Characterizing and Improving the Service Level Agreement at Amazon

by

Alberto Luna

B.S. Mechanical Engineering, University of Puerto Rico, Mayagüez Campus, 2001
M.S. Mechanical Engineering, The Ohio State University, 2009

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

Master of Business Administration and
Master of Science in Engineering Systems
in conjunction with the Leaders for Global Operations Program
at the

Massachusetts Institute of Technology

June 2015

Signature of Author Signature redacted

MIT Engineering Systems Division, MIT Sloan School of Management,
May 8, 2015

Certified by Signature redacted

Juan Pablo Vielma, Thesis Supervisor
Richard S. Leghorn (1939) Career Development Assistant Professor
MIT Sloan School of Management

Certified by Signature redacted

Bruce Cameron, Thesis Supervisor
Director, System Architecture Lab, and Lecturer, MIT Engineering Systems Division

Accepted by Signature redacted

Munther A. Dahleh, William A. Coolidge Professor of Electrical Engineering and Computer Science Chair, Engineering Systems Division Education Committee

Accepted by Signature redacted

Maura Herson, Director of MIT Sloan MBA Program
MIT Sloan School of Management

Characterizing and Improving the Service Level Agreement at Amazon

by

Alberto Luna

Submitted to the MIT Sloan School of Management and the MIT Engineering Systems Division on May 8, 2015 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Engineering Systems

Abstract

Amazon's Service Level Agreement (SLA) is a promise to its customers that they will receive their orders on time. At the Fulfillment Center (FC) level, the SLA is based on the capability to fulfill open orders scheduled to ship at each departure time. Each center's capability depends on a complex interaction between fluctuating product demand and time-dependent processes. By lowering SLA, Amazon could provide an enhanced the customer experience, especially for same day delivery (SDD). However, providing additional time to the customer also means that the FCs have less time available to fulfill open orders, placing the customer experience of those orders at an increased risk of a missed delivery.

This thesis explores cycle time reductions and throughput adjustments required to reduce the SLA at one of Amazon's Fulfillment Centers. First, a method to analyze time-dependent cycle time is used to evaluate the individual truck departure times, revealing that the current process conditions have difficulty meeting current demand. Then, using lean principles, process changes are tested to assess their ability to improve the current processes and allow for an SLA reduction. Although a 1% increase in capacity is possible by improving the processes, system constraints make the changes impractical for full implementation. Consequently, a capacity analysis method reveals that an additional capacity of up to 9.38% is needed to improve the current process conditions and meet current demand. The capacity analysis also reveals that reducing the SLA from its current state requires up to 13.79% more capacity to achieve a 50% reduction in SLA.

Through capacity adjustments, the added cost of late orders is mitigated, resulting in a reduced incidence of orders late to schedule and a reduced risk of missed deliveries. The methods utilized in this thesis are applicable to other Amazon FC's, providing a common capability and capacity analysis to aid in fulfillment operations.

Thesis Supervisor: Juan Pablo Vielma
Title: Richard S. Leghorn (1939) Career Development Assistant,
MIT Sloan School of Management

Thesis Supervisor: Bruce Cameron
Title: Director, System Architecture Lab, and Lecturer,
MIT Engineering Systems Division

Acknowledgments

I would like to thank everyone at Amazon's Fulfillment Center in Chattanooga, TN for giving me a glimpse into what it means to be an Amazonian. I owe thanks to all the Ambassadors who took the time to train me in the Amazon ways during my first week and who later provided valuable information and insight to the inner workings of a fulfillment center. I would like to thank the numerous managers at Chattanooga, in particular Travis Duea, who met with me every time I needed to gage my ideas and kept me on the right track, and to Erik Kurtz, who gave me his time, resources, and guidance to test new setups. Special thanks go to Ben Mullican, one of the best managers I have ever worked for.

I must also thank and acknowledge the Leaders of Global Operations Program for their support of this work and the opportunity to belong to this amazing and challenging community. Additionally, thank you to my academic advisors, Bruce Cameron and Juan Pablo Vielma, for their guidance and support.

I owe a great debt of gratitude to my fellow LGO, fellow Amazon intern, and friend, Hugh Churchill, whom I used as a sounding board more times than I can remember. Thank you Hugh, for listening to my ideas and helping me figure out which ones made sense. Thank you as well to the rest of my fellow LGOs for making these 2 years of my life a great and adventurous journey.

Finally, I would like to thank my family for their support and guidance throughout my entire life. The values and beliefs that my family has imparted on me have made me into the person I am today, and I would not be here if it were not for them.

Table of Contents

List of Figures

List of Tables

1 Introduction

As an online seller, Amazon constantly works on ways to improve the customer experience for its clients. The subject of this thesis stems from these efforts, focusing on the service level agreement (SLA), Amazon's on-time delivery promise to its customer, as a conduit to a better customer experience. The service level agreement takes the form of the words "Order within 2hr 44min to get it Wed Free" on Amazon's website, and is directly linked to the performance of the Fulfillment Center, Amazon's version of a distribution center, to deliver an order on time. By improving the performance at the fulfillment center, Amazon improves the customer experience for its clients.

This chapter introduces the subject of this thesis: to characterize and improve the service level agreement at Amazon's fulfillment centers. First, it covers the motivation behind improving the service level agreement, followed by the problem to be solved: how to improve the SLA while minimizing the risk of missing a scheduled shipment departure time. Then the hypothesis of this thesis is stated, followed by the approach used to test the hypothesis. The chapter closes by providing the outline for this thesis.

1.1 Project Motivation: Continuous Improvement at Amazon Fulfillment Centers

Since its incorporation in 1996, Amazon has grown from an online book seller to a global company offering a multitude of products and services. Amazon offers web services, manufactures and sells electronic devices, and provides software to develop, publish, and sell content to a client base that ranges from consumers to sellers, and from content creators to enterprises.

Still true to its origins, Amazon remains an online seller, offering millions of unique items to its customers worldwide. At the heart of online retail operations lie Amazon's 96 fulfillment centers (FCs) [1, p. 5], its version of distribution centers, where it stores and ships items purchased from vendors as well as items offered by third party sellers. Once an order arrives to a fulfillment center, dozens of activities start immediately, and follow a sequence to complete the order accurately and ship it on time so that it arrives at the customer when promised.

Amazon fulfilment centers continuously seek to improve their selection and service, not only to stay ahead of the competition, but also to reinvent 'normal' for its customers. It is in this spirit that Amazon seeks to enhance fast order delivery, especially for its 2nd day, next day, and same day deliveries, while maintaining the highest standard in customer satisfaction.

1.2 Problem Statement: Enhancing the Service Level Agreement

The Service Level Agreement (SLA) is a promise made to the customer about order delivery. At Amazon, this promise takes its form with the words "Order within 2hr 44min to get it Wed Free" on Amazon's website. This promise is tied to the fulfilment center from which the order will ship. The "order within" time given to the customer at the website depends on when the order has to depart the FC to be delivered on the promise date. It takes into consideration how much time it takes the FC to prepare the order and place it on a truck and how much time it takes the order to arrive at the customer from the FC. At the fulfilment center level, thousands of orders have to be ready to ship at each truck departure time to get to their destinations on time. The SLA then becomes an aggregate promise to hundreds to thousands of customers each time a truck leaves the FC.

At Amazon's fulfilment centers, the SLA is a time setting used to stop accepting orders that ship at the upcoming, scheduled truck departure time. The FC stops accepting orders so that it has enough time to finish processing all the open orders leaving at the next departure time. Each order is assigned a scheduled departure time when the order arrives based on algorithms that look for the lowest delivery costs to Amazon, among other criteria. Except for some delivery options (like same day delivery), if an order is not ready by the scheduled time, it gets upgraded to a new departure time, usually at an added cost absorbed by Amazon. Orders normally have at least one upgrade option in case of a missed scheduled departure time, and the risk of missing the promised delivery date is increased every time the FC misses a departure time.

The problem for Amazon then becomes how to improve the service level agreement, in particular for same day, next day, and 2nd day delivery, while curbing the increased risk of missing the scheduled departure time.

1.3 Project Hypothesis:

The hypothesis for this project is that reductions in cycle time and process variation through process improvements will allow Amazon to reduce the time setting of its service level agreement and reduce the risk of missing scheduled departure times.

1.4 Project Approach

Reductions cycle time and process variation stem from areas of opportunity in Amazon's FC operations. Identifying these areas of opportunity requires two phases. The approach for this project is segmented into these two phases, plus an additional phase that addresses the case when process improvement is insufficient to reduce the SLA and the risk of missing scheduled departure times. Due to physical and time constraints, only one fulfillment center is analyzed for this project.

In the first phase, the current process conditions must be well understood to determine the current performance. Amazon's collected data provides the information needed to assess the current conditions. At this step, the conditions are determined by two parameters: cycle time and volume. By segmenting orders into their respective scheduled departure times, each departure time can be characterized individually. The trends for both cycle times and volumes define the capacity of each departure time, and thus define the current state. Once defined, the current state of the individual scheduled departure times is compared to determine the departure times that can run on a shorter SLA and the ones that cannot.

In the second phase, the scheduled departure times that do not support a shorter SLA provide the opportunities to improve the processes and reduce cycle time. Using lean principles, the activities with the longest cycle times and highest waiting times are the best candidates for process improvements. After performing tests in the selected areas, the resulting cycle time is compared to the current cycle time to assess if the process improvement supports a shorter SLA.

In the case that process improvement alone is insufficient to support a shorter SLA, then the last phase for the project involves adding the labor force to the two parameters from phase 1. With the labor force in the mix, the analysis needs to determine how many associates are needed to support shorter SLAs.

1.5 Thesis Outline

The next chapter summarizes Amazon Fulfillment Centers and their operations. The chapter reviews the outbound operations, the operations responsible for preparing and shipping orders, in detail. Outbound operations are the operations of interest to this thesis. The chapter describes the customer order flow, order prioritization, and order upgrades.

Chapter 3 contains the literature review for the concepts and methods used in this thesis. It first reviews process capacity analysis used to determine the maximum capacity of a process through the use of process maps, input and output constrains, and utilization analysis.

Afterwards, the seven wastes associated with lean are identified and describes. Lastly, the topic of simple linear regression is described.

Chapter 4 delves into the analysis of each phase described in this chapter. It begins with the data collection and analysis of the current state to determine the cycle time for order completion. After the cycle time analysis of the current state, the chapter describes the selection process for the process improvement activities. The chapter closes with the capacity analysis.

Chapter 5 uses the analysis from Chapter 4 to identify the systems state where the fulfillment center ships orders on time and utilizes shorter SLAs. It first describes the process changes tested at the fulfillment center to reduce cycle time, followed by discussion about its benefits and limitations. The chapter then digs into the throughput capacity and how many worker hours are needed both for the current SLA and for reduced SLAs.

Chapter 6 summarizes the results of the analyses, and provides recommendations for Amazon's fulfillment centers as well as recommendations for future areas of research and analysis.

2 Operations at Amazon Fulfillment Centers

Amazon's customer orders are prepared and shipped from their fulfillment centers. Fulfillment centers are divided into two operations, inbound and outbound operations. Of these, outbound operations are responsible for shipping customer orders. Inbound operations receive inventory to fulfill orders, and are outside of the scope of this thesis.

This chapter covers the operations at a fulfillment center's outbound operations, specifically for customer orders. The path followed by all customer orders involves multiple activities, each described in detail. Order prioritization and late orders are also discussed.

2.1 Fulfillment Center Operations

Amazon's fulfilment centers serve two purposes. The first purpose they serve is to prepare and store the millions of items available through Amazon.com. Having these items at hand helps minimize the customer's waiting time to receive their order. The second purpose that the fulfilment centers serve is to prepare and ship orders to their customers and to other fulfilment centers. These two purposes naturally divide the FC's operations into two sub-business: Inbound and Outbound operations. These two operations are interlaced by the goal to place the customer first and to provide the best prices and the widest selection of items.

Inbound operations are responsible of stocking the millions of items made available to Amazon's customers. Amazon purchases these items from thousands of different vendors and receives them at the FC's dock. From there the items are received, unpacked, inspected for quantity and quality, added to inventory, and then routed to be placed on the hundreds of shelves that hold the FC's inventory. Once the items are placed on the shelves, any online purchases will get matched with the inventory and the outbound operations begin.

Outbound operations are responsible for picking ordered items from the shelves and getting them packed and shipped for delivery. These are the operations of interest for this thesis. The next section goes into the details of the multiple activities for preparing orders.

2.2 Outbound Operations

At Amazon's fulfilment centers, the outbound operations are composed of the activities that prepare and ship all customer and transship orders, plus all support activities. All outbound operations activities share the same resources, namely equipment and associates. Managers continuously monitor incoming, in-process, and outgoing workflow to manage resources and maintain a balanced workflow. When needed, managers activate or deactivate equipment and move associates around, giving them flexibility to match demand.

Outbound operations are segmented into three groups: customer orders, transship orders, and quality control. Customer orders are those placed on Amazon's website by its customers. Transship orders are requests for items by other fulfillment centers. These requests are necessary when a fulfillment center lacks the inventory needed to complete customer orders. The quality control group supports all outbound operations. They audit inventory levels and location to reduce and correct discrepancies, evaluate performance against process guidelines, implement corrective actions when inventory or process deviations are identified, and analyze customer-facing metrics [2]. Of these three groups, the focus of this thesis is in customer orders.

2.2.1 Customer order flow

When an order arrives at a fulfillment center, it triggers a sequence of decisions and steps based on multiple parameters. Orders get sorted by order type, by item size, and by priority. While the type and size determine where the order gets routed within the FC and how many steps it goes through, while the order priority determines when it gets processed. Although multiple systems determine an order's priority, one of the main priority factors is the scheduled departure time. Orders that ship sooner are given priority over orders that ship later. Combined with the profile of incoming orders, these three parameters dictate how the orders flow through the fulfillment center.

At the time an order is received at the fulfillment center, the order is categorized and marked as one of two main types. The first order type is singles, which are orders that contain only one item. The second order type is then orders that contain more than one item, called multis. Orders of the same type are grouped together by computer systems into groups of orders. Typically, the order items within a group are physically close together from each other to facilitate picking from the shelves.

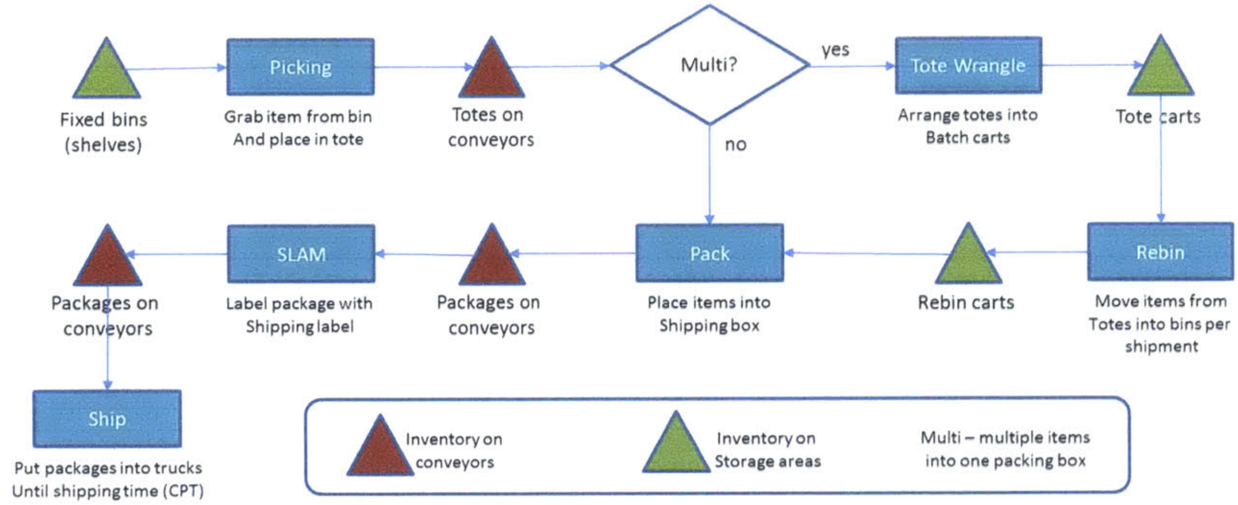

Fixed bins (shelves)
Picking
Grab item from bin And place in tote
Totes on conveyors
Multi?
yes
no
Tote Wrangle
Arrange totes into Batch carts
Tote carts
Rebin
Move items from Totes into bins per shipment
Rebin carts
Pack
Place items into Shipping box
Packages on conveyors
SLAM
Label package with Shipping label
Packages on conveyors
Ship
Put packages into trucks Until shipping time (CPT)
Inventory on conveyors
Inventory on Storage areas
Multi – multiple items into one packing box

Figure 1. Process map for outbound operations at a fulfillment center.

Picking is the first activity in the outbound process ('Picking' in Figure 1), and it consists of associates armed with a handheld scanner and a small cart with a tote. The handheld scanner shows the associate where it needs to go in the FC to find the next item to pick. Once at the location, the associate looks in the scanner indicated bin on the shelves and grabs, or picks, the item and places it in the tote. The scanner keeps track of how many items are in a tote, and, once the tote is full, the scanner relays the information to the computer systems and the associates place the tote in one of the conveyors located throughout the picking area. The conveyor system at Amazon's fulfillment centers are highly automated, and since the order information is constantly updated, the conveyor system knows where to route the totes from Pick (see Figure 2). Based on order type, the totes are routed to a singles packing area or to a multis sorting area. This point is where the order types go through different activities.

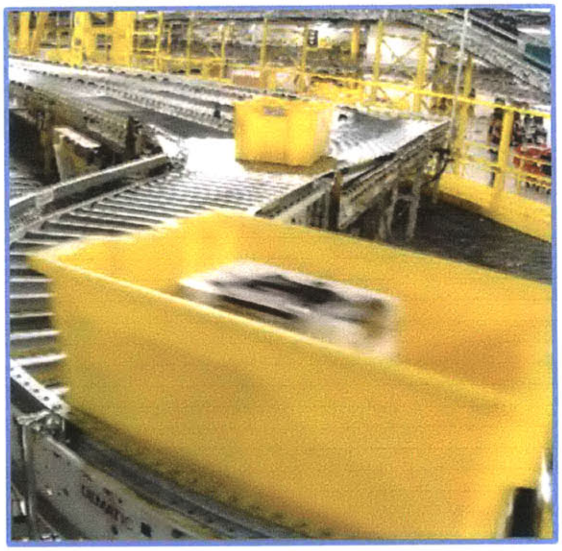


Figure 2. Totes with picked items travelling on the conveyor system.

Multis orders have the extra complication of having multiple items spread throughout the fulfilment center that need to come together and get packed together. At picking, the items from one order may be picked by one picker or by as many pickers as there are items in the order, placing each item in the order into different totes throughout different locations in the FC. Because the computer systems grouped orders of the same type together, they keep track of the totes that contains all the items from the group of orders, which, for multis, are called batches. The conveyor system routes all totes in a batch to the same location, where the next activity takes place. The next activity is called tote wrangling ('Tote Wrangling' in Figure 1), and it consists of associates with handheld scanners pulling totes from the conveyors and placing them on batch carts, effectively regrouping the all the items from all the orders in one batch. Once all totes from a batch are collected, the cart is moved to the next activity.

At this point, a batch cart has numerous items from numerous orders on different totes. The items from the different orders need to be segregated from multiple totes into each individual order. This activity is called rebin, and it consists of an associate at a rebin computer station sorting items from the numerous totes into bins on a cart, hence the activity's name, rebin ('Rebin' in Figure 1). The associate scans each item at the computer station and the computer station indicates the bin to place the item. After all items are processed, each bin on the cart contains all the items from one order. Afterwards the rebin cart is moved to the next activity, packing.

Packing for mutis takes place at a packing station, where an associate moves the items from a rebin cart bin into a packing box ('Packing' in Figure 1). In this activity, the computer system tells the associate which box to use for each order and which order from the rebin cart to pack next. One by one the associate packs all orders from the cart and sends the packed and taped boxes down a conveyor. Each box is identified with a barcode that is scanned prior to placing on the conveyor, updating the computer systems. Packages from multis orders are joined by packages from singles orders on the same conveyor.

Singles packing is different from multis packing in three ways. First, the totes travel straight from picking to the singles packing area without going through tote wrangling and rebin, therefore taking less time between picking and packing than it takes for multis. Second, since items do not need to be sorted, items from singles orders are packed straight from the tote. And third, because the items are for singles orders, there is only one item per packing box. Packing takes place at packing computer stations where associates pull a tote from the conveyor. One by one, each item of the tote is scanned, and the computer indicates the box to use for packing. Once the order is packed and taped, each package is scanned prior to placing it on the conveyor.

At this point in the process each individual package on the conveyor represents one individual order.

The conveyor moves all packages through an automated system for label application (SLAM in Figure 1). The automated station scans each individual package and applies the shipping label. From this point on the conveyor system starts routing the packages to different dock stations. Each dock station represents a different shipping company departing at different times. Associates grab packages from the conveyor and place them in the trucks. Thousands of packages may go into each truck, and the associates arrange the boxes to maximize the amount of packages in each truck. Packages accumulate at the trucks until it is time to depart ('Ship' in Figure 1).

2.2.2 Order Prioritization

As mentioned previously, customer orders get sorted by priority, and priority determines when an order gets processed. In general, orders that are scheduled to ship earlier get priority over orders that are scheduled to ship later.

Most prioritization is controlled automatically. Computer systems determine which orders get processed and when they get processed, like the items at pick. Other systems, like at the packing station, let the associates know if an order has priority over another so that it can get packed first. The rest of prioritization requires human intervention. For example, two multis batches start at the same time, get through rebin at the same time, and are in queue for packing. One of the batches has an earlier departure time than the other batch. The next associate that goes fetch a rebin cart will grab the former before the latter. These priority decisions are made throughout the day for each departure time.

2.2.3 Orders Late to Schedule

Orders that are not completed prior to their scheduled departure time are deemed late. This does not necessarily mean that all late shipments will arrive late to the customer. What it means is that the order missed the lowest cost shipping option available to the FC, and Amazon may have to incur extra shipping costs. It also means that, by missing the original scheduled departure time, the risk of a late customer delivery is higher. The risk increases because the order now has less time to arrive at its destination and fewer shipping options to get there.

When an order misses it scheduled departure time it gets rescheduled to another departure time, and typically at a cost to Amazon. Depending on the shipping option, the order may have several opportunities to be delivered on time, or it may have only one opportunity. A same day delivery option may have only one departure time available to get to its destination, and missing it would mean arriving late at a customer's address. The scheduled departure time are assigned to take advantage of the lowest-cost, ground transportation option. When scheduled departure times are missed, the shipping options escalate, and Amazon has to incur additional costs to ship via more expensive options, like air transport.

2.3 Chapter Summary

Customer orders are completed at Amazon fulfillment centers, which are tied to the service level agreement. Each center has inbound operations, responsible for receiving inventory, and outbound operations, responsible for preparing and shipping orders. When they arrive, customer orders are classified into orders of a single item or of multiple items. This classification dictates the path orders follow through the fulfillment center, sending them through picking, sorting, and packing at their respective stations. These activities control how an order gets prepared, while order priority controls when it gets prepared. If an order missed its scheduled departure time, Amazon provides a new departure time, typically at an increased shipping cost and at an increased risk of arriving late at the customer's doorstep. Meeting the service level agreement is paramount for Amazon's level of customer experience.

The next chapter summarizes the concepts and methods used to analyze outbound operations.

3 Literature Review

Amazon's service level agreement performance can be analyzed in numerous ways. This chapter discussed the concepts and methods used to analyze a fulfillment center's outbound operations. This chapter first describes process capacity analysis and its three main components: process flow diagrams, bottlenecks, and utilization. Then it gives a brief description of lean and of the seven wastes that lean seeks to eradicate from operations. Finally, it defines linear regression and identifies the mathematical expressions used in linear regression analysis.

3.1 Process Capacity

Process capacity is the measure which indicates the maximum product that a process can make in a given time period [3, p. 32]. This measure is determined by analyzing the operation in detail, and understanding the activities involved in the production or provision of a good or service. The analysis entails the preparation of a process diagram, the calculation of the capacity for each resource, and identifying the bottleneck. The analysis is limited to the process of interest, considering its inputs and outputs, and all the steps in between. The analysis of a process can be as simple as one input, one activity, and one output, or as complicated as a whole supply chain.

3.1.1 Process Flow Diagram

A process flow diagram is a graphical representation of the materials, steps, and flow of a process. Process flow diagrams help identify and organize the information needed for the process analysis, and they define the process boundaries, the inputs and outputs, the process activities and their sequence, the flow units, and the buffers that collect inventory.

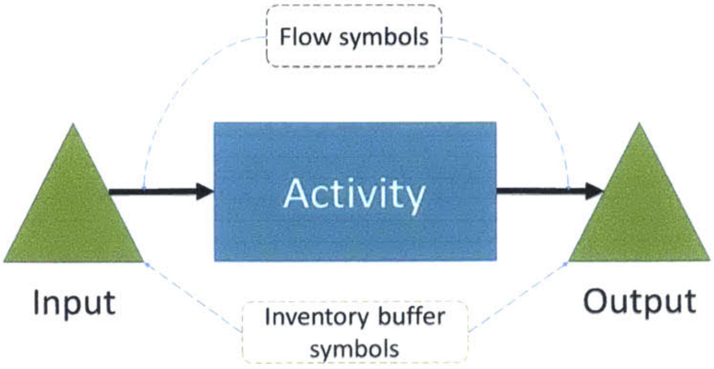


Figure 3. Elements of a process flow diagram.

The process boundaries determine the start and end of the process. The boundary definition depends on the project and the analysis that needs to be performed. For example, in the manufacture of turbine airfoils for jet engines there might be multiple manufacturing lines processing different parts. A project that looks at the incoming raw material for all parts would set the boundary as the entire site, while a project looking at the manufacturing capability of one part would set the boundary as the manufacturing line for said part and ignore the rest of the site.

Once the boundary is defined, the process is mapped out with a series of boxes, arrows, and triangles representing the activities, the flows, and the buffers that define the process (see Figure 3). The items that move through the process are called flow units, and they represent the object or information that is moved through the mapped process. The flow units can be the turbine blades in the example above, patients in a hospital, or data packets in a data system. The flow units are transformed by the activities in the process. Activities add value to the flow units and have a limit to how many units can flow through the activity in a given time frame. Buffers, on the other hand, do not add value to the flow unit, and they can accumulate units. Because they can accumulate units, there may be a limit to how many units a buffer can hold. Finally, the flows indicate the progression of units between activities and buffers. Flows can be different flow units within the process, like an inflow of components into an activity that outputs an assembly.

Process flow maps can be as simple or as complicated as necessary. Their purpose is to help make business decisions based on capacity and performance of a process.

3.1.2 Capacity Calculation and Bottleneck Identification

The capacity of a process is the maximum amount of output that can be produced in a given time frame. Capacity is a flow rate, or the rate at which the flow unit moves through the process in the given time frame. The time frame depends on the desired analysis and can be as short or as long as needed (microseconds or years). If the process has only one activity then the activity limits the capacity of the process. If the process has more than one activity then the activity with the smallest capacity limits the capacity of the entire process. An example of process capacity would be a process able to manufacture a maximum of 200 turbine blades in one day even in the case when there is ample input and demand. In this example, one of the activities in the process can only process 200 blades per day, with the rest having a higher capacity. As the example suggests, the input or demand could limit the amount of blades manufactured each day. If the input or demand only allow 100 blades per day to be produced, the process still has the capacity to produce 200 blades per day. This last example illustrates that, aside from capacity, input and demand are flow rates [3, p. 38].

Any one of the three flow rates can limit the output from the process (see

Figure 4). When demand is lower than the available capacity and there is ample input, the process is demand-constrained. This would be the case when 150 blades are needed per day and the manufacturing site would limit production to 150 even though it can make 200 blades per day. When demand exceeds the process output the process is supply-constrained, and in this case the limiting factor can be either the input or the capacity (input-constrained or capacity-constrained, respectively). An example of an input-constrain is when the process runs out of raw material and can only produce what the limited raw material allows. On the other hand, when there is enough raw material and demand to produce 250 blades per day, the process is capacity-constrained because the manufacturing site can only produce 200 blades per day.

Capacity constraints are determined by the activity in the process with the lowest capacity [3, p. 40]. No matter how much more capacity other activities have, the lowest capacity activity does not allow a higher flow rate through the process. In the turbine blade example, if one activities has a flow rate of 250 blades per day, the maximum output is still 200 blades per day. This lowest capacity activity is known as a bottleneck, analogous to a bottle, where the neck of the bottle only allows a maximum amount of fluid to flow out regardless of how much fluid

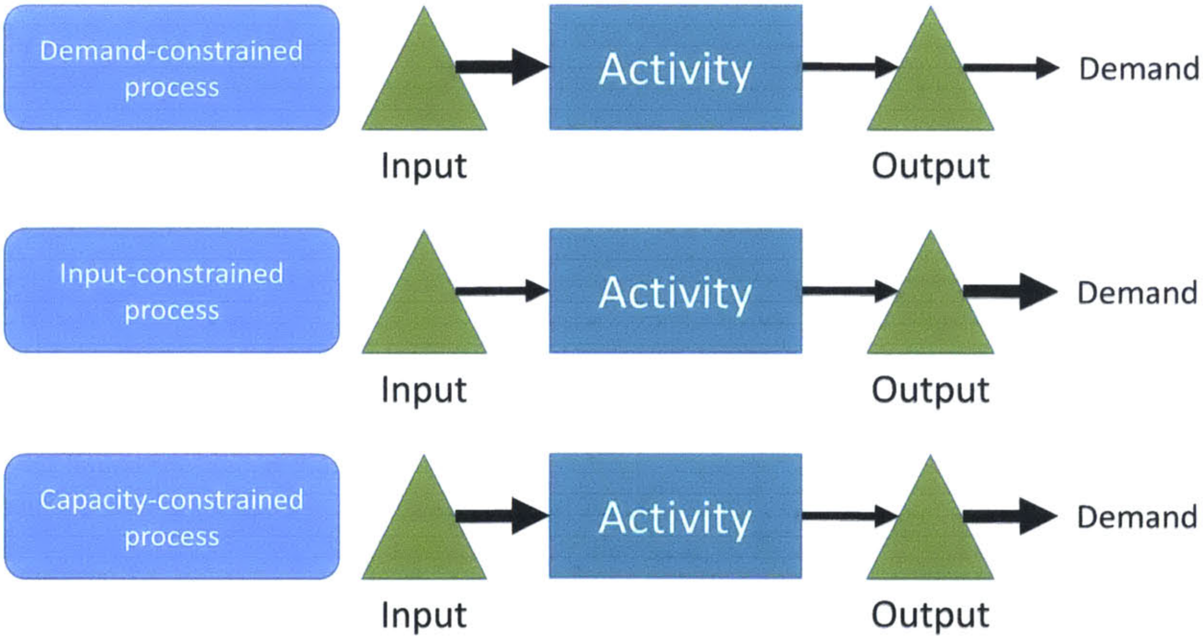

Demand-constrained process
Activity
Demand
Input
Output
Input-constrained process
Activity
Demand
Input
Output
Capacity-constrained process
Activity
Demand
Input
Output

Figure 4. The three types of process constraints. The thickness of the arrows are proportional to the flow. The demand-constrained process (top) has enough input material and process capacity, but since demand is lower than the possible output, production is limited to demand. The input-constrained process (middle) has ample demand, but not enough input to supply the process and meet demand. The capacity-constrained process (bottom) has ample input and demand, but the process cannot output more product due to its capacity limitations.

the bottle may hold. Identifying the activity with the lowest capacity is one way to identify the bottleneck of a process.

3.1.3 Utilization

Utilization is the ratio of the amount of output produced to the amount of output that can be produced [3, p. 41]. In other words, it is the ratio of actual output to maximum capacity, and it is always less than or equal to 100%. Going back to the turbine blade example operating at full process capacity of 200 blades per day, the bottleneck process runs at 100% utilization (200 blades per day/200 blades per day), while the second process runs at 80% utilization (200 blades per day/250 blades per day). The activity with the highest utilization is the process's bottleneck.

Another useful definition is implied utilization, which is similar to utilization but with the caveat that it uses the desired amount of output to be produced rather than the actual amount of output produced [3, p. 43]. With implied utilization the ratio can be higher than 100%, which indicates that the process does not have the capacity to meet demand. If the customers wanted 300 blades per day, the implied utilization of the manufacturing process is 150% (300 blades per day/200 blades per day) and 120% (300 blades per day/250 blades per day) for both activities in the process. Like with utilization, the highest implied utilization points to the bottleneck.

3.2 Lean Operations

Lean is a discipline which focuses efforts in eliminating non-value added activities from a process. A value added activity is one that adds value to the good or service as the customer sees it. Examples of value added activities are fastening an assembly together or applying paint to a car. These activities increases the value of the good or service for the customer and their willingness to pay for it. Lean eliminates the activities that do not add value for the customer. Some non-value added activities are moving goods from one location to another or accumulating items in a buffer. As lean principles are applied, the removal of non-value added activities leads to increased productivity, reduced defects and reduced costs, among other benefits.

Non-value added activities are referred to as wastes, or muda. Different sources list different types of wastes. Typically, there are seven sources of wastes associated with lean [3, p. 226]:

1. Overproduction – producing more output than needed.
2. Waiting – waiting can take two forms: a resource waiting for flow units and flow units waiting for a resource.
3. Transport – moving flow units from one location to another.
4. Overprocessing – spending more time than necessary on a flow unit.
5. Inventory – accumulating flow units without processing them or turning them over to the customer.
6. Rework – processing a flow unit through an activity more than once.
7. Motion – movement associated with performing a task.

By removing wastes, the process will take less time to complete and may add capacity to the operations. Removing wastes may also reduce cost associated with losses from rework and defects.

3.3 Linear Regression Analysis

Regression is the analysis of the relationship between variables. More specifically, regression analyzes how some variables affect other variables. The variables that affect others are called independent, predictor, or regressor variables, while the affected variables are called dependent or response variables [4, p. 373]. Regression analysis is used to model this relationship between variables based on data obtained through observations and then used for predicting how the response variables will behave. In the case of linear regression, the variables follow a linear relationship between themselves.

3.3.1 Simple Linear Regression [5, Ch. 6]

A simple linear regressions is a type of linear regression where there is only one independent and one dependent variable. The linear relationship between the two variables, defined as

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i,$$

linearly models the independent variable, x_i, and its effect on the dependent variable, y_i. The other terms in the equation define the y-intercept of the linear equation, β_0, the slope, β_1, and a random error, ε_i. The slope represents the ratio of change between the variables. If the slope is zero then there is no relationship between the two variables.

There are several methods available to calculate the estimates of the slope and the intercept. Of these methods, the least squares estimation minimizes the sum of the squares of the residuals, where the residuals are the difference between the actual dependent variable and the predicted dependent variable. The resulting equations for the intercept and slope are

$$b_0 = \bar{y} - b_1 * \bar{x}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

where $\bar{x}$ is the average for the independent variable values, $\bar{y}$ is the average of the dependent variable values, and

$$SS_{xy} = \sum_{i=1}^{n} x_i * y_i - \frac{1}{n} * \left(\sum_{i=1}^{n} x_i\right) * \left(\sum_{i=1}^{n} y_i\right)$$

$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} * \left(\sum_{i=1}^{n} x_i\right)^2$$

The linear regression equation is complete once the intercept and the slope are calculated. However, the linear regression equation might predict a value that is too different from the actual value. One way to determine if the model is adequate is calculating the coefficient of determination, commonly known as R^2. This coefficient compares the variability explained by the model to the total variability (which is the variability explained by the model plus the variability that is left unexplained). The model explains the data well when the coefficient is closer to 1. The formula for the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

where

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

$$SS_{total} = SS_{reg} + SS_{res} = \sum_{i=1}^{n} {y_i}^2 - \frac{1}{n} * \left(\sum_{i=1}^{n} y_i\right)^2$$

and $\hat{y}$ is the predicted value of the dependent variable. A high coefficient of determination does not necessarily mean that the linear regression is adequate. However, when the plotted data looks linear and the coefficient of variation is close to 1, the model is adequate to accurately predict the dependent variable.

3.4 Chapter Summary

The analysis of Amazon's performance requires the use of the three concepts covered in this chapter. Process capacity for the fulfillment center's outbound operations is critical to understand how many orders the center can process. Lean concepts and the seven wastes aide in the identification of activities with opportunity for removing non-value added steps. Linear regression provides the analysis needed to condense hours on information into a handful of variables, facilitating interpretation.

Chapter 4 uses the concepts presented in this chapter to analyze outbound operations. These concepts allow the cycle time analysis, the analysis for process improvement, and the capacity analysis.

4 Current State Analysis

Amazon's priority is to improve the customer experience. In this case, they seek to improve the customer experience by shortening the service level agreement through process improvement. However, the performance of Amazon's fulfillment centers needs to be evaluated to determine if they meet the customer's expectations with the current SLA. The current SLA needs to be characterized.

The characterization of the current state begins with the analysis of cycle time. The analysis of the cycle time defines the performance of outbound operations and defines the baseline, which is later used for before and after process change comparisons. Once the cycle time and its baseline are set, a lean analysis of the individual activities and their cycle times aids in identifying areas of process improvement opportunities. Finally, capacity analysis determines the current throughput capacity which sets the limits to Amazon's shipped units per day. These three analyses pave the way for identifying the future state with a shorter SLA.

4.1 Cycle Time Analysis

There are many metrics maintained by Amazon to describe the performance of the different individual activities within outbound operations and of outbound operations as an integrated system, from unit flow per hour to average cycle times. However, these metrics are an aggregate of all orders averaged over all departure times, rather than orders segmented per individual departure time. Also, the metric that tracks the quantity of late orders only focuses in one day's performance in aggregate. With current metrics, it is unclear beforehand whether an order is more likely to be completed on time than to be completed late, and vice versa. With this in mind, this analysis focuses on identifying a practical way to clearly diagnose how the outbound operations perform at each departure time relative to on-time shipping.

4.1.1 Data Collection for the Cycle Time Analysis and Process Improvement Analysis

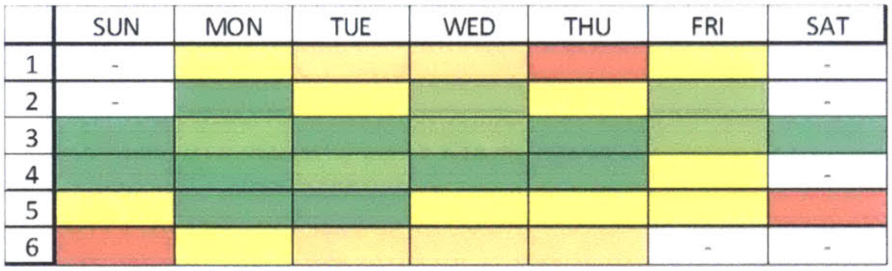
Amazon continuously collects data for all of its outbound operations. Data is transferred daily to Amazon's Data Warehouse for long-term storage and for use by all Amazon employees. In general, the collected data is marked with different tags to indicate the operations and their date and time stamps. With this information cycle times are calculated for the different operations at their different times of the day, of the week, or of the year.

The information extracted from the data warehouse is consolidated into several parameters that are useful to characterize the current state. These parameters help categorize the orders' characteristics on an individual order basis given their assigned departure time:

- Order identification – Each order is assigned a unique number that identifies the order in Amazon's systems.
- Quantity – The number of items in each order is important to determine the order type and the order density (number of items per order).
- Order type – Orders are separated into two main categories: orders that only contain one item (singles), and orders that contain more than one item (multis).
- Cycle time – Using date and time stamps for each order at each step of the process, the cycle time is calculated by subtracting stamps at the beginning of a process from stamps at the end of a process. Cycle times for individual activities and for the overall process (from order receive to order ship) help analyze the system and its parts.
- Time until assigned departure time – Orders are assigned a departure time when they arrive. The time to departure is calculated by subtracting the order's arrival time from its assigned departure time. This parameter illustrates how orders arrive relative to their departure time.

These parameters are pulled from the data warehouse for each departure time on a specific date using Structured Query Language, or SQL. The data is pulled per departure time and per day because it provides a practical way to segment orders and analyze the data. Table 1 illustrates the six departure times used to pull the data. Other alternatives, like pulling data on a per day basis, prove difficult to analyze due to the large volume of orders and transactions recorded in one day.

Table 1[1]. Table illustrating six scheduled departure times for an FC. The numbers on the left margin identify the individual departure times. The colors indicate the departure times which ship orders on time (green color) escalating towards the departure times that have frequent late shipments (red color). The dashed lines indicates that the departure time is not active on that specific day of the week.



	SUN	MON	TUE	WED	THU	FRI	SAT
1	-	🟨	🟧	🟧	🟥	🟨	-
2	-	🟩	🟨	🟢	🟨	🟢	-
3	🟩	🟢	🟩	🟢	🟩	🟢	🟩
4	🟩	🟩	🟢	🟩	🟩	🟨	-
5	🟨	🟩	🟩	🟨	🟨	🟨	🟥
6	🟥	🟨	🟧	🟧	🟧	-	-

4.1.2 Data Analysis for Average Cycle Times and Standard Deviations

The scatter plot of order cycle time against the time until departure, depicted in Figure 5, reveals the trend of decreasing cycle time as the departure time approaches. Orders that arrive farther from their scheduled departure time take longer to complete than orders arriving closer to their scheduled departure time. There is a minimum amount of cycle time required to complete an order, indicated by the gap at the bottom of the plotted data. This minimum time is the fastest time that the FC can fulfill an order. The diagonal gap in the plot indicates one of the work breaks during the shift, like lunch or shift change, where the associates are not working on the lines. The gaps are in different locations of the plot depending on the times of the work breaks

[1] The data used to build this table is undisclosed due to its proprietary nature.

relative to the scheduled departure times. After segregating the data into fractions of an hour to improve the statistical analysis, the box plot of the data shows two key characteristics of the data (see Figure 6). The first one is that even though the spread of cycle time per order occupies almost all the available time, the bulk of the orders' cycle times are concentrated closer to the top of the graph. This indicates that most orders take more time to complete rather than less time, regardless of when they arrive relative to their departure time. The second key characteristic of the data is that the median declines as the departure time approaches.

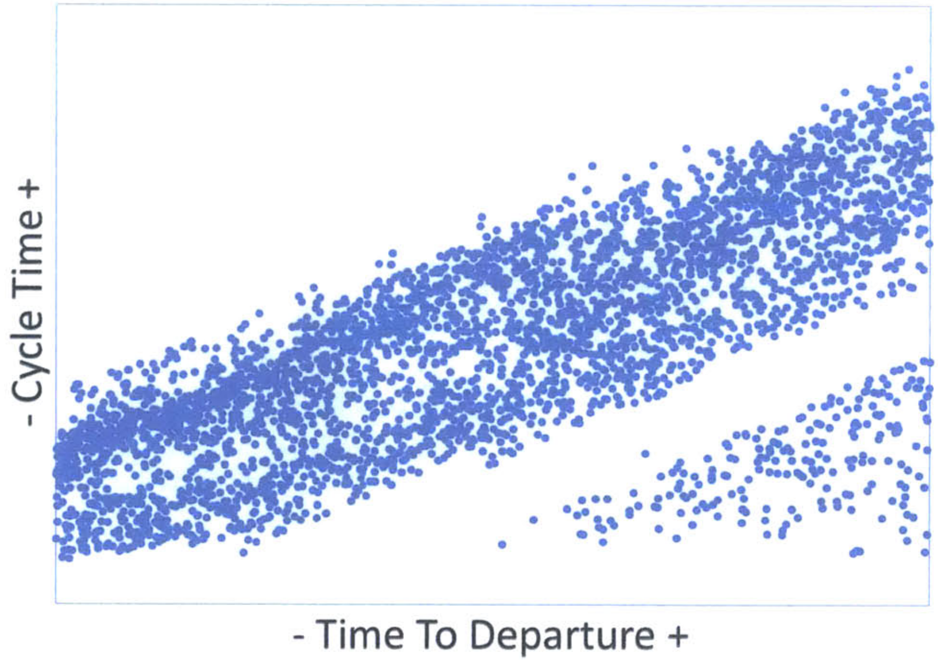


Figure 5. Scatter plot of cycle times for each order relative to departure time.
Each dot represents the time the order arrived relative to its departure time (x-axis) and the time it took to complete the order (y-axis). Data points to the right of the plot have more time available prior to departure than data points to the left of the plot. Data points towards the top of the plot have a longer cycle time than the data points towards the bottom of the plot. The diagonal gap indicates a work break during the shift.

The trend in the data suggests a linear relationship of decreasing cycle time as departure time approaches, which is supported by the boxplot in Figure 6. The calculated averages and standard deviations visually exhibit a linear relationship relative to time left to departure, shown in Figure 7. Linear regression verifies the linear relationship of these two parameters. Visually, the difference between the data points and the regressed line is small, and the coefficient of determination (R^2) is higher than 0.99 for both parameters. Therefore, the linear regression is adequate to describe and predict the average cycle time and its standard deviation based on the time to departure.

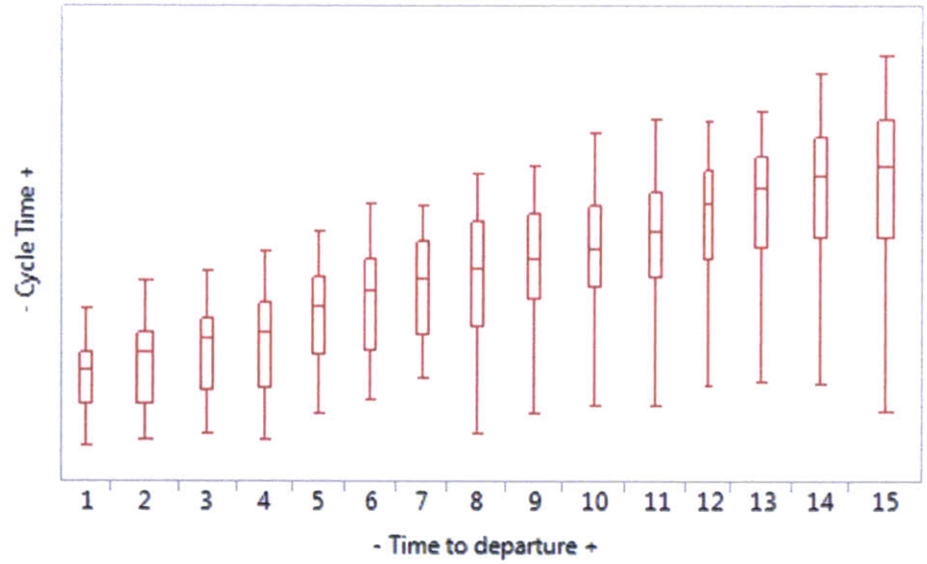


Figure 6. Boxplot of cycle time as orders approach departure time. The x-axis represents the time to departure segmented into fractions of an hour.

The regression analysis of the average cycle time and its standard deviation consolidates the information from all the data points in Figure 5 into four numbers: a slope and an intercept for the average cycle time and its standard deviation. Alternatively, the data points can be expressed graphically in two lines as illustrated in Figure 7. This allows either a visual comparison of process performance using figures or a numerical comparison using the slopes and intercepts from the regression analysis.

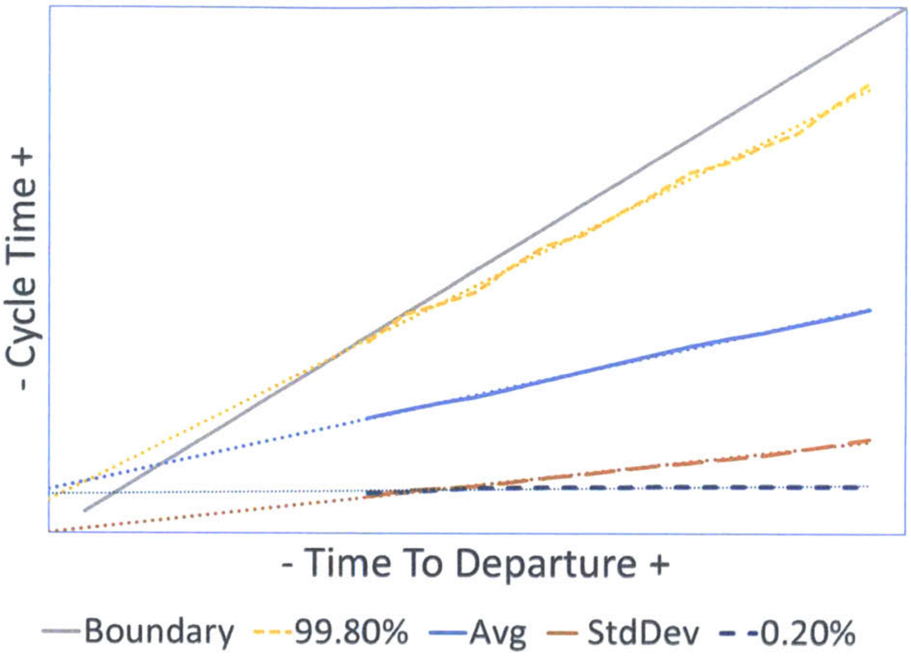


Figure 7. Average cycle time, standard deviation, and percentile as orders approach departure time. The solid and dashed lines represent the calculated parameters. The area near the middle of the x-axis where the solid and dashed lines stop indicates the SLA, where the FC stops accepting orders for the current departure time. The dotted lines represent the linear regression. The boundary line splits the on-time orders from the late orders. Anything above the boundary line is a late order.

In addition to the average cycle time and its standard deviation, the data allows the calculation and prediction of percentiles. Figure 7 depicts the 0.20 and the 99.80 percentile lines and their regression line. These lines represent the cycle time for completing the 0.020 and the 99.80 of all orders relative to the departure time. This information proves crucial since it establishes a direct measure of how quickly orders can be processed (0.20 percentile) and how many orders are shipped on time per departure time (99.8 percentile). A 0.20 percentile means that 2,000 out of one million orders will be processed at or below the cycle time depicted by the percentile plot. A 99.80 percentile means that 2,000 out of one million orders will be late to their scheduled departure time. The 0.20 percentile provides an idea of how fast orders can be processed at low volumes. The 99.8 percentile provides a target for dividing a well performing departure time from one that frequently misses scheduled departure times.

The baseline analysis applied on the six departure times shown on Table 1 characterizes the performance of SLA at each departure time for one fulfillment center. The results indicate that:

- Two of the departure times consistently ship orders as scheduled. Figure 7 illustrates one of these departure times. These are represented by departure times 3 and 4 on Table 1. The regression analysis predicts that these departure times are capable of running shorter SLAs than the current SLA. This prediction is supported by two key facts about the departure times. The first is that these two departure times have a volume several times smaller than the other four departure times. The second fact is that these two departure times occur in the same shift as one of the higher volume departure times. This means that they have the same amount of workers and thus the same capacity. If workers within a shift can process the higher volume and occasionally ship orders late to schedule at other departure times, then it is likely that they will be able to process the smaller volume with a shorter SLA at these departure times.
- Two of the departure times occasionally complete some of the orders late with respect to their schedule. These are represented by departure times 2 and 5 on Table 1. The regression analysis suggests that slightly longer SLAs would improve on-schedule shipping. Given that these two departure times are on the fringe between being able to manage the volume and occasionally shipping orders late to schedule, it is unclear if these two departure times could handle a shorter SLA.
- The two remaining departure times consistently ship orders late to schedule. These are represented by departure times 1 and 6 on Table 1. These two departure times

require significantly longer SLAs for on-schedule shipping. Based on the performance of these two departure times, they need additional capacity to reduce the incidence of shipping orders late to schedule.

These results indicate that the performance of the current state is not sufficient to meet demand at the current SLA for two out of the six departure times, with an additional two departure times that occasionally ship orders late to schedule. Consistent late shipments at these departure times are likely due to lack of capacity rather than last minute preparation or unexpected variability for three reasons. First, prioritization is controlled automatically, with priority increasing as orders get closer to departure time (see Section 2.2.2 for reference) which minimizes last minute preparation. Second, of all the departure times, only departure time 1 gets a high volume of orders within the last few hours prior to departure time (see Figure 19). The high volume at this departure time is consistent week after week, so unexpected variability has little impact on its volume. Lastly, since departure times 1, 2, and 6 are in the same shift (night shift) and have the same capacity, their consistent and occasional late shipments indicates a lack of capacity. This is accentuated by the better performing departure times, 3, 4, and 5, only have occasional orders shipped late to schedule. Without a discernible difference in arrival volume (see Figure 19), it is likely that the day shift is better staffed for demand than the night shift.

The results bring into question whether reducing the SLA is a viable option to improve the customer experience. While orders late to schedule do not mean that they are late for the customer, decreasing the SLA setting will exacerbate the situation and increase the risk of orders late deliveries at customers' doorsteps. Process improvements must achieve a significant cycle time reduction to consider decreasing the SLA setting.

4.2 Process Improvement Analysis

The process improvement efforts to reduce cycle time and variation start with identifying the process or processes to analyze and modify. With six distinct activities that directly affect cycle time, it is important to apply efforts at the right location to obtain the largest benefit. The analysis of the individual activities, based on lean waste identification, indicates that the sorting area, where tote wrangle and rebin take place for multis orders, provides the largest opportunity for improvement.

This section describes the selection process for the areas of process improvement.

4.2.1 Process Cycle Times at Outbound Operations

Outbound operations consists of four or six activities in sequence that process customer orders from the time of their arrival to the time of their departure. The factor that determines the number of activities is the order type. Singles orders go through four activities while multis orders go through six. Section 2.2.1 describes these activities in detail. Each activity consumes a certain amount of available time, and each activity must be completed prior to starting the next activity. The time each activity takes to complete decreases as orders approach departure time. Figure 8 illustrates the sequence of activities and the relative cycle time differences between the activities for multis orders. The width of the bands are proportional to the length of the activities' cycle time.

The longest cycle time in Figure 8 is the time between applying the shipping label and departing in a truck, which corresponds to the top layer in the figure (thickest layer). However, this activity is where packages get delivered to the dock automatically via conveyors and then get placed into the trucks by associates. Although there is a lot of waiting time involved in this part of the process, it is the end of the outbound processes, and it is irrelevant if a package waits hours or minutes for the truck to depart. At this point of the process the orders are ready to leave and there are no other value added activities to perform.

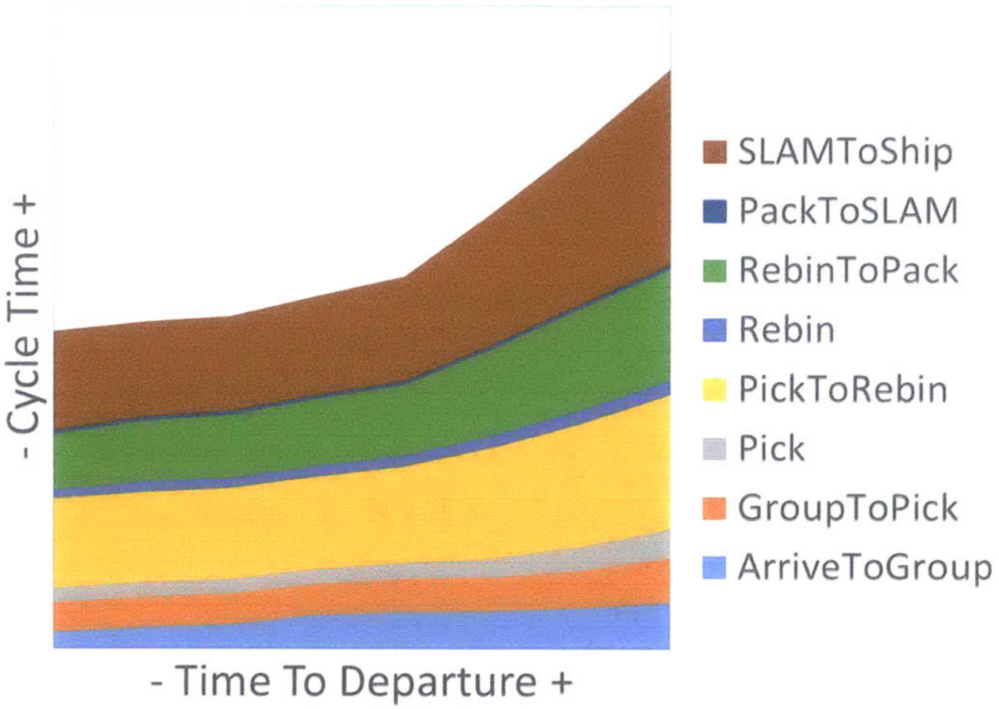


Figure 8. Outbound operations cycle time per activity in a multis order.

The second longest cycle time in Figure 8 is the time after pick and up to rebin. At the end of pick, the totes filled with order items are placed in the conveyor system, which delivers them automatically to the right location in the FC to get sorted into batches, then delivered to the rebin queue to wait for an associate to become available so that they can rebin the batch. Although some of the cycle time for this activity is attributed to movement on the conveyor system, it still has a longer cycle time than the next longest activity, rebin to pack, even after compensating for the time spent on the conveyor.

The long cycle time of this portion of the process can be attributed to two conditions of the processes. The first one is that grouping totes into batches strongly depends on the pick activity.

A multis batch has the items from its group of orders spread all over the FC. Different pickers pick the different items at different times, which means that the totes carrying the items are placed in the conveyors at different times and arrive to the sorting station over a period of time. Although the average time to pick all the items in a batch declines as the departure time approaches, the associates at the sorting area wait a significant amount of time for all the totes from one batch to arrive. To compensate for the time it takes for all totes from a batch to arrive, associates in this area sort through numerous batches in parallel. The second condition that characterizes the long cycle time of the activities between pick and rebin is that once a batch is completed it is transferred to a queuing area. Depending on the time of the day, batches may be snatched from the queue as soon as they become available, or they may wait in queue for an associate to become available. These two characteristics contribute to waste in the form of (a) resources waiting for units and (b) units waiting for resources (see Section 3.2 for details), which makes this section of outbound operations attractive for process improvement.

4.2.2 Process Steps at Sorting Areas

The current state of the sorting area involves a series of activities consisting of tote wrangling and rebin. Section 2.2.1 contains details on the sequence of operations. Figure 9 depicts the schematic for the current layout at the sorting area. The sequence of events for the current processes is as follows:

1. Totes arrive at conveyor buffer lane where they wait for an associate to process them. Totes from multiple batches arrive at random times, depending on pick activities.
2. An associate scans the tote with a handheld scanner, which tells the associate the buffer number in which to place the tote (see Figure 9 for an illustration of

buffer numbers and locations). The buffer numbers indicate the identification and location of the designated areas where totes are accumulated per batch. There can be one or more associates assigned to one of the multiple sorting areas.

3. The associate picks up the tote and moves it to the indicated buffer, where a batch cart accumulates all the totes in the batch.
4. The associate repeats steps 2 and 3 for all batches in their sorting area until one of the batches is complete.
5. When a batch is complete, the associate goes to a nearby shelf to pick up the paperwork for the batch and performs additional transactions on the scanner to close the batch.
6. The associate matches the paperwork to the batch and moves the cart over to the rebin area buffer. This is the end of the tote wrangling activity for the batch. The associate returns to the conveyor buffer lane to pick totes and repeat these steps from step 2.
7. The batch cart waits in the rebin area buffer until a rebin associate becomes available.
8. When available, a rebin associate grabs a batch cart from the rebin area buffer and moves it to their active rebin station.
9. The rebin associate scans the batch into the computer station.
10. The associate scans an item from the batch cart, and the rebin computer indicates the bin on the rebin cart where to be place the item.
11. The associate places the item in the indicated bin.

12. The associate repeats steps 10 and 11 until all items in the batch are in the rebin cart.
13. The associate closes the batch and moves the rebin cart to the packing area. This is the end of the rebin activity. The rebin associate goes back to the rebin area buffer to grab a new batch cart and repeat these steps from step 8.

The time that these 13 steps take are indicated by the 'PickToRebin' and 'Rebin' areas of the graph in Figure 8, minus the time it takes the totes to arrive via conveyor after associates at Pick place them on the conveyors by the Pick area.

The steps targeted for process improvement are steps 4 through 8. These steps involve waiting for all the totes from a batch to arrive to complete a batch. The associate picks totes for multiple batches, therefore their utilization is high. However, from the point of view of an order, orders are sitting on a cart waiting for the remaining totes in the batch to arrive (step 3). The steps also involve batch carts waiting for a rebin associate to become available (step 7). Again, orders are waiting, this time for resources to become available. The overall cycle time of outbound operations benefits from reducing the time wasted waiting for totes and for associates.

It is important to highlight that improving these steps will only benefit multis orders since singles orders do not go through the sort activities. Because the ratio of multis items to singles items is significantly more than 1:1, process improvements for multis steps affect more items. Figure 10 shows the average cycle times for singles and multis orders from the graph on Figure 7. The plot shows that the average for all orders is closer to the average of the cycle time for multi orders. Cycle time reduction on multis orders would bring the overall average cycle time down towards the average cycle time for singles orders. Therefore, the limit of cycle time reduction for multi orders is the cycle time for single orders.

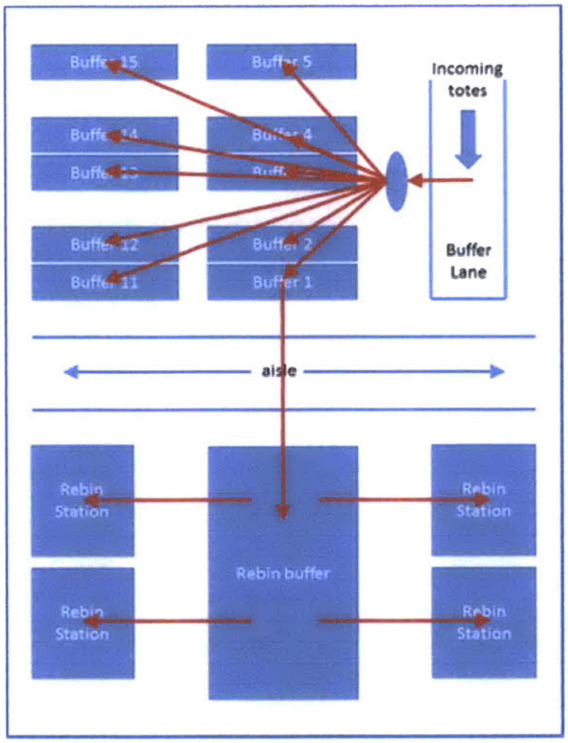


Figure 9. Schematic of the current layout for tote wrangling (upper area on schematic) and rebin (lower portion of schematic).

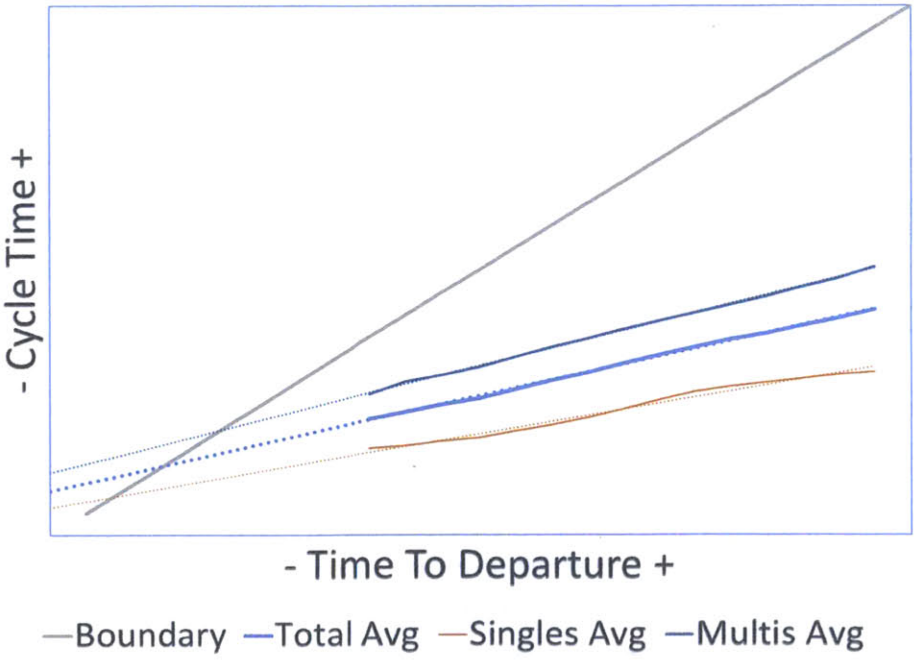


Figure 10. Average cycle times for single, multi, and all orders.

4.3 Capacity Analysis

Amazon's supply chain strategies at the FC level simplify the capacity limit for outbound operations. Section 3.1.2 identifies the three limiting factors for capacity as input, output, and process capacity. An FC will not get an order assigned unless it has the inventory to fulfill it, eliminating the possibility of an input-constraint. Outbound operations are also a make-to-order, meaning they do not start working on inventory until an order has been received. This eliminates the possibility of demand flow rate constraint, since they make orders to meet demand at the FC level. This leaves the maximum production capacity constraint as the scenario that limits an FC to meet its demand.

As another part of its supply chain strategy, Amazon builds its FCs to meet the expected demand during the end of the year, where Holiday shopping causes a surge on orders. The Holiday order surge is several times higher than the average order volume for the rest of the year, earning it its nickname, peak. The FC is built with enough equipment to have capacity for peak, and throughout the year it controls its maximum capacity with personnel.

This section discusses the analysis of the FC's maximum capacity based on past performance. Since the FC facility is equipped for peak demand, the analysis focuses on the number of associates and how many are needed to meet average demand excluding peak.

4.3.1 Data Collection for the Capacity Analysis

The data needed for capacity analysis involves the number of units completed in a given time frame and the number of hours used to complete those units. In this case, aggregating data by individual days helps visualize day to day performance. Since daily performance aggregates all orders whether they were completed on time or not, segregating the orders by departure time identifies the orders that are completed late. This data is pulled from the databases that Amazon uses to store data on its processes (see Section 4.1.1 for more information).

The following parameters help segregate the data spatially by day or per departure time:

- Date information – The month, day, week number, and day of the week provide multiple ways to group the data.
- Process – The data can be pulled by individual outbound process, which can be analyzed individually or collectively. The outbound operations available are pick, sort (tote wrangle and rebin activities combined), and pack.
- Quantity – The number of units scanned through each process in one day is critical co calculating capacity. In this case, units are individual items rather than orders.
- Incoming and Outgoing units – The number of units that arrive (orders placed) and depart (orders shipped) helps determine their rates and identify how the units accumulate at different times of the day.
- Total Hours – Each process takes a certain number of worker hours to complete. A worker hour is one hour of labor by one associate. When more than one associate work in one process, the total number of worker hours multiplies by the number of associates. For example, 100 associates working the same 10 hour shift provide 1000 worker hours in one shift.

Once this information is pulled, it is manipulated depending on the desired granularity. In addition, it is used to compute new parameters, like throughput, or how many units are completed per worker hour.

4.3.2 Current Capacity Performance at Outbound Operations

Capacity is defined in Section 3.1 as the measure which indicates the maximum product that a process can make in a given time period. For outbound operations overall, it is the number of units that are shipped every day. The daily capacity at outbound operations can then be obtained by combining the total units produced with the total number of hours invested in shipping the units each day.

However, since completed orders can wait in a truck for a long period of time (see Section 4.2 for more information), it is impractical to mark units as complete when the truck leaves for the purpose of this analysis. A more practical boundary is when an order is marked as packed. At this point in the process, the packages start moving through the conveyor system to get to its designated dock. The time accrued in the conveyor is automatic time and it is not attributed to associates. It is more valuable to evaluate how many units are packed (a simplified view of a complete order) and how many worker hours are invested in getting all the units from picking through packing operations.

The data selected for this analysis comprises daily unit outputs and total worker hours for a time span of 28 weeks. This time span is valuable for its consistency, which helps identify changes from this baseline. The data of the daily totals for units and worker hours are plotted on Figure 12 and Figure 13.

As the figures illustrate, there is a cyclical pattern that repeats every week. This oscillation is due to the demand variation throughout the week, which is heavier at the beginning of the week than at the end of the week (see Figure 11). Although there is a delay from the time orders arrive to the time that they are shipped, more than half of all orders are scheduled to ship within a day of their arrival. Thus the shipment output follows a similar weekly profile. It is

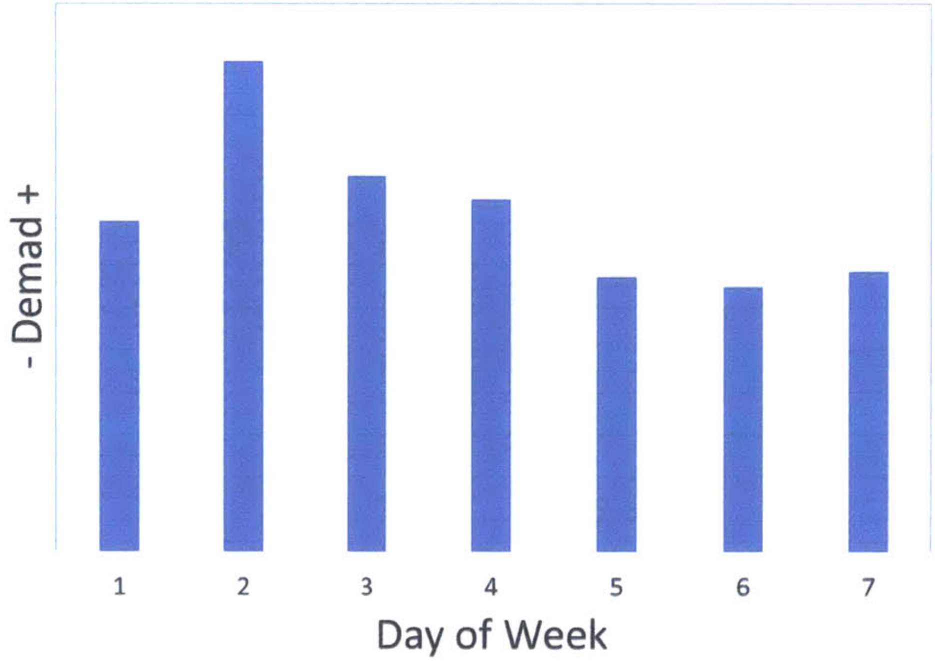


Figure 11. Example of incoming demand during the week. The first day of the week is Sunday.

important to note that the number of units completed include orders late to the scheduled departure time. These late orders are shipped at later departure times, in general either within the same day or early the next day of its original scheduled departure time.

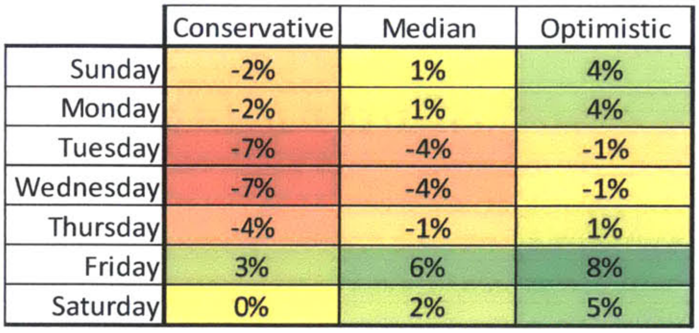
The throughput, or how many units are produced per time unit, is calculated by dividing the daily total number of units completed by the daily total number of worker hours used to complete the units. Typically, throughput is calculated on an activity basis, and, if further granularity is desired, on a machine basis. In this analysis, the throughput is obtained on an Amazon associate basis, referred to as worker during this discussion. The main reason for selecting workers as the basis for throughput is the low level of automation in the analyzed FC. The largest automation is in the form of the conveyor system, which automatically decides where to moves totes and boxes from their point of origin to their determined destination. The average time that items spend on conveyors is small when compared to the rest of the activities, making manual labor the predominant driver for the cycle time.

The plot on Figure 14 shows the throughput per worker for each hour worked. The throughput shows the same day-of-the-week cyclical variation as the daily totals for output and worker hours, but, regardless of the day of the week, the throughput has a coefficient of variation (standard deviation divided by the average) of 5% versus 25% and 24% for the daily output and daily hours, respectively (Figure 12 and Figure 13). This means that the dispersion of the data is much less for the throughput than for its individual constituents. This allows its use for analysis without segmenting the data by day of the week, simplifying the analysis. The plot also shows a trend that indicates the throughput has decreased slightly from the beginning of the sample time span to the end of the span. After normalizing the data to compensate for the declining throughput trend, a histogram plot of the data suggests that the underlying distribution for the throughput is slightly bimodal and skewed. The quantile-quantile plot confirms the distribution is slightly bimodal and slightly left skewed. Both plots are shown in

Figure 15. Given this information, the median rather than the sample mean is used for the capacity analysis to take the skew into consideration. In addition, the first and third quartiles are used for conservative and optimistic estimates, respectively. The conservative and optimistic estimates consider the days when the throughput is lower or higher than the majority of the time.

The number of worker hours needed per day to complete the actual number of shipped orders is then calculated by dividing the units per day by the median and the quartiles. This is necessary to gage how the estimated worker hours compare to the actual worker hours per day from the historical data. The results are summarized in Table 2. The data suggests that on average there is a capacity surplus on two of days of the week and a deficiency in two days of the week, with the other three days at the right capacity. When the throughput is lower than average, more worker hours are needed, while the opposite is true for higher throughput days. However, Fridays and Saturdays are estimated to at least have the right amount of worker hours, suggesting that there are more associates on these two days than needed.

Table 2. Estimates of worker hours relative to conservative, median, and optimistic worker hours. The estimates are calculated using conservative, median, and optimistic throughput calculations obtained from historical data. The red values indicate that more worker hours are needed, while the green values indicate that worker hours are available.



	Conservative	Median	Optimistic
Sunday	-2%	1%	4%
Monday	-2%	1%	4%
Tuesday	-7%	-4%	-1%
Wednesday	-7%	-4%	-1%
Thursday	-4%	-1%	1%
Friday	3%	6%	8%
Saturday	0%	2%	5%

This analysis sheds light on the results from the cycle time analysis (see Section 4.1 for more information). The cycle time analysis concluded that four out of six departure times have, at least, occasional orders that are late to scheduled departure time. The results from the capacity analysis indicate that there are certain days where more worker hours are needed. These results suggest that orders late to schedule are due to lack of capacity. If the process improvement efforts do not provide sufficient cycle time reductions, shifting associates from days with excess capacity to days with deficient capacity would provide better performance on the departure times that see orders late to schedule.

4.4 Chapter Summary

The analysis of the scheduled departure times reveals that the system state does not perform well with the current SLA setting in use. The cycle time analysis shows that four of the six departure times process orders late to schedule. It also demonstrates that the SLAs would have to be extended to improve their performance. The capacity analysis indicates that, based on average throughput, some days of the week lack capacity in the form of worker hours, while other days of the week have excess capacity. The analysis of outbound operations identifies the sorting operations, tote wrangling and rebin, to have the largest potential for cycle time reduction.

The next chapter discusses the process improvement efforts to reduce cycle time at the sorting operations. Based on the results of the cycle times, the chapter also completes the capacity analysis to identify the future state where the fulfillment center would ship all orders on schedule and at the required capacity if the SLA setting is shortened.

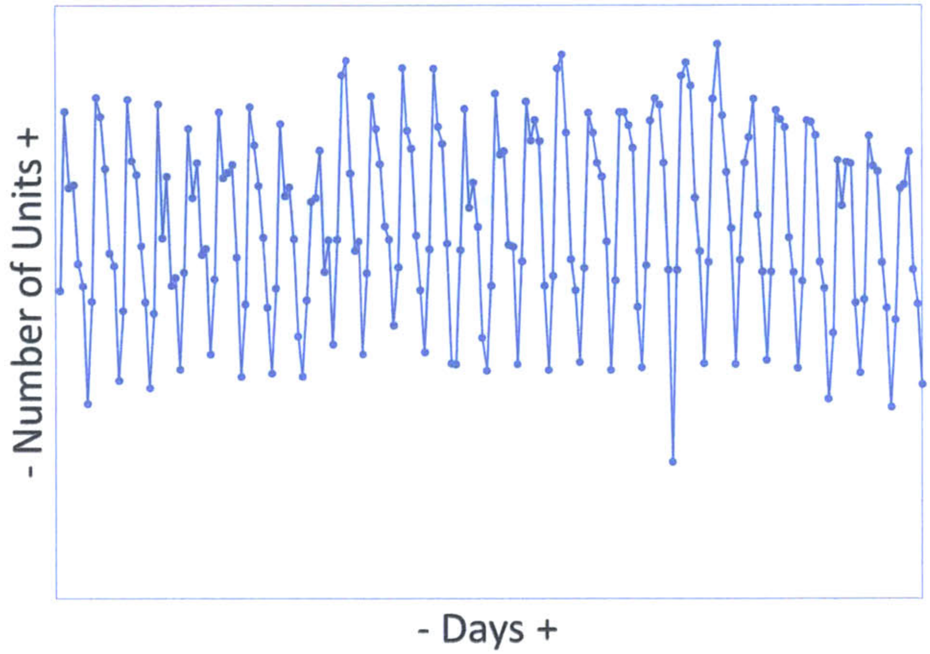


Figure 12. Daily total number of units for outbound operations over 28 weeks.

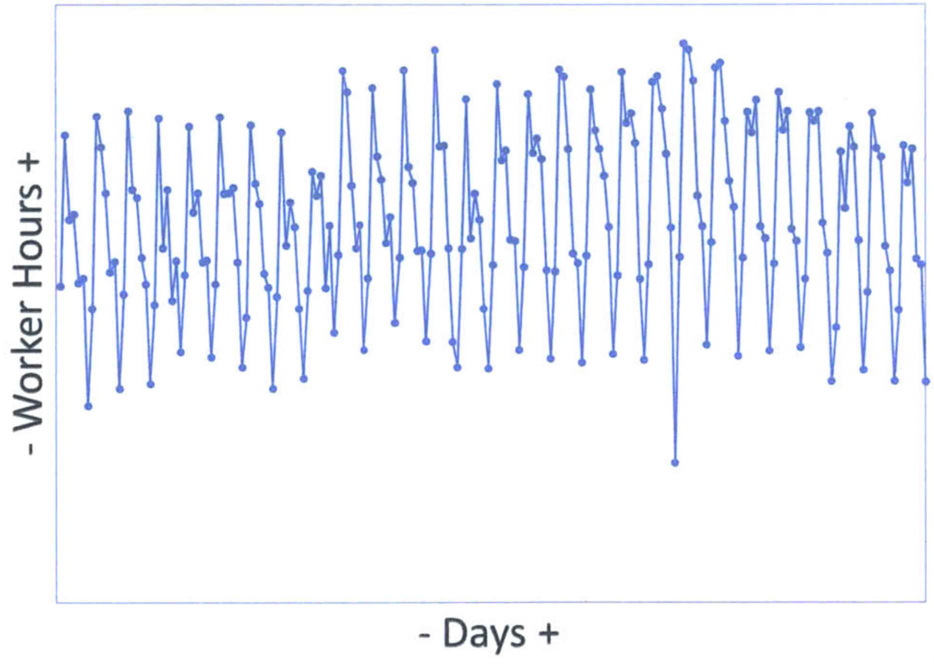


Figure 13. Daily total number of worker hours at outbound operations over 28 weeks.

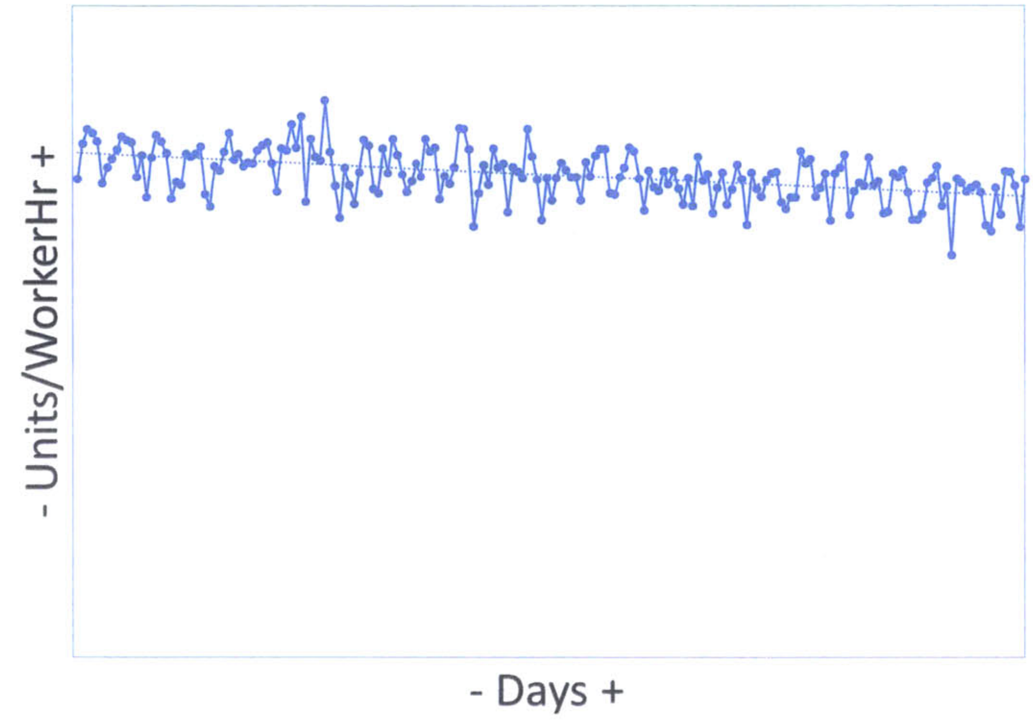


Figure 14. Daily total number of units completed per worker hour.

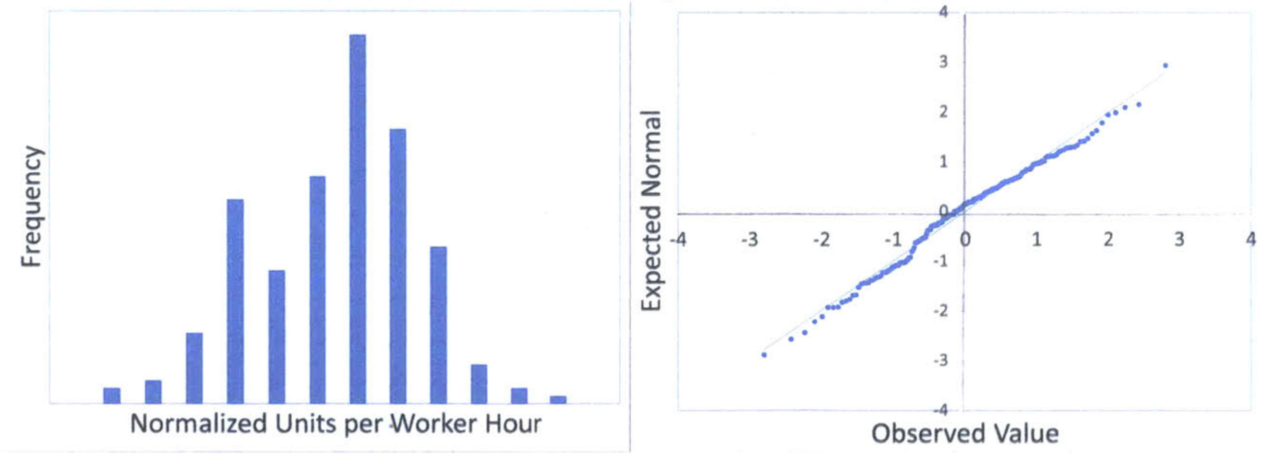


Figure 15. Histogram of the normalized units per worker hour (left) and the Q-Q plot for the data (right).

5 Future State Analysis

Based on the results from the current state analysis, the fulfillment center is not meeting all the required demand in terms of shipping orders as originally scheduled. With this issue in the current state, any reduction in the SLA exacerbates the current condition and would result in additional late orders. Cycle time improvements and capacity adjustments are needed to raise the performance level to better meet the current demand, and then beyond to use shorter SLAs.

The first move towards the future state, cycle time improvement, involves physical setup changes and changes in the sequence of operations. The end result of process improvement is a projected reduction of the overall cycle time by 6% if the changes are fully implemented throughout the entire shop and used 100% of the time. However, this cycle time reduction is not sufficient to significantly improve performance, therefore capacity adjustments are needed. Capacity calculations are performed using average and conservative estimates of how the units are processed and the number of units per worker hour calculated in Section 4.3.2. The capacity, in terms of worker hours, has to be increased a minimum of 1.96%, and up to 9.38% for a more conservative approach. To reduce the SLA from the current state, capacity needs to be increased up to around 13.79% for an SLA reduction of up to 50%.

These analyses help understand the overall performance of the FC, and provides the groundwork for further improvements in both cycle times and capacity.

5.1 Process Improvement Implementation Analysis

The analysis discussed in Section 4.2 indicates that the tote wrangling and rebin activities are attractive for process improvement efforts. The elimination or reduction of resources waiting for orders, and of orders waiting for resources, reclaims cycle time from non-value added activities, increasing throughput capacity. This section covers the modifications to the sort areas (tote wrangle and rebin) and their resulting performance.

5.1.1 Process Improvement Implementation Analysis

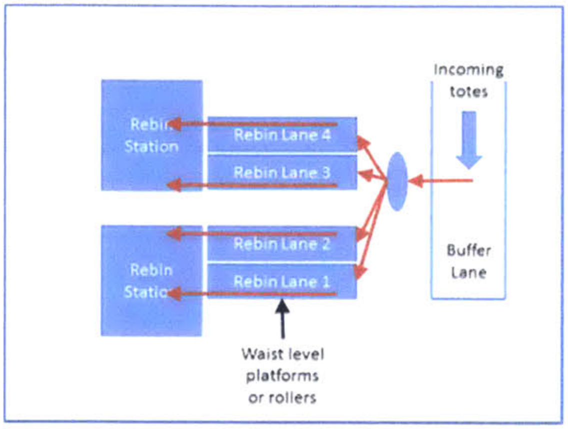


Figure 16. Schematic of the proposed layout for the tote wrangling and rebin combined operation.

The main goal for improving the tote wrangling and rebin activities is to eliminate or reduce the waiting time inherent to the current setup. One way to reduce waiting time during tote wrangling is to rebin totes as soon as they arrive at that station without waiting for the other totes in the batch to arrive. Further reductions are achieved if the totes do not have to be moved from the buffer lane to the batch cart and then to the rebin cart. Physical modification of the sort area helps achieve these reductions.

A reconfigured station setup that combines the tote wrangling and rebin activities into one (see Figure 9 for the current setup) provides the basis for cycle time reduction. The reconfiguration places the rebin station closer to the conveyor buffer lane and have the totes feed directly into the rebin station. The reconfigured layout is shown in Figure 16. The time waiting for all the totes to arrive decreases when totes are processed as they arrive while the remaining totes from the batch are still on their way. In addition, distances are shortened so the associates travel less during their shift. Finally, since rebin begins as soon as the first tote from a batch arrives, the time waiting for rebin associates in the current setup is also reduced. Based on the reconfigured station, the new sequence of events is as follows:

1. Totes arrive at conveyor buffer lane where they wait for an associate to process them. Totes from multiple batches arrive at random times, depending on pick activities.
2. An associate scans the tote with a handheld scanner, which tells the associate the rebin lane in which to place the tote. There can be one or more associates assigned to one of the multiple sorting areas.
3. The associate picks up the tote and moves it to the rebin lane. The rebin lane can be either a platform or a small conveyor at waist height to move or roll the totes towards the rebin station. The associate can alternate between picking totes from the conveyor rebin lane or rebin the items in the totes that have arrived.
4. When the first tote of a batch arrives, the associate grabs its paperwork from a nearby shelf and matches the paperwork to the batch.
5. The rebin associate scans the batch into the computer station.
6. The associate scans an item from the rebin lane, and the rebin computer indicates the bin on a rebin cart where to be place the item.
7. The associate places the item in the indicated bin.
8. The associate repeats steps 10 and 11 until all items in the tote are in the rebin cart. The associate also has the ability to pause rebin and pick more totes from the conveyor buffer lane to keep the lane clear, then continue rebin.
9. The associate repeats steps 2 through 9 until a batch is complete.
10. When all the totes in a batch have arrived and been placed in the rebin cart, the associate closes the batch. This is the end of the proposed process.

These sequence of steps result in a minimum of 15% cycle time reduction from the current process (see Figure 17). This process improvement only affects multis orders. When scaled up to all orders, the overall cycle time reduction is 6%. The cycle time reduction is calculated with the following equation:

$$\%CT\ reduction = \frac{current\ CT - proposed\ process\ CT}{current\ CT} * 100\%$$

where CT is average cycle time and the current and proposed process cycle times refer to the average cycle times from the time items are assigned to pick to the time the items are through the end of the rebin process. The overall cycle time reduction is calculated with the same equation but using the average cycle times from the time the orders arrive to the time the orders are through the end of the pack process. To compensate for the differences between multis and single orders, the average cycle time was decomposed into the average cycle time for multis and the average cycle time for singles based on the number of multis and single orders. Since the average cycle times for multis and singles are based on the number of orders from each type, they are proportional to the contribution that each order type has on the overall cycle time (see Figure 10 for an illustration of the multis, single, and overall average cycle times). These two averages are then used to calculate the contribution of multis orders to the overall average cycle time,

$$\%multis = \frac{multis\ avg\ CT - overall\ avg\ CT}{multis\ avg\ CT - single\ avg\ CT}$$

The average cycle time for multi orders for the proposed process is calculated by subtracting the minimum cycle time reduction, obtained from the numerator in the %CT reduction equation above,

$$\begin{aligned}&\mathit{proposed\ process\ multis\ avg\ CT}\\&\qquad= \mathit{multis\ avg\ CT} - (\mathit{current\ CT} - \mathit{proposed\ process\ CT})\end{aligned}$$

which allows the calculation of the overall cycle time reduction,

$$\begin{aligned}&\mathit{proposed\ process\ overall\ CT}\\&\qquad= (1 - \%\mathit{multis}) * \mathit{singles\ avg\ CT} + \%\mathit{multis}\\&\qquad\quad * \mathit{proposed\ process\ multis\ avg\ CT}\end{aligned}$$

Finally, the overall cycle time reduction is calculated with the following equation:

$$\begin{aligned}&\%\mathit{proposed\ process\ overall\ CT\ reduction}\\&\qquad= \frac{\mathit{current\ overall\ CT} - \mathit{proposed\ process\ overall\ CT}}{\mathit{current\ overall\ CT}} * 100\%\end{aligned}$$

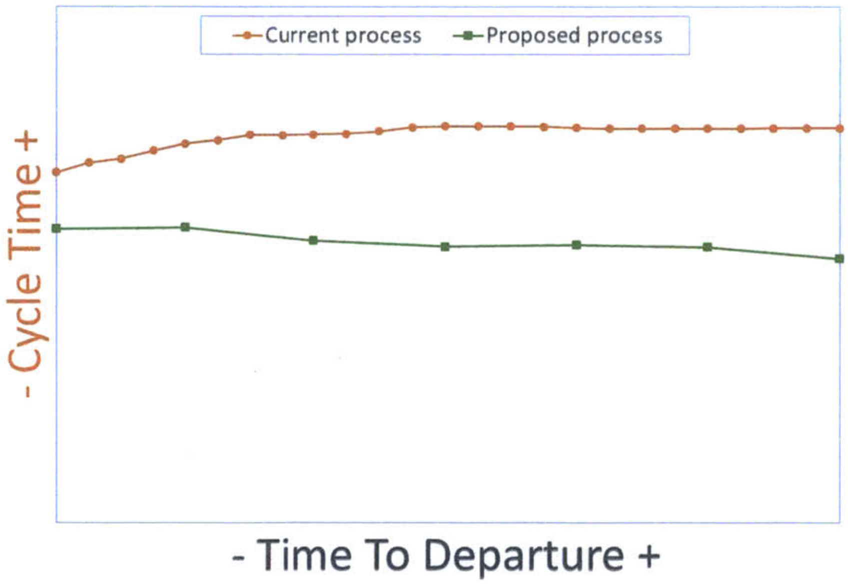


Figure 17. Plot of cycle time for current sorting operations and proposed process.

Although the process improvement achieves a 15% reduction in cycle time for these two activities, the overall cycle time reduction for operations is only 6%. There are two reasons for the lower impact. First, the cycle time reduction is part of only two of the five activities in the overall process for multi orders. A 15% cycle time reduction from the current cycle time for tote wrangling and rebin is a smaller reduction from the overall cycle time from order receive to order ship. And second, the distribution of single and multi orders is not even, therefore this cycle reduction does not impact 100% of the orders. The former dilutes the benefit when scaled up to the full cycle time. The latter further dilutes the benefit because not all orders are multi orders.

Taking these results into account, the cycle time reduction increases daily throughput capacity by a maximum of approximately one percent. This estimate assumes that this process change is implemented in the entire FC and that it is in use 100% of the time. The estimate are based on the performance of the last few hours[2] prior to departure time, taking into account the average cycle times calculated in Section 4.1.2 and the remaining units to ship in the specified time frame. These last few hours are selected because they are the most influential for late orders. Orders that arrive prior to these hours are less likely to ship late. The number of estimated worker hours is calculated by multiplying the number of orders left to ship by the average cycle time (calculated in Section 4.1.2) for the departure times with the largest number of late shipments (see Table 1),

$$current\ worker\ hours = \#\ of\ orders\ left\ to\ ship * avg\ CT/order$$

This value is then compared to the worker hours for same number of units but at the reduced cycle time (current cycle time reduced by 6%). Since the reduced cycle time allows the same

[2] The number of hours used in the analysis is undisclosed due to their proprietary nature.

number of orders to be processed in less time, the difference in worker hours between the current process and the proposed process translates to additional orders when divided by the highest average cycle time within the last few hours prior to departure time,

$$extra\ orders = \frac{current\ worker\ hours - reduced\ worker\ hours}{avg\ CT/order}$$

The added capacity would be able to reduce number of days with late orders by at approximately 66% (see Figure 18 for reference). This optimistically assumes that historically the days with late shipments were constrained by capacity and did not have any other issues. Other issues affecting on-time shipment include misplaced packages and machine breakdowns.

5.1.2 Process Improvement Limitations

Given these estimates, the process improvement would have a significant impact on the number of late orders for the fulfillment center, leaving only the days with the largest amount of late orders. However, these remaining days account for 76% of all late orders in the analyzed time period. Therefore, even if the process changes are implemented on the entire FC and used 100% of the time, the cycle time reduction will not be sufficient to improve the current state and reduce the SLA. In addition, the process changes have other physical and operational limitations. The combination of insufficient capacity and process limitations require alternative means to address capacity.

The physical limitations stem from the physical sort area layout in the fulfillment center. The design of the FC gave a specific amount of floor space to these activities, making the floor spaces physically constrained by conveyors, columns, aisles, and equipment. Within the available area, the process changes discussed in this chapter would require more space than available. The current setup allows more than a dozen batches to be at tote wrangle simultaneously, while the new layout allows no more than a dozen rebin lanes in the same space.

This would make full implementation impractical given the volume of totes in process at any given time.

The operational limitations are less impactful than the physical limitations. The software system that controls the tote wrangling and rebin operations is not setup for the process modifications. The software has safeguards that prevent rebin prior to processing the entire batch through tote wrangle. The software needs to be modified to allow rebin while tote wrangle is in process.

Given these results and limitations, the best application for the proposed process changes is its use on a limited basis, like a few hours prior to departure time, to speed up throughput at the most critical time for the orders. Given the physical limitations, the modifications could be applied to a few areas rather than applied to the whole FC, and the software could be modified to allow rebin during tote wrangle only at the specified times and only at the modified locations. Even so, the capacity limitations need to be addressed by alternative means.

5.2 Capacity to Support Current and Future State

The analysis so far points to the need for worker hours at key times of the day. The cycle time analysis indicates that the current state has late orders on four of the six departure times. Any reduction in SLA would exacerbate the situation. The process improvement analysis shows that, even though the number of days with late orders would be significantly reduced, the reduction in cycle time is insufficient to meet customer expectations. The capacity analysis shows that, on average, the calculated throughput indicates times during the week where extra capacity is needed and where there is a capacity surplus. These results point towards incrementing worker hours.

This section first looks at the days with late orders and the capacity needed to eliminate the late orders. Afterwards, the analysis shifts focus to the capacity needed for a shorter SLA.

5.2.1 Capacity for On-Time Orders at Current SLA

The analysis in the previous section indicates that process improvement on the activities with the largest opportunities for cycle time reduction have little impact on the number of late orders in the current state. Since cycle time reduction is insufficient to eliminate late orders, additional worker hours are needed to increase throughput. Based on the historical performance for late orders and the throughput per worker, the analysis for the current state indicates that at least 1.96% and no more than 9.38% additional work hours are needed to meet average demand.

5.2.1.1 Assumptions

When the data for late orders is analyzed, it reveals that the majority of days have at least one late order. Some of the late orders may be for reasons other than throughput limits, but the data does not provide this level of detail. Therefore, the data is filtered to look at the days with a number of late orders greater than 0.2 % of the daily output to simplify the analysis. This filter reduces the number of days with late orders from more than half to less than half in the analyzed time frame. With this filter in place, the days with late orders are extracted from the total unit output data to work out the percentage of late orders per day and the additional capacity needed to eliminate late orders.

Figure 18 shows the cumulative percentage of days with late orders, relative to the total units processed per day, for every day that had late orders. The first half of the days with late orders account for 20% of the volume of all late items, with the second half accounts for the remaining 80%. The improvement efforts need to focus on the percentage of late units per day to have a significant impact on reducing late orders. The capacity increase needs to be sufficient to reduce the per-day percentage of late units by 40% to impact more than 80% of the number of late units.

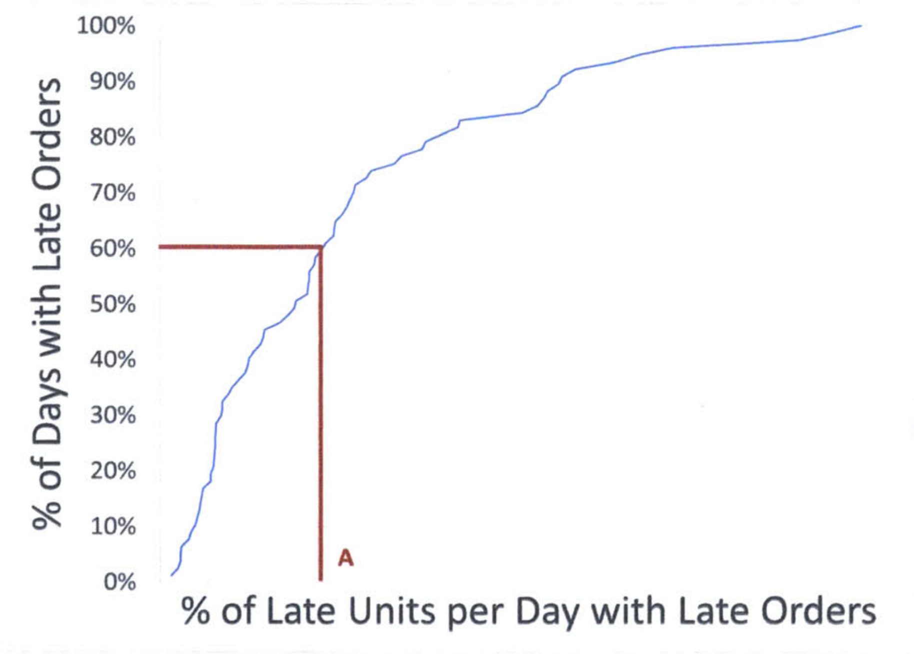


Figure 18. Cumulative percent of late units for each day with late orders. The red lines illustrate an example at 60% of the days with late orders. On 60% of all the days that ship late orders, there are no more than A% of units shipped late to schedule. The A% is the percentage of units that ship late out of all the units that need to be shipped on that particular day.

The data for daily orders, both total orders and late orders per day, is limited by the way the data is segmented. There are six departure time on most days (see Table 1 for a summary), and the data daily data does not have the granularity needed to identify which departure times are responsible for late orders. On the other hand, the data per departure time gives an hourly breakdown of all incoming and outgoing items up to 24 hours prior to departure, and the total number of orders arrived and shipped prior to the last 24 hours. However, the data per departure time does not provide information on when the late items are shipped, just the number that are late. Given the characteristics of both data sets, they will be used to generate several assumptions.

The first assumption is that all late orders are shipped within the same calendar day of their original departure time. This assumptions simplifies the complexity of how outbound operations handles late orders, and is reasonable for this analysis given that most orders are scheduled to be shipped at most with a 2nd day delivery. This means that most late orders will be rescheduled to ship within the following 24 hours after their original departure time to make an on-time delivery to the customer.

The second assumption involves shifts. The shift structure at an FC is complex and involves multiple schedules throughout the week. Regardless of the different shifts, the main break down between shifts is day and night shift, each with 10 hours per shift, and each in charge of three of the six departure times involved in this analysis. The second assumption is then that the shifts are simplified into day and night shifts, with a fraction of the maximum available time of 20 hours per day. The resulting fraction considers the total working time of 20 hours per day and then takes away mandatory breaks and time spent in daily meetings.

The last two assumptions involve capacity calculations. Capacity calculations are approached in two ways. The first method assumes that orders are processed evenly through the day. Although this method simplifies the actual rate of incoming orders (see Figure 19 for an example of the incoming order profile), it provides an optimistic approach that captures the minimum capacity increase needed. The second method assumes that the late orders on any given day only have the SLA as the time available to process those orders. In reality, the orders will be processed at different times throughout the day, and assuming that late orders start getting processed at SLA provides a conservative approach that captures the maximum capacity increase needed.

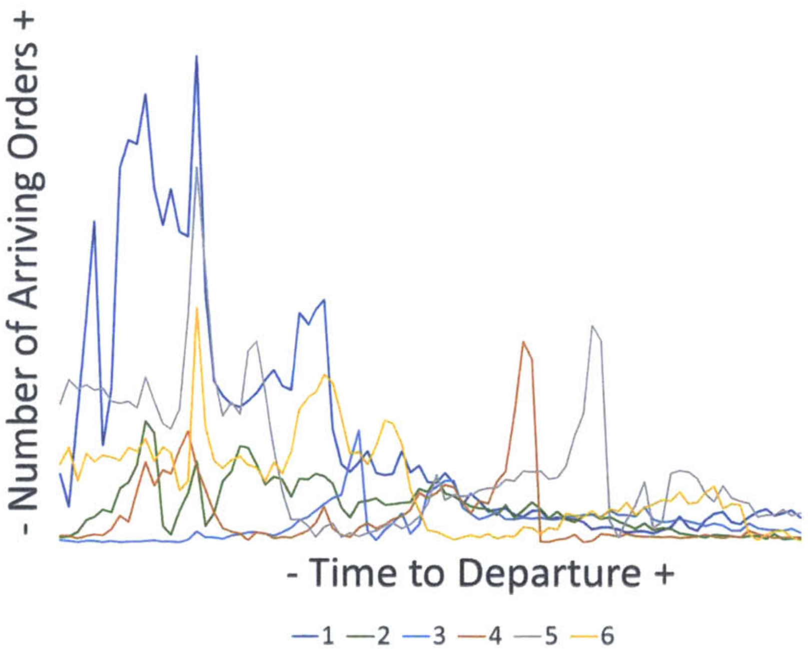


Figure 19. Example of incoming orders during the last 24 hours prior to departure time up to SLA. The colors and numbers represent the individual departure times indicated on Table 1.

5.2.1.2 Analysis

Based on these assumptions and using the average units per worker hour obtained in Section 4.3.2, the estimates for minimum increase in capacity (in worker hours) is between 1.96% and 6.03%. These values consider the conservative and average throughput and ignores the optimistic throughput. The optimistic throughput is ignored because it is likely to happen a maximum of 25% of the time. These results are calculated by dividing the number of completed and late orders by the conservative and average throughput to obtain the needed worker hours, and then comparing the resulting hours to the ones from the daily data set.

The estimates for the maximum increase in capacity (in worker hours) is between 7.14% and 9.38%. These results are obtained by first calculating the required throughput rate to process the late orders in the available time dictated by the SLA. This throughput rate is then divided by the conservative and average throughput rates from Section 4.3.2 to obtain the number of additional workers needed to process the late orders during the SLA. Finally, the number of workers are multiplied by the time available per worker to obtain the additional hours needed, which are then added to the actual hours used per analyzed day.

Table 3[3]. Example of late orders per departure time and day of the week based on shift start day. This table shows the same data as in Table 1 after the departure times are rearranged by day shift (top three departure times, 3, 4, and 5) and night shift (bottom three departure times, 6, 1, and 2). Departure times 1 and 2 start on the day indicated and roll over to the next day. For example, on Sunday, departure times 1 and 2 are actually departure times on Monday, but they are completed by associates starting their shift on Sunday night and continuing through Monday morning.

	SUN	MON	TUE	WED	THU	FRI	SAT
3	🟩	🟩	🟩	🟩	🟩	🟩	🟩
4	🟩	🟩	🟩	🟩	🟩	🟨	-
5	🟨	🟩	🟩	🟨	🟨	🟨	🟥
6	🟥	🟨	🟧	🟧	🟧	-	-
1	🟨	🟧	🟧	🟥	🟨	-	-
2	🟩	🟨	🟩	🟨	🟩	-	-

Both of these calculations are for day to day operations, and do not take the shifts into consideration. The data needs to be broken down into the departure times and then to the corresponding shift, day or night, to identify the first distribution level for the worker hours. Analysis of the data indicates that the number of late items is not evenly distributed amongst the shifts. Table 3 shows the same data shown in Table 1 but rearranged to align departure times within a shift. The top three departure times shown in Table 3 are for the day shift while the bottom three are for the night shift. The color coding corresponds to the number of late orders, with green indicating zero or few late orders to red indicating the maximum amount of late orders for the weekly data. Visual inspection reveals that the night shift has more late orders than the day shift. Closer inspection of the number of late orders breaks them down as follows: 20%

[3] The data used to build this table is undisclosed due to its proprietary nature.

of all late orders occur during the day shift, with the remaining 80% occurring during the night shift. When combined with the conservative and average capacity calculations, the additional worker hours needed to ship orders on time are broken down into day and night shift to distribute the added capacity accordingly.

5.2.2 Capacity for On-Time Orders at Reduced SLA

The analysis in the previous section determined the percent increase needed for the current state to ship orders on time. This section follows up the preceding analysis by calculating the necessary capacity to reduce the SLA from its current setting.

For the minimum capacity estimates, the assumption that the units will be processed evenly throughout the day means that, regardless of the SLA length, the additional capacity will remain the same as calculated in the previous section. In reality, reducing the SLA will begin having issues when the SLA length approaches the minimum time required to process an order. As the SLA approaches this minimum processing time, orders will begin to wait for resources.

For the maximum capacity estimates, more worker hours are needed as the SLA is shortened. Using the current SLA as the baseline (100%), Figure 20 shows the effect of reducing the SLA. The graph illustrates the increase in capacity required to meet the shorter SLA. The relationship between the percent reduction in SLA and the capacity is not linear, and, incrementally, more capacity is required as the SLA gets smaller.

5.3 Chapter Summary

The analysis for the future state of the fulfillment center seeks to determine the throughput needed to reduce or eliminate late orders. Testing the process improvements opportunities identified in Section 4.2, the results from the process changes indicate that the new setup provides added capacity, but the level of improvement is not sufficient to handle the volume seen

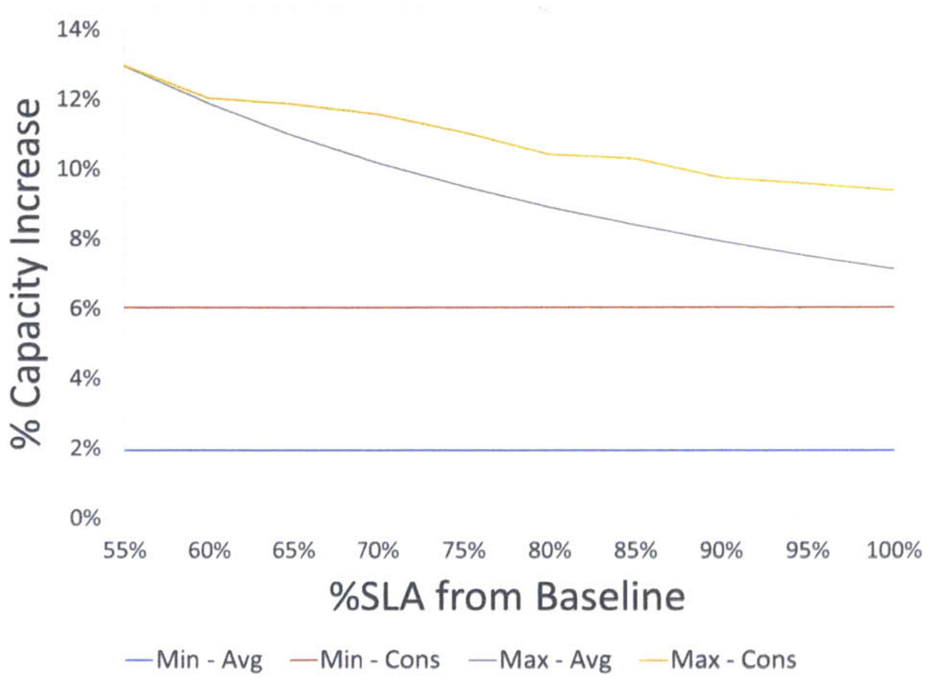


Figure 20. Capacity effect from SLA reduction.

by the FC. In addition, the limits to the process changes make it more suited for selective use that for plant-wide implementation. To address the capacity gap, an analysis conducted using historical data of the current state reveals that a capacity increase of worker hours between 1.96% and 9.38% is needed to reduce late items by at least 80%. If the FC decides to reduce the SLA from its current setting, then the capacity increase has to be higher than 9.36% and will depend on the desired SLA reduction.

The next chapter summarizes the conclusions obtained through the analyses performed on the fulfillment center. In addition, recommendations for future improvement efforts are discussed.

6 Conclusions and Recommendations

As an online seller, Amazon constantly works to improve the customer experience of its clients to provide anything anybody could want and at the best possible prices. The subject of this thesis stems from these efforts, focusing on the service level agreement (SLA), Amazon's on-time delivery promise to its customer, as a conduit to a better customer experience. By improving the performance at the fulfillment center, in this case through the reduction of the SLA, Amazon can improve the customer experience for its clients.

The analysis of the current state of operations reveals that the system does not perform well with the current SLA. The cycle time analysis shows that four of the six analyzed departure times process orders late to schedule. It also demonstrates that the SLAs would have to be extended to improve their performance. The capacity analysis indicates that, based on average throughput, some days of the week lack capacity in the form of worker hours, while other days of the week have excess capacity. The analysis of outbound operations identifies the sorting operations, tote wrangling and rebin, to have the largest potential for cycle time reduction.

The analysis for the future state of the fulfillment center determines the throughput needed to reduce or eliminate late orders. New setups stemming from process improvement analysis provide added capacity, but the level of improvement is not sufficient to handle the volume seen by the FC. In addition, the limits to the process changes make it more suited for selective use that for plant-wide implementation. To address the capacity gap not fulfilled by process improvements, an analysis conducted using historical data of the current state reveals that a capacity increase of worker hours between 1.96% and 9.38% is needed to reduce late items by at least 80%. If the FC decides to reduce the SLA from its current setting, then the capacity increase has to be higher than 9.36% up to 13.79%, depending on the desired SLA reduction from the current state.

. These analyses help understand the overall performance of the FC, and provides the groundwork for further improvements in both cycle times and capacity.

6.1 Recommendations for Improving the SLA

Based on the results of the analyses conducted for this thesis, the customer experience is best enhanced by first adjusting the staff levels to address the demand profiles, and then utilize process improvement benefits to complement the staff adjustments during higher demand periods at key points during the week. Furthermore, adjusting the SLA by individual departure times adds flexibility to the different departure times. These changes offset Amazon's expenses incurred when rescheduling late orders.

Before a reduction in SLA is implemented, the staff levels need to be balanced to meet the demand profile. Even though demand changes throughout the week, the daily demand profiles are similar in a day to day basis. The analysis of staff levels on a weekly basis revealed the need for more capacity in the night shift. With the exception of the weekend, where only a day shift is needed, the analysis of staff levels on a daily basis reaches the same conclusion. In addition, capacity can be further increased by combining staff adjustments with process improvement implementation at key points during the day. However, staff levels need to be increased to achieve a shorter SLA.

The FC needs to add even more capacity for shorter SLAs. This stems from the volume of orders that have to be processed from the time the SLA is triggered to the time the next truck departs the FC. In other words, the same volume of orders has to be processed in a shorter period of time, requiring more capacity. Given that the SLA setting applies to all departure times equally, the departure times with the highest volume will need the largest portion of the added capacity. An alternative would be to selectively change the SLA for each individual departure time. This alternative adds flexibility, allowing the FC to shorten SLA at departure times that already have the right capacity, while allowing other SLA lengths for departure times with high volume. With this flexibility, the level of capacity can be customized to address the needs of the FC and enhance the customer experience at a lower cost.

These recommendations would allow Amazon to meet their demand profile with customized added capacity at selected departure times. The flexibility added by these recommendations increase the throughput of outbound operations while reducing upgrades to late orders and improving the customer experience.

6.2 Future Research

The analysis followed in this thesis focused on system-wide performance rather than on individual activities and how they interact with each other. The benefit for this type of analysis is obtaining good estimates based on daily numbers, which are readily available to Amazon employees. Even though some of the analysis looked into finer details, like orders per departure time, more data is needed to break down outbound operations into its individual activities and how they interact with each other based on the time of the day.

One follow up to this thesis is breaking down the capacity analysis by activity. Data is available to understand how many units are processed through most activities and how many worker hours are invested in completing the activities. However, the data available is on a daily basis. The granularity of the time of day or even the shift needs to be extracted from Amazon's data warehouse to properly characterize and analyze each activity. The next step to this approach would be to break down each activity into hourly events, like incoming orders, outgoing orders, and number of items processed. This next step allows understanding the capacity requirements by hour for each activity based on how they interact with each other and how they interact with the next departure time.

Another area of opportunity is the picking operation relative to batches for multis orders. During the process improvement analysis, data for batch completion time revealed a distinctive pattern during batch picking. Most batches get assigned to pick as soon as they are compiled by the computer systems. However, the pick pattern showed that some items are assigned to pickers within minutes of batch compilation, followed by a gap where the remaining items are not assigned, and finally the last items in the batch are assigned to pickers again. This gap means a significant amount of waiting time, and it appears on most batches. With numerous factors affecting the pick process and batch assignment, further study would identify the reasons for this patter, and help determine different options for picking item assignment.

Although the results of the analyses in this thesis increased the knowledge base on outbound operations and how it performs at a system level, there are more opportunities to expand the knowledge base. The data made available through Amazon's data warehouse contains the key to understanding the intricacies of the activities within outbound operations. Using this work as a foundation, using the analyses at the activity level will pinpoint where capacity is needed.

6.3 Chapter Conclusion

Amazon's fulfillment center operations involve a complex interaction amongst activities, manual labor, automated and manual decisions, all depending on what order needs to be processed next. All these interactions depend the customer and when they want their delivery.

This thesis looked into the activities that prepare an order for delivery and what it takes to improve the experience for the customer in the form of the service level agreement. Although the original hypothesis theorized that process improvements are sufficient for reducing the SLA, the data and results demonstrate that more capacity is needed. The determined capacity needed to complete orders as originally scheduled reduces the risk of late order arrivals at the customer. The analysis also provides the basis for reducing the SLA from its current setting.

Further work on capacity requirements would improve the accuracy of the results presented in this thesis by analyzing the individual activities within outbound operations and their interactions with each other and the departure times. The schedule priority system used by the FC determines which orders get processed when, and their breakdown by activity would improve the knowledge base for capacity requirements at the fulfillment centers.

Bibliography

[1] "Amazon.com Investor Relations: Annual Reports and Proxies." [Online]. Available: http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-reportsannual. [Accessed: 22-Feb-2015].

[2] "ICQA description," *Amazon Careers*. [Online]. Available: http://amazon-operations.co.uk/business-areas/profile?profile=5. [Accessed: 27-Jan-2015].

[3] G. Cachon, *Matching supply with demand: an introduction to operations management*, 3rd ed. New York, NY: McGraw-Hill, 2013.

[4] G. G. Vining, *Statistical methods for engineers*, 3rd ed. Boston, Mass: Cengage Learning, 2011.

[5] D. Bertsimas and R. M. Freund, *Data, models, and decisions: the fundamentals of management science*. Belmont, MA: Dynamic Ideas, 2004.