Cycle-time Analysis and Improvement Using Lean Methods within a Retail Distribution Center

by

Hugh Churchill
M.S. Biomedical Engineering, University of Michigan, 2009
B.S. Mechanical Engineering, University of Michigan, 2008

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 2015

Signature of Author Signature redacted
MIT Sloan School of Management, MIT Department of Mechanical Engineering
May 8, 2015

Certified by Signature redacted
Stephen Graves, Thesis Supervisor
Professor of Management Science, MIT Sloan School of Management

Certified by Signature redacted
Brian Anthony, Thesis Supervisor
Principle Research Scientist, Department of Mechanical Engineering, MIT

Accepted by Signature redacted
David E. Hardt, Chairman, Committee on Graduate Students
Department of Mechanical Engineering, MIT

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

This page intentionally left blank

Cycle-time Analysis and Improvement Using Lean Methods within a Retail Distribution Center

by

Hugh Churchill

Submitted to the MIT Sloan School of Management and the Mechanical Engineering Department on May 8, 2015 in Partial Fulfillment of the Requirements for the Degree of Master of Business Administration and Master of Science in Mechanical Engineering.

Abstract

Fulfillment cycle-time, or the time it takes to pick an item from inventory, pack it into a box, and load it on a truck for shipment, is one of the main inputs in determining how quickly an online retailer can promise customer order delivery. The faster the fulfillment cycle-time, the later an order can be received and still make the appropriate truck for guaranteed, on-time arrival (e.g. same-day, next day, 3-5 business days). Thus, the customer experience is improved, as they are allowed to place an order later and still receive their purchases quickly. To take advantage of this, the retailer must first be able to measure cycle-time appropriately within their facility.

This thesis examines the outbound fulfillment process within an underperforming Amazon fulfillment center (Site A) with the purpose of fully characterizing and measuring fulfillment cycle-time. Comparisons are drawn with like Amazon facilities, and a lean operations approach is taken to identify and eliminate major forms of waste in an effort to shorten cycle-time.

The baseline analysis within this thesis provides evidence that current-state cycle-time at Site A is in fact 15% faster than originally thought. However, process improvements were still needed to bring cycle-time in line with the network standard. The remainder of the work within this thesis focuses on these process improvements and develops the following recommendations:

1. Standardize the pick process with a move closer to single piece flow.
2. Reduce and control queue length prior to the pack process in order to reduce non-value-added wait time.
3. Reduce batch size for critical items that must move through the facility the fastest.
4. Rearrange process steps to allow completion in parallel rather than series.

The method for evaluating cycle-time and the implementation of lean solutions introduced throughout this thesis are useful as a template for similar analyses throughout the Amazon FC network, as well as within other warehousing and online retailer operations.

Thesis Supervisor: Stephen Graves
Title: Professor of Management Science, MIT Sloan School of Management

Thesis Supervisor: Brian Anthony
Title: Principle Research Scientist, Department of Mechanical Engineering, MIT

This page intentionally left blank

Acknowledgements

The completion of this project work and thesis writing would not have been possible without the generous help and continued support of many individuals along the way. I would like to thank Amazon for this fantastic opportunity and for their continued support of the Leaders for Global Operations program over the years. Thanks goes to those individuals who served as sounding boards and mentors throughout the process while at the Amazon fulfillment center, especially my fellow classmate, Alberto Luna.

I also wish to acknowledge the Leaders for Global Operations program for its support of this work. I would also like to thank my academic advisors, Stephen Graves and Brian Anthony, for their guidance and input throughout the internship and thesis writing process.

Finally, and most importantly, I owe the biggest debt of gratitude to my loving family and fiancé. Their kind words of encouragement, editing help, and continued support have been invaluable resources throughout my entire MIT journey.

This page intentionally left blank

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Amazon.com

Amazon.com, started by Jeff Bezos in 1995, holds the mission "to be Earth's most customer-centric company where people can find and discover virtually anything they want to buy online"("Amazon Media Room: Overview," n.d.). Beginning as an online bookstore shipping out of a Seattle garage, Amazon now offers millions of unique items that ship worldwide from an extensive network of U.S. based and international fulfillment centers (distribution centers). Amazon continues to drive toward its customer focus by offering this ever-growing item selection at faster delivery speeds. Amazon Prime offers customers the option of free two-day shipping on many purchases, and Prime Now, introduced in December of 2014, promises tens of thousands of daily essentials delivered within an hour("Amazon Media Room: History & Timeline," n.d.).

1.2 Project Motivation

Through the acquisition of Shopbop.com (2006) and Zappos.com (2009) and the creation of MYHABIT.com (2011) ("Amazon Media Room: History & Timeline," n.d.), Amazon has made a continuous effort to become a leader in online fashion sales and expand its product offering of soft-lines (apparel, shoes, jewelry, and watches). Soft-lines fulfillment poses a number of operational challenges including inventory placement, storage, packaging, product mix, and seasonality that are not present when fulfilling more traditional items like books and movies. These challenges are thought to slow the soft-lines fulfillment cycle-time (time between customer purchase and finished package ready for shipping) and necessitate storage within a dedicated facility. One such dedicated facility will be the focus of this thesis, and will herein be called Site A.

It is important for Amazon to offer the same fulfillment speeds for soft-lines as those which customers grew accustom to with other products. To keep up with these demands, Amazon must continue to drive efficiencies within its fulfillment centers. Site A provides the perfect environment for soft-lines research and testing that can be further rolled out across the network at a later date. This thesis examines, in detail, the fulfillment process and cycle-time within a soft-lines dedicated facility, with a focus on eliminating wastes while minimizing risk to customer.

1.3 Problem Statement

Fulfillment cycle-time directly impacts the online promise that Amazon can provide to its customers. When a customer is shopping for an item online, Amazon provides the customer with a time promise that, if they order within a specific amount of time and choose next day delivery, they can receive their items tomorrow. The algorithm running behind the scenes to help estimate this time promise includes a very important input from the fulfillment center (FC) where the items are located. This piece of information is known as the pick agreement time and is directly correlated with an FC's fulfillment cycle-time. The shorter the fulfillment cycle-time, the shorter the pick agreement time, and thus the larger the period of time a customer has to order an item for immediate delivery. Pick agreement values are set individually by FC and are left as a static number year-round. The majority of Amazon's FCs have pick agreements set to a network standard value, defined herein as 1 normalized time unit. This means that customers may order for next day delivery up until approximately 1 normalized time unit prior to the last available truck departure that day.

The pick agreement time at Site A is the longest in the Amazon network, set at 1.56 normalized time units (i.e. 1.56 times the network standard value for pick agreement time). This

larger time setting shortens the customer purchase window for soft-lines products, thus negatively affecting the customer experience and limiting the potential for additional soft-line product orders. This pick agreement setting remains this way at Site A due to the following:

1. The soft-lines related operational challenges mentioned in section 1.2 Project Motivation, and their perceived negative effect on outbound cycle-time.

2. Sub-optimal outbound processes (Site A has only been in operation for two years).

3. Lack of an appropriate measure of fulfillment cycle-time, and thus no data to support changing the standing value.

These three factors limit the volume of orders fulfilled by Site A and, ultimately, negatively affect customer experience.

1.4 Project Goals

The main goal of this project is to create a better way to set pick agreement time at Site A, thus improving customer experience and potentially increasing soft-lines volumes without increasing risk to customer promise. Internally, this manifests as an evaluation of Site A's fulfillment process, with focus on first measuring and then reducing fulfillment cycle-time. This thesis examines the current-state fulfillment cycle-time for Site A and compares it to other sites within the network. It also identifies the key drivers behind this cycle-time, and develops solutions to eliminate wastes and implement lean improvements. These methods for measuring and improving fulfillment cycle-time provide Amazon with a template for future analysis and the proper setting of the pick agreement time at Site A, and other sites.

1.5 Thesis Overview

The discussion in this thesis proceeds as follows:

Chapter 2: Amazon Fulfillment

This chapter describes relevant background information on the promise that Amazon provides to its customers and the internal fulfillment process that delivers on that promise.

Chapter 3: Literature Review

This chapter provides background research on the theory employed to complete the work throughout this thesis.

Chapter 4: Problem Analysis

This chapter provides a baseline analysis of cycle-time within Site A and two comparison sites. The analysis in this chapter creates a clear picture of the current state at Site A and sets up the waste elimination and improvement work discussed in Chapter 5.

Chapter 5: Putting Lean and Operations Management Theory into Practice

This chapter presents a discussion of the approach utilized to study sources of lean waste within Site A. Three separate cases walk through the process of observation, hypothesis, testing, analysis, and suggested improvements.

Chapter 6: Conclusions and Recommendations

This chapter reflects back on the project work presented throughout the thesis and compiles the lessons learned along the way. These lessons learned set up a number of recommendations for cycle-time reduction and proper pick agreement time setting at Amazon. This chapter also provides recommendations for future work and additional projects.

2 Amazon Fulfillment

Amazon receives, stores, and ships all of its available items using a network of U.S. and international distribution centers known as Fulfillment Centers (FCs). These FCs have two main operational functions: 1) inbound, and 2) outbound. Inbound operations involve the logistics of ordering, receiving, and stowing products that will be available for customer purchase. Outbound operations focus on fulfilling customer orders, including all the steps from customer purchase to completed package on a truck. The analysis within this thesis focuses entirely on outbound operations and the fulfillment process, which is defined in detail within this chapter.

2.1 The Customer Experience

When an Amazon customer decides to make a purchase, he or she first visits the website and browses through the selections. A decision is then made and the purchase button is clicked. This information is immediately sent to the appropriate FC, and the fulfillment process begins. Once the order is packaged and placed on a truck, it makes its journey via third-party carrier to the customer's front door (as seen in Figure 1).

Figure 1: Diagram of customer experience with Amazon

Keeping in mind the mission of a customer-centric company, Amazon puts customer experience above profitability. When a purchase is made, Amazon makes a promise to the customer that the order will arrive on or before the date indicated by the choice of shipping method. Faster methods, upgrades, are costly and generally paid for by the customer upon checkout. Every effort is made to fulfill this delivery promise, i.e. Amazon will absorb the cost of an unplanned shipment upgrade to get a customer order to his or her door on time. Thus, it is of critical importance that the FC fulfillment process runs smoothly and predictably.

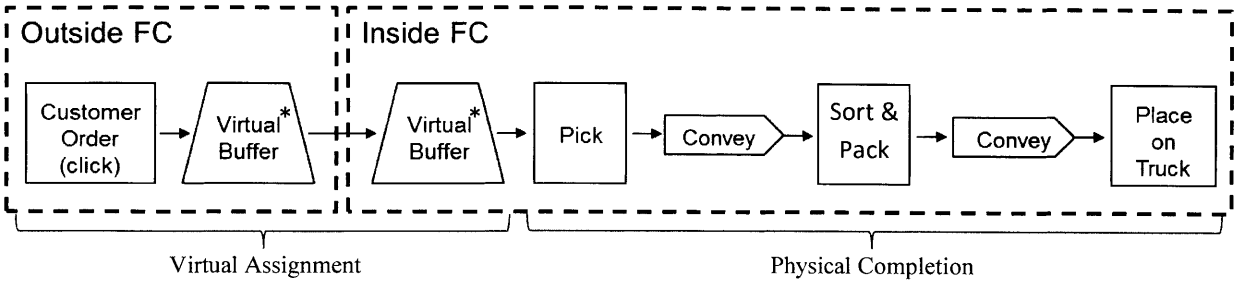
2.2 Fulfillment Process Overview

The fulfillment process at Amazon consists of two steps: 1) virtual assignment, and 2) physical completion within an FC. Standardized computer algorithms control the virtual assignment process. First, the algorithms run a credit check and locate the ordered items at the FC from which shipping is the cheapest from a system standpoint (often the closest to the particular customer). Once the credit green light is received, the order is passed virtually to the appropriate FC where it is then placed in a virtual queue, waiting to get physically completed. The physical completion process is taken care of by the outbound operations team within an Amazon FC. This process consists of three main physical steps: pick, pack, and truck loading.

Depending on the FC, these physical fulfillment steps can be performed in different ways. Most sites utilize manual labor to fulfill items that are small enough to be shipped via standard carrier methods (i.e. a standard sortable site). Some new sites employ the help of robots, while other small FCs fulfill low-volume, large items (e.g. televisions and canoes) which require very specific procedures. The analysis within this thesis centers on Site A, which is a standard sortable site with one additional non-standard procedure (i.e. a special case sortable site). The following sub-sections provide further detail on sortable versus special case sortable.

2.2.1 Standard Sortable Site

A standard sortable site within the Amazon FC network fulfills items that are small enough to be shipped via standard carrier methods. Sortable refers to the fact that these facilities fulfill customer orders for more than one item (multi-item orders). These items are picked by different associates within the FC and later grouped (or "sorted") together prior to packing. See Figure 2 for a visual representation of the outbound fulfillment process at a standard sortable FC.

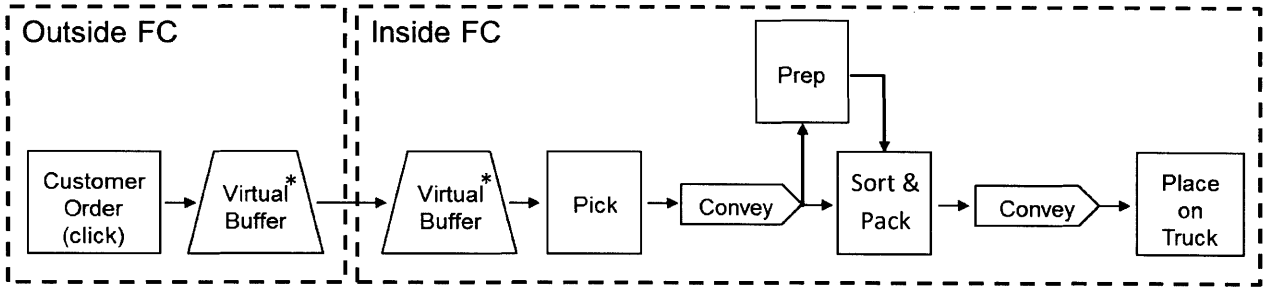


*Virtual Buffer = virtual list of customer orders waiting to be physically fulfilled

Figure 2: Fulfillment process at a standard sortable FC

2.2.2 Special Case Sortable (Site A)

Site A is also a sortable site, set up to fulfill soft-lines items with sizes as mentioned in the previous standard sortable section. In addition, Site A also fulfills jewelry orders that require special "prep" steps prior to packaging. The "prep" work involves placing rings, necklaces, bracelets, earrings, etc. in a traditional jewelry box like those received from a brick and mortar jewelry store. This step, indicated in the augmented process map in Figure 3, is what makes Site A a special-case sortable site.



*Virtual Buffer = virtual list of customer orders waiting to be physically fulfilled

Figure 3: Fulfillment process at Site A showing addition of non-standard prep step

Note that standard sortable sites and special case sortable sites also fulfill customer orders for single items. After the pick process, items follow different fulfillment "paths" through the site depending on the category under which the order falls (i.e. single-item or multi-item). Information on these separate fulfillment paths is presented in further detail in Chapter 4, Section 4.1.1.

2.3 Pick Agreement Time

As stated previously, Amazon takes its customer experience very seriously. The customer facing goal is to ensure that all orders arrive on or before the date promised (and sometimes paid for by a customer) at checkout. Internally, Amazon strives to achieve this goal at minimum company cost, meaning as few shipment upgrades as possible. This is mainly a concern for next-day and two-day shipments. In order to fulfill this promise, Amazon must ensure that each order has time to complete the fulfillment process and make it on to a truck that can actually deliver it to the customer on time. This is accomplished using a value known as the pick agreement time.

Pick agreement time is an algorithm input set individually at each Amazon FC. It corresponds to a rough estimate of the fulfillment cycle-time, or the time it takes for a customer order to move through the fulfillment process detailed in Figure 2 and Figure 3 above. This value is utilized in the algorithm that provides each customer with an option such as the following while browsing:

Figure 4: Example customer promise prior to purchase

As shown in Figure 4, if a customer orders within the specified amount of hours/minutes and pays to upgrade shipping, he or she can receive their items as soon as possible. This only works if Amazon can be certain that a customer can order up until the "order within" calculator reaches 0 hrs 0 mins, and an FC can still fulfill that order and get it on the last truck that can deliver the order on time. Thus, the 0 hrs 0 mins mark must come at a specific time prior to this final truck departure. This specific time is the pick agreement time. It is the time that each FC guarantees it can move an order through the building and onto a truck.

Barring lengthening adjustments to guard against increased holiday volumes at the end of the year, each FC sets pick agreement once and leaves it as a static number for the remainder of the year. There is a network standard value of 1 normalized time unit that most FCs try to set; however, there is no standard way of calculating fulfillment cycle-time, upon which pick agreement time is based. Therefore, Amazon lacks the necessary data to support lowering pick agreement time, even if a site could support it. This thesis addresses this issue.

2.3.1 Site Comparison

Throughout this thesis, three different FCs are compared in terms of pick agreement time and fulfillment cycle-time. These three FCs are as follows:

Site A: Special case sortable site that fulfills soft-lines products, including some jewelry that requires an additional "prep" step.

Site B: Standard sortable site that fulfills soft-lines products.

Site C: Standard sortable site that fulfills typical Amazon products like movies and books.

Currently, Site B and Site C both operate with the network standard pick agreement time. Site A, on the other hand, has a pick agreement that is approximately 1.56 times as long. This shortens the window in which customers can shop for the soft-lines products available at Site A, and limits the volume the site could be processing.

3 Literature Review

This chapter provides a review of the relevant literature that inspired and supported the work performed within this thesis. It covers the steps involved in academically assessing a manufacturing/operations environment from a lean and waste perspective. It also reviews the statistics, principles, and modeling techniques utilized to analyze the data produced. This theory lays the groundwork to properly map the process at Site A, identify issues, and address the problematic areas.

3.1 Lean

The philosophy of lean is built upon a foundation of principles developed and utilized for decades within various manufacturing companies, supermarket delivery and inventory control systems, and even the U.S. military (Alukal & Manos, 2006, p.xiii). It was first packaged into a single system known as the Toyota Production System (TPS) by Eiji Toyoda, Taiichi Ohno, and Shigeo Shingo at Japan's Toyota Motor Company just after World War II. The title "lean manufacturing" became popular when it was introduced to the United States by James P. Womack, Daniel T. Jones, and their group out of MIT (Alukal & Manos, 2006, p.2). At its core, "[l]ean is a manufacturing or management philosophy that shortens the lead time between a customer order and the shipment of the parts or services ordered through the elimination of all forms of waste. Lean helps firms in the reduction of costs, cycle times, and non-value-added activities, thus resulting in a more competitive, agile, and market-responsive company" (Alukal & Manos, 2006, p.1). As Womack and Jones put it, lean gets its name because it allows companies to do "more and more with less and less" (Womack & Jones, 2003, p.9).

Although it began in manufacturing, the lean philosophy can be and has been applied within multiple industries (e.g. finance and healthcare) and throughout multiple departments,

including sales, customer service, accounting, and engineering, among others (Alukal & Manos, 2006, p.xiii). The reach and applicability of lean is wide. This shows that lean is an adaptable approach to help Amazon reduce fulfillment cycle-time and keep up with growing consumer demands.

Amazon competes in a world where fast, on-time delivery is expected. Consumers want the convenience of online shopping and the immediate satisfaction of having their purchases in hand that they get with traditional brick-and-mortar purchasing. This is especially true in the world of fashion (soft-lines), where consumers like to try on purchases and utilize them as soon as possible. Fashion fulfillment also brings along style trends and seasonal changes. Lean provides Amazon with the tools necessary to eliminate wastes, thus allowing them to fulfill orders faster with the same amount of resources, as well as be more reactive to ever changing and last minute demand.

3.1.1 Wastes

As discussed above, lean employs the reduction of wastes (or *muda* in Japanese) to aid in overall process improvement and cycle-time reduction. Lean thinking identifies eight main wastes, and Alukal and Manos describe them as follows (2006, p.3 - 4):

1. *Overproduction.* Making more, earlier, or faster than is required by the next process.
2. *Inventory.* Excess materials or more information than is needed.
3. *Defective product or service.* Product requiring inspection, sorting, scrapping, downgrading, replacement, or repair. This also affects information, if it is not accurate and complete.
4. *Overprocessing.* Extra effort that adds no value to the product (or service) from the customer's point of view.

5. *Waiting*. Idle time for staff, materials, machinery, measurement, and information.
6. *People*. The waste of not fully using people's abilities (mental, creative, skills, experience, and so on).
7. *Motion*. Any movement of people (or tooling/equipment) that does not add value to the product or service.
8. *Transportation*. Transporting information, parts, or materials around the facility.

This thesis focuses mainly on the wastes of overproduction, waiting, and motion, all of which contribute to extensive work in process and long idle wait times.

3.1.2 Tools

Lean is built on a number of tools that allow a user to systematically address non-value-added steps and the root cause of waste in a process. These tools are numerous, and the following subset was utilized, in part, to complete the project work associated with this thesis:

1. "*Standard work*: Consistent performance of a task, according to prescribed methods, without waste and focused on human movement (ergonomics)" (Alukal & Manos, 2006, p.7).
2. "*Batch-size reduction*: The best batch size is one-piece flow, or make one and move one! If one-piece flow is not appropriate, reduce the batch to the smallest size possible" (Alukal & Manos, 2006, p.7).

3.2 Process Flow diagram

A popular way to begin the analysis of complex operational problem is to create a process flow diagram. A process flow diagram is a graphical way to describe a process, utilizing a system of boxes, triangles, and arrows. In this system, boxes represent process activities, triangles represent buffers holding inventory, and arrows represent the route the material (e.g.

consumer goods at Amazon) takes through the process steps (Cachon & Terwiesch, 2013, p.33-36).

When creating a process flow diagram, the first step is to define the process boundaries. These boundaries define the area of main focus. Further analysis only concerns inputs and outputs to the system encompassed by these boundaries. The second step is to define the "flow units, which are the entities flowing through the process" (Cachon & Terwiesch, 2013, p.35). Finally, the boxes, triangles, and arrows are connected in order of the process being evaluated. Figure 5 shows a small example.

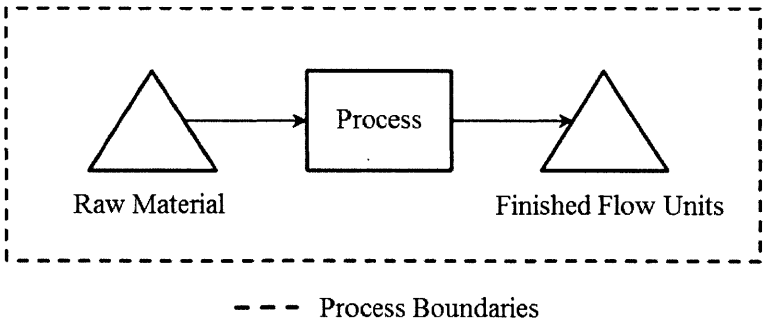


Figure 5: Example process flow diagram

Process activities add value and are required for completion of an order, while buffers/inventories do not add value and a flow unit does not need to spend time in them (Cachon & Terwiesch, 2013, p.35). By helping to identify buffers and representing a complex system in an easily understandable format, process flow diagrams create a useful first step in identifying lean wastes.

3.3 Monte Carlo Method

The Monte Carlo method is a simulation tool used to model complex, real-world problems. It employs the use of random numbers and probability distributions to obtain estimates of solutions to mathematical problems. The random numbers are generated using a roulette-like

machine like those at the casinos in the Monte Carlo Principate, hence the name origin (Zio, 2012, p.1). Today, a number of commercial software packages exist to run complete Monte Carlo simulations, but it is also possible to generate random variables and run simple models using Microsoft Excel (McKee & McKee, 2014, p.47).

A Monte Carlo simulation begins with a formulated model of a system that includes a number of input and output variables, and a series of algorithms. The input variables are defined as random (or stochastic) and conform to a specific probability distribution. The model is run many times to generate a series of random outcomes that could occur given the defined inputs. A computer simulation can be run as many times as necessary to help generate statistically significant results (Thomopoulos, 2013, p.1).

Amazon's FC environment provides the perfect environment to utilize this method of simulation. Humans perform many of the steps of the fulfillment process; thus, a natural randomness occurs in the system. A simple Microsoft Excel based Monte Carlo simulation provides a way to model this randomness and help predict the effects of input changes. The subsections in the remainder of this chapter outline random variable choice, simulation, and statistical methods used specifically for the project work on this thesis.

3.3.1 Random Variables

Monte Carlo simulations use random variables as inputs. These random variables can be one of two classifications: discrete or continuous. Discrete random variables can take on any one of a specified list of values, while continuous random variables can take on any value in a specified interval. The value of these random variables is related with a probability by a mathematical function known as a probability distribution. Some examples include the normal distribution and uniform distribution, among others (Thomopoulos, 2013, p.15).

When the value of an input variable is randomly chosen using a probability distribution, it is called a random variate. One method for choosing a random variate from a probability distribution is called the inverse transform method (Thomopoulos, 2013, p.15). This method is commonly used in Microsoft Excel to generate random variates from both discrete and continuous probability distribution functions. The following subsections highlight the process.

3.3.1.1 Generating a Random Variate from a Discrete Distribution

Thomopoulus describes the process clearly by telling the reader to "...consider a discrete random variable, x_i, where i = 1, 2, ..., with probability distribution $P(x_i)$ for i = 1, 2, ... The cumulative distribution function of x_i is $F(x_i) = P(x \leq x_i)$" (Thomopoulos, 2013, p.16). The algorithm implemented in Excel is as follows:

1. Generate a uniform random variate ($u \sim U(0,1)$) using the RAND()[1] function.
2. From $F(x_i)$, find the minimum i where $u < F(x_i)$.
3. Return x_i, where i is the minimum value found in step 2 (Thomopoulos, 2013, p.16).

3.3.1.2 Generating a Random Variate from a Continuous Distribution

Thomopoulus also describes this process clearly by first telling the reader to "[s]uppose x is a continuous random variable with probability density f(x) for $a \leq x \leq b$. The cumulative distribution function (cdf) of x becomes $F(x) = \int_a^x f(x)dx$ where $0 \leq F(x) \leq 1$" (Thomopoulos, 2013, p.16). The algorithm implemented in Excel is as follows:

1. Generate a uniform random variate ($u \sim U(0,1)$) using the RAND() function.
2. Set F(x) = u.

[1] RAND() = "Returns an evenly distributed random real number greater than or equal to 0 and less than 1" ("RAND function," n.d.).

3. Find the value of x that satisfies F(x) = u using an inverse function ($x = F^{-1}(u)$). The appropriate inverse function ($F^{-1}(u)$) to use depends upon the desired continuous probability distribution. The normal distribution is used for the purpose of this thesis. The proper inverse function to use in this case is NORM.INV[2].
4. Return the value of x (Thomopoulos, 2013, p.16).

3.4 Queueing Systems

Operations management includes a field of study called queueing theory which focuses on queueing systems and how they process entities (e.g. people, orders, and materials). Three elements make up a queueing system: 1) an arrival process, 2) a service process, and 3) a queue. The arrival process describes how entities enter the system of study, and the service process describes how entities exit the system of study. The queue is the line or buffer that entities wait in prior to being serviced, and the entities waiting in the queue are known as work in process (WIP). Queues can grow large, stay in a steady state, or shrink to zero depending on the variability and speed of both the arrival process and the service process (Chhajed & Lowe, 2008, p.53-56). Under steady state conditions, a rule known as Little's Law can be used to better understand certain characteristics of the queueing system. The following subsection looks at this a bit deeper.

3.4.1 Little's Law

Little's Law applies to queueing systems in a steady state condition, where arrival rate and service rate are stable and not changing over time. The law says that the average number of items in a queueing system equals the average rate at which entities enter the system multiplied by the

[2] NORM.INV(p, mu, sigma) = "returns the value x such that, with probability p, a normal random variable with mean mu and standard deviation sigma takes on a value less than or equal to x" ("Excel statistical functions: NORMINV," n.d.).

average time that an entity remains in the system. The variables and equation are as follows (Chhajed & Lowe, 2008, p.82):

L = average number of entities in the queuing system,

W = average waiting time in the system for an entity, and

λ = average number of entities arriving (entering) per unit time

$$L = \lambda * W$$

A useful feature of Little's Law is that by knowing two of the three parameters listed above, the third can be calculated, given that the assumption of steady state holds true (Chhajed & Lowe, 2008, p.84).

4 Problem Analysis

This chapter sets the foundation for the process improvement work discussed in later chapters of this thesis. The introduction of lean first necessitates an understanding of the current state of the process under study. The following walks through a detailed, baseline analysis of the outbound fulfillment process and corresponding cycle-time at Site A, beginning with a process flow diagram and ending with a comparison with like sites. Some initial recommendations and testing are proposed based on this analysis, process wastes are identified, and, finally, some hypotheses on lean implementation and improvements are proposed.

4.1 Baselining the Process

This section details where Site A stands with regards to its current state outbound fulfillment process (time from online order placement to completed package on truck). A comparison is drawn with like sites within the Amazon FC network.

4.1.1 Process Flow Diagram

As described in the literature review, creating a process flow diagram is an important first step in analyzing a complex operational process. The process flow diagram in Figure 6 represents the flow of an order from online purchase, on the far left, through package in a truck, on the far right. This is the outbound fulfillment process, which consists of a virtual component and a physical component. The virtual component takes place in two phases. First, the order is assigned to the appropriate FC. Second, each item in an order is assigned to a specific associate for picking. Now in the physical component, the order goes through a series of physical movements, represented by the process steps, buffers, and routes of Figure 6. In short, this series of steps can be described as the following:

Pick: Process by which an item is located in inventory and moved toward a packing station. This is represented in Figure 6 as the two steps to the left of the Single/Multi split.

Pack: Process by which an item is placed into a box and labelled for shipping to a customer. At Site A, this also includes the additional prep steps performed on specialty jewelry items. This is represented in Figure 6 by all steps and buffers within the dotted rectangles.

Truck Load: Process by which a completed order is placed on a truck for shipping, represented in Figure 6 as the final step on the far right.

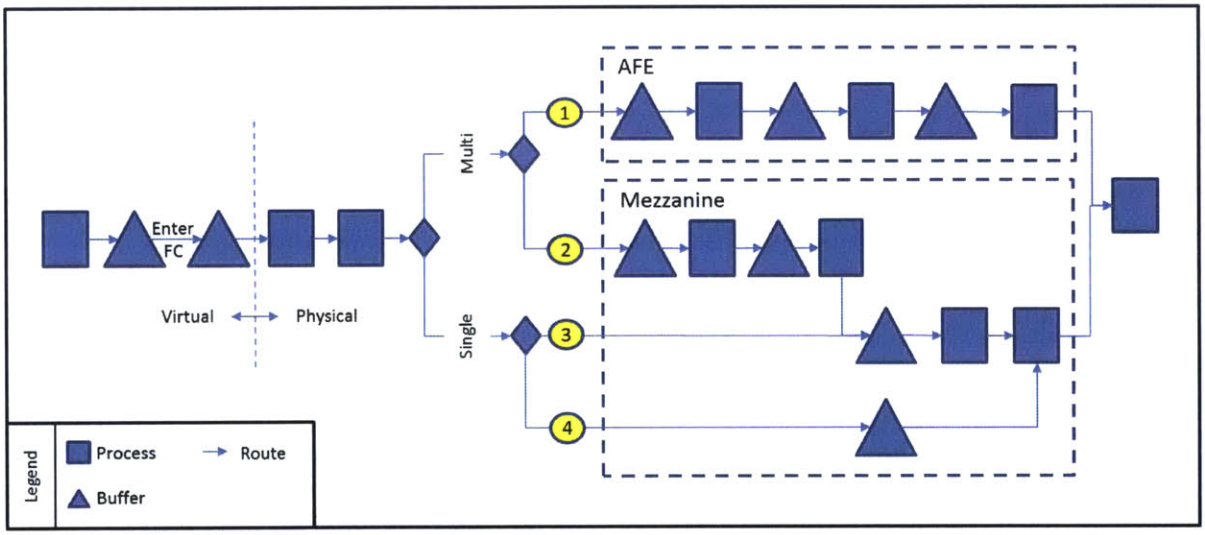


Figure 6: Detailed process flow diagram of the Amazon fulfillment process through Site A

At Site A, customer orders fall into two main categories, single or multi, referring to the number of items in the order. These categories are then subdivided into prep or non-prep, referring to the necessity of an additional jewelry prep step prior to final packaging. Items follow different fulfillment paths through the site depending on the category and subcategory under which the order falls. The process flow diagram (Figure 6) identifies two pack areas and four main fulfillment paths, described as follows:

Pack area 1: Amazon Fulfillment Engine (AFE) – semi-automated process

Fulfillment path 1: Orders of two or more items where no jewelry prep is necessary. Herein referred to as "multis – no prep".

Pack area 2: Mezzanine – fully manual processes

Fulfillment path 2: Orders of two or more items where at least one item is jewelry that requires additional prep work. Herein referred to as "multis – prep".

Fulfillment path 3: Orders of one jewelry item that requires additional prep work. Herein referred to as "singles – prep".

Fulfillment path 4: Orders of one item where no jewelry prep is necessary. Herein referred to as "singles – no prep".

Note: Gift-wrapping also takes place at this site, but is handled in small volumes through a different process. This process is not considered for analysis within this thesis.

The practice of creating a process flow diagram and visualizing the operation in this way helps form a mental model of item flow through a facility such as Site A. Areas of potential concern become observable, and hypotheses about which fulfillment path has the longest cycle-time can be made. For example, fulfillment paths one and two have the most steps; therefore, these can be initially hypothesized to take the longest amount of time. These hypotheses are then easily testable through the collection of time data. At Amazon, a massive data-warehouse stores all kinds of time data for each item moved through an FC. The following section walks through the process of data selection.

4.1.2 Data Selection

Building off of the process flow diagram, an analysis of fulfillment cycle-time can be pieced together by tracking time stamps as an item moves through each process step and buffer. Amazon's data storage system captures and saves time stamps associated with a few key points in the fulfillment process. These critical time stamps include:

Customer order time: Time stamp associated with the moment a customer places an order.

FC assignment time: Time stamp associated with the moment a customer order is virtually transferred into the computer system at the appropriate FC.

Pick ready time: Time stamp associated with the moment a customer order is assigned to an associate to go and pick the item or items from inventory.

Ship label apply time: Time stamp associated with the moment a shipping label is applied to a completed customer order box, just prior to truck loading.

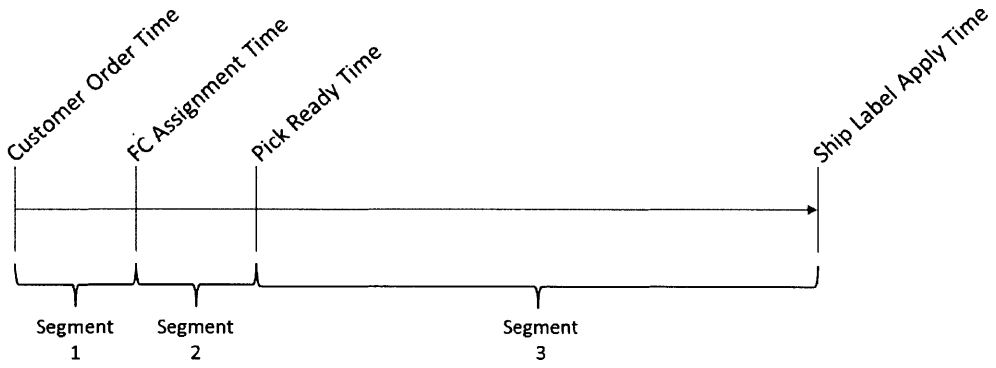


Figure 7: Customer order fulfillment timeline including key time stamps

Figure 7 shows a timeline of the key time stamps captured during the fulfillment of a customer order. A series of data queries written in SQL allows access to years of historical data stored in Amazon's data-warehouse. Quick analysis using Microsoft Excel shows that Segments 1 and 2 (Figure 7) are on the order of seconds as customer orders approach the pick agreement

time cutoff. These segments can be approximated as zero. Therefore, the critical fulfillment cycle-time needed for a lean analysis is defined as Segment 3 in Figure 7, the time interval from Pick Ready Time through Ship Label Apply Time. This cycle-time definition is used for all analyses from this point on in the thesis.

4.1.3 Current State Cycle-time

The process flow diagram and clear definition of cycle-time, developed in the previous subsections, provide a foundation to assess the current state of outbound fulfillment at Site A. With the framework in place, historical data is queried, downloaded, and analyzed. Amazon's systems are set up to provide easy access to thousands of data points. To create a more manageable data set, this analysis only considers orders processed from start to finish within the last few hours[3] prior to their scheduled truck departure (i.e. approximately 4% of all orders processed). These items are the worst case scenario and must move through the facility as fast as possible; thus, the speed at which they move through the Amazon FC depicts the true cycle-time capabilities. Figure 8, on the next page, shows cycle-time frequency plots for each of the four fulfillment paths through Site A. The data represent a typical week worth of fulfillment. Each frequency plot utilizes thousands of observations to create the corresponding distribution.

[3] Last Few Hours – The specific value is left out for proprietary reasons. However, this "last few hours" is the time period prior to each truck departure during which orders due out on that specific truck are prioritized over others at each step of the fulfillment process (i.e. pick and pack).

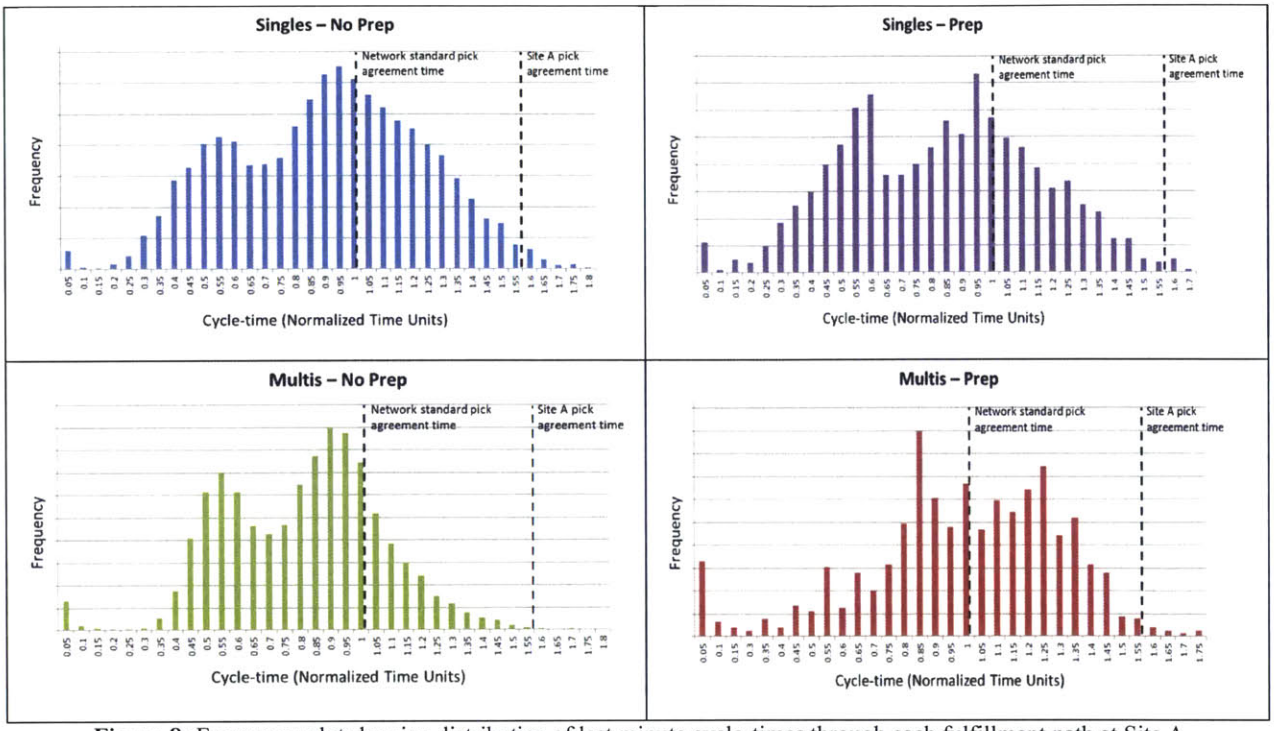


Figure 8: Frequency plot showing distribution of last minute cycle-times through each fulfillment path at Site A

The cycle-times are normalized using the network standard value for pick agreement time; therefore, a value of 1 represents a cycle-time that corresponds to the network standard pick agreement time. Study of each of the four graphs reveals some interesting observations. First, a large proportion of orders move through Site A slower than the network standard. In fact, up to 50% of the orders move slower through the Multis-Prep fulfillment process than the network standard target. Second, a noticeable percentage of orders move though Site A in almost no time at all. A quick look into this data reveals that these orders went through an additional problem solve process. Each critical time stamp gets reset to approximately the same value at the point of problem solve and stored in the Amazon data-warehouse, thus the improbable cycle-time calculations. No further attention is given to these data points. Third, Site A has its pick agreement time set at a point where only a small percentage of the last minute orders take longer to process. These longer cycle-times are seen as the data to the right of the far right dotted line in Figure 8.

One final observation can be made concerning the shape of the distribution. The data points trace out a double peak. A double peaked, or bimodal, distribution often indicates data collected from two processes with two different distributions ("Typical Histogram Shapes and What They Mean - ASQ," n.d.). Attempts were made to sort the data by shift, time of day, and day of week with no success. Further study is appropriate and will be discussed further in the next steps of this thesis. For the baseline analysis and recommendations herein, no specific distribution is assumed, and the data is analyzed empirically.

4.1.4 Comparison to Like Sites

In a similar fashion as for Site A, the process flow diagram and cycle-time definition can be used to analyze outbound operations at other sites within the Amazon network. The method of cycle-time baseline analysis described throughout this chapter applied to Site B and Site C reveals the results shown in Figure 9, on the next page. As a quick reminder, Site B fulfills soft-line products but does not prep any jewelry, while Site C fulfills standard items other than soft-lines, such as movies and books. Also, without the necessity for prep, both Site B and Site C have two fulfillment paths (Singles and Multis) rather than the four described for Site A.

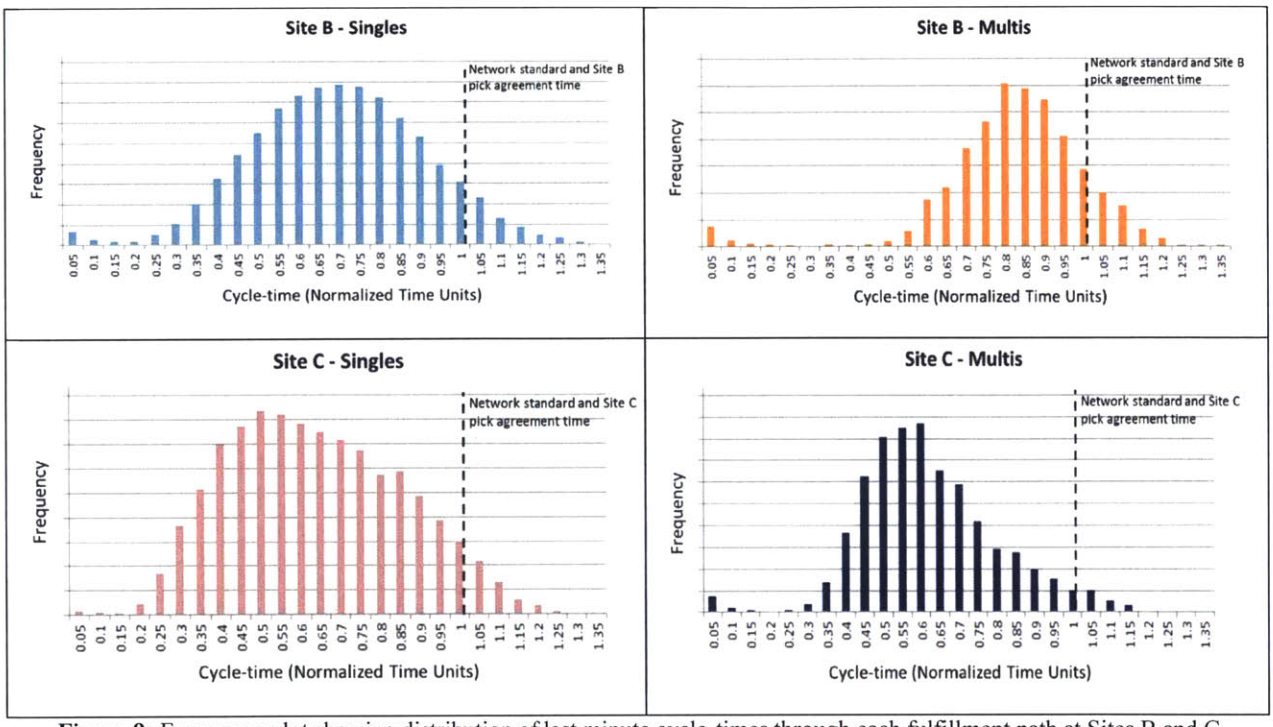


Figure 9: Frequency plot showing distribution of last minute cycle-times through each fulfillment path at Sites B and C

Again, a value of 1 normalized time unit in Figure 9 corresponds to the network standard pick agreement time. Study of each of the four graphs reveals that, like Site A, items that go through a problem solve step show up as a very small cycle-time (bar on the far left of each graph). On the other hand, both Site B and Site C have shorter fulfillment cycle-times than Site A, and the cycle-time distribution is smoother. Also, both of these sites have a pick agreement time set at the network standard value; however, this value is set less conservatively than at Site A. Table 1, on the next page, shows the percentage of orders that take longer to process than the pick agreement time setting. These numbers represent the data that falls to the right of the far right dotted black lines in Figure 8 ad Figure 9.

Site	Fulfillment Path	Percentage of orders with cycle-times longer than the pick agreement time setting
Site A	Singles – No Prep	1.24%
	Singles – Prep	0.62%
	Multis – No Prep	0.20%
	Multis – Prep	1.01%
Site B	Singles	6.22%
	Multis	10.41%
Site C	Singles	4.78%
	Multis	3.00%

Table 1: Percentage of orders that take longer to fulfill than the set pick agreement time

4.2 Initial Results/Recommendations

The current state baseline shows that fulfillment cycle-time at Site A is, as assumed, slower than fulfillment cycle-time at Site B and Site C. This means that, on average, Site A takes a longer amount of time to move an item from its place in inventory to a shipping box and onto a truck; therefore, Site A is justified in setting the pick agreement time higher than the network standard. However, Site A does not need to set pick agreement time as high as it currently does, even without making any changes to its current process setup. This is hinted at in the data depicted in Table 1. A different look at the cycle-time data shows why this is.

Figure 9, above, is a frequency plot showing the distribution of cycle-times for all orders that get assigned to Sites B and C within the last 1.35 normalized time units prior to a truck departure. Some items take the entire 1.35 normalized time units to make it through the building, but the majority take less time than the network standard of 1 normalized time unit. Figure 10 plots this same data as a cumulative percentage, from 0 to 100%. In other words, it shows what

percentage of orders is processed under each possible cycle-time from 0 to 1.35 normalized time units. The multi orders at Site B get fulfilled in the slowest manner.

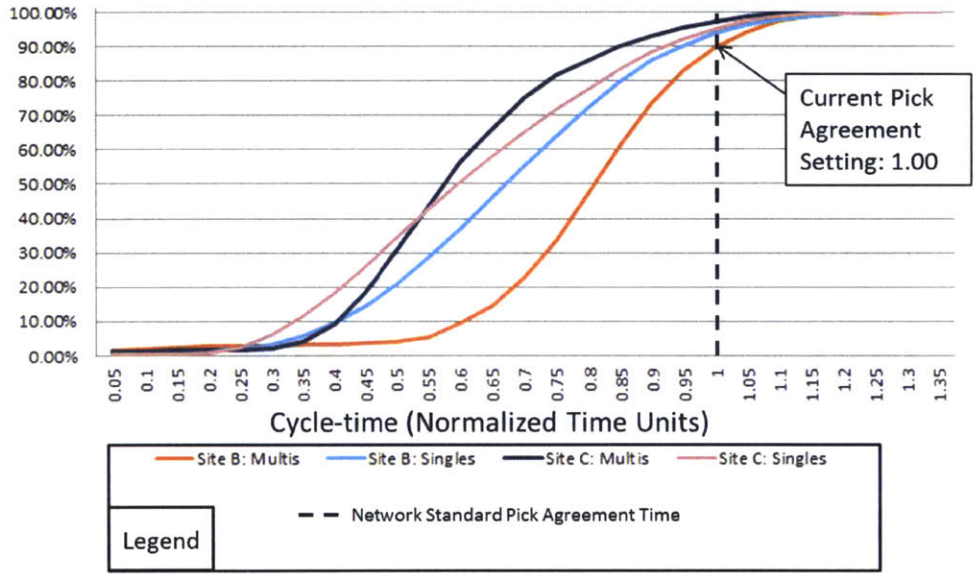


Figure 10: Cumulative percentage plot of cycle-times over a week at Site B and C

The cumulative percentage line associated with the multis fulfillment path at Site B crosses the setting for pick agreement time at the 90% mark. This means that, of all the orders that get assigned to Site B within 1.35 normalized time units of their scheduled truck departure, 90% of those orders are processed in one normalized time unit or less. Only 10% necessitate up to an additional 0.35 normalized time units to complete the outbound fulfillment process. Amazon FCs are well equipped to prioritize and properly handle this small volume; therefore, this pick agreement time setting does not pose a problem for Site B, as the operations team is able to comfortably fulfill all the orders they receive, under standard conditions. Extending this logic, Site A could set its pick agreement time to the point where its slowest fulfillment path reaches the 90% mark. Figure 11 plots Site A's data from Figure 8 as a cumulative percentage.

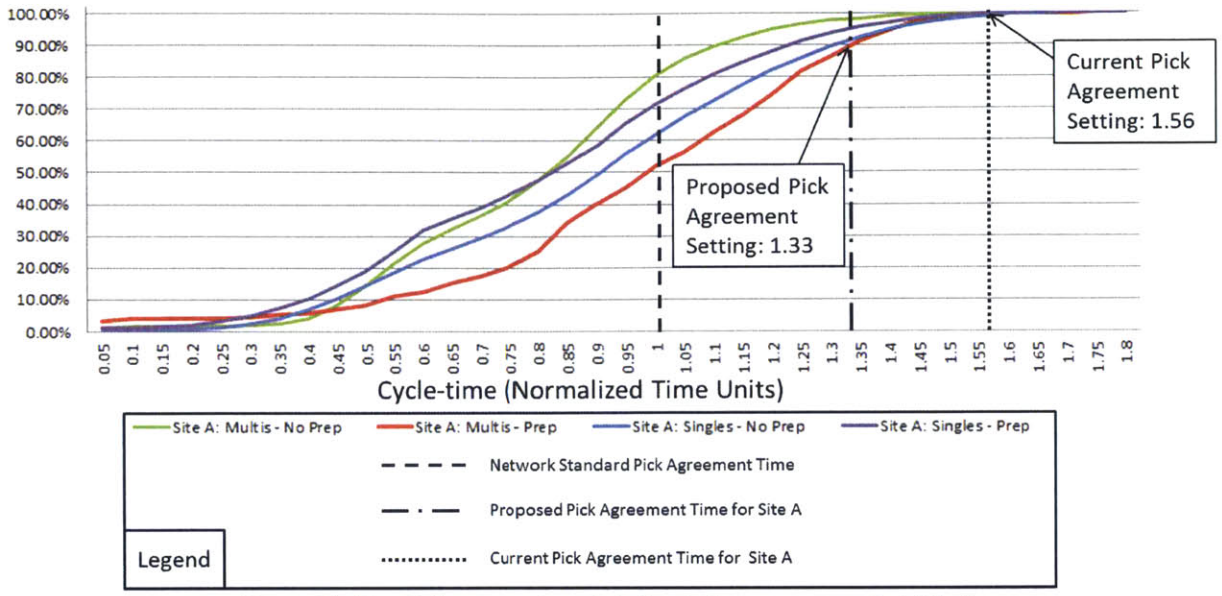


Figure 11: Cumulative percentage plot of cycle-times over a week at Site A

Site A's worst fulfillment path is the multi – prep path, and it reaches the 90% complete mark at a cycle-time of 1.33 normalized time units. A live test was run for a week with Site A's pick agreement time set to this value of 1.33 normalized time units, and Site A fulfilled all orders without issue (i.e. no unusual prioritization, no increase in shipment upgrades, and no missed truck departures), lending support to this analysis.

Thus, Site A can withstand a reduction in pick agreement time by up to 0.23 normalized time units (or 15%) without any changes to its process.

However, Site A is still not close to having a fulfillment cycle-time that supports reducing pick agreement time to the network standard. The fulfillment process must be improved for this to happen. The next subsection of this chapter identifies areas for improvement, and the remainder of the thesis walks through the work associated with developing, testing, and implementing these improvements.

4.3 Opportunity Identification

The baseline analysis detailed throughout this chapter confirms that soft-lines fulfillment at Site A is slower than fulfillment at other sites within the Amazon FC network. Fulfillment cycle-time must be reduced before Site A can set its pick agreement time in line with the network standard. This thesis addresses this cycle-time issue through the application of lean methods and operations management theory.

Looking through this lens, the main causes of lengthy cycle-time are system wastes. Study of the fulfillment process at Site A and a comparison with the data shown in Figure 6, Figure 8, and Figure 11 identifies the following observations and opportunities for waste elimination:

1. The process of picking is not standardized.
2. The singles – no prep fulfillment path has the fewest process steps and yet one of the longest cycle-times.
3. Items that move through the longest fulfillment path (multis – prep) are grouped into large batches.
4. The multis – prep fulfillment path contains a large amount of buffer waiting time during which zero value is added.

Chapter 5 goes into full detail on how these four observations were addressed. A complete identification of problem, observations, hypothesis, testing, and results is provided in detail. The project work uses applications of standard work, Little's Law, and batch size reduction to address three lean wastes of overproduction, waiting, and motion. Modeling techniques and analysis is provided as needed.

5 Putting Lean and Operations Management Theory into Practice

The work contained in this chapter consists of three cases, each with a different problem and different method of analysis. Case 1 addresses the issue of a non-standardized pick process and presents a method of improvement. Case 2 addresses why the shortest fulfillment path has one of the longest cycle-times and presents a method of correction and control. And finally, Case 3 addresses the batch size and order of operations issues in the multis – prep process path and presents a few options for improvement. Case 1 includes process modeling, hypothesis testing, and full analysis, Case 2 includes some simple modeling and analysis, and Case 3 walks through a conceptual analysis of the problem and possible solutions.

5.1 Case 1: The Pick Process

Current Process

The first step in the physical fulfillment process is referred to as pick. Pick involves an associate walking to the location where an inventory item is located, physically selecting the item from storage, and placing that item into a plastic carrying tote. The associate then moves on to another item, and then another item, until the tote is full. Associates are equipped with a cart that can hold up to three of these plastic totes. When a tote is full, the associate can switch to a second or third tote and continue picking items, or the associate can return to a conveyor and drop the tote for transportation to the downstream pack process. Associates are also given the freedom to drop an incomplete tote on a conveyor if the tote is "sufficiently" full and the associate is close to a conveyor.

Problem

There is no standard to how often associates drop totes on a conveyor. There are five unofficial protocols being followed: 1) pick into one tote and drop when full, 2) pick into two totes and drop when both are full, 3) pick into three totes and drop when all are full, 4) drop a

tote whenever close to a conveyor regardless of fill, and 5) switch between the four previous options throughout a shift. This introduces a large amount of variation to the pick process step, and it also introduces a point of unnecessary *waiting* waste. Any amount of time that an item spends in a tote on a pick cart is time that the item is not moving toward a customer, thus it is non-value-added time. Due to the current layout of the FC, a certain amount of wait time is inevitable, as it would be impractical to have an associate retrieve one item at a time and return to the conveyor. However, there is no need for a completed tote to sit full of items on the bottom of a pick cart while items are picked into one or two more totes. The conveyors are located in a manner that an associate conceivably will pass by them every few picks. The time a tote sits on the bottom of a cart is pure wasted time spent *waiting*, and this time adds minutes to the fulfillment path cycle-time. But how big of a problem is this at the moment?

The first step in analyzing the size of a problem is to create a way to measure the issue in question. In this case, the issue is item wait time in a pick tote. Due to data collection limitations, item wait time cannot be measured for each individual item. Wait time is, instead, measured for each tote. Therefore, wait time is defined as the time interval from the moment the last item is placed into a pick tote until the moment a tote arrives downstream at packing.

$$Wait\ Time = Time_{arrive\ downstream} - Time_{last\ item\ in\ tote}$$

Figure 12 depicts a distribution of wait times associated with the non-standard process of picking into one, two, three, or a combination of totes. This data is empirical, collected via direct database query for all pickers over the course of one shift.

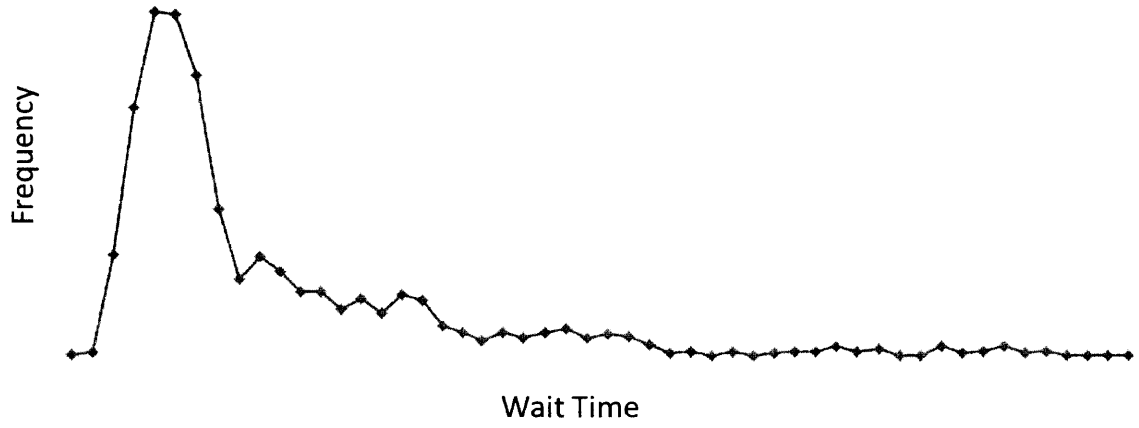


Figure 12: Distribution of tote wait times

Efforts to reduce the depicted wait time and variation involve a twofold process: 1) modelling the pick system for basic understanding, and 2) hypothesis testing based on this model. The next sections go into further detail.

Modelling

Model development helps create an understanding of the basic mechanics driving the distribution of wait times seen in Figure 12. The created model uses basic mathematical techniques to relate the system inputs with the output. In this way, changes to the system output can be observed theoretically by manipulating inputs to the model prior to testing in the field. The use of a computer allows for a simple Monte Carlo simulation to run thousands of data outputs. Model building begins with a definition of the inputs and output to the system under study.

Inputs:

Pick Rate: Defined as the number of items an associate can pick over the period of an hour. This value helps determine how much time it takes to completely fill a tote.

Number of Items in a Tote: Defined as the number of items picked into a tote before considered "full" by the associate. This value helps determine how much time it takes to completely fill a tote.

Associate Travel Time: Defined as the time it takes an associate to carry completed totes to a conveyor after the final item has been picked.

Conveyor Travel Time: Defined as the time it takes for a tote to make it from the point at which the associate dropped it off to the downstream pack station.

Number of Totes Used: Defined as the number of totes used by an associate before they return to a conveyor and drop them on the line.

Output:

Wait Time: Defined previously as the time interval from the moment the last item is placed into a pick tote until the moment a tote arrives downstream at packing.

Each input variable takes on a random value from a given distribution. Each of the inputs is studied for the appropriate distribution to approximate the stochastic process. Due to data collection limitations, Associate Travel Time and Conveyor Travel Time are not able to be measured separately; therefore, both inputs are captured by Conveyor Travel Time. Empirical data analysis determines the following:

Pick Rate: $PR \sim N(\mu, \sigma^2)$

Number of Items in Tote: $NI \sim N(\mu, \sigma^2)$

Conveyor Travel Time: $T_{Conv} \sim N(\mu, \sigma^2)$

Number of Totes Used: NT with range (1, 2, 3) and probabilities $p(1) = P_1$, $p(2) = P_2$, and $p(3) = P_3$

There are three different outcomes based on the number of totes (NT) chosen to pick into. In the first outcome (NT = 1), wait time will always equal the conveyor travel time. In the second outcome (NT = 2), wait time for the first tote will equal the time it takes to completely fill the second tote plus the conveyor travel time, and wait time for the second tote will equal just the conveyor travel time. In the third outcome (NT = 3), wait time for the first tote will equal the time it takes to completely fill the second and third totes plus the conveyor travel time, wait time for the second tote will equal the time it takes to completely fill the third tote plus the conveyor travel time, and wait time for the third tote will equal just the conveyor travel time. Thus, the model output variable is defined as follows:

if NT = 1

Tote 1: $Wait\ Time = T_{Conv}$

if NT = 2

Tote 1: $Wait\ Time = T_{Conv} + \left(\frac{1}{PR_2} * NI_2\right)$

Tote 2: $Wait\ Time = T_{Conv}$

if NT = 3

Tote 1: $Wait\ Time = T_{Conv} + \left(\frac{1}{PR_2} * NI_2\right) + \left(\frac{1}{PR_3} * NI_3\right)$

Tote 2: $Wait\ Time = T_{Conv} + \left(\frac{1}{PR_3} * NI_3\right)$

Tote 3: $Wait\ Time = T_{Conv}$

The subscripts in the above equations represent the tote number to which the particular variable refers. The Monte Carlo simulation is run in Microsoft Excel, utilizing the built in random number generator (RAND) and inverse transform method (NORM.INV) to choose numbers for the random variables in the above equations, as described in the Literature Review section titled 3.3 Monte Carlo Method (p.22). The exact proportion of associates who pick into one tote, versus two totes, versus three totes is unknown; however, informal discussions and observations point to an approximate split of $p(1) = 0.25$, $p(2) = 0.5$, and $p(3) = 0.25$. The results of three simulations[4] with these parameters and 1000 trials each are seen in Figure 13.

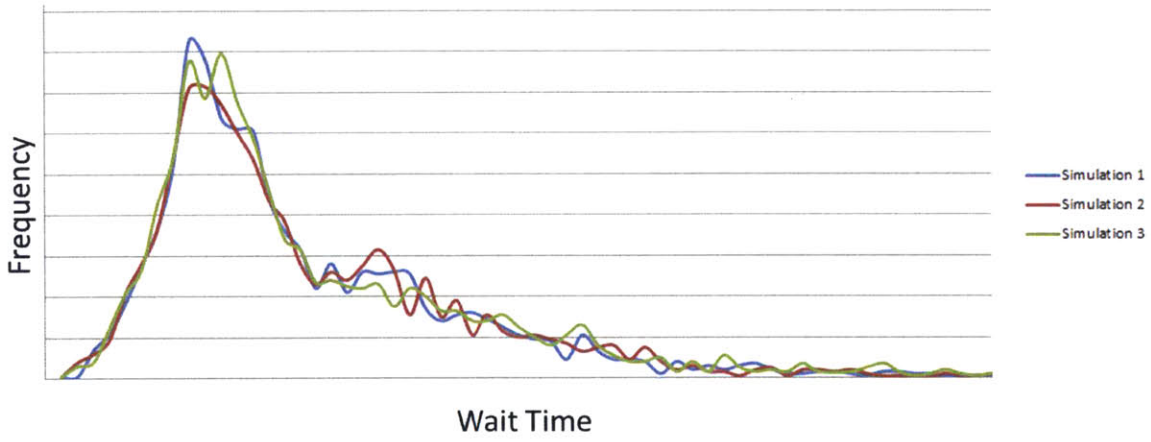


Figure 13: Monte Carlo simulation output for model of current state tote wait time

The values associated with the simulation output are left out for proprietary reasons, but the shape of the modeled distribution matches the original data in Figure 12 well enough for hypothesis formulation and testing purposes. Time that picked items sit waiting, in a tote, on the bottom of a pick cart is pure wasted time. From a lean standpoint, no value is being added. Also, these items are effectively being batched together prior to moving on to the next step of pack. Lean tools dictate that batch size should be reduced as small as possible if single-piece-flow is not possible. Reducing the number of totes picked into prior to returning to a conveyor

[4] Three replicates of the simulation were run to verify consistency and stability of the model results.

effectively lowers the batch size. Therefore, the analysis thus far on the pick process informs the following hypothesis:

> **Hypothesis**: Reducing and standardizing the number of totes that associates can pick into prior to returning to a conveyor will reduce average wait time for totes, thus reducing average overall cycle-time.

Two possible scenarios for standardization and reduction include 1) all associates pick into two totes, or 2) all associates pick into one tote. The developed model predicts the following trend in average wait time and corresponding standard deviation for both of these scenarios:

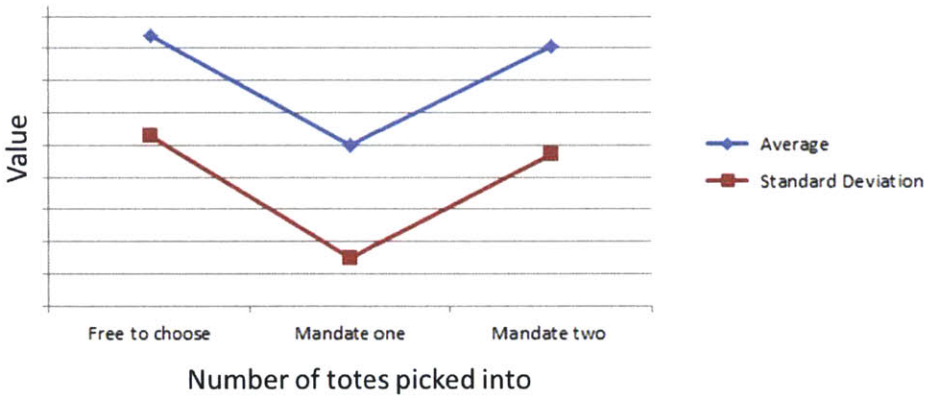


Figure 14: Predicted trend in average wait time and standard deviation based on simulation model

Again, Figure 14 eliminates values for proprietary reasons. The trends predict that having everyone pick into two totes may reduce wait time a small amount, but picking into one tote could reduce wait time average and standard deviation by up to 40% and 72%, respectively. This large reduction makes sense based on the way in which the model is defined above. Picking solely into one tote would create a wait time distribution equivalent to the conveyor travel time distribution used as an input. This is in fact reflected appropriately in the Figure 14 results. These predictions can now be compared with real numbers through a controlled experiment at Site A.

Testing

Designed experimentation allows for testing of the model predictions and consequent hypothesis under true FC operating conditions. Designing the appropriate experiment involves controlling for confounding variables (Foster, Stine, & Waterman, 1998, p. 171 - 184) that are not under study, such as time of day, day of the week, and specific associates. These variables pose a threat to unintentionally bias the results of the testing. The nature of the fulfillment process at Site A makes a completely controlled experiment near impossible, but spreading data collection out over several days and randomly choosing participants helps avoid some bias.

An experiment to test the proposed hypothesis was designed and executed in October, 2014. This experiment involved 28 randomly chosen associates picking for an entire shift on two separate days. Prior to testing, pick data was collected for each of the 28 associates to provide baseline information to compare test results against. On the first day of testing, fourteen associates were told to pick into one tote for the entire shift, while the remaining fourteen were told to pick into two totes for the entire shift. On the second day, the two groups switched instructions (i.e. picked into the number of totes that the other group was picking into on the day prior). See Appendix A for full protocol. The results of the experiment are shown in Figure 15.

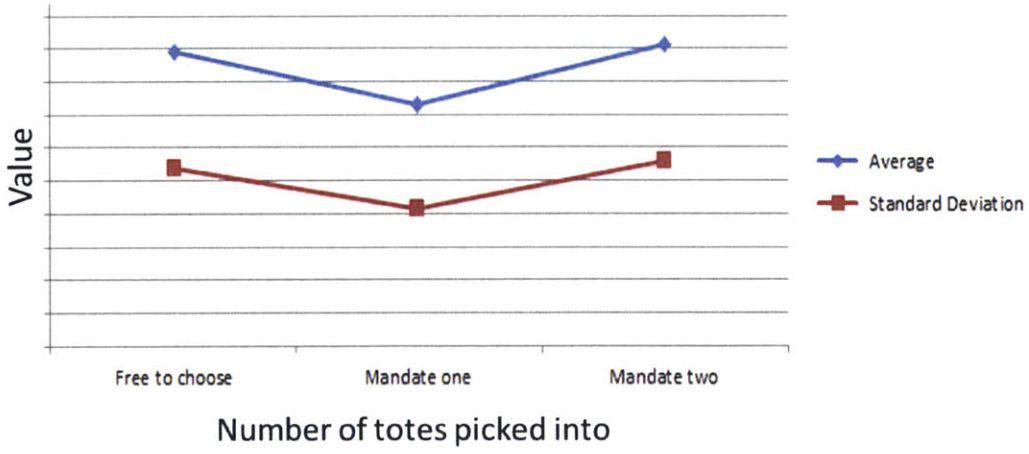


Figure 15: Testing outputs for average wait time and standard deviation

The data shown in Figure 15 represent the average and standard deviation of tote wait time associated with the experiment described above. As seen, the testing results match the modeled results in basic trend. Mandating pick into two totes causes very little change in wait time, while mandating pick into one tote makes a significant reduction in wait time. The model is off in magnitude of change, however. The test results show a slight increase in wait time for two tote picking, and single tote picking reduces wait time average by 18% and standard deviation by 23%, rather than 40% and 72% as predicted.

Discussion of Model and Test Results

The test data supports the general trend put forth by the model results, but the outcomes are different in two main ways. This section explores some possible reasoning behind these differences.

First, there is a slight increase rather than decrease in wait time when associates were asked to pick into two totes. The discrepancy between model and test regarding a mandate of two totes is not very large. Both show a small change, one in a positive direction, and the other in a negative direction. Also, it is possible that the model used incorrect estimates for the probability of picking into one, two, or three totes. If more people than predicted already picked into one tote versus three, asking associates to pick into two may increase the number of totes waiting on the bottom of pick carts. This could explain an increase in wait time as seen in the test data.

Second, the reduction in wait time seen during testing is not as large in magnitude as predicted by the model. The model assumes normally distributed data for most of the inputs. Most importantly, it assumes normally distributed data for conveyor travel time. As noted above, the model converges on the conveyor travel time as a prediction of tote wait time when picking into one tote. Although this normality assumption fits the data fairly well, the real distribution is

not this perfect all of the time. There are a number of reasons that conveyors get backed up and run slower than usual. Associates are also able to take breaks and extend the wait time of some totes, which is not taken into consideration in the model. The model is a tool for predicting trends, but, due to the discussed limitations, it does not predict the exact change in output magnitude based on input manipulation.

Precise numbers are difficult to predict without a tightly controlled experiment and a perfect model of the system. But, in general, both model and testing confirm the hypothesis that implementation of lean principles through reduction in batch size (one tote) and process standardization does reduce wait time and overall cycle-time. To put this into larger context, however, the reduction in tote wait time is on the order of minutes while the overall cycle-time is on the order of hours. Although beneficial, larger reductions in cycle-time are necessary in order to truly help reduce pick agreement time. The real benefit of single tote picking and standardization may actually be in smoothing of the downstream processes.

The smoothing effects due to single tote picking have a lot to do with the manner in which associates return totes to the conveyor system after they are full of items. If it is assumed that each associate picks at an average rate and the number of items per tote remains close to an average value[5], then filling a single tote takes approximately the same amount of time for each associate. Therefore, if the decision to pick into two or three totes is removed, all items should move from the inventory shelves to the conveyor in a predictable, repeatable amount of time. With every associate picking into one tote and immediately dropping it onto a conveyor once full, all totes destined for the downstream pack process would begin to arrive at the conveyor at

[5] Both pick rate and number of items per tote were shown to be normally distributed about an average value earlier in this section. This assumption of all associates operating at an average value is appropriate for this hypothetical future-state analysis.

consistent intervals, thus arriving at pack at consistent intervals (i.e. singularly, rather than in batches of two or three). This eliminates the ebb and flow of tote arrival at the downstream pack process associated with the unpredictable nature of picking into multiple totes at one time. The downstream pack process currently holds a number of totes in a buffer to prevent process starvation due to the ebb and flow of totes. Smoothing of this process thus could reduce the amount of buffer needed. A complete, in-depth analysis of this smoothing effect is not considered in this thesis, but the next steps section revisits the idea for future work.

5.2 Case 2: The Singles – No Prep Fulfillment Path

Current Process

Site A has four main fulfillment paths depending on the item type and item quantity of a customer order. Single-item, non-jewelry orders follow the fulfillment path with the fewest number of physical steps (see path 4 in Figure 6). After these single-item orders are picked from inventory and placed into totes, each tote is placed onto a conveyor that carries it directly into one of seven queue lines in the packing area specifically set up for single-item orders. These queue lines are approximately 100 feet long and each contains over twenty pack stations, situated on either side of the line as shown in Figure 16. Pack associates are assigned to a specific pack station, based on need, where they pull totes off of the conveyor and place each single-order into a box for shipping. Each queue line and every pack station on either side of the queue line are available for pack associates to be stationed, and the Amazon operations team must decide each day how many open lines and manned pack stations are necessary to meet the day's throughput target. Under standard (off-peak) conditions, Site A runs with four of the seven queue lines open and three left empty for additional capacity. Amazon's main concern in staffing these open lines is to minimize indirect support roles, thus associates are placed up the "insides" of the four lines as seen in Figure 16. In this way, only two indirect laborers are needed to replenish pack stations

with boxes and respond to other issues. It is believed to be inefficient for one indirect laborer to manage two sides of the same line as this would necessitate walking completely around the 100 foot line and back.

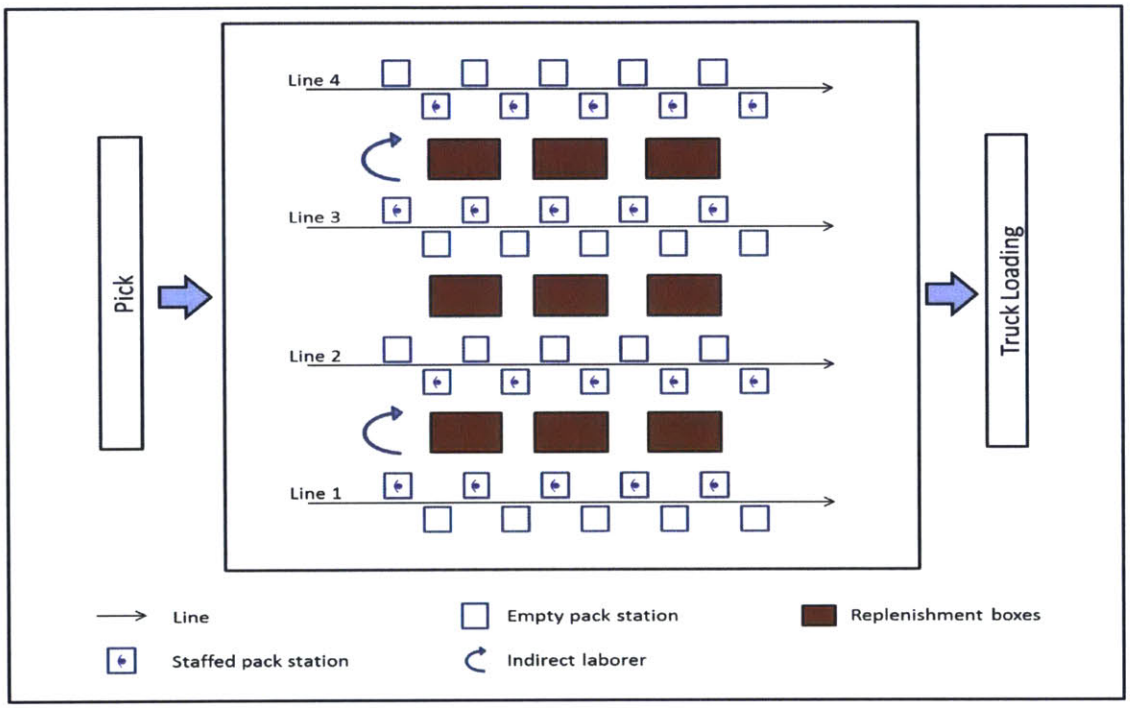


Figure 16: Current state staffing of four pack lanes

Amazon likes to keep work available for all pack associates 100% of the time. This means that a tote must be available for an associate immediately upon completing the pack out of the prior tote. To ensure this is the case, Site A targets filling each open lane with completed[6] totes, as seen in Figure 17, by balancing pick and pack rates. Each item sitting in a tote is known as work-in-process (WIP). After picking and conveyance time, these WIP items now sit, waiting in queue to be packed out by the next available associate. It is also important to note that, under normal conditions, associates fill approximately 1/3 of the available pack stations in a queue line,

[6] Completed – This means that the tote is full of multiple single-item orders, not just one item per tote.

as demonstrated in Figure 17. The WIP target of completely filling a line with totes remains constant regardless of the number of filled pack stations.

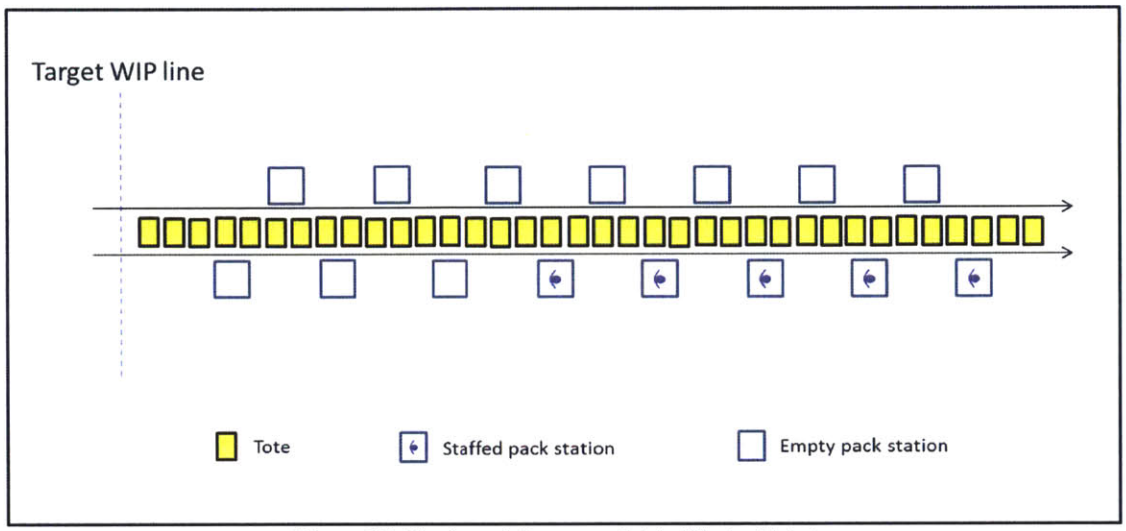


Figure 17: Single pack lane completely filled with totes

Although the WIP level tends to stay fairly consistent, natural fluctuations do occur as pick and pack rates shift throughout the day. Each queue line is equipped with a photo-eye at the end closest to the conveyance and furthest from the truck. If the totes in queue stack up beyond the photo-eye, the system stops sending totes to that line and recirculates[7] them until the queue subsides. This recirculation can cause conveyance overloading; therefore, indirect laborers are instructed to keep totes from reaching the photo-eye. Indirect laborers accomplish this task by removing totes from the queue line and stacking them in a designated buffer. The totes are then returned to a line when the queue returns to normal. By targeting a WIP level that fills out each queue line, this process of removing and replacing totes occurs regularly throughout a shift.

Wait time in queue is defined as the time period beginning with the arrival of a tote at a singles pack queue and ending when an associate grabs that tote to begin packing the items

[7] Recirculate – The outbound conveyance system at an Amazon FC is one large closed loop that connects the inventory shelves with the pack queue lines. If a tote does not get diverted to a pack queue line due to large queue length, the tote is sent on a loop about the facility until the tote queue reduces. The tote is then diverted onto the appropriate pack queue line upon its next pass.

contained within. This wait time does not include time in recirculation, but it does include any time spent sitting on the floor after being removed from the line by an indirect laborer. Figure 18 shows the distribution of wait times in queue at singles pack over the period of a typical day at Site A. Wait time is shown in normalized time units, which, as described previously, is a transformation of the data by dividing each data point by the network standard pick agreement time value. The distribution of data shows that a large number of totes are processed through the queue relatively quickly; however, there is a long tail of wait times extending up to 1.3 normalized time units. This means that some customer orders spend the majority of their total fulfillment cycle-time waiting in queue to be processed at pack.

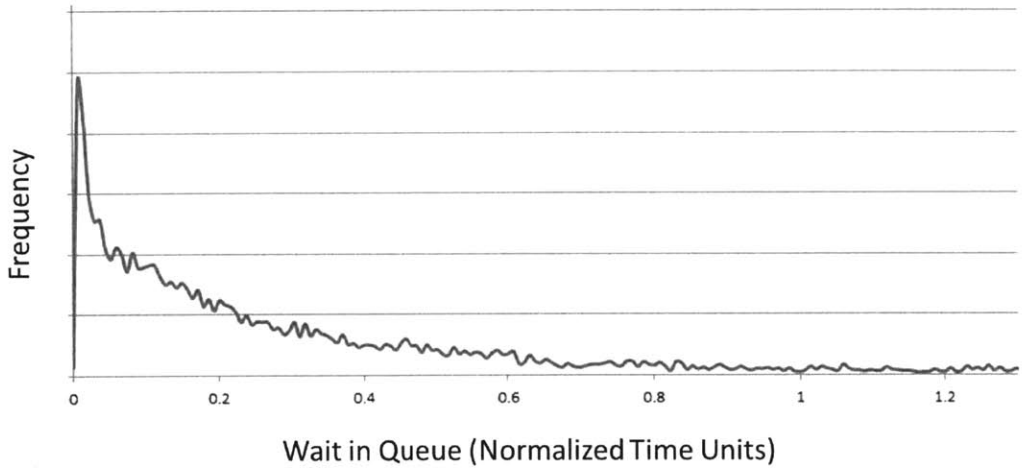


Figure 18: Distribution of wait time in queue at singles pack lines over a typical day at Site A

The Problem

The singles – no prep fulfillment path, with the fewest number of steps at Site A, should take the least amount of time to move an item from virtual order to completed customer package; however, Figure 8 and Figure 11 both show this not to be the case. In fact, this fulfillment path is one of the slowest in the building. Customer orders need to move through this path in an optimal manner in order to reduce fulfillment cycle-time. Based on the description of the current state

process above, three lean wastes are identified: 1) *overproduction* (large amounts of WIP), *waiting* in queue, and *motion* (additional non-value-added touches).

The current staffing model and WIP level targets play a large part in creating these lean wastes and ultimately increasing wait time in queue. Currently, staffing on the insides of four pack lines necessitates four completely full queues. This model minimizes indirect labor, but does not consider how cycle-time is affected. There are two main issues contributing to lengthy cycle-time. First, too many lines are open for the number of associates being utilized. And second, the WIP target is too high and should be adjusted based on the number of associates working in pack rather than the number of open pack lines.

Proposed Solutions

Addressing the first issue, the number of open pack lines can be changed from four to three. This new solution has the same number of associates working in pack, and, thus the same throughput. The main difference is that the second of three lanes requires double the amount of associates as the other two (see Figure 19 on the next page). This solution eliminates the need for an entire queue line, and immediately lowers the required WIP level. It also begins to lower the amount of *overproduction* waste needed to support four open lanes full of WIP.

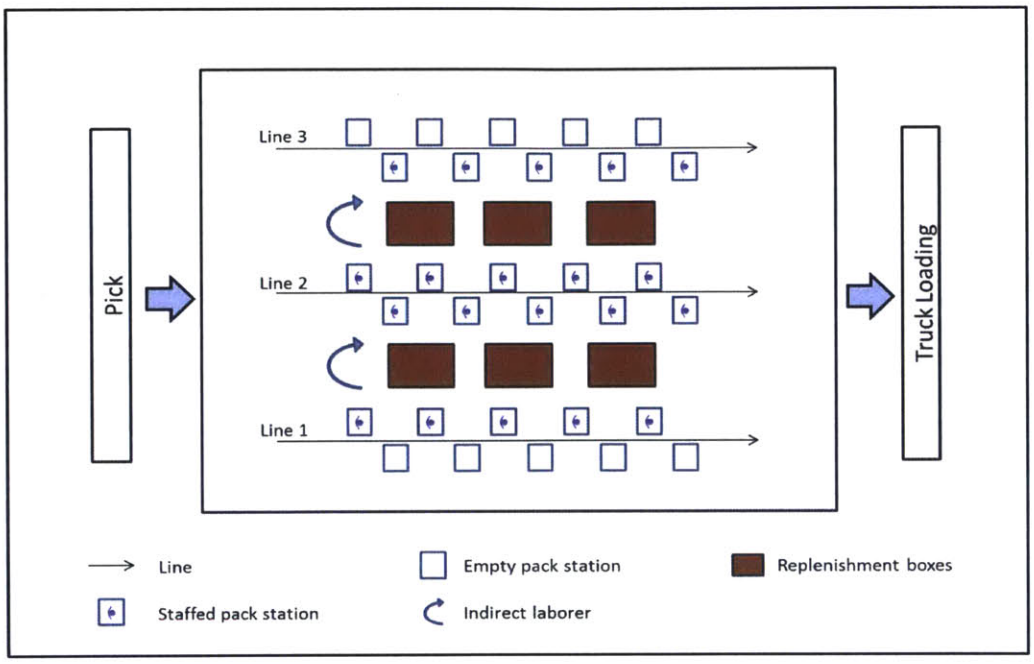


Figure 19: Proposed staffing of three lines

Addressing the second issue, the target WIP level can be lowered for each open pack line. The current state process aims to fill each line completely with totes regardless of the number of associates working on that line. The proposed process aims, instead, to fill each pack line up with totes so that the WIP target line is just before the first station occupied by an associate (see Figure 20 on the next page). There is still an appropriate amount of buffer to guard against any brief reduction in WIP level due to a fluctuation in pick versus pack rate. This ensures that associates still have work present at all time. Lowering the WIP target also has the benefit of guarding against tote buildup beyond the photo-eye. Slight increases in WIP are less likely to require the indirect process of removing totes from a pack line, thus reducing the occurrence of waste *motion* described above.

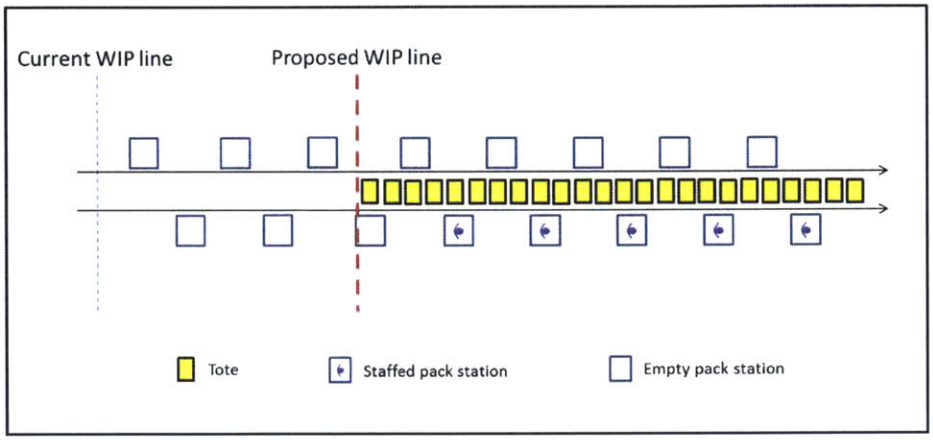


Figure 20: Singles pack lane showing proposed WIP target versus current state

Both solutions eliminate the waste of *waiting*. The next section uses an application of Little's Law to describe just how much wait time may be reduced.

Little's Law Analysis

A theoretical analysis of wait time in queue is possible using an application of queueing theory known as Little's Law. Little's Law equates WIP level (L) to wait time in queue (W) multiplied by arrival rate (λ). Solving for wait time in queue gives the following formula:

$$W = \frac{L}{\lambda}$$

To apply Little's Law to the singles – no prep pack system, a number of assumptions must be made. These assumptions include:

1. The system is stable in that the average arrival rate and service rate are equal and not changing over time.
2. Since only the service rate is available for measure, the service rate can be used as a proxy for arrival rate (λ) under the first assumption.

3. WIP level (L) is approximately held constant at the target level. We assume that the system is highly utilized and that it is able to keep the WIP in each line at its target level. Thus, the average WIP is approximately equal to the target.
4. The average number of items in a tote does not vary much over the course of a day. This assures that the elimination of each tote worth of WIP has the same reductive effect on total WIP level of individual items that must be packed out.

Figure 21 confirms the validity of these assumptions and the applicability of Little's Law to this system. This figure shows the average wait time in queue in each of four pack lines on a typical day at Site A. The Little's Law approximation values are calculated assuming four lines full of totes that each has the same number of items in them (number approximated as average number of items in a tote for that day).

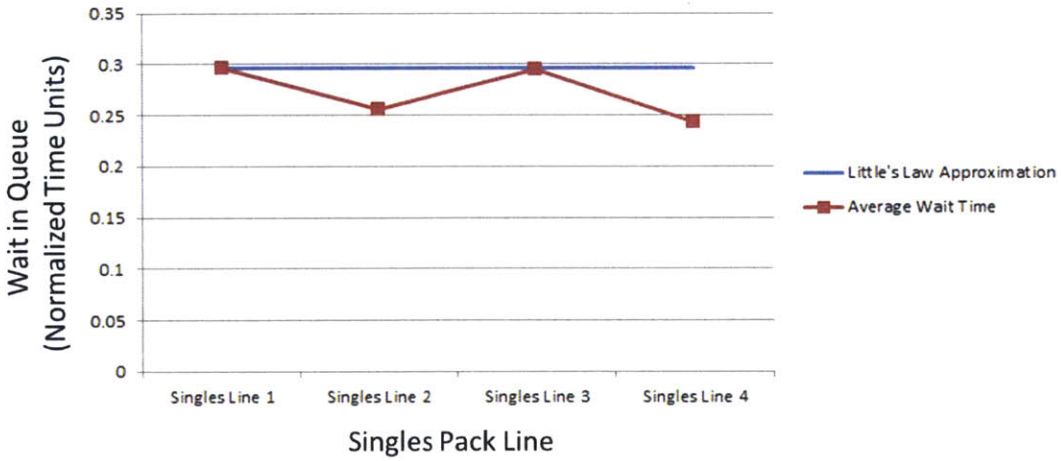


Figure 21: Comparison of actual average wait time in queue over a day versus Little's Law approximation

The Little's Law calculations overestimate the wait time for the true data by an average of 8% on the day the data was collected. While not perfect, this small overestimate is close and validates the use of Little's Law to calculate approximate changes in wait time resulting from the two solutions discussed above. The first solution is moving from four pack lines to three pack

lines. For the purposes of this discussion, the length of the lines in Figure 16 and Figure 19 will be measured as L. By calculating current WIP (4L) and proposed WIP (3L), an average wait time (W) in queue can be evaluated, assuming the same arrival rate (λ). The wait times calculate to:

Current staffing: $W = 4L / \lambda$ Proposed staffing: $W = 3L / \lambda$

This leaves the proposed staffing model with an average wait time in the pack lines that is ¾ that of the current staffing model. The second solution involves reducing the WIP level target based on the number of associates working in a line rather than completely filling that line. To make the analysis easy, WIP level in the current WIP line in Figure 20 is measured as L, and the proposed WIP level in the same figure is measured as 2/3 L. The wait times calculate to:

Current staffing: $W = L / \lambda$ Proposed staffing: $W = 2/3\ L / \lambda$

This leaves the proposed staffing model with a wait time in the pack lanes that is 2/3 that of the current staffing model.

This theoretical analysis of the pack process for singles – no prep items demonstrates the significant cycle-time improvements achieved through WIP reduction. Both solution one and solution two describe concrete, easy to implement ways in which Amazon can reduce wait time in queue. Also important to note, solution two describes a method to shorten target queue length, making it less likely that totes will stack up beyond the photo-eye when the queue line experiences a large influx of totes. This, in turn, reduces the need for indirect laborers to remove totes from a queue line, and also reduces the amount of tote recirculation due to over-full queue lines.

The true benefit of minimizing the level of WIP at pack needs to be tested through a controlled experiment within Site A. This will be discussed in the next steps portion of this thesis.

5.3 Case 3: The Multi – Prep Fulfillment Path

The work in this section did not have the full benefit of time for proper testing and scientific analysis. This case is presented as observations and suggestions for potential improvement through batch size reduction and better use of wasted wait time.

Current State

The multi – prep fulfillment path (path 2 in Figure 6) contains the largest number of process steps at Site A. As such, it should and does (see Figure 11) have the longest overall cycle-time. Multiple customer orders are virtually grouped together based on common shipping requirements (e.g. departure time, currier), and then the items are batched together through the physical steps of the fulfillment process. Figure 22 gives a time-series overview, from left to right, of the physical steps in the process.

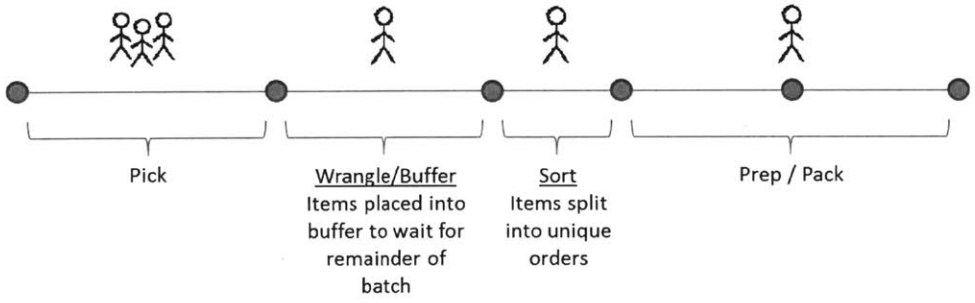


Figure 22: Time series look at the physical steps of the multi - prep fulfillment path

Multiple associates complete the pick process throughout various regions of the FC and then send the items along conveyors to converge at a wrangle buffer[8], where the items are set

[8] Wrangle buffer – Dedicated floor space, near pack lines, where customer orders sit and wait to be sorted and packed.

aside to wait on the remainder of the batch. All items in a batch are then sorted into their individual orders, followed by prepping and packing. Wrangle[9], sort, and prep/pack are completed in series; thus, every order in the batch must complete a step before any of the orders move to the next step. For sort and prep/pack, one batch is worked on by a single associate, and each associate works on one batch at a time.

Discussion of Problem and Proposed Solutions

In addition to the wrangle and sort steps that add cycle-time to multi – prep orders, the multi – prep fulfillment path includes a significant amount of waiting. This waiting unnecessarily increases overall cycle-time for customer orders, and, from a lean perspective, is waste in the system. Based on observation, this wait time has two main causes: batch size and order of operations.

Batch size: The batch size is equivalent to the number of customer orders grouped together and moved through the fulfillment process as one. As mentioned, no customer order moves on to the sort step or prep/pack step without the entire batch completing the previous fulfillment step first. This increases cycle-time for each order in the batch. For example, if the prep/pack step takes X minutes of physical work per order to complete, then the second order must wait X minutes while the first order is processed, and then it is processed for X minutes itself. The total time at prep/pack is thus 2*X minutes for the second order. The third order waits for completion of the first two orders prior to processing, and then is processed itself for a total of 3*X minutes spent at prep/pack. This pattern continues through batch complete, where the final order processed takes N*X minutes (assuming N orders in the batch) to completely move through the

[9] Wrangle – The process of physically moving totes full of customer items from a conveyor and into the appropriate wrangle buffer.

fulfillment step. Every minute spent *waiting* to be processed is wasted time in the overall cycle-time of a customer order.

Batch size for multi – prep orders is set to a high value at Site A. Amazon's algorithms that determine which inventory item an associate should pick next are affected by batch size. Essentially, the larger the batch size, the more items to pick, and thus the greater likelihood that items are located close together in warehouse inventory. This increases time on task (efficiency) for picking associates, as they do not spend a lot of time walking between picks. When there is a considerable amount of time left prior to a shipping truck departure, large batch sizes and corresponding lengthy cycle-times do not pose a threat to customer promise. However, as truck departure times approach and it is critical that orders move through the FC as quickly as possible, large batch sizes pose a significant problem and extend batch completion time through the downstream fulfillment processes.

Solution: A suggested solution to this issue is to appropriately reduce batch size as truck departure times near. In this way, pick efficiencies can be maximized for the majority of the day, and cycle-time can be reduced during critical periods. Little's Law can be used to help illustrate the benefit of reducing batch size through a process. The batch size is effectively the queue length (L) prior to a process. The process has a set rate (λ) at which it completes each order in the batch. Solving Little's Law equation for wait time (W) in queue, $W = \frac{L}{\lambda}$, and assuming a constant rate, it is seen that any reduction in batch size results in a proportional reduction in wait time.

Note: Amazon already does reduce batch size to some degree as truck departures near, but the reduction is standardized across all sites and all batch process paths. The multi – prep

fulfillment path at Site A includes the additional jewelry prep step that needs to be taken into special consideration. Also, the reduction in batch size is done without knowledge of the actual effect to overall cycle-time. Essentially, the reduction is arbitrary and only occurs once. Little's Law, here, can be used to set an appropriate batch size. Knowledge of time remaining (W) until orders must be on a truck and the rate (λ) at which items are processed can be used to calculate a proper batch size for the remaining orders due out. .

Order of Operations: The question of interest here is whether or not the jewelry prep step is properly sequenced in the current state multi – prep fulfillment process (see Figure 22). Prep activities are completed along with packing. One associate per batch selects an order, preps the necessary jewelry, and then places the order contents into a completed box. This process is repeated until the batch is complete, as mentioned previously. In this way, the prep and pack steps are coupled to each other. This coupling, along with the batch process described above, extends wait time for orders in queue and lengthens overall item cycle-time.

Solution: A suggested solution to this issue is to decouple prep from pack, and complete all or most prep activities while orders are waiting in the wrangle buffer. During the current state process, plastic totes carrying inventory items are dropped off individually by pick associates throughout the FC and then traverse a conveyor to arrive at a wrangle buffer location. Multiple associates pick orders for a single batch, thus multiple totes must arrive at wrangle before the entire batch is complete. These totes do not arrive all at once. In fact, many totes sit in a wrangle buffer waiting on the remainder of the totes in that batch for quite some time. This time spent waiting is pure wasted time during which value could be added. Prepping in parallel with wrangle utilizes this wasted wait time to complete value-added prep work. Once a tote full of items arrives at the wrangle buffer, prep work should

immediately begin. Figure 23 compares the current state with the proposed state for prep work.

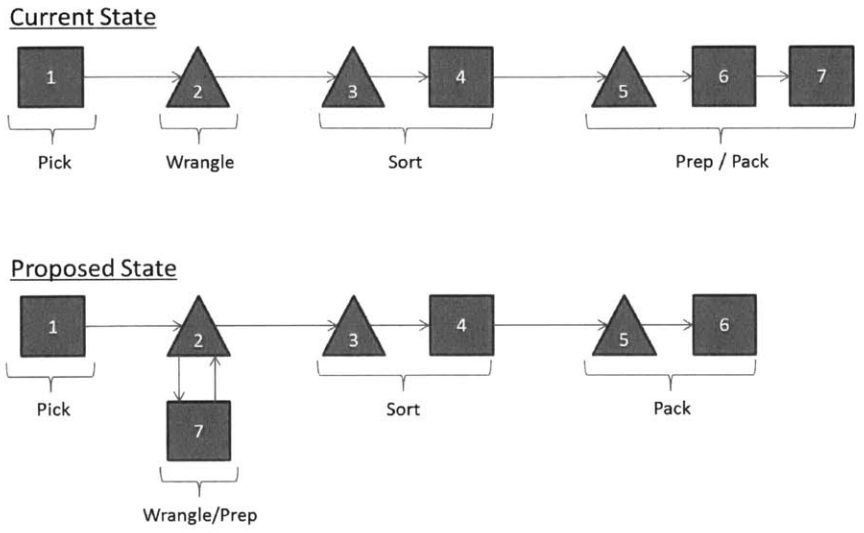


Figure 23: Current versus proposed state order of operations for multi - prep fulfillment path

Step 7 represents the jewelry prep step, shown in parallel with wrangle in the proposed state fulfillment path. The proposed state utilizes wasted wait time at wrangle, and it shortens time spent in the pack buffer because orders no longer need to be prepped, thus reducing overall cycle-time for each item.

The observations and suggested improvements in this section provide further ways for Site A to reduce overall cycle-time utilizing wait time reduction and improved use of buffer time. The exact cycle-time benefit of batch size reduction and prep during wrangle still needs to be tested through a controlled experiment within Site A. This will be discussed in the next steps portion of this thesis.

6 Conclusions and Recommendations

The goal of this research was to evaluate the current state cycle-time for Amazon fashion items as they move through the outbound fulfillment process of an Amazon FC (Site A) and determine if that cycle-time supported a lower setting of pick agreement time. Secondarily, this research aimed to develop suggestions for an improved future state outbound fulfillment process that reduced overall item cycle-time and supported further reductions in pick agreement time. This chapter begins with a discussion and review of the steps taken to achieve these stated goals, and it concludes with recommendations for Amazon going forward.

6.1 Discussion

The project work began with a baseline analysis of the outbound fulfillment process at Site A, a fashion fulfillment center. Detailed process flow diagraming enabled a complete end-to-end understanding of the steps necessary to move a customer order from virtual purchase to physical package. Due to the special prep activities involved in jewelry fulfillment, Site A includes four primary process paths as opposed to the standard two (multis and singles) at similar sites. These paths include multis – no prep, multis – prep, singles – no prep, and singles – prep. With an understanding of how the current process works, critical cycle-time data was queried over several days at Site A and compared to similar data queried over the same time period at a separate fashion fulfillment site and a standard sortable site, Site B and Site C respectively. Data analysis revealed two main conclusions. First, Site A sets pick agreement time much more conservatively than other sites in the Amazon network. And second, cycle-time at Site A is slowed by operational inefficiencies above and beyond the natural occurring hindrance caused by fashion and jewelry fulfillment nuances.

Conclusion 1: A comparison of the outbound cycle-time distributions at Sites A, B, and C with their respective pick agreement time settings indicates that Site A sets pick agreement time in a more conservative way than other sites in the network. One could speculate that Site A's pick agreement time was set this way to accommodate "peak" end of year conditions the year it opened, and then it was never re-evaluated during the following standard selling season. Based upon the analysis in Chapter 4, pick agreement time could be lowered by 15% for the first three quarters of the year (January through September), as a pilot run in August 2014 confirmed. In fact, this method of changing pick agreement setting is common among other fulfillment centers in the network. Outbound cycle-times vary throughout the year and from site to site for a variety of reasons. The method of cycle-time analysis utilized in this thesis provides a standard method to help evaluate cycle-time throughout the year and properly set pick agreement time. To improve upon the current state and thus allow even lower pick agreement settings, the second conclusion needs to be examined further.

Conclusion 2: Site A operates less efficiently than other fulfillment centers in the network. This is evidenced through the comparison of cycle-times for single item orders as shown in Figure 10 and Figure 11. Fulfillment of singles is most similar across all Amazon fulfillment centers and thus is easiest to compare and draw conclusions. The nature (e.g. packaging type, storage needs, and size) of soft-lines products naturally contributes some slowing effects to overall cycle-time, evidenced by longer average cycle-times for singles fulfillment at both Site A and Site B than at Site C. On top of this natural soft-lines slowing, though, Site A exhibits additional cycle-time slowing in comparison with the other soft-lines facility (Site B), even when jewelry prep items are not considered. This thesis identifies the root cause of this additional cycle-time slowing as operational inefficiencies and attempts to address these inefficiencies

through lean method utilization and waste elimination. The identified wastes were studied, tested, and analyzed via three main cases. The following provides a brief recap of the process and highlights the conclusions from each case study:

Case 1: The Pick Process

The pick process at Site A lacks standardization, causing a large variation in cycle-time and extended item wait times during which items sit, with no value added, on the bottom of pick carts. Currently, associates are given the freedom to fill one, two, or three totes with customer items prior to returning the totes to a conveyor for delivery to the downstream packing process. Mathematical modeling predicted and small pilot testing confirmed that picking into a single tote prior to conveyor drop off reduces both wait time average and standard deviation. Interestingly, modeling and testing also confirmed that standardizing pick into two totes prior to conveyor drop off does not significantly change wait time. In fact, two tote picking may actually increase the average time that items spend on the bottom of pick carts. The most likely explanation for this is that, on average, most associates pick into two totes prior to returning to a conveyor. From a lean standpoint, picking into a single tote reduces batch size, standardizes work, and eliminates unnecessary, non-value-added wait time on top of contributing to a reduced overall item cycle-time.

Case 2: The Singles – No Prep Fulfillment Path

The singles – no prep fulfillment path has the fewest number of physical steps at Site A, and yet it has one of the longest average cycle-times. Site A purposely builds large buffer queues full of work in process just prior to the singles pack step with the goal of minimizing indirect labor and ensuring the pack process is never starved of work. By

assuming the process to be in steady state, Little's Law can be used to show that the lengthy cycle-times through singles – no prep are almost entirely caused by wait time in these queues. Furthermore, Little's Law can also be used to highlight the potential for large reductions in cycle-time through simple, planned work in process reduction. The theoretical analysis in this thesis highlights the cycle-time reduction benefits of tighter work in process control, and demonstrates that goals of minimum indirect labor and zero work starvation are still achievable.

Case 3: The Multi – Prep Fulfillment Path

The multi – prep fulfillment path is the most complicated process path that was studied at Site A. Only about 5% of the total order volume fulfilled at Site A goes through this path, yet the lengthy cycle-time hinders the entire building and necessitates a high pick agreement setting. Limiting the scope of work to solutions the FC could implement without external team involvement, this thesis does not pursue the possibility of implementing a separate pick agreement time for preppable jewelry items. Instead, the analysis walks through a theoretical discussion of the benefits of batch size reduction and a more appropriate use of buffer wait times. As with Case 2, Little's Law provides a way to estimate the approximate cycle-time benefit of batch size reduction. Also, the discussion shows that prepping jewelry as soon as it arrives in a wrangle buffer utilizes previously wasted time to add value to customer orders, thus reducing overall cycle-time.

This thesis work successfully measured and analyzed current state cycle-times for soft-lines products as they move through the outbound fulfillment process at one of Amazon's fashion fulfillment centers. The work also highlighted several cycle-time benefits of the implementation of lean and reduction of process waste. The following sections provide Amazon with

recommendations for an improved future state outbound process and potential ideas for continued research and improvement.

6.2 Recommendations

The work completed at Site A and detailed throughout this thesis supports a few recommendations that Amazon can implement immediately. First, Site A should set a lower pick agreement time during the first three quarters of the calendar year prior to "peak" yearend conditions. Other sites in the network already practice this method of adjusting pick agreement time for standard versus "peak" time periods. Based on analysis and a short pilot test during August of 2014, Site A could stand to reduce pick agreement time by up to 15%. This recommendation is possible even under current conditions without any process improvements.

Second, Site A should standardize the way in which pick associates fill totes prior to returning to a conveyor. Ideally, pick associates should pick into a single tote, but this may not be practical 100% of the time (e.g. when picking in parts of the warehouse a significant distance away from a conveyor). The modified recommendation that has been adopted at Site A is to pick into a single tote but carry a second for instances when returning to a conveyor would significantly decrease work rate.

Third, Site A should staff three pack lines rather than four during non-"peak" conditions. The middle pack line should be staffed on both sides and will consequently have twice the throughput rate as the other two lines, effectively equaling the throughput associated with four pack lines. This recommendation reduces wait time in queue by approximately 25%, while keeping indirect labor requirements the same and protecting against work starvation.

Finally, Site A should utilize non-value-added wait time during the wrangle process to also prep jewelry items. Also, Site A should reduce batch size by an additional 50% as critical truck departure times approach. A reduction in batch size has a proportional reduction in buffer wait time, thus contributing to an overall reduction in cycle-time.

Once these process improvements have been implemented, Site A should utilize the cycle-time baseline analysis methods developed in Chapter 4 to assess the new current state cycle-time and adjust pick agreement time setting accordingly. This method of improvement, re-assessment of cycle-time, and adjustment of pick agreement time can be used iteratively to optimize the benefit for both Amazon and Amazon's customers.

6.3 Future Work

Going forward, there are a few projects that would benefit Site A in terms of cycle-time reduction and efficiency gains in outbound operations. The following are areas of focus that arise directly from the work in this thesis:

1. Single tote picking: Pick standardization and single tote picking may actually provide more benefit than just initially reducing wait times for items on pick carts. Picking in a standard, predictable manner could smooth out of the arrival of totes at downstream processes. Balance between picking and packing could then be more tightly controlled, eliminating the need for large WIP buffers to hedge against work starvation at pack. A potential future project should attempt to model the entire picking system and run controlled experiments to capture this smoothing benefit. Additionally, this project should address the potential tradeoff between picking efficiencies (# of picks per hour) and wait time reductions / downstream smoothing that arise from always picking into a single tote.

2. WIP reduction at pack: A future project at Site A should evaluate the optimum WIP level setting that prevents pack work starvation but minimizes overall cycle-time. The ideal system would adjust dynamically based on the following inputs: number of pack associates, pack rate, number of pick associates, pick rate, and average number of items in a tote.
3. Further study of double peaked (bimodal) cycle-time data: As discussed in Chapter 4, the distribution of cycle-times at Site A takes on a double peaked shape (see Figure 8). Attempts were made to sort the data by shift, time of day, and day of week with no success. A future project should focus solely on discerning what causes this double peak behavior, as it could produce further reductions in overall item cycle-time.
4. Batch size optimization: A final future project at Site A that emerges from this research is to evaluate the tradeoff between pick efficiency (# of picks per hour) and cycle-time as batch size is varied for the multi – prep fulfillment path. Reduction of batch size by 50%, and thus a reduction in cycle-time, is advisable as truck departure times approach, but larger batch sizes allow for optimal picking throughout the rest of the day. The exact tradeoff should be studied in further detail with the goal of optimally setting batch size throughout the day for minimum cycle-time and maximum pick efficiency.

7 References

Alukal, G., & Manos, A. (2006). *Lean Kaizen: A Simplified Approach to Process Improvements*. Milwaukee, WI: ASQ Quality Press.

Amazon Media Room: History & Timeline. (n.d.). Retrieved January 14, 2015, from http://phx.corporate-ir.net/phoenix.zhtml?c=176060&p=irol-corporateTimeline

Amazon Media Room: Overview. (n.d.). Retrieved January 14, 2015, from http://phx.corporate-ir.net/phoenix.zhtml?c=176060&p=irol-mediaKit

Cachon, G., & Terwiesch, C. (2013). *Matching Supply with Demand : An Introduction to Operations Management*. New York, NY: McGraw-Hill.

Chhajed, D., & Lowe, T. J. (2008). *Building Intuition: Insights from Basic Operations Management Models and Principles*. New York, NY: Springer Science & Business Media. Retrieved from http://www.springer.com/gp/book/9780387736983

Excel statistical functions: NORMINV. (n.d.). Retrieved March 15, 2015, from http://support.microsoft.com/en-us/kb/827358

Foster, D. P., Stine, R. A., & Waterman, R. P. (1998). *Basic Business Statistics: A Casebook.* New York, NY: Springer.

McKee, T. E., & McKee, L. J. B. (2014). Using Excel to Perform Monte Carlo Simulations. *Strategic Finance*, *96*(12), 47–51.

RAND function. (n.d.). Retrieved March 15, 2015, from https://support.office.com/en-sg/article/RAND-function-e98f1011-127d-4815-96f5-a26850ca1866

Thomopoulos, N. T. (2013). *Essentials of Monte Carlo Simulation: Statistical Methods for Building Simulation Models*. New York, NY: Springer. Retrieved from http://www.springer.com/us/book/9781461460213

Typical Histogram Shapes and What They Mean - ASQ. (n.d.). Retrieved January 23, 2015, from http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram2.html

Womack, J. P., & Jones, D. T. (2003). *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*. New York, NY: Free Press.

Zio, E. (2012). *The Monte Carlo Simulation Method for System Reliability and Risk Analysis*. London, UK: Springer. Retrieved from http://link.springer.com/book/10.1007%2F978-1-4471-4588-2

8 Appendix

8.1 Appendix A: Pick Test Protocol

Goal

Measure item dwell times when pick instructions are standardized, mandating pick into one or two totes.

Test Description

A subset of picking associates will be asked to pick and drop totes on the conveyors in a more standardized manner. Associates will not have the choice of how many totes (1, 2, or 3) that they want to pick into before dropping them on the conveyors. Instead, associates will be told to either pick into one tote or pick into two totes until full and then return to the conveyor to drop the tote(s). This test will run over two consecutive day shifts (ideally). The details are as follows:

- Number of days: 2
- Shift: Day
- Number of pickers: 28 (two groups of 14)

Protocol

1. 28 pickers chosen at random (including good, bad, and average) – document log-in IDs
2. Split chosen pickers into two groups of 14 each
3. Explain test/directions to pickers, pass out instructions
4. Run test
    - Day 1
        - Group 1: Pick into 1 tote
        - Group 2: Pick into 2 totes
    - Day 2
        - Group 1: Pick into 2 totes
        - Group 2: Pick into 1 tote

Directions for Pickers

1. Log-in to system as normal
2. Grab pick cart and fill with specified number of totes (1 or 2) only
3. Pick until totes are full – according to current standards on reasonably "full"

a. If picking into one tote: pickers allowed to drop totes if near a conveyor and tote is reasonably full
b. If picking into two totes: pickers allowed to drop totes if near a conveyor and second tote is reasonably full. Two totes should be dropped every time. Do not drop single tote if it is full and the second has not been picked into yet.

4. Return to conveyor from wherever you are in MOD and drop totes
5. Repeat until end of shift

Variables to Measure

Need to measure the following variables for each picker over the two testing days and compare with standard values prior to the tests:

1. Units picked per hour versus theoretical goal
2. Units picked into each tote (average)
3. Tote dwell times