valid target for delivery window is 2 hours for a certain order. The timing of this delivery window is communicated prior to the delivery date, at least 24 hours prior.

2.1.4 Current Performance Overview and Context

Wave Cycle Time

Over the past year our example DC has undertaken significant continuous improvement projects to meet the shifting demand landscape. In line with these projects, the target for wave completion has also been lowered. Figure 3 displays the average cycle time per week for normal waves traveling through the DC. In this chart, the values have been normalized, it can be interpreted for this thesis that a value of 1.0 on the y-axis represents the target, 12 hours.

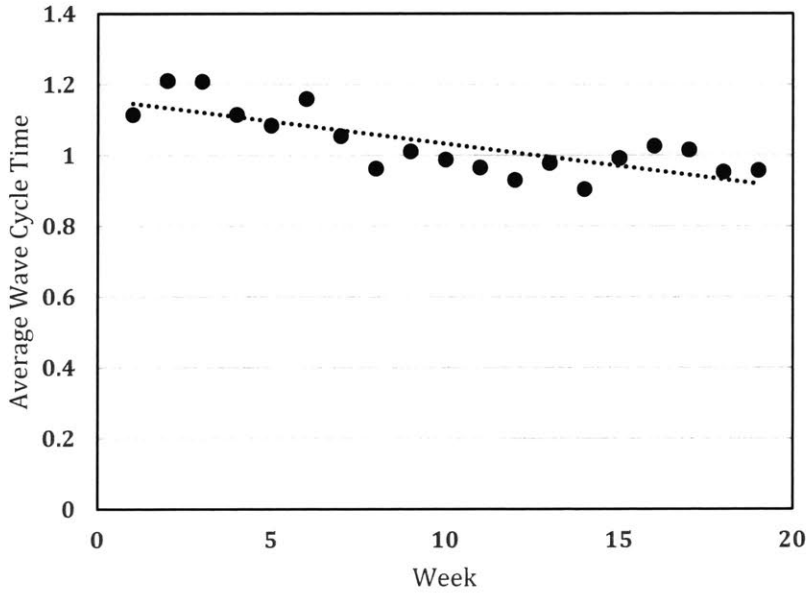

Figure 3. Average weekly normal wave cycle time (Normalized)

SIFOT

As stated before, SIFOT cannot be measured in real-time. Therefore, for the purposes of this thesis we will consider the percentage of volume routed to chase waves as a proxy for quality. Figure 4 shows the percentage of volume sent via chase waves over the same time period as listed in Figure 3.

This metric is creeping up over time in the same way that the average cycle time for normal waves is decreasing. This suggests a systematic reason for the elevated chase wave volume. This will be further explored in the following chapter of this thesis.

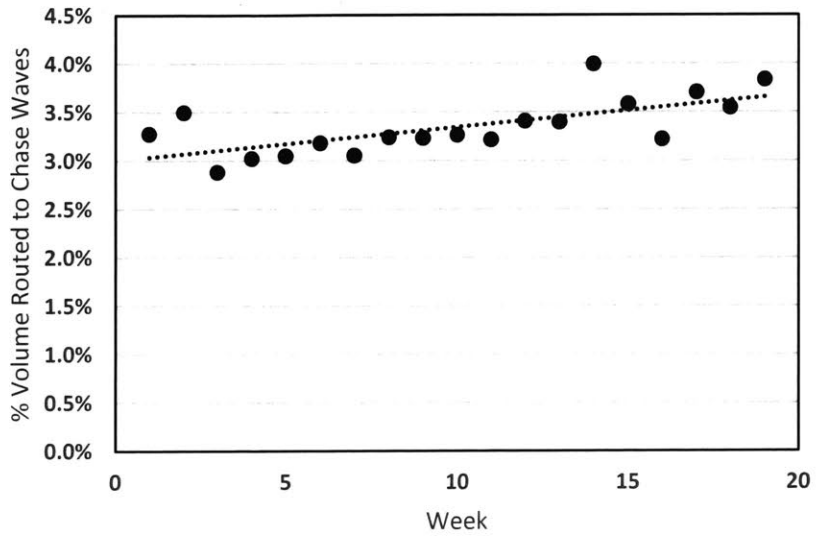

Figure 4. Percentage of volume being routed through chase waves

Capacity

The DC consistently meets the capacity demands placed on it from headquarters. As a result, the output ~100,000 units per shift, or 200,000 units per day. Hour-to-hour, however, there can be fluctuations in capacity based on work availability, staffing level fluctuations and work mix. Figure 5 and Figure 6 show the relative capacity for a typical day (two shifts) in SRS and Picking.

While the DC meets all required demand, it is difficult to assign a single number to maximum capacity due to the effects listed above. At its base, the charts in these figures suggests that capacity is highly dependent on the conditions in which the workers are operating.

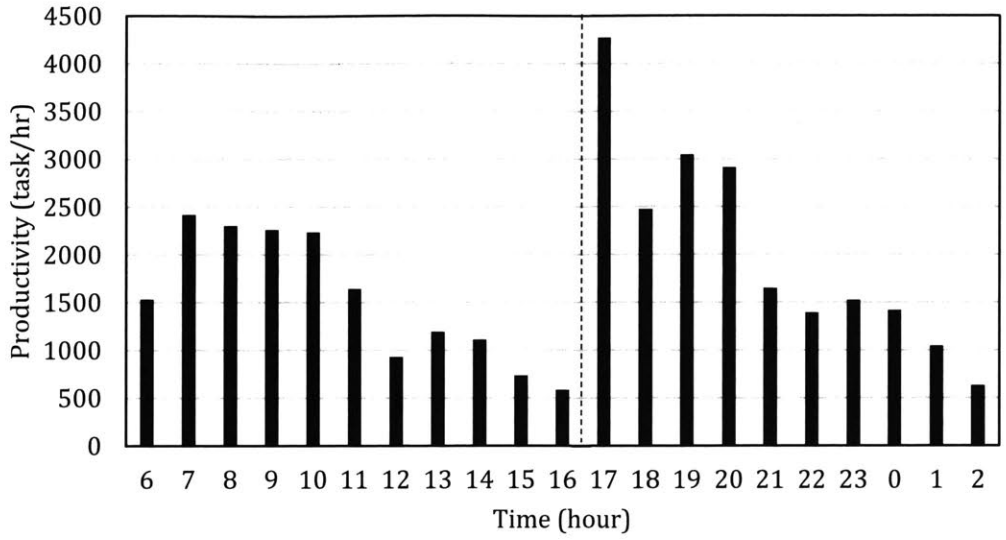

Figure 5. Picking hourly performance. Shift change at 17:00 and daily maintenance between 02:00 and 06:00 for this shift

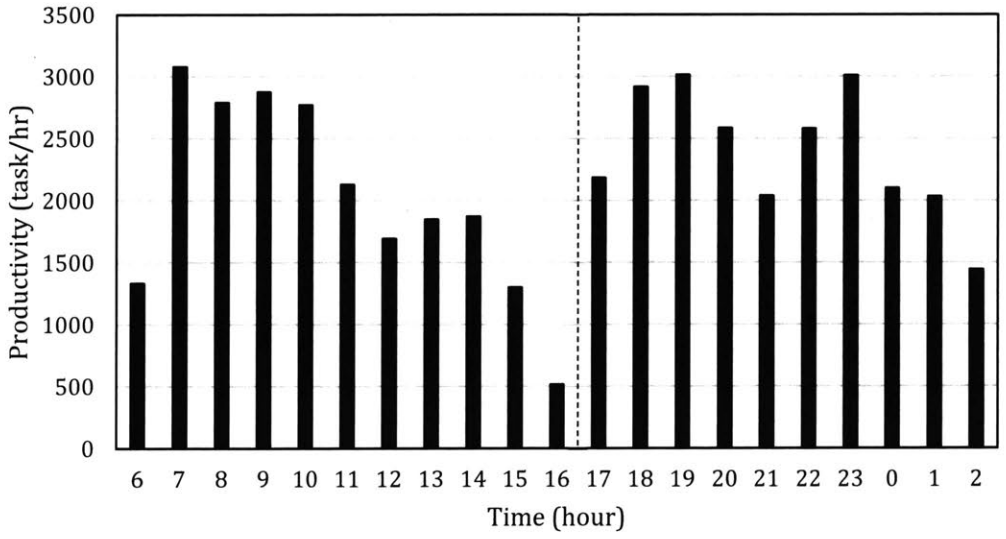

Figure 6. SRS hourly performance. Shift change at 17:00 and daily maintenance between 02:00 and 06:00 for this shift

3. Literature Review and Analysis

To begin an analysis of the current state of the distribution center, it is helpful to refer to existing literature concerning dynamics of work completion within these facilities. This chapter will document relevant publications and discuss their implications within the facility and the greater distribution network. The learnings described in this chapter will help give context to the challenges and opportunities in our example DC and be referenced throughout the root cause analysis and system recommendations.

3.1. Zoning, Batching and Waving Strategies

Due to the size of these large DCs, coordination of workers and their movements can have large repercussions on productivity and efficiency throughout the shift. Parikh and Meller (2007) take an analytical view of these decisions and their potential impact on operations within large DCs with medium to high volume requirements (Selecting between batch and zone order picking strategies in a distribution center, 2007). Their research focuses on how wave tasks are picked in the facility and how the basic strategy for selecting and distributing tasks can influence workflow on the floor.

At a high level, Parikh and Meller group DCs into two primary categories, each with two subcategories. The primary categories are based on the DC's decision to pick tasks based on Batching or on zones:

- **Zone Picking –** Workers are assigned to a specific zone in the DC and responsible for picking all units in that area. Can be problematic if zones have higher workload than others for any given wave
- **Batch Picking –** Orders are grouped together and distributed to a single worker. Orders are not confined to a specific area. If specific picking locations have multiple tasks assigned, workers could block each other, leading to delays in processing time.

Both strategies are performed in a waving environment, where orders are distributed and organized by a central control team and orders are coordinated according to their parent waves. Within these picking strategies, the DC can perform the sorting step for an order on the pick floor

or construct a sorting machine to handle this step downstream. The title of each of these systems is summarized in Figure 7.

	Batch Pickcing	Zone Picking
Sorter	Sort-while-Pick	Progressive
No Sorter	Pick-and-Sort	Synchronized

Figure 7. Picking Strategies employed in DCs

The requirements of every one of these systems leads to tradeoffs between efficiency, speed and capital investment. As described earlier, batch picking and zone picking have problems that can arise with work distribution in the facility, with workers blocking each other's paths and work imbalance from zone to zone across the facility. Sorters enable workers to pick product in the most efficient way possible but requires costly investment and another step in processing time. Overall, selecting an order picking process in a DC requires careful consideration of future needs regarding cycle times, throughput, and cost structures.

3.1.1 Picking Strategies Within the DC

Due to our example DC's size and complexity, the facility's strategy for picking is more complex than the four options laid out in the article by Parikh and Meller. While the strategy is predominantly zone picking, there are many elements of batch picking employed.

The DC is organized into multiple work zones, each consisting of multiple aisles, within SRS and Picking. Workers who are 'logged in' to a work zone can only read, accept and complete tasks in that area. Workers will not switch zones until prompted by a supervisor or until they run out of work in a zone. In this way, operations run the facility as a zone pick strategy, which could help ensure complete coverage of the facility, but means that the capacity can be heavily influenced by workload imbalance if one zone has a higher work amount than another.

While the system is organized into different work zones and run on a zone-based picking strategy, there are also many elements of batch picking strategies employed at the DC. Workers within the facility are not constrained to working only in a single zone. Within SRS, up to three workers can

be 'logged in' to a particular zone, and Picking doesn't have a capacity limit. Workers are free to work in the areas that have the most work available, which allows work areas to better respond to a potential situation of work imbalance, but also leads to a potential for workers blocking each other. This is a larger problem within SRS, where safety concerns restrict workers to having only one worker in an aisle at a time, as forklifts heighten risk factors for collisions.

Overall, the 'hybrid' system employed in the example DC for picking tasks is meant to allow workers to adapt to workload imbalance throughout the facility. This is especially important considering the method by which work is assigned in the facility. As a rule, the WMS of a DC has an order of preference for assigning work in the facility. The WMS will first look at zone 1, aisle 1 for the items needed to fulfill an order and move to zone 1 aisle 2 and so on. This inherently causes a workload imbalance between work zones, where work zone 1 will tend to have higher traffic than later zones. This effect is best demonstrated through the charts in Figure 8, where the zones within Picking and SRS show the distribution of work for an 'average' wave.

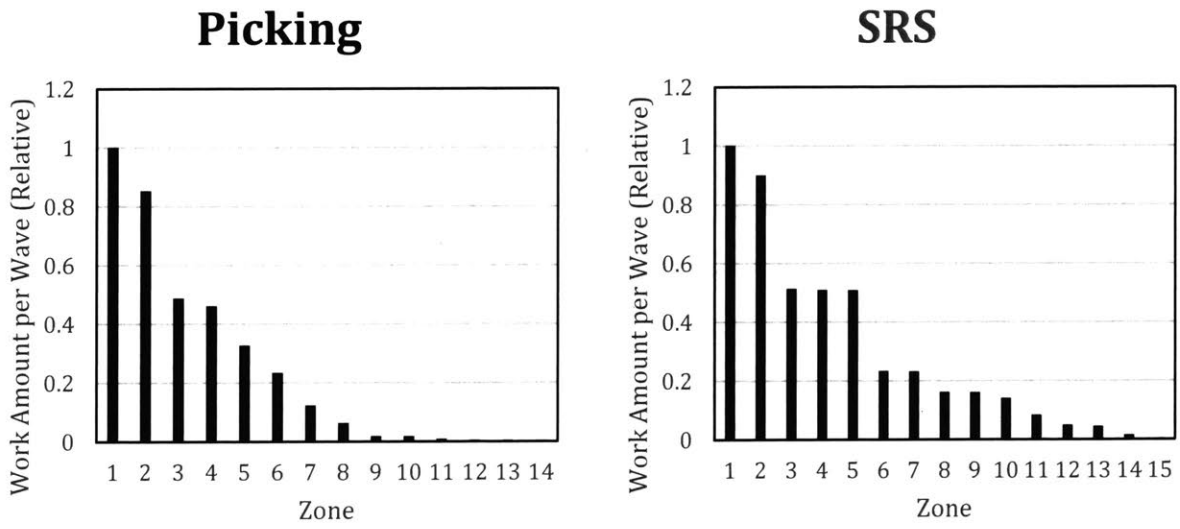

Figure 8. Work amounts distributed across work zones in Picking and SRS

Considering how WMS assigns tasks to locations, it is imperative that the DC allows flexibility for the workers to shift zones to counteract the workload imbalance. WMS also requires workers to spend valuable time shifting between zones throughout their day to ensure waves are completed

in a timely manner. In the same passage, if a zone remains unstaffed for an extended period (due to low work content), it could have major impacts on the completion profile of the wave.

3.2. Lean Systems, Push vs Pull

When analyzing any sequential system in operations, it is relevant to consider the frameworks of Push and Pull systems in a Lean context. While typically seen as mutually exclusive options when considering operations in an academic setting, David Pyke and Morris Cohen (1990) argue that real operations exist on a spectrum between the two extremes in their paper "Push and Pull in Manufacturing and Distribution Systems".

The concept of Push and Pull is not how an entire warehouse is run; rather, it is the core of multiple decisions that are implemented at every step within the operation. At its core, a push system bases a workload on anticipated results (such as an order forecast) and works to fill the expected demand in the most efficient way possible. Pull system, by comparison, reacts to an 'existing' demand, and behaves in a much more flexible manner. Pyke and Cohen argue the decisions of implementing push or pull at every step within an operation can be grouped into one of four categories:

- **Batch Size –** To put another way, how large will each production run be in the facility. Push systems determine batch sizes early in the process flow to optimize efficiency metrics locally on each work center. Pull systems determine their batch sizes based on steps much later in the process flow, usually an unedited demand stream. This keeps Pull systems from potentially running efficient batches but leaves the facility much more able to respond to consumer demand.
- **Timing of Production Requests –** Timing is one of the key components of Push vs. Pull and is primarily concerned with which work center sparks a work request. In a push system, the upstream work center will complete work, which sends a work request to the next process step. In a pull system, the downstream process step submits a need for material, which triggers a work request upstream. In short, push systems authorize production in advance of demand, while pull systems authorize work in response to demand.

- **Dispatch or Allocation Rules (Priorities) –** This category is concerned with the rules by which work is selected for completion in the queue. When given a list of action items to complete, a push system will typically have the work center in which the work resides decide which order to complete tasks. In a pull system, the downstream work center has authority over which item is completed next and communicates those priorities via the work request order.
- **Handling of Exceptions –** This can be one of the key elements in the performance of a workflow, as the flexibility of operations typically limits the overall metrics displayed in day to day operation. In a push system, if an exception arises (such an increased demand or special job), the work area will need to decide whether to stop, change and revise the work plan in the moment. Many times, it may be inefficient to respond to the exception, so extra time is wasted. In a pull system, the work area schedule is already closely linked to the work center downstream, so an exception request will have minimal effect on the minute-to-minute operation.
- **Information Source –** This is primarily concerned with what information is used to determine the release and completion of work in the facility. Push systems tend to use more global information, as the hypothetical case of 'we need to get x units through the facility, so give x tasks to work center A'. Pull systems, on the other hand, work with more local information, or by the demands of an individual work center.

When outfitting an operation, each decision made can usually be included in one of these categories. Each work center will inherently have aspects that are more push oriented, and others that tend to give the overall system a pull approach.

3.2.1. Push vs Pull Simulation for the DC

To join the method of work within a large DC and the foundational elements of Push and Pull as expressed in this literature, a small demonstration was generated using Python. In this demonstration a simplistic view of a wave-based DC was modeled under two sets of standardized work processes. The following chapter will describe the basis of the simulation, the initial results, and how we can use this simulation to better understand Lean Systems and Distribution Centers.

Modeling Parameters and Assumptions

The system generated has three primary work areas and a flow center that assigns work to each area based on conditions in the DC. All these work centers are sequential, but all work centers can receive work independently, as some tasks will only need the last step to be completed. Figure 9 shows the basic layout of the facility in the simulation.

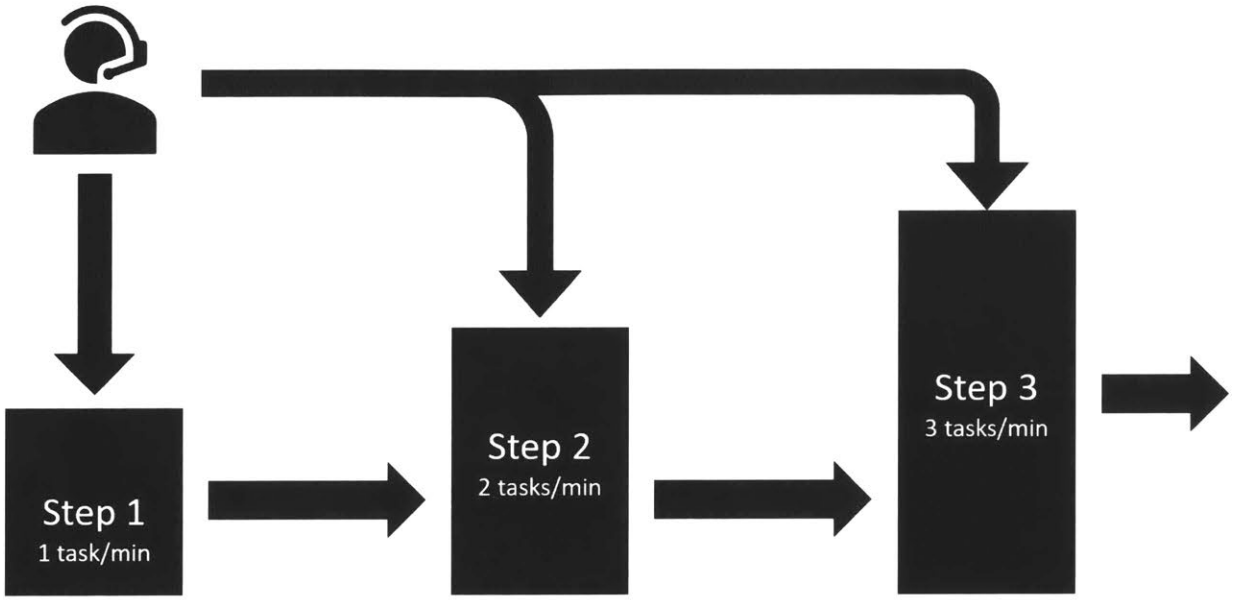

Figure 9. Simplified DC layout

On average, each wave generated by the flow center will have 100 tasks starting in every work area. Effectively, this would mean every wave has 100 tasks for Step 1, 200 tasks for step two (after receiving 100 completed tasks from Step 1) and 300 tasks for Step 3 (after receiving the 200 completed tasks from Step 2).

In a perfect system, where there is no variability in wave size, this would mean every work center would complete a wave's batch of work in 100 minutes and work would progress fluidly through the facility. Figure 10 shows this optimal situation. Each shape represents a batch of work for a given wave (i.e. the circle wave) and stacked shapes show that work has been received from upstream.

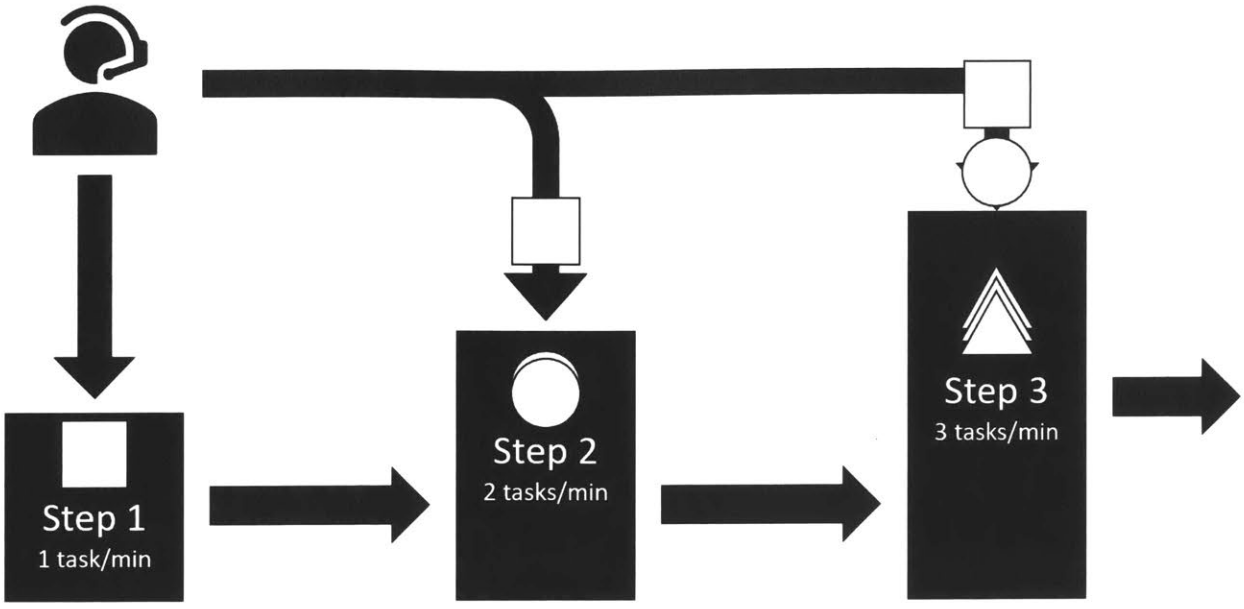

Figure 10. System initial setup

As it can be seen, each Work Area is completing a different wave, and the downstream steps have work batches in the queue. At the next takt, each work center will send the work downstream where it will be joined with the other work on the wave and completed in sequence. In response to the tact time, the work center will release another wave which will immediately be worked on in Work Area one, and the DC will continue to operate.

Under ideal conditions, where there is no variability of wave size, the plant runs smoothly. Work Area utilization is 100% and the wave cycle time is 300 minutes for all waves. Every 100 minutes each work center will complete their work, pass it downstream and receive another batch of work.

Once variability is presented to this system, a series of decisions must be made as to how work will be progressed and distributed in the facility. The next two chapters will explain the concepts of push and pull, and how they appear in this DC. In simulations, the starting point for every system will be the starting point described in Figure 10.

DCs in a 'Push' System

As wave size variability leaks into the system, work centers will receive variable amounts of work and finish waves at different times. The Flow Center must decide when to release the next wave,

as a clear tact time is no longer evident. In a push system, the flow center will release work whenever a work center becomes idle.

To demonstrate this system, follow along in Figure 11 with the steps below:

1. Initial starting point is same as above
2. Work Area 2 receives a smaller amount of work than the others. Work Area 2 completes the batch and sends it to Work Area 3 (who is not able to begin work on this wave). Work Area 2 begins work on new batch (without waiting on same wave coming from Work Area 1)
3. Work Area 2 finishes another small batch of work and sends work to Work Area 3
4. Work Area 2 is now idle, so Flow Center releases a wave to ensure Area 2 continues to work

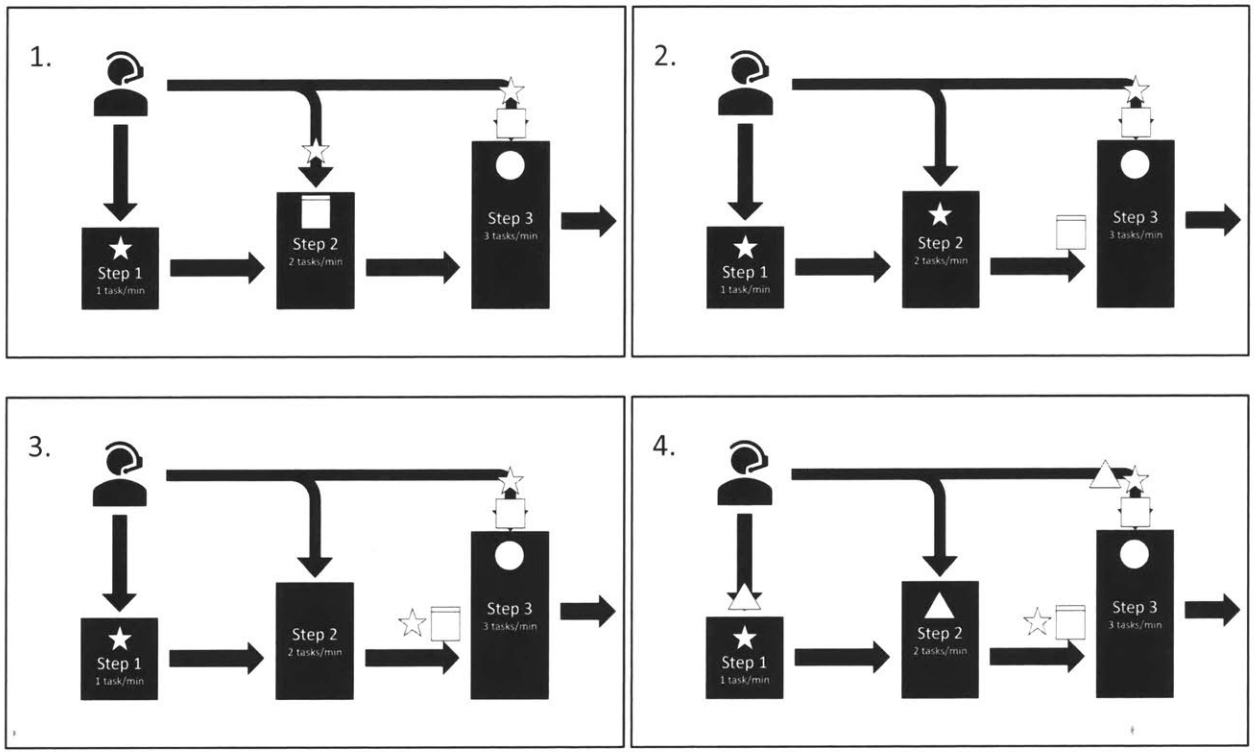

Figure 11. Simulated facility operated under Push conditions

The assumption in this case is that the Flow Center prioritizes utilization numbers and, by association, throughput of the facility. This is potentially at the expense of wave cycle time, as

traffic in the system is high, especially around Step 3. The code used to simulate this facility was constructed using Python.

DCs in a Pull System

While the Push system is focused on increasing utilization and throughput of the facility, a Pull system's focus is to ensure waves are kept together and traffic is reduced through the facility. In this way, it could reduce the wave cycle time for an average wave.

In the Pull system, the flow center will only release work when Work Area 1 is idle and has an empty queue. Work centers will only pass work downstream if the following Work Area is sitting idle. A work area will sit idle until the work center downstream is ready to accept their work.

To demonstrate this system, follow along in Figure 12 with the steps below:

1. Initial starting point is same as before
2. Work Area 2 receives a smaller amount of work than the others. Work Area 2 completes the batch and, with Work Area 3 unable to receive the work, sits idle
3. Work Area 1 finishes the batch and sits idle, as Work Area 2 is still waiting for Work Area 3
4. Work Area 3 finishes the batch, accepts work from Work Area 2. Work Area 2 accepts work from Work Area 1
5. With Work Area 1 idle and empty, the Flow Center releases another wave

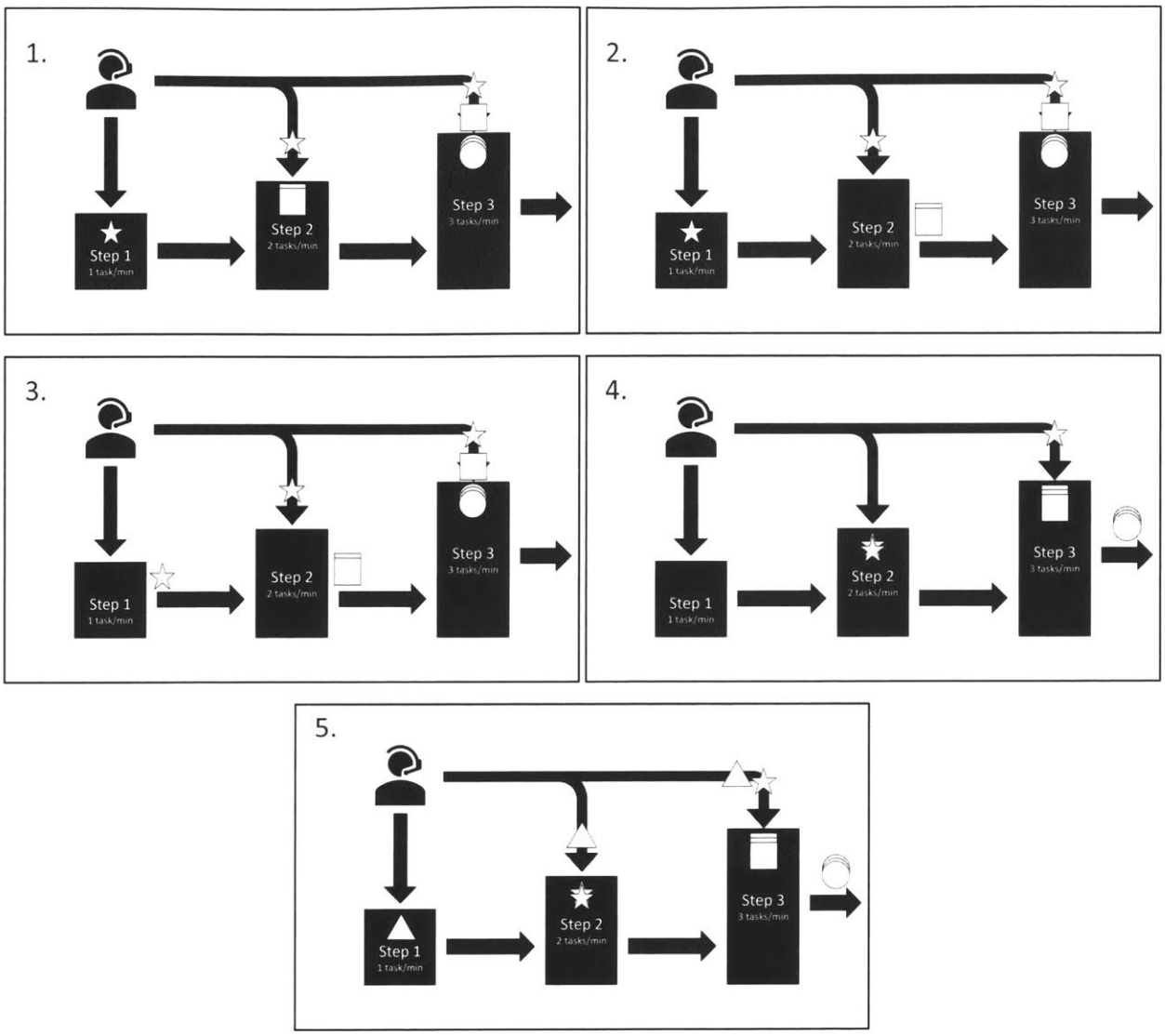

Figure 12. Simulated facility operated under Pull conditions

As it can be expected, the tradeoff of a facility run in a Pull system sacrifices utilization numbers, and potentially throughput, for the sake of decreasing cycle time. The code used to simulate this facility was constructed using Python.

Simulating Utilization and Wave Cycle Time

For the simulation of each system the maximum capacity was constant for each Work Area. The waves produced by the Flow Center were 100 tasks for each work center on average but

subjected to a coefficient of variation of 0.4[10]. As predicted, the Push System and the Pull System displayed significantly different behaviors.

The results of the simulation for the Push System and the Pull System are shown in Figure 13.

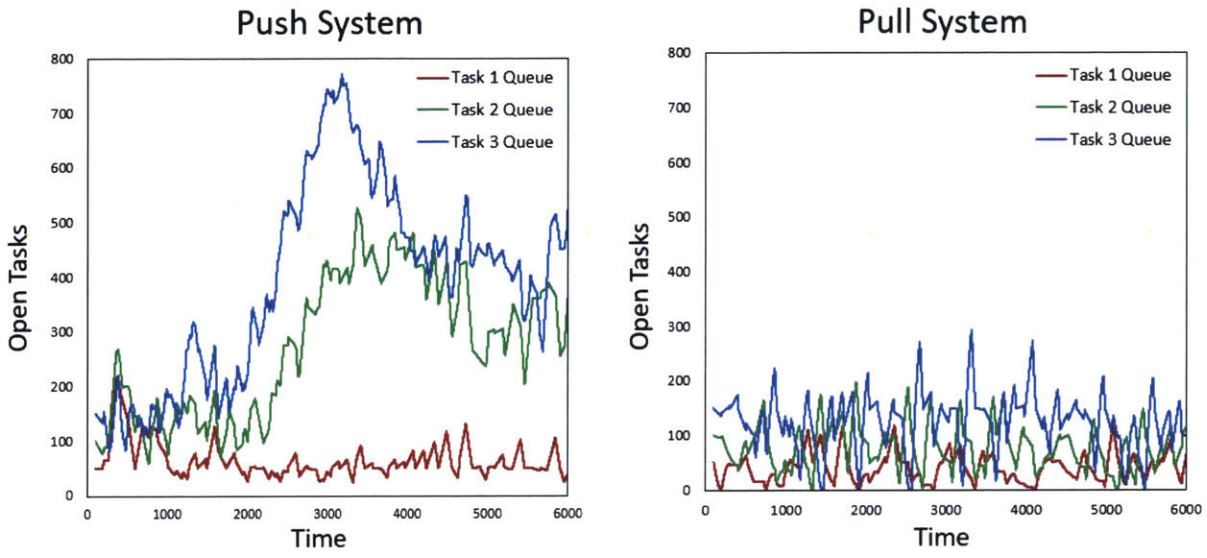

Figure 13. WIP over time for the Push simulation and Pull simulation. Results are smoothed for ease of interpretation

The Push system never reaches a point where work centers are idle, but the volume of tasks in the queue can be large. Given constant maximum capacity of each work center, this would increase the average wave cycle time (as per Little's Law). In contrast, the Pull System has much less WIP at any given time, but there are periods where each work center sits idle, impacting the utilization numbers of the facility. A summary of the important numbers in each system are given in Table 1.

Table 1. Simulation results for various factories

System	Ideal*	Push	Pull
Avg Wave Cycle Time	300 min	436 min	306 min
Utilization	100%	100%	79%

* Ideal system has no variability in wave size

[10] This coefficient of variation was chosen because it is similar to the variability of Normal Waves generated in DCs.

In this case, the intuition behind the simulations agrees with the results. The Push system has an average wave cycle time that is 45% higher than the ideal case while maintaining utilization numbers. The Pull system sustains an average wave cycle time near to the ideal case, but at the expense of the utilization of each work center[11].

Impacts of Variability on Utilization in Pull Systems

While the above comparison shows the relative performance of a push system versus a pull system, we can give more understanding to the effects of variability on the metrics by performing a sensitivity analysis. Multiple renditions of the Pull system simulation were performed with coefficients of variations ranging from 0 to 0.4 (as in the previous simulation) and the results were summarized in Figure 14.

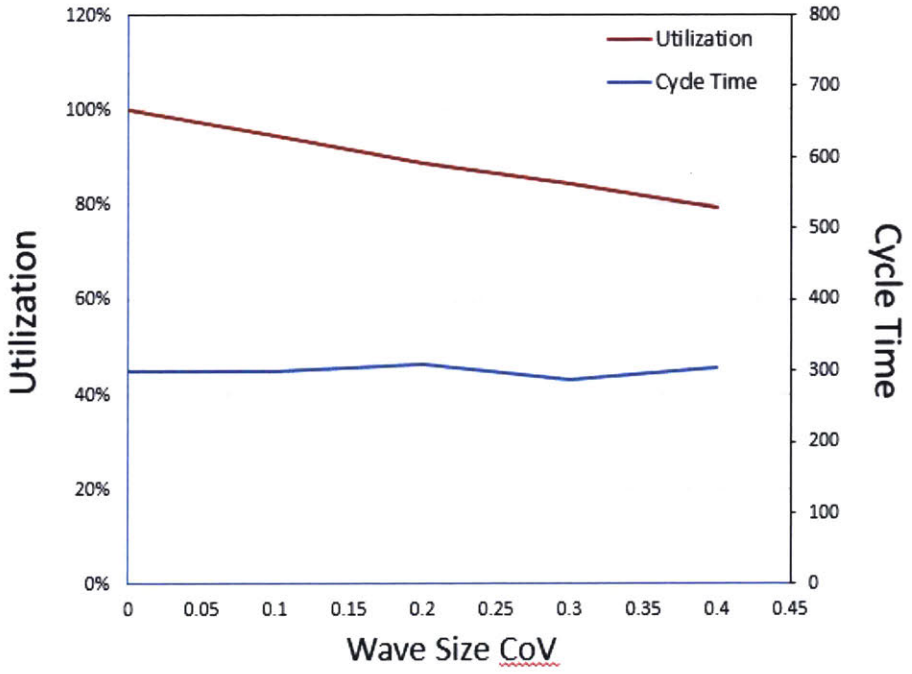

Figure 14. Effects of variability on a pull system

This result suggests that Pull systems, if run effectively, will allow DCs to run with a near-optimal cycle times, but with sub-optimal utilization numbers. This is of no concern if the facility still meets demand, but if utilization drops too low it could impact the performance of the DC.

[11] This system is an ideal system where demand is infinite, so throughput of this system behaves the same as utilization. In real systems, we must ensure demand does not become the bottleneck to limit throughput

Reducing variability of wave size, however, curtails the effects of this drop in utilization and can restore capacity.

3.3. Wave Jumping

Wallach (2018) describes his experiences working within a DC and a series of hypotheses that were tested to understand the internal dynamics of the production system. This document is an excellent prequel to this thesis, as many of the topics and trends are still relevant to the work within these operations.

Wave Jumping

Perhaps one of the most important factors Wallach's thesis describes is the tendency of workers to 'jump' waves throughout their day. Wallach describes wave jumping as the workforce changing from completing work on the 'current' wave to a 'younger' wave without finishing all tasks on the current wave[12]. This can be due to multiple factors, but as Wallach argues, it typically happens when priority (i.e. Digital) waves are released to a work center. This causes a re-prioritization of work in the WMS and redirects all workers away from the current wave to start work on another. This eliminates the ability for the DC to work in FIFO and extends the cycle time of the wave that had been jumped. During the hypothesis testing portion of the document, Wallach tests this theory by generating the plot in Figure 15.

[12] In conversation, personnel would use the term in the sentence "We had a fresh Digital wave released in SRS, so we jumped to that one because it's priority"

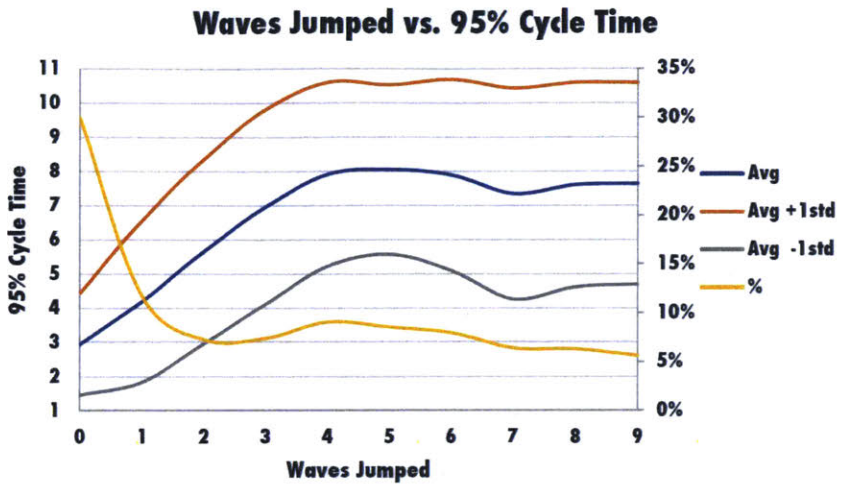

Figure 15. Relationship between number of times a wave was jumped (for a given wave) versus the cycle time for the wave

According to the message of this plot, the more times a DC 'jumps' to work on a new wave, the longer the resulting wave tends to be. This follows the logic that starting and stopping waves during completion will extend the cycle time of the wave. To counteract this, the recommendation of the thesis was to generate larger digital waves at a lower frequency, thereby reducing the number of waves in the system and reducing the impact of 'wave jumping' by eliminating the opportunity to jump to another wave.

Processing Multiple Waves at a Time

Wallach also alludes to another important aspect of the operation and the drivers behind work completion. While a DC has specific priorities assigned to every task pushed onto the floor, waves do not tend to be completed in a sequential manner. Figure 16 shows a snapshot taken during the time of this thesis where Wallach shows that, for a specific Task type, multiple waves are open and being completed at the same time.

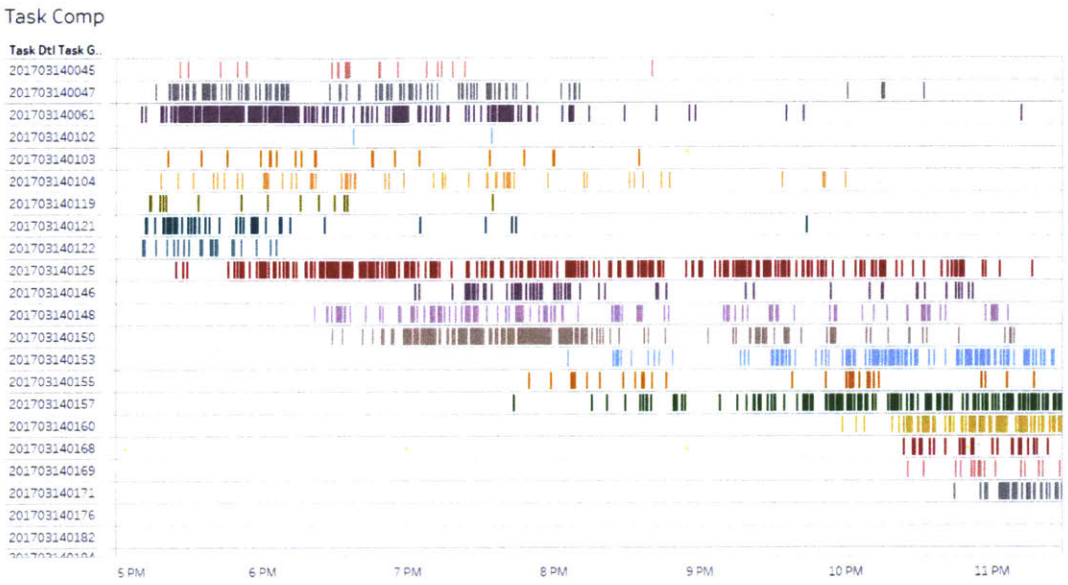

Figure 16. Completion times for Task As over multiple waves. Each row shows a wave and each tick mark shows a task completion time

Here it is even more evident that waves are not completed in FIFO and, consequentially, not completed solely based on priority. The author suggests multiple causes for this effect, including:

- Reordering of tasks at the throw line (intermediate step)
- Conveyor length leading to jumbled arrival times for cartons
- Work density and location driving worker behavior
- WMS work assignment tendencies

Relevance to Current Research

Many of the trends shown in this research are universal to DC operations for large retailers. For instance, the average number of waves open in our example DC at any time is ~35, showing that multiple waves are open at every work center. As already stated, the primary learning from this research was to consolidate digital waves to limit the amount of opportunities for wave jumping. This has been implemented into standard practices of our DC, but wave jumping is still present on the floor, even between normal waves. This suggests that the root causes enabling wave jumping have not yet been addressed. As a result, this behavior persists within the operation.

The following chapter will dissect the current operation to reveal the main drivers behind critical metrics in the distribution center and reveal fundamental root causes that explain the behavior of the DC identified by Wallach and this thesis.

4. Root Cause Analysis

Previous efforts in continuous improvement in this operation have resolved major problems such as buffer overloading, low SIFOT, and reduced the average wave cycle time to what it is today. Despite this, the targets set by headquarters and the demands of the new marketplace require the DC to make significant changes to the current work system. This root cause analysis will explore the three main topics that are keeping the distribution center from fulfilling the future demands of the industry. These three topics are listed below:

- Average wave cycle time is higher than desired
- SIFOT increases with a decreasing trend in average wave cycle time
- Capacity (throughput) is variable and seemingly unpredictable

This root cause analysis will explore each of these qualities and drill down to a set of actions and conditions that represent universal opportunities for improvement within normal DCs.

4.1 Average Wave Cycle Time

The primary goal of this thesis is to reduce the average wave cycle time for the distribution center. In this chapter, we will be analyzing the reasons for what determines the length of a wave. This focus will be set primarily on the completion percentage profile for tasks in a wave as well as the determination for number of outstanding waves in the workflow.

Wave Completion Profile

To understand what determines the length of a wave, it is helpful to see the completion profile of a wave broken down by task type. Figure 17 shows the completion profile of an 'average' normal wave. This behavior is typical for all types of waves.

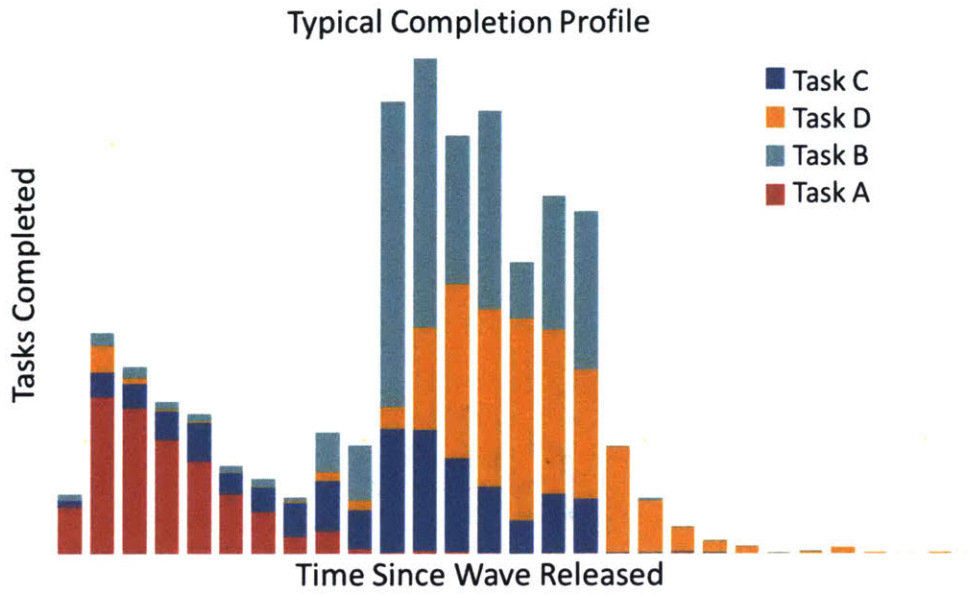

Figure 17. Completion profile for a 'average' normal wave. Axis labels removed to protect confidential information

The chart in Figure 17 is generated by combining all the normal waves from several months into one 'average' wave. As is apparent, there is an inherent order in how tasks are completed. Task As are completed first to enable Task Bs. Task Cs are performed by SRS while Picking is working on Task Bs. And Task Ds are released and completed last to finish the wave. As this diagram is a conglomerate of many waves, small quantities of tasks are distributed throughout the timing of a wave. These can be considered outliers.

The most important item to note in the chart is the relatively long 'tails' of the Task As and Task Ds. These tails are one of the primary reasons for the length of a wave. For Task As, ~80% of the tasks will be completed in the first 40% of the time of the wave and the last 40% of the time in the wave only accounts for ~4% of the tasks completed. A similar story is told for Task Ds, with ~90% of the tasks complete in the first 40% of time and <5% of tasks completed in the last 40% of time. Table 2 summarized this data. In the table, the time period measured is from the release of the first task to the completion for the last task of a specific task type on a specific wave. For example; when the flow center releases task Ds to the floor, 90% of the tasks are completed within the first 40% of the total time these tasks are left outstanding.

Table 2. Comparison of progress for different time periods during task completion

	Time from Task Release to Task Closure	
	First 40%	Last 40%
Task As	80%	4%
Task Ds	90%	5%

Task Bs and Task Cs also display this behavior. Unfortunately, the WMS only records the time from when the task is released to when the product leaves the buffer. Not only does this artificially delay the completion time of each task, but it eliminates the tail of the wave for viewing.

This quality suggests that almost half of the time taken to complete a wave exists for a small portion of the volume. The problem of tails can be attributed to two main sources:

- Decreased productivity for waves with low work density
- Worker pursuit of increased productivity numbers and work center focus on hourly capacity

4.1.1 Work Density vs Productivity

This DC represents a massive facility, a majority of which is taken up by SRS. If tasks were released to the floor without organization, it could result in workers traveling thousands of feet between tasks. As a result, the WMS groups tasks into specific work areas, and workers must sign into each area to receive tasks to complete. For this thesis, we can assume that the operation has 15 work zones in SRS and another 14 in Picking, and each work zone can host up to three workers at a time. A work zone consists of ~5 aisles in SRS.

The DC distributes work through SRS based on zone priority. That is, the WMS will first look in Zone 1 to see if product exists to fill an order. If the product doesn't exist in the zone, it will move to Zone 2, and so on. As a result, work zones with a higher priority will receive a higher density of work per wave on average. Work zones with low priority (i.e. zones 10-15) will only be assigned a few tasks per wave.

Consider Figure 18 which shows a high work density area compared to a low work density area. On each aisle, a worker can travel 100 m/min and pick 3 tasks/min. In the higher populated aisle,

the worker effectively completes 2 tasks per minute, while the worker on the lower populated aisle completes only 1.5 tasks per minute.

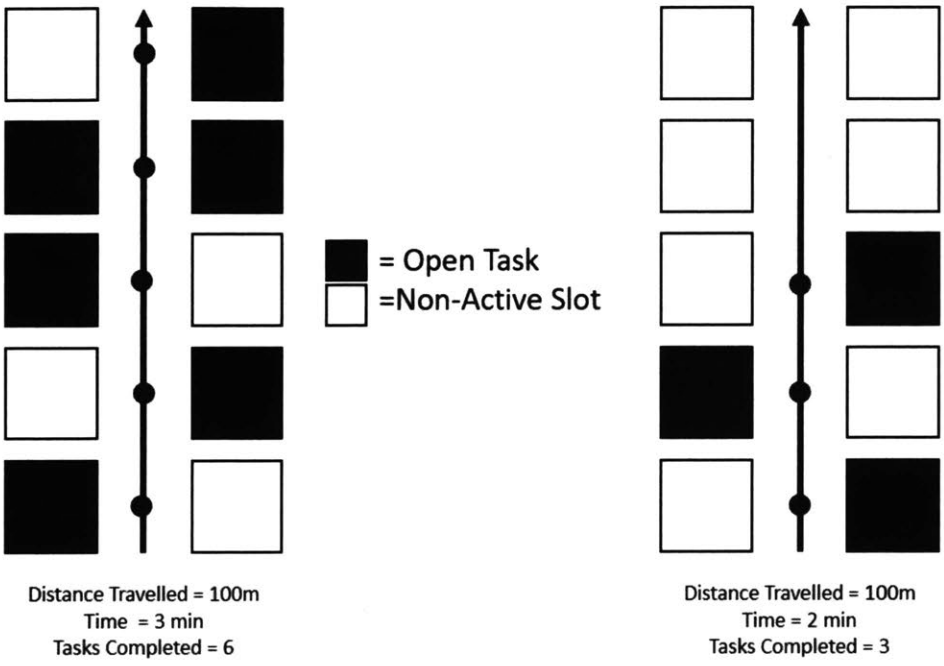

Figure 18. Effects of work density on completed tasks

To solidify this concept and relate it to work within the DC, the average distribution of work for a large sample of waves was generated and compared. Figure 19 shows a measure of this relative work density by showing (on average) how many tasks are available per aisle.

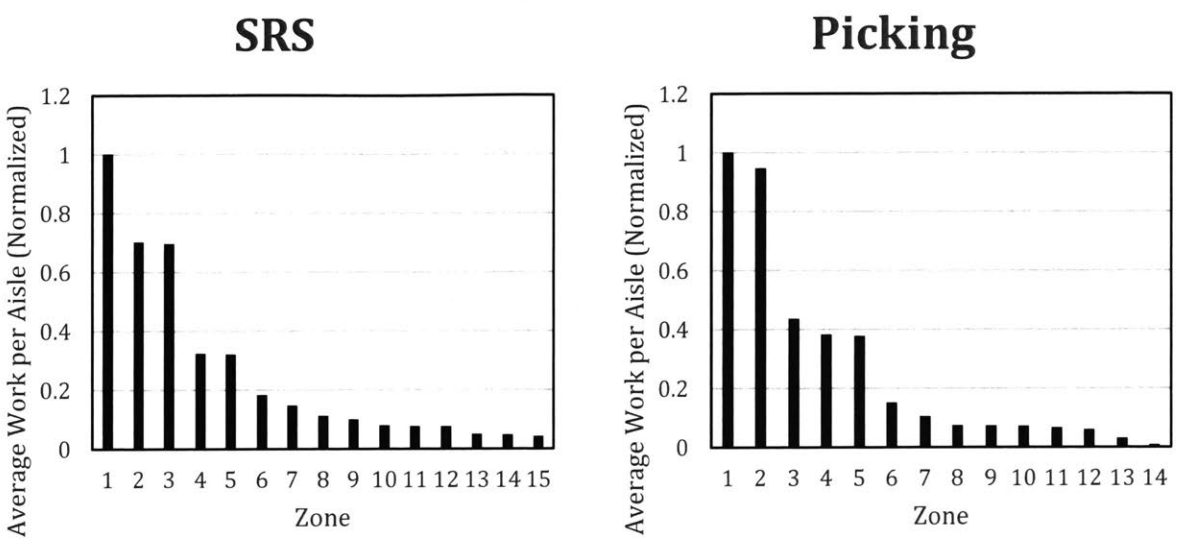

Figure 19. Work Density Distribution for an 'Average' Wave

Aisles in DCs are not typically identical across the facility. From a worker perspective the relative density of a given aisle is important as switching aisles and zones can take more time than moving to an adjacent holding bay on an aisle. As a result, their maximum productivity at any time is influenced by which zone they are in, as well as the density of work in that aisle.

Productivity Numbers

Individual workers and work area managers are evaluated based on their productivity, or overall task completion rate. As a result, workers are incentivized to populate areas with higher work density to take advantage of the shorter distances between tasks. This behavior is at the expense of the lower-density work areas, which will only receive workers when the area builds up enough work to incentivize a worker to sign in, or the Flow Center issues a special request for fulfilling specific tasks.

Throughout the shift, work areas complete tasks in the densest work areas until the flow center stops issuing waves for the shift. At this point, work areas move into 'clean-up' mode where workers move to the less densely populated zones to pick tasks that have been left outstanding. This clean up period typically accounts for the last 20-30% of the end of the shift, and accounts for a smaller productivity. To ensure individual workers maintain favorable utilization and productivity numbers, workers may be sent home early.

Overall, the relationship between work density and productivity combined with the focus on individual productivity numbers means that a small number of tasks will always be left outstanding for an extended period for each shift. Figure 20 shows this path of completion. Notice that the wave never reaches 100%, as the flow center will close waves and route volume to the hospital to finish a wave. This decision of when to close a wave is subjective to the discretion of the flow center, and not a standardized process.

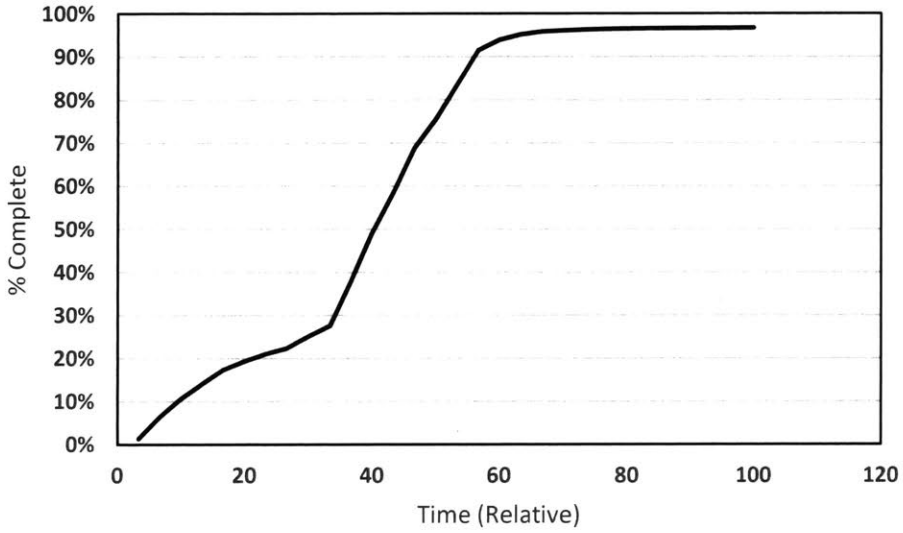

Figure 20. Completion progression of a wave

Summary and Impact

Waves have long tails in the completion curve due to the effects of work density distribution coupled with the desire to maintain high productivity. Wave tails account for a significant portion of the lifespan of the wave and attenuating the tails could greatly reduce the average wave cycle time. As workers and work areas reach the end of a wave, they tend to request additional waves to refresh the work availability on the floor. This allows workers to take advantage of higher work density and maintain productivity.

4.1.2 Wave Variability and Employee Utilization

The previous sections have discussed the interaction between work density and worker desire for productivity. In the next section we will touch on work in the DC from a higher level.

Specifically, we will be discussing the variability of our inputs to the workflow (in terms of work content) and their impact on workforce utilization.

Wave Construction Methods and Variability

Recall the methods by which waves are constructed in the flow center. Wave planners pull all open orders available to the facility and determine their priority based on customer, delivery date and SKU type (e.g. footwear, apparel or equipment). Using this information, the wave planners will build a wave by pulling orders in order of priority to 'fill up' the wave. A wave is determined to be 'full' once a sorter has been filled to maximum capacity.

While this ensures that waves will not overflow the sorters, it does little to estimate the workload that will be included on the wave across all tracked tasks. Figure 21 displays a visualization of which outbound cartons determine work content in a wave.

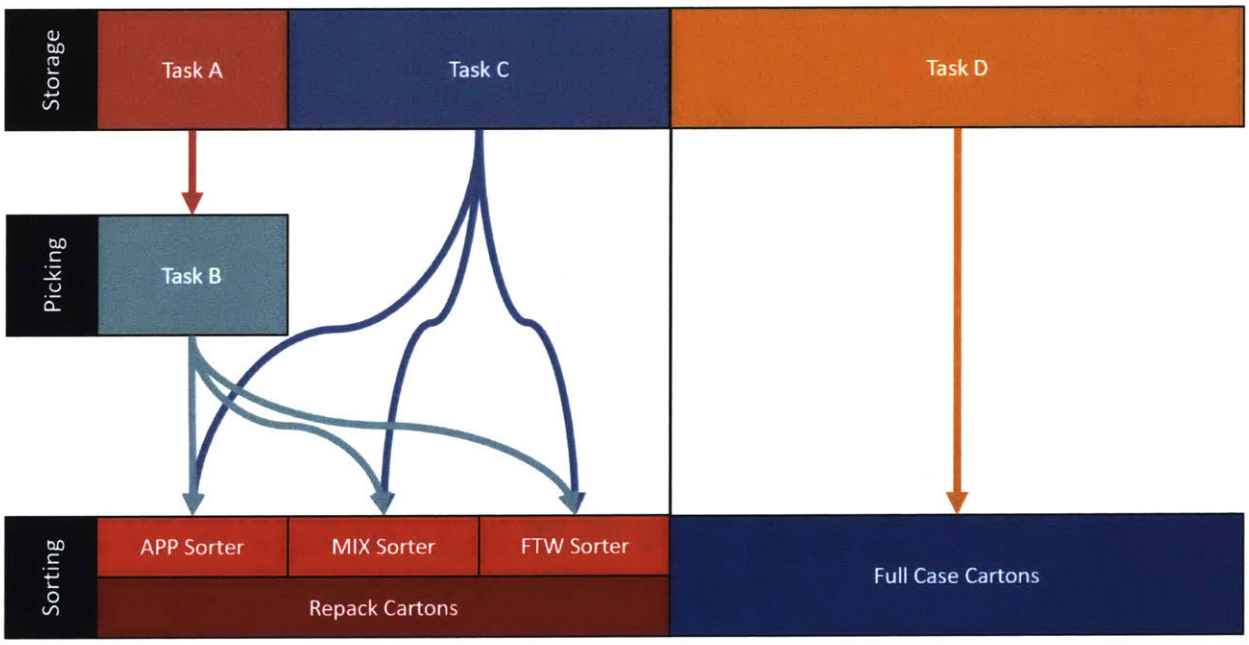

Figure 21. Visualization of which cartons determine work content

During wave construction, the wave planner will reach maximum capacity of either the footwear (FTW) sorter or the mixed (MIX) sorter and finish the wave. That single node builds the wave to a single metric: outbound cartons on this sorter. This gives the wave planner limited control of

multiple aspects of the work content of the wave. Table 3 describes these variables and the control a wave planner has over them when planning a wave.

Table 3. Summary of Controlled Variables in Waves

High Control	Limited Control	Minimal Control	No Control
Outbound Cartons on Filled Sorter	Outbound Cartons on Secondary Sorters Total Units to Sorters	Total Tasks to Sorters Task As	Ratio of Task B to Task C Task Ds

While wave planners have developed loose heuristics to control the work content of a wave, they are largely constrained by the work they have available in the order pool. As a result, the waves that are distributed to the facility throughout the day have high variability in work content and work balance across work tasks and work centers. Figure 22 displays the relative spread of work content for a wave, measured in tasks, for an extended period at the DC.

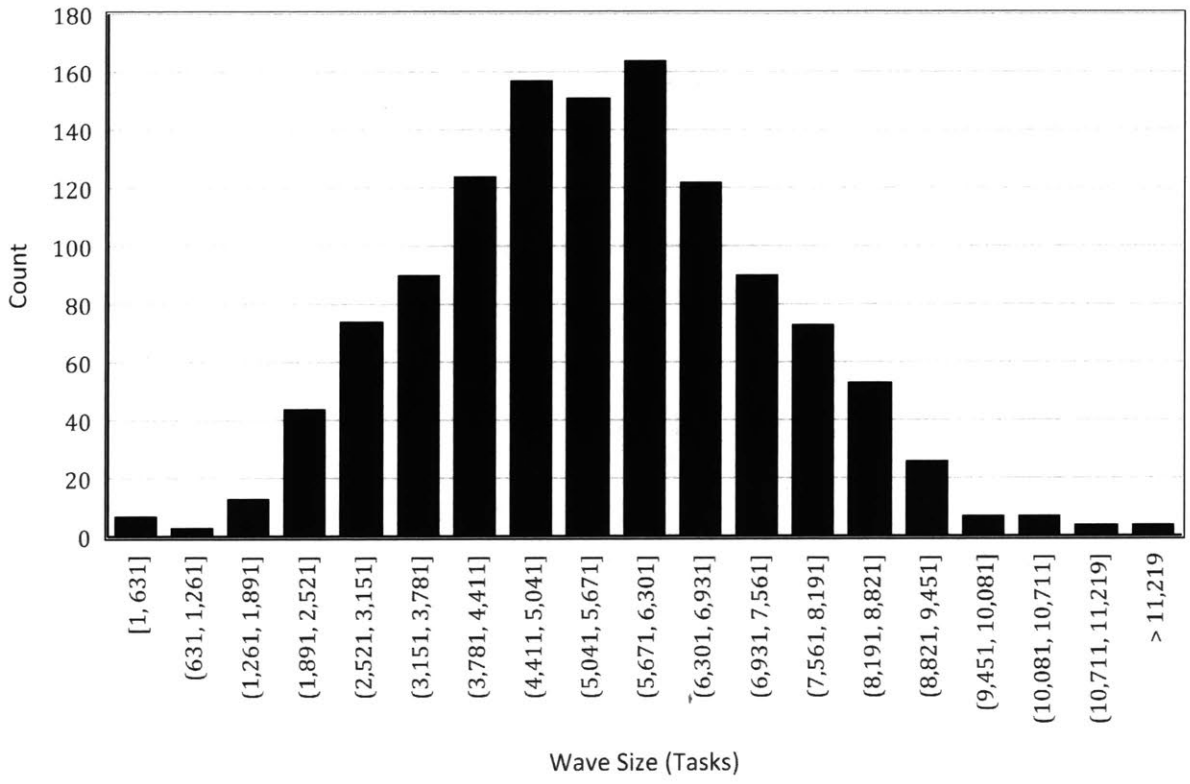

Figure 22. Histogram of normal wave sizes taken over a 5 month period. CoV for wave sizes is measured at 0.35

Furthermore, if we begin to piece apart the relative work content of each wave, we see even greater variation in work that is being distributed to work centers on a wavely[13] basis. Figure 23 shows the variation in work split for every wave between full-case shipments (Task Ds) and repack shipments (Task Bs and Cs) in the wave[14]. The shape of the histogram shows that there is little control over the amount of work that would be sent via full-case shipment (work fulfilled only in SRS) and work fulfilled by the repack tasks (work within SRS, Picking and Sorting). The coefficient of variation for this ratio is 0.31.

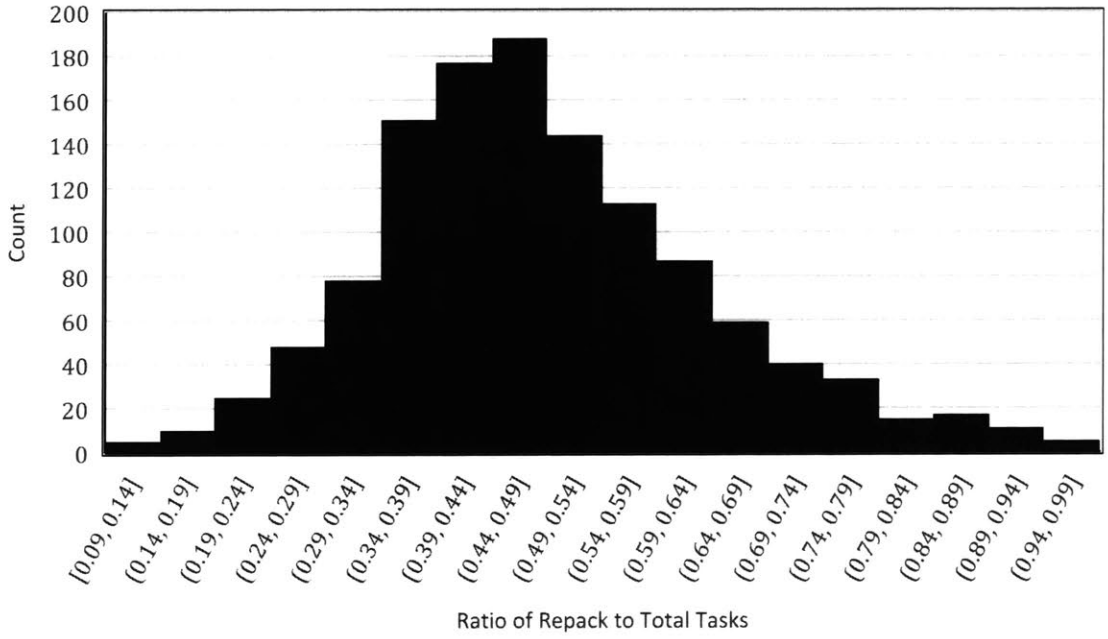

Figure 23. Histogram describing the ratio of repack tasks to total tasks or, effectively, high level work balance between SRS, Picking and the Sorters

The ratio histogram in Figure 23 shows the relative relationship between work in SRS and work shared across the facility. The ratio of repack tasks to total tasks shows how much work is

[13] Wave-by-wave

[14] This ratio is important, as it determines the relative staffing demand for the work center. Given a set throughput of units, work content can cause work areas to have inadequate or over-staffing. Wave-by-wave, it can cause major demand shifts throughout a work day

available to Picking and Sorting for any given wave. Waves with a high ratio (right of the histogram) have lots of work for picking while low ratio waves have minimal picking work. Viewing this information in raw form can give more clarity to the potential problems that could originate from this variation. Figure 24 shows the raw data behind the histogram shown in Figure 23.

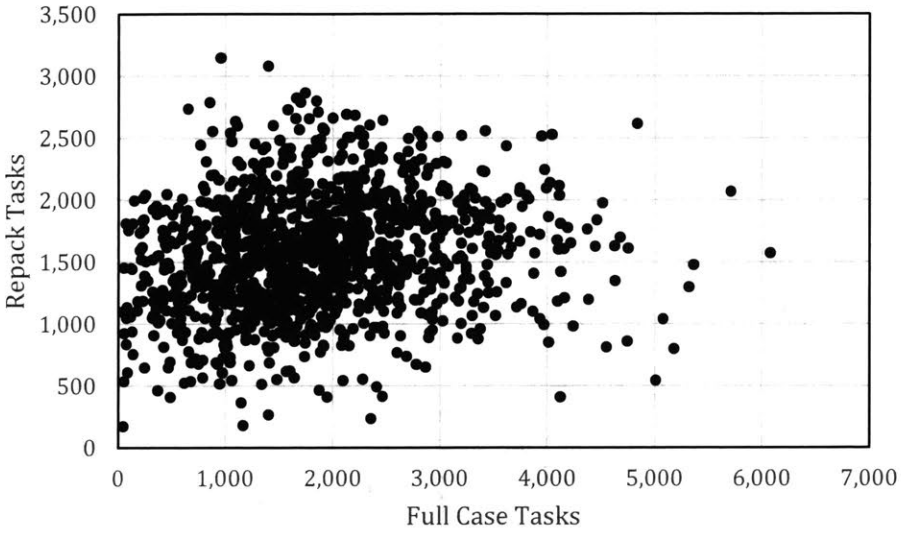

Figure 24. Full-Case Tasks (Task Ds) Versus Repack Tasks (Task A, B and C) for the sample set of normal waves

From this data, the magnitude of the problems associated with wave variation begins to take shape. Not only are waves of massively different sizes when they reach the production floor, but each wave lacks the ability to ensure that work centers are receiving equal amounts of work for the course of the next wave. There are many waves shown in Figure 24 that have a high amount of full case tasks, but relatively little repack tasks. This would effectively starve the work centers in Picking and Sorting while ensuring a large amount of work is available in Storage.

Lastly, it is important to discuss the amount of work available for Picking on a wave-by-wave basis. Due to the presence of Task Cs (cartons sent directly to the sorters from Storage) it is not possible to predict the amount of work content that will be allocated to the picking with the

current system. This is shown in Figure 25, the coefficient of variation for this distribution is 0.2, showing high variation in this work balance.

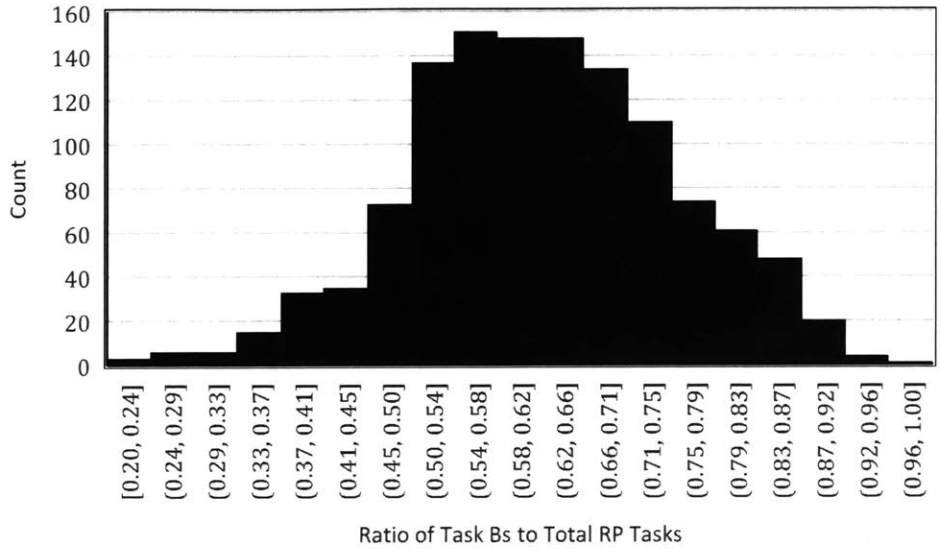

Figure 25. Ratio of Picking Tasks (Task Bs) to total Repack Tasks (Task Bs + Task Cs)

While work variation within Storage is spread out across three different task types, Picking is reliant on a single task type within normal waves to provide work. During waves that have low work content available for Picking, it is probable that the work area will finish work prior to others and employees will be idle or shifted to another work area.

Utilization, Summary and Impact

As workers are incentivized to maintain high productivity throughout the day, work areas and the flow center have the same desires with utilization metrics. If a work area is low on work, workers may become idle due to lack of tasks. This will then be reported in the weekly metrics and be scrutinized by the staffing and financing departments. Utilization numbers that are too low will result in lower staffing levels in the DC for future operations.

As a result, when work areas become low on available work tasks, the flow center and work zones are incentivized to release another wave to the floor to ensure tasks are available. When waves are as variable as listed in the previous section, then the flow center may not have the ability to construct and deliver a wave that satisfies the work area with low work availability. As a result, it may take multiple waves to satisfy the work demands for the facility. Furthermore, as work areas

are unable to rely on the next wave satisfying their needs, they tend to ask for new waves sooner than needed as determined by their current utilization status.

4.1.3 Wave Jumping and Extended Cycle Times

From the previous two chapters, we saw that workers, work areas and the flow center all benefit from having higher work in process (WIP) and, consequentially, more waves open in the system. The effects of having more waves open in the system was well documented in Wallach's thesis in the wave jumping section. After solidifying these concepts, we can now determine a root cause for the wave jumping and, sequentially, the primary behaviors that drive wave cycle times:

- As work availability depletes in a work area, workers will request new waves to increase work density and boost productivity
- As work availability depletes in a work area, work area supervisors and the flow center will request new waves to increase work availability and boost utilization[15]
- When a new wave is released, workers will move to the area of highest work density to complete the most efficient tasks
- Inability to tailor work content in waves eliminates the ability to plan waves for even work distribution across work centers

The behavior manifests in visible ways, most apparently with increased WIP throughout the system. Figure 26 shows the WIP over the course of two shifts at the DC. As shown below, workers will enter their shift with minimal workload in the system and request work to start the shift. Due to a combination of all factors listed above, it appears that multiple waves are released early in the shift.

[15] The first two bullets have the same outcome, but from different origins. It is important to note that the current system incentivizes higher work availability from a worker perspective as well as a managerial perspective

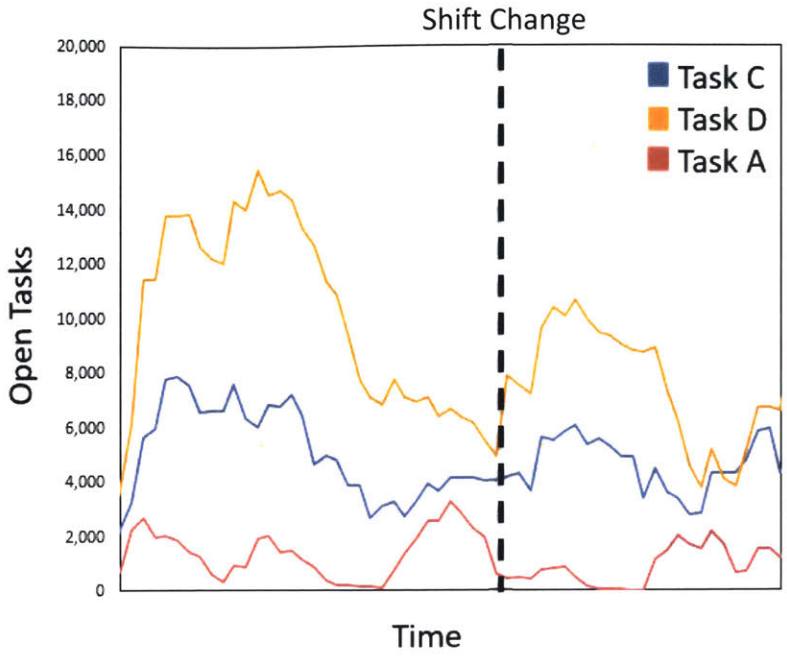

Figure 26. WIP at the DC over two shifts

As the distribution center finishes allocating the daily work to the floor, there's a sudden stop in work input, and overall WIP drops at the end of the day. This has been referred to as the 'clean up' period. The last few hours of any shift are typically dedicated to finishing waves and completing tasks that have been left outstanding throughout the day. In response to the limited work selection and low work density, many temporary employees are sent home early.

Summary

This behavior listed in this chapter drives wave cycle time due to these effects:

- Waves are naturally completed less each hour due to work density effects
- Wave releases are pulled forward to ensure high productivity and utilization
- First-in-first-out is not enforced as tasks are delayed until the 'clean up' period

While the large spike in WIP may lead readers to think of the implications of Little's Law in closed system, the primary reason waves have extended cycle times is due to the lack of FIFO in the operation. The inability to incentivize wave completion over productivity and utilization metrics causes wave tails to be left outstanding until the cleanup period of a shift.

In the next chapter, we will look at this behavior through another lens to analyze the impact on quality within the operation.

4.2 Quality

Simply put, high quality in a supply chain means every customer and consumer receives their order on time and in full. Retailers track this metric through customer chargebacks, which are received after a customer receives a late or incomplete shipment. As a result, it is difficult to directly track the quality of the operation in the moment. The best indication of the quality of the operation is tracking the number of tasks that are diverted to a chase wave.

4.2.1 Chase Waves and Quality

As previously described, chase waves are waves that consist of tasks that are "cancelled" in their original wave. There are two ways that a task can be cancelled:

- Inventory accuracy issues result in a picker being tasked to pick a unit that does not exist in the proper location
- Wave planners will force close a set of tasks to progress the wave or close the wave

Every half hour the WMS will collect all outstanding cancelled tasks and build a chase wave. This chase wave is sent to the hospital where a team of workers will work to expedite these tasks and have the units 'catch up' with the rest of the wave. Tasks that have been cancelled due to inventory accuracy could then be resolved 'in the moment' while tasks that have been cancelled due to wave closure are the final outstanding tasks of the wave that is usually waiting at the shipping docks.

Tasks attributed to a chase wave do not necessarily catch back up to their parent wave. Waves that are near complete and waiting at the shipping docks have low tolerance for waiting on single units to ship. Tasks that are in chase waves due to inventory accuracy are considered lower priority, as the waves are still active. This delays their completion until their parent normal waves are waiting at the dock doors and the Hospital has a limited time to complete the resolution tasks.

Lastly, the evening shift uses the workforce in the hospital to fulfill high priority digital orders (e.g. Digital Next-Day-Air). The workforce is consumed with this task and is taken off chase wave

reconciliation. Tasks appearing on chase waves at this time are ignored until the high priority tasks are complete for the shift.

Regardless of the success rate of reconciliation tasks in chase waves, a high volume of reconciliation work indicates that the general process of a distribution center is failing too many of its tasks. Recall the figure first shown in Chapter 2 that shows the total volume of tasks routed to chase waves, expressed in % of total, over time.

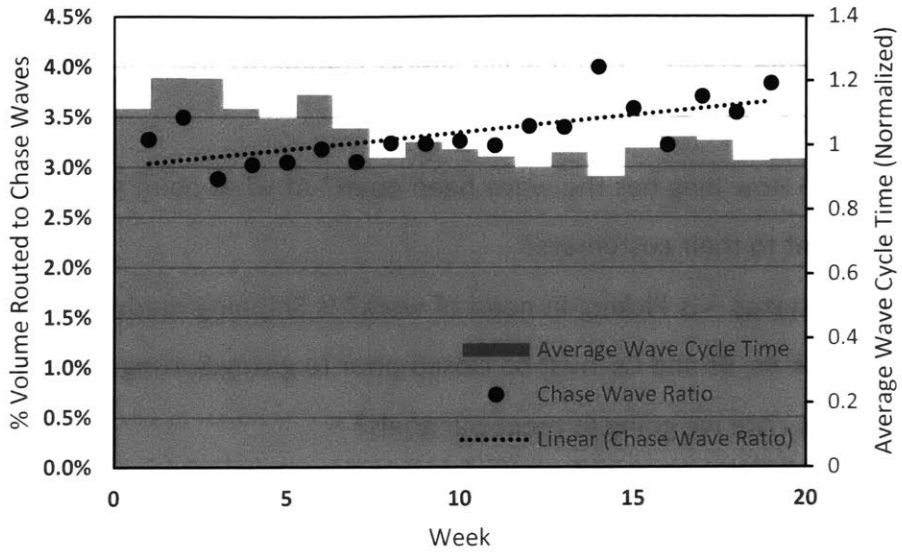

Figure 27. Volume routed to chase waves and Average Wave Cycle Time

Figure 27 displays a clear trend of increasing volume routed to chase waves over time[16], thus a steadily lowering of quality. During this period, inventory accuracy remains constant, fluctuating between 97.5% and 98.5% without consistent direction. This suggests that chase wave volume is growing due to a systematic trend in the behavior behind wave closures. The next chapter will identify this behavior, the driver behind it, and the reason for increased prevalence of impact in the operation.

[16] The increased volume at weeks 2 and 14 represent the end of fiscal quarters where shipped volume reaches a maximum. Lower inventory inaccuracy is associated with these high-volume periods

4.2.2 Flow Center Decisions and Metrics

Due to the extended tails for waves, it is uncommon for any large DC to close waves that are fully complete. Instead, the flow center has heuristics on when waves can be called 'complete' and closed, thereby sending outstanding tasks to chase waves. Generally, waves are considered ready to close after they have reached 95% completion status. From then on, the actual time from when the flow center calls closure is largely subjective and dependent on the status of the operation. Some of the factors that are considered when closing waves include:

- **Wave completion status** – Is the wave almost complete? Are a few tasks holding up a large amount of orders?
- **Age of wave –** How long has the wave been open? At what point should we release the truck to be sent to their customers?
- **Work center status –** Is Picking in need of work? Is Shipping ready to receive outbound packages? Task As, Bs and Cs must be closed prior to giving Sorting more work.
- **Metric targets -** Are we going to meet our goals?

With the emphasis on increasing the speed of order fulfillment, the operation has been steadily decreasing their targets for average wave cycle time. The operators have seemingly had success in hitting the goals set out by management, as shown by Figure 28.

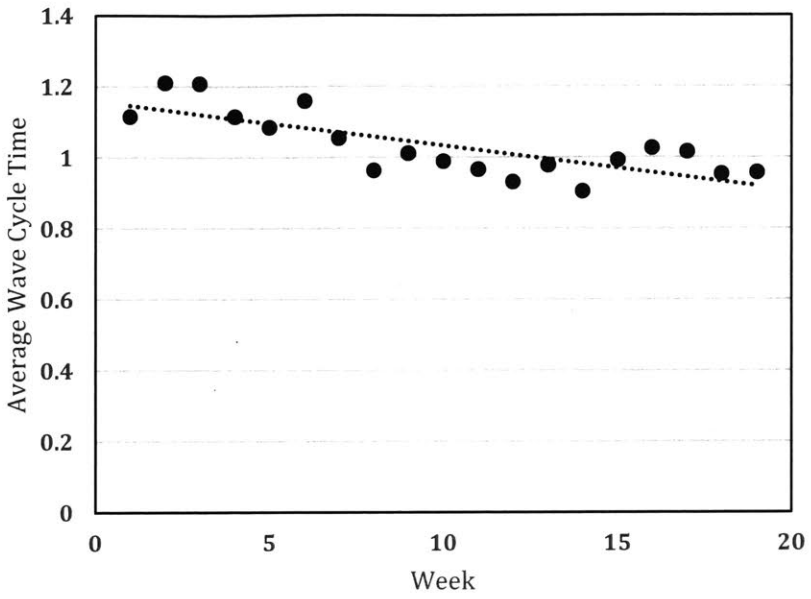

Figure 28. Average weekly wave cycle time (Normalized)

If we consider that the wave cycle time determined by the time that a wave is closed, and waves are typically forced closed by the flow center, a driver behind lower quality begins to emerge.

Flow Center Priorities

For this section, we will consider the wave holistically, and include completion status of all tasks in the wave and consider the decision to finish a wave. If we refer to Figure 20 again of the typical wave completion profile, we can begin to understand what is driving quality. Figure 29 shows the same information as Figure 20, but as a completion percentage over time. The wave represented in this figure is an 'average' wave, which is closed at ~97%.

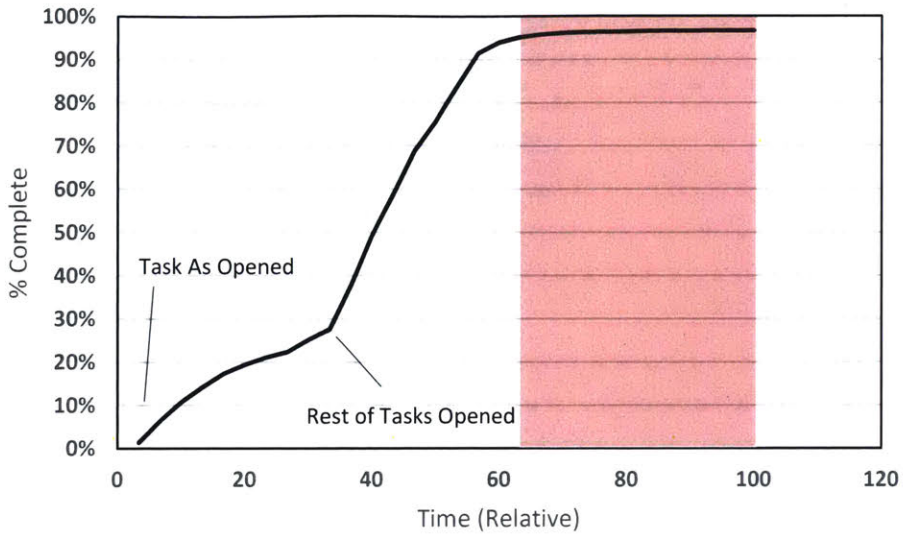

Figure 29. Window of 'appropriate' wave closures as per the listed heuristics

The red box in this figure represents the time period in which the wave is available for closure according to the standard of wave closure after 95% of tasks are completed. If we claim that an average wave cycle time is 12 hours, a typical wave would be available for closure after 7.5 hours have elapsed (first ~60% of the wave). Over the last 4.5 hours of this aggregated wave, 2% of volume for the wave is completed.

From the flow center perspective, this allows significant flexibility in decision making as they are free to adapt to the needs of the operation. Problematically, however, the continued decrease of target wave cycle time incentivizes the flow center to close waves earlier, effectively 'clipping' the wave tails and routing more volume to the hospital. This behavior is best described in Figure 30. According to worker interviews within the flow centers, this is a major reason why the operation can meet weekly wave cycle time goals as they decrease.

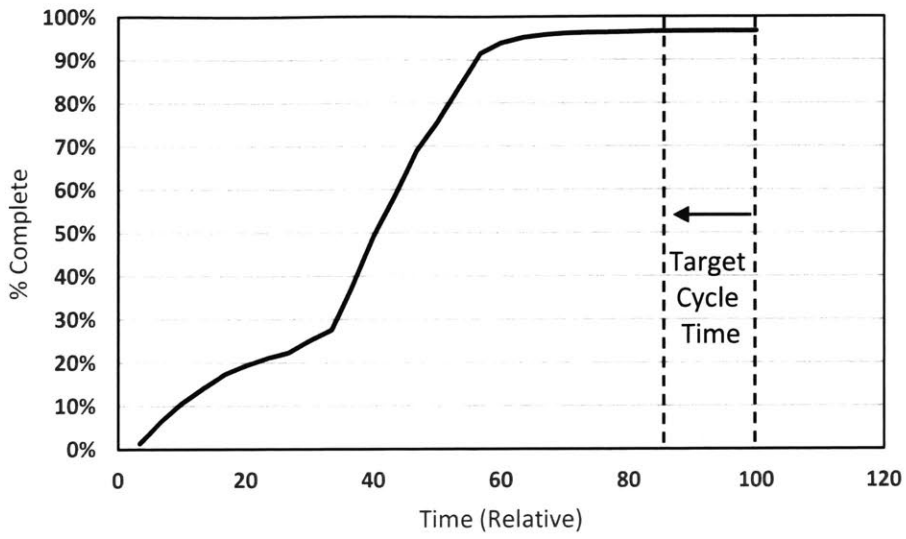

Figure 30. Wave closure times as target wave cycle time decreases

Impact on the Hospital

While the decision to move the wave closure time earlier seems to have a seemingly minor impact on the completion percentage of the wave, earlier closures can have a drastic effect on the work within the hospital. Within the tail of the wave, we can assume that ~0.2% of work is completed on the wave per hour. While this work content is marginal for the larger operation, it represents over 6% of the work normally attributed to the hospital.

When we look at the trend in Figure 27, we can see that this increase in Hospital workload is not sustainable given the current infrastructure. Over the sampling period, the hospital saw an increase in volume over 20%. The infrastructure of the hospital has a limited number of physical slots for reconciliation units, and this increase in volume has strained the ability of the DC to keep these spaces open. If this trend were to continue, the hospital would need to respond with increased staff and footprint.

Furthermore, the increased load on the Hospital inhibits the ability of this work center to provide fast service for task reconciliation. Every task that enters the Hospital must be worked and resolved quickly enough to catch up with the rest of the wave or meet the order on the outbound shipping container before the truck leaves. Increasing the arrival rate of reconciliation tasks by 20% would increase the average amount of open tasks in the hospital by a similar amount,

potentially exceeding the capacity of physical slots in the infrastructure. In addition, the increased tasks could flood the employees with work, straining their ability to complete individual tasks in a timely manner for each wave. The resulting increase in reconciliation time could reduce the ability of the hospitals to provide meaningful work to the operation, and SIFOT could suffer.

Summary

The behavior described in this chapter drives quality due to these effects:

- Wave closure criteria is largely subjective, allowing for flexibility to respond to current operation
- Consistently shortening the target for wave cycle time incentivizes the flow center to close waves earlier in their life cycle
- Wave tails are getting clipped earlier, pushing more work to the hospital and overloading their capacity

Overall, this behavior suggests that the base operation is unable to improve on productivity, efficiency, or wave completion. Instead, the operation is trading off cycle time for quality to meet the targets set out by management. To meet the eventual target of 2/3 of current cycle time, the underlying work system must be changed to break this link between quality and cycle time and allow both metrics to be improved together.

4.2.3 Capacity

As previously described, the DC can meet all current demand (100,000 units per shift). It is important to understand the primary drivers of capacity, however, as any solution that is implemented to shorten cycle time cannot do so at the expense of capacity. This chapter will describe the behavior of capacity within the DC and allow us to better predict the impacts of any solutions we propose.

Work Availability and Productivity

As previously described in Chapter 4.2.1, worker productivity is closely linked with work density and availability. As a result, the flow center releases waves earlier in the shift until they have run out of work for the shift. This behavior was described in Figure 26. With this understanding, we

can now look at the hourly productivity to understand how this influences the throughput of each work center. To begin, we will separate a shift into two halves to characterize periods of high WIP and low WIP. When we plot the hourly throughput of SRS and Picking for two shifts, a trend appears.

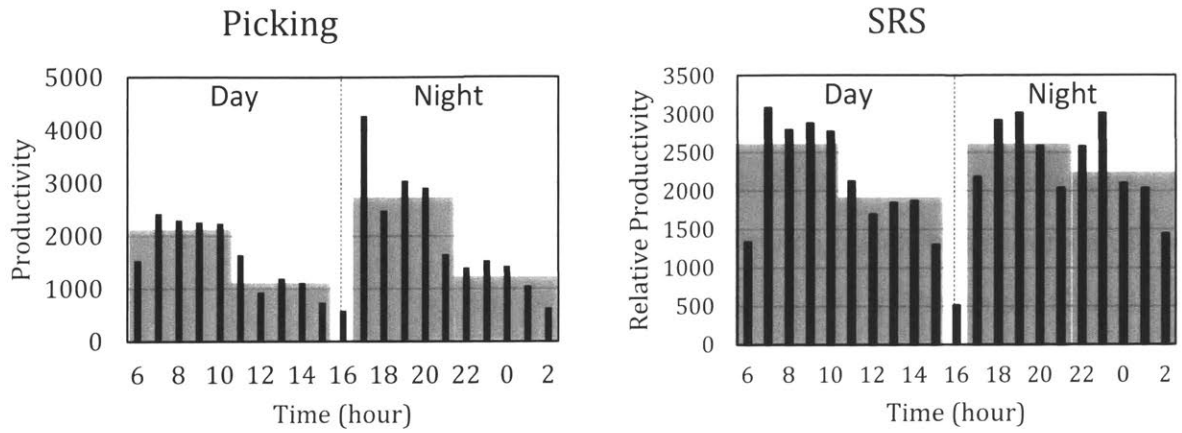

Figure 31. Average hourly throughput of the day and night shift with average bars for first and second half of shifts

On average, Picking has over double the throughput of tasks in the first half of the shift as the second half, and SRS has over 25% better performance in the first half. These throughputs correspond to a similar distribution in the average WIP for the first and second half of these shift. Table 4 shows the data for the comparison of WIP to throughput for SRS and Picking.

Table 4. Average WIP and Throughput for first and second halves of shifts

		Average WIP (tasks)		Average Throughput (task/hr)	
		First Half	Second Half	First Half	Second Half
Day Shift	SRS	21,000	10,500	2,600	1,800
	Picking	4,200	2,100	2,100	1,100
Night Shift	SRS	16,000	11,500	2,600	2,250
	Picking	7,500	1,500	2,700	1,200

As expected, the levels of WIP within each work center correlate with the throughput of the work center. Considering the drivers described in the analysis of wave cycle time, we can assert that throughput of the work centers is determined by work density and work availability. As the

DC is currently meeting all demands of capacity, however, we cannot claim that total capacity of work centers is determined by these factors. Any attempts to quantify this relationship would ignore the specific behaviors and qualities that cause this relationship to begin with. These factors include:

- Worker idleness based on work availability
- Temp workers sent home after half-shift
- Stalling wave releases to prepare for digital waves (especially on night shift)
- Time spent on workers shifting work zones, aisles and work centers
- Break times and startup/shutdown activities
- WMS time reporting of tasks (especially task As and Cs)

When determining a system to decrease the cycle time of the DC, it is important to consider the effects of this system on capacity. Unfortunately, the high number of confounding variables mean that we can make recommendations but cannot fully quantify an impact on the overall system.

Summary

As the current operation currently meets all required demand, it is best to speak about throughput within the facility. Throughput is driven by these effects:

- Imbalance of WIP as work is pulled forward during the day and runs out at the second half of the shift
- WIP determines work availability and work density, which directly influences worker throughput

4.2.4 Summary and Integration of Drivers

To understand the work completion characteristics of the DC, it is important to begin with the inherent capabilities of the work system and layer in context to the work environment. Figure 32 shows a high-level map that links the inherent system qualities to their observed outcomes.

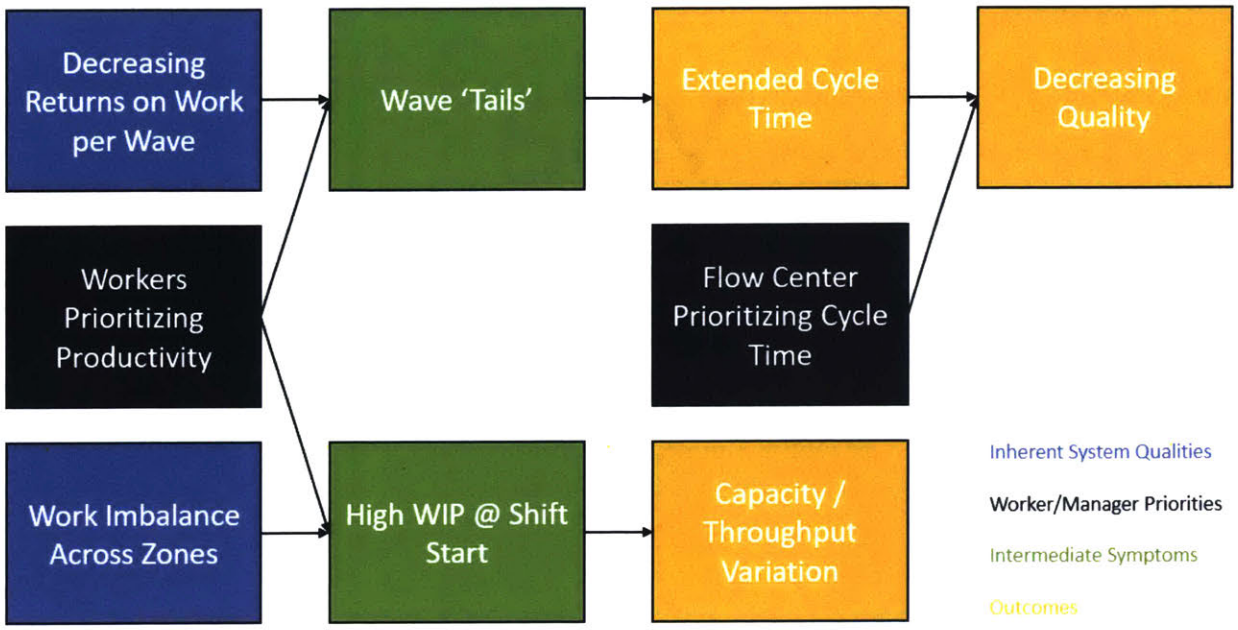

Figure 32. High level causal relationships that take place in the DC

Any system that attempts to change the outcomes for distribution centers will not be able to fundamentally change the inherent system qualities of the operation (listed in blue). The solutions may mitigate their effects but cannot eliminate their presence. The items listed in black can be altered, but this must be agreed on by management and workers alike.

The next chapter of this thesis will describe the proposed changes the DC's workflow to break or improve some of the relationships described in the root cause analysis.

5. Recommendations

After assembling much of the previous data and concluding interviews with the key players in the operation, a kaizen was held to determine the best course of action to reduce the cycle time of the DC. The primary recommendation is a schedule that will plan the release and completion of waves for the entire facility. All the recommendations in this first section do not require capital investment to realize the benefits of this system. Secondary recommendations consist of projects and tools that will enable this system to run smoothly and efficiently but are an additional investment to this new waving system. The capital requirements for these recommendations will be discussed.

5.1. Scheduling Workflow

5.1.1. Solution Identification

In previous chapters we have described the workflow in the DC as a predominantly 'Push' operation. Within the literature review, we determined that while push allows the workers to have high productivity and utilization numbers, it does this at the expense of cycle time. This aligns with worker emphasis on these metrics, but not with retailer's overall strategy and needs.

To remediate this, the DC must commit to implementing more 'Pull' mechanisms in the operation. As the operation is complex and difficult to control, a schedule of releases and closures was created. This schedule assumes a standard wave 'takt time' at every work center. Figure 33 shows this schedule for a single wave as it progresses through the DC while assuming a standard takt of two hours.

	06:00	08:00	10:00	12:00	14:00
SRS	Task A	Task C	Task D		
Picking		Task E, B			
Sorting			Sorting		
Shipping			Shipping		

Figure 33. Schedule of task releases and closures for a single wave

To describe the operation, we will walk through the workflow of this single wave:

6:00 - Wave is created and Task As released to SRS (sending cartons to Picking)

8:00 - All outstanding Task As closed and sent to Hospital for resolution

Task Cs are released to SRS (sending cartons to the Buffer)

Task Bs and Es[17] are released to Picking (sending units to the Buffer)

Task Es are completed in Singles Area and placed on outbound trailer

Digital Singles Orders Complete

10:00 - All outstanding Task Bs, Cs and Es are closed and sent to Hospital

Shipping confirms docks are prepared for wave (trailers in place)

Buffer releases wave to the sorters

Task Ds released to SRS (sending cartons to Shipping)

Shipping receives all cartons from Sorters and SRS

12:00 - All outstanding Task Ds closed and sent to Hospital

Shipping finalizing outbound trailer, waiting on Task Ds from Hospital

[17] Recall Task Es are Digital Singles Tasks. These represent a minority of the tasks in DCs, but are important to include in planning for this operation

13:00 - Outbound trailers packed and routed to final destination

Overall, this schedule would constrain the wave to have a cycle time of seven hours. Orders normally included on a digital singles wave would have a cycle time of four hours. Furthermore, as the DC is a continuously run operation, a new wave will be released into the system every two hours to ensure even loading of all work centers throughout the shift. This system waterfall is shown in Figure 34.

	06:00	08:00	10:00	12:00	14:00
SRS	1	1 2	1 2 3	2 3 4	3 4 5
Picking		1	2	3	4
Sorting			1	2	3
Shipping			1	2	3

Figure 34. Waterfall of waves released to the workflow

After the system has been initialized as in Figure 34, the facility will commit to releasing waves every standard takt time to keep the system in balance. If a shift change were to occur at 14:00 hours, workers in Sorting would arrive at the start of a shift with wave 3 in progress ready to be worked. Under normal operation, the system would never have gaps in the schedule for work areas.

5.1.2 System Implications

While this solution seems simple, enacting it in the operation requires careful consideration of system needs and tradeoffs. The following list gives details on implementing this schedule in the operation.

Wave Types and Task Priority

The DC has five different types of waves: Normal, Digital, Digital-Single, Digital NDA and Chase. Under the new system, there would only be normal waves and chase waves. Wave planners would generate waves by selecting all available digital orders and adding orders from general customers until the wave is appropriately sized for work content[18]. Each wave would then be entered into the system every two hours in accordance with the schedule and standard takt time. This eliminates multiple types of waves and the associated issues regarding different wave priorities, such as wave jumping (Wallach, 2018).

At any point of the day, SRS will have only three waves open, one working on Task As, one working on Task Cs and the last working on Task Ds. While SRS is responsible for completing all tasks in accordance with the waving schedule, the tasks themselves will have a priority order in the WMS. Simply put, the Task As that are open in the system will have highest priority due to the need for Picking to complete the task. The next priority will be Task Cs, as failure to deliver all necessary units to the Sorters can result in frozen slots and exceed the capacity of the work center. Finally, as there is an inherent delay in units coming from sorting to shipping, Task Ds will be the last priority within a single takt time in SRS.

Similarly, Picking will have only one wave open at a time with two task types open. The priority will be the tasks associated with digital orders, as speed on these orders can directly unlock cheaper, faster deliveries to consumers. Beyond Picking, there will only be one wave open at any given point (with flexibility in Shipping for delivery drivers).

Ensuring Pull and Releasing Waves

This system is designed to set strict rules on workflow within the DC to better direct the efforts of the workers. As this schedule represents an ideal workflow, it is critical to establish a set of operating guidelines to be used when the workplan must deviate from the ideal. Using the learning from our literature review, specifically within Chapter 3.3 Push vs Pull, we can create

[18] As digital orders represent a small amount of the volume, the inclusion of these orders on Normal waves would have miniscule impact of the typical wave profile already established in this thesis

rules that will ensure the system will maintain a sense of 'Pull' throughout any point in the operation. These rules are listed below:

- Only one set of tasks open in any work center at a time (but could include different waves)
- All work centers must cycle their work content together – for example, SRS cannot open a new wave without also opening a new wave in Picking
- If all work areas are ready to complete the waves and open new tasks, they may do so together. The new wave released to the floor must account for the total two-hour wave plus the amount of time needed to get back on a two-hour schedule

Ensuring that these rules are followed will ensure that WIP is low and waves will be able to cleanly move through the system without obstruction. These rules also eliminate the ability of workers to request additional waves to boost productivity numbers and will ensure even workloads throughout the day.

Wave Closure Rules and 'Last Task Agony'

Every wave will reach a point where completing the last tasks will be a major detriment to the productivity of the operation. At this point, the DC will be waiting on the final few tasks of a wave that are likely in control of only a few workers on the floor. Meanwhile, workers will be idle, and momentum will halt in the work areas. To keep this 'last task agony' from impacting the greater operation, it is imperative that margins of error for wave completion are established for when the flow center can force close a wave. These margins will be refined with experience working in the new system, but initial values proposed for the first trial are below:

- Before the takt time for the wave, completion of 98% of the wave in all work areas allows forced wave closure
- After the takt time for the wave, completion of 96% of the wave in all work areas allows forced wave closure

These completion percentages are based on the performance capabilities of the Hospital. In an ideal case, 2% of the DC's volume would be routed through the Hospital for resolution. In the worst case, 4% would be routed through the Hospital, a value which is currently regarded as

nearing capacity for this work center. If the system is working as expected, most waves will be completed at the takt time for the wave and the Hospital will receive a volume between 2%-4% of the DC's volume[19].

From the Hospital's perspective, this system allows much greater flexibility in task reconciliation times. As waves are currently closed while the wave is on the outbound shipping dock, the Hospital has only a short time to reconcile the tasks in the wave. In the new system, the Hospital will be able to reconcile tasks 'in the moment' to help all tasks reach shipping in the same time window.

Picking Strategies

Due to the reduced WIP, the DC will maintain lower density of tasks in the overall system. This will cause the current system of picking orders (where workers can move to higher density areas to be more productive) to be less effective in completing waves on time. We recommend that the DC moves to a more zone-based picking strategy that requires workers to stay within areas of lower density tasks and clean the area throughout the day. This will result in smaller wave tails and allow waves to reach the 96%-98% threshold much faster than waves currently completed in the operation.

Inventory Storage Strategies

Recall the WMS looks to assign work by indexing through a hierarchy of zones to check for inventory. As a result, zones higher on the index tend to have higher throughput volume than zones lower on the list, causing work imbalance and many other effects described in Chapter 4. To create a more balanced system, we recommend changing the work assignment algorithm from the current zone-based approach to a row-based approach.

A row-based work assignment algorithm would look to assign work that exists on the most accessible row in every zone and then moves to the second most accessible row in every zone. In SRS, the algorithm would first look to the bottom row across all of SRS, as forklifts can travel fastest when low to the ground. The next row in SRS would be the row immediately above it, and

[19] Given a standardized method of wave closure, future continuous improvement initiatives will be able to identify potential root causes and design experiments to reduce the overall amount of volume routed to the Hospital

so on. In Picking, the algorithm would first look to the row at waist level of a typical worker for ergonomic purposes.

This strategy would serve two purposes. First, the work would be more evenly distributed across aisles and work zones. Second, 'optimal' rows will have inherent preference in work assignment, allowing workers to work faster and safer.

Implementation and Metric Recommendations

This system will upset many of the metrics currently used to track progress and efficiency at a typical DC. As a result, it is easy to understand that this system could be met with resistance if it was not introduced in the proper manner. First, considerations of the current metrics must be anticipated to set expectations for the new system. Below is a list of items that will be 'negatively' impacted with this new system:

- **Utilization** – probabilistically, there will be a lower utilization from workers as workers wait on new work throughout the day. Additionally, work areas will not be able to send home workers before the end of the shift to alleviate utilization numbers
- **(Hourly) Productivity** – The shifts typically begin with a boom in WIP, and a corresponding boom in productivity. This will be eliminated in the new system, but sustained productivity will allow the facility to meet its capacity demands throughout the shift. As workers will not be sent home prior to the end of shift, per-worker productivity will decrease
- **Cost per Unit (CPU)** – As workers stay for the entire shift, they will have more hours charged to the operation. This will result in a higher effective CPU per product. It is worth noting that this will be a minor increase in the overall CPU as workers are typically paid for at least half of their shift regardless of how much time they spend working

Workers will be sensitive to the decreased hourly productivity as this has been their primary method of evaluating employee performance. Work Center supervisors and the flow center will be sensitive to the utilization numbers as this is their primary performance metric, and management and finance will be resistant to systems that will increase overall CPU.

While the above metrics are considerations when running any operation, they are secondary to the goals of many retailers. To properly implement this system, all stakeholders must be reminded of the goals (reduction of cycle time) and challenges that operators currently face. Instead of the metrics listed above, we propose these alternate metrics that will align management, work centers and individual workers:

- **Average Wave Cycle Time** – In this system, wave cycle times are constrained to be completed in three standard takt times. This metric will be the primary indicator that the system is functioning as expected
- **Average Wave Completion %, or % Volume outed to Chase Waves** – This will be the primary metric used to track performance. An average wave completion of 96% will be considered an area for improvement while a high functioning system would have completion amounts of 98%. Continued performance at either of these levels would justify moving the bounds of acceptable completion rates.
- **Throughput** – It is critical that the operation meets all demand assigned to the system. As throughput is not equivalent to capacity of the DC, it is better to track the throughput through every work center to better understand how the system performs to shifting workloads throughout the day.

To ensure workers have security in their continued employment, the current performance tracking system must be overhauled to account for the changes to the new work system. Workers who are assigned to less dense areas will be unable to meet the performance metrics of their peers in the higher density aisles and zones. For the first trials of this system, it would be reasonable to 'disable' the current performance metrics and track the performance of the system, rather than the employees. Once the system is running, it will be easier to determine and implement metric tracking that more closely align the goals of the workers with the goals of the operation.

5.1.3. Expected Outcomes

This new system is designed to ensure a level of performance in wave cycle time while motivating workers to work towards high quality. By constraining waves to spend two hours in every work

center and focusing all effort on completing a single wave, we can ensure that the tails of the wave will be shortened, and waves will move through the facility in under a third of the time taken on normal waves today. Furthermore, as tasks are forced closed within the wave, the Hospital will have ample time to reconcile tasks while the wave is still active within the workflow, rather than when the wave is at the shipping docks ready to leave.

While the current performance metrics tracked in the DC will decline under the new system, new metrics will ensure that the goals of the operation better align with the goals of the business. As retailers push to increase speed across their supply chain, this system represents a low cost and high performing alternative to other investments such as automation or building new infrastructure.

5.2 Secondary Tools and Recommendations

This chapter collects multiple recommendations that will improve upon the primary set of recommendations and allow easy adoption of the new system. Each of these recommendations will require investment for implementation, but the overall capital required will be modest compared to other alternatives.

5.2.1 Wave Planning Tools

A key element in the new system is the ability to reliably plan a wave that constitutes a set work content for each work area. Wave planners currently do not have the ability to plan waves for a prescribed work content. Instead, planners assemble waves based on order amounts and sorter slots. As described in Chapter 4.1.2, this does not reliably determine the quantity of tasks available to each work center. If wave planners are unable to tailor work content of waves to meet the needs of the schedule, utilization will decrease due to the imbalance of work across work centers and this could impact the capacity of the DC.

To give wave planners the ability to custom tailor waves, a set of tools were developed to predict likely work content of a wave based on key order characteristics. This tool is meant to resemble the current tools wave planners use to construct waves and is meant to give more information on what to expect in the resulting wave composition.

Predicting Work Content

The developed tool for predicting work content in a new wave was created to provide proof of concept for a wider implementation across the distribution center. The tool allows wave planners to pick orders from the pool of all available orders as they would with their current system. Once the initial wave has been constructed, wave planners can generate a preview of the wave through Excel. A snapshot of the preview page is shown in Figure 35. Using this wave preview, planners would be able to manipulate the content of the wave prior to entering the wave into the WMS. This allows changes to be implemented faster and eliminates the effort of deconstructing waves.

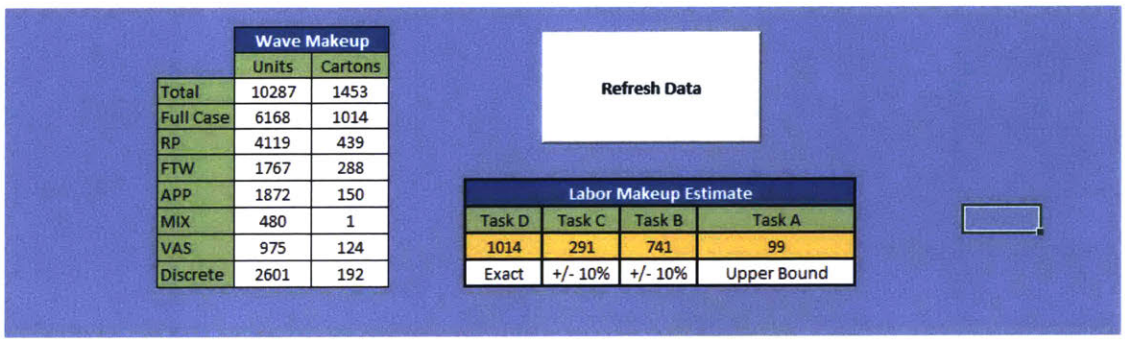

Figure 35. Front Dashboard of Work Prediction Tool

As seen on the dashboard, the tool gives estimates of work content for the wave rather than exact values. This is due to the inability of the tool to account for the inventory levels of work areas in real time and account for which sub-sorters units are routed to. For instance, if Picking is well stocked and a wave is entered with a small repack amount, then the wave will have a small number of Task As. Similarly, a larger Repack wave will have more Task Cs in the wave, but only if those outbound units are routed to the same sorter and sub-sorter. The algorithm behind this prediction tool is meant to simulate the logic that the WMS uses to assign work within the facility while making statistical estimates in areas where the tool is missing important information.

Tool Precision

Using a selection of waves gathered from two months, the tool was calibrated and tested to prepare it for implementation. Overall, the ability of the tool to predict work content is listed below for each task type:

- Task A – Tool can predict an upper bound for the wave that is accurate for 97% of tested waves. Upper bound was used as the tool is unable to account for inventory staged in Picking
- Task B – Tool can predict the value +/- 10% for tested waves. Error in this tool is derived from sub-sorters, which tend to increase the number of Task Cs as the WMS cannot split Task Cs over sub-sorters. This is an unbiased prediction of work content
- Task C – Tool can predict the value +/- 10% for tested waves. Error derived from the same source as the error for Task Bs. This is an unbiased prediction of work content
- Task D – Tool can predict exact values of Task Ds as there is no dependence on inventory or sorters
- Task E – Tool lists Task Es in the Discrete tab on the tool. This number is exact.

The large number of degrees of freedom[20] in work assignment lead to low covariance of errors between these predictions.

Next Steps

While this tool can give planners some insight to waves prior to WMS generation, it could be greatly improved by increasing accuracy and giving additional information to the planner prior to wave creation. This would further reduce the variability in workflow and increase the reliability of predictions from the tool. A future tool must be developed that will have the following content:

- Ability to precisely predict work content for a wave prior to generation
- Guidance for which orders account for what work content
- Provide guidelines on how to re-route and re-package orders within the DC to optimize work content and balance
- Predict work zones where work would be located, allowing work areas to flex capacity in different areas to respond to shifting work density

[20] Including sorter assignments, order preferences, inventory representation

Ideally this tool would be generated and implemented by the contractors who design and maintain the current WMS system. After initial talks with providers of WMS solutions, it appears that this tool can be developed and implemented on request.

5.2.2 Hospital Decentralization

The Hospital is a single work center that is accountable for reconciliation of all chase wave tasks. Because this center needs to travel to all areas of the facility, it is centrally located in the operation. Due to the size of large DCs, workers in the Hospital have extended travel times on routine tasks. To reconcile a task, a worker must leave the Hospital to travel to the work zone, collect the item, and travel back to the Hospital to store the unit in a temporary slot. Another worker will come to the hospital to collect the task and travel back to the parent work center in order to reconcile the task. The amount of distance traveled by these workers leads to a significant amount of time and effort to solve a single work task in a chase wave.

To remediate this problem, a DC can split this single Hospital into a network of decentralized work areas (called Urgent Care Centers). Under this system, reconciliation tasks will be completed within the work center, drastically reducing travel time. Once the unit is collected, the conveyor system can be used like every other task completed within the DC. The reduced effort for task reconciliation will expand the capacity of the Hospital network and allow for faster task reconciliation in line with the needs of the new work scheduling system.

5.3 Future Recommendations and Investments

The following recommendations require further testing or larger investment for implementation. While they are not critical to the success of the new work scheduling system, they can unlock new capabilities for retailers to leverage as it speeds up its supply chain.

5.3.1 Full SKU Representation in Picking

DCs are constrained to a specific order of task completion based on the inventory staging standards in Picking. To ensure that all Task Bs and Es are available for Picking, all the Task As for the wave must be completed. Picking areas may have difficulty providing full SKU representation on the picking floor due to the high number of SKUs in their portfolio. If a DC can commit to full

SKU representation within Picking[21], the system could begin work on Digital orders immediately on wave creation and release to expedite these units through the facility.

Updated Schedule –Single Full Case SKU Representation in Picking

With at least one full case available, digital orders would be able to be fulfilled using only the inventory represented on the Picking floor. This allows the facility to release all tasks associated with digital orders to Picking when the wave is generated and released to the floor. Task As would be completed as normal and ensure that the next wave would still have full representation of SKUs on the Picking floor. Figure 36 shows this schedule.

	06:00	08:00	10:00	12:00	14:00
SRS	Task A Task C Digital	Task C	Task D		
Picking	Task E Task B Digital	Task B Wholesale			
Sorting			Sorting		
Shipping			Shipping		

Figure 36. Waving schedule with 100% SKU availability on the Picking floor

It is important to note that in this system, all digital single orders would be fulfilled by Task Es and FC cartons to the digital singles processing areas. All digital Task Bs must be sent to the same Mixed Sorter to ensure that high-volume SKUs on the digital selection can consolidate orders onto Task Cs. If multiple sorters are used, the WMS could split a Task C into multiple Task Bs, which would require volume from picking larger than a single FC carton.

[21] Full SKU representation assumes at least one full-case amount of a unit is always present on the Picking floor

This schedule would allow digital singles orders, representing over 80% of the digital volume, to be fulfilled within a two-hour wave cycle time. Non-singles orders would have a wave cycle time closer to four hours due to the sorting requirement (though time in sorting is not tracked).

Updated Schedule – Multiple Full Case Representation in Picking

Storing more inventory in Picking reduces the dependence of Task Bs and Es relying on Task A completion. If a DC were to store enough inventory on the picking floor to ensure all Task Bs and Es are available for Picking at the start of a wave, these tasks could be released at the beginning of the wave along with Task As. In this system, Task As are replenishment tasks more closely associated with a Kanban system. Figure 37 shows the resulting schedule of wave completion under these conditions. Each wave would have a trackable wave cycle time of four hours in this system.

	06:00	08:00	10:00	12:00	14:00
SRS	Task A Task C	Task D			
Picking	Task E, B				
Sorting		Sorting			
Shipping		Shipping			

Figure 37. Wave completion schedule with high SKU inventory in Picking

Implementation and Considerations

This system would require any typically sized distribution center to dedicate a large amount of space to Picking to accommodate the additional storage capacity. Dedicating this inventory capacity for single case quantities for all SKUs is feasible considering the footprint available and the current slot utilization, allowing the schedule in Figure 36 to be a realistic goal. The amount of inventory capacity required to enable the schedule in Figure 37, however, would require

significant footprint of Picking in relation to the entire operation. Retailers must evaluate the cost of this footprint and weigh it to the benefit of gaining additional reductions in takt time.

5.3.2 Continuous Wave Takt Time Reduction

With all continuous improvement initiatives, it is important to have a clear path for future improvements. The easiest way to reduce the overall wave cycle time is to decrease the wave takt time for every work center. A two-hour wave was selected (as described by this thesis) for closely resembling the current system's wave amounts and wave sizes. In the future, it would be a simple task to move to lower takt times, such as a single-hour takt.

Implementation and Considerations

A single-hour takt wave would result in a three-hour wave cycle time. This would drastically speed up the response time of the operation, but could lead to issues such as:

- Higher workload imbalance – leading to lower utilization
- Increased wave tails if workload imbalance not properly managed
- Erratic workload hour-by-hour making workload planning difficult for work center managers
- Magnification of problems arising from inventory accuracy issues

Ultimately, reducing the cycle time will need to be done gradually to ensure the operation maintains efficacy. Additionally, the retailer will need to determine if further reducing the cycle time of the system is needed for better fulfilling the demands of customers and consumers.

5.4 Recommendations Summary

This chapter has outlined a new system constructed to address the problems described in the root cause analysis. Overall, the proposed system is established to have workers, work areas and the Flow Center all work towards a common priority of wave completion. This schedule creates this effect through the following reasons:

A. **Cycle Time is pre-determined through waving schedule –** With clear guidelines on the workflow for the day, the DC will be enabled to plan work accordingly and maintain consistent staffing levels.

B. **Workers unable to request new waves –** Workers will still prioritize productivity, but without the ability to release waves forward in the schedule, this productivity will be directed at finishing the tail of the wave.
C. **Flow center has clear wave closure guidelines –** The flow center will still be concerned with the cycle time of a wave but must achieve this through wave completion. Without the ability to close a wave early, they can better focus their efforts on ensuring outstanding task completion on the floor.

These changes will have many effects, especially on the level of WIP in the system throughout the shift. To optimize the current workplace processes within a DC for this new system, the following recommendations are proposed.

D. **Workers move to a zone-based strategy for task assignment –** Decreased WIP will exacerbate the effects of work imbalance across zones. Moving towards a zone-based strategy for picking will allow workers to ensure coverage of tasks across the DC.
E. **Work assignment algorithm switches to row priority over zone priority –** Switching to row preferences from zoning preferences will spread work across the facility more evenly and prioritize locations that are easier to reach for workers.
F. **Completion margins for wave closure –** To avoid 'last task agony', the Flow center will be able to close waves before they are fully complete, allowing systematic, but manageable tail clipping within the operation.

This system will mitigate the impact of inherent tradeoffs in the working environment of a typical DC while constraining the metrics that are important to the success of the operations. These constraints will allow workers to work towards improving metrics locally without negatively affecting cycle time and quality. To enact this work planning system, it is important for all stakeholders to understand that tracked performance metrics of the operation (productivity and utilization) will have less meaning in the new work assignment system.

6. Conclusions

To better serve customers and fulfill consumer demand, large retailers must speed up its supply chain. This means each of their distribution centers must reduce the time it takes to fulfill customer orders within the four walls of the operation. Currently, these systems tend to have tradeoffs between quality and speed, and improvement initiatives have not been able to improve underlying performance in order fulfillment. By implementing a new system for completing work, we can break the tradeoffs that currently define the performance and align the goals of workers, work centers and management.

This system begins with a standardized waving schedule with standardized work distributed throughout the facility. By reducing the variability of the workday, we can allow workers to better plan their days and empower them to achieve higher performance without tradeoffs. This new system unlocks future continuous improvement initiatives that can have an even greater impact on the speed of the distribution network.

The improvements proposed in this thesis represent low-cost but high-impact solutions. Alternatives to these solutions include automation, and projects that require massive infrastructure investments. The benefits of these systems are well known, as the industry is continually turning to reduction of workforce and increasing automation. Overall, this system will be able to deliver fast turnaround times for a much lower cost and minimal disruption to the ongoing operation.

As distribution centers move closer to the retail industry's vision for future supply chains, businesses must ensure that the objectives and priorities of the workers within the operations align with those of the overall business strategy. In this way, retailers will be able to respond to the developing retail landscape to adapt performance to suit the needs of customers. The solutions outlined in this thesis can boost the current performance of a distribution center; in the future it will also ensure the flexibility within the operation to respond to unknown future demands.

7. References

Nike. 2017. NIKE, Inc. Announces New Consumer Direct Offense. *Nike News.* [Online] June 15, 2017. https://news.nike.com/news/nike-consumer-direct-offense.

Push and Pull in Manufacturing and Distribution Systems. **Pyke, David F. and Cohen, Morris A. 1990.** 1, s.l. : Journal of Operations Management, 1990, Vol. 9.

Reference for Business. 2019. NIKE Inc. *Reference for Business.* [Online] 1 14, 2019. https://www.referenceforbusiness.com/history2/99/NIKE-Inc.html.

Selecting between batch and zone order picking strategies in a distribution center. **Pratik J. Parikh, Russell D. Meller. 2007.** 2007, Transportation Research Part E.

Statista. 2018. Nike Stores Worldwide. *Statista.* [Online] 1 14, 2018. https://www.statista.com/statistics/250287/total-number-of-nike-retail-stores-worldwide/.

Wallach, Matthew. 2018. Reducing Wave Cycle Time at a Multi-Channel Distribution Center. Cambridge, MA : MIT, May 11, 2018.