Optimal Staffing Recommendation for Inbound Operations

by

Carla Li-Carrillo
B.S. Chemical Engineering, Stanford University, 2012
B.A. Art Studio, Stanford University, 2012

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

MASTER OF SCIENCE IN MECHANICAL ENGINEERING
AND
MASTER OF BUSINESS ADMINISTRATION

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

JUNE 2017

Signature of Author _____ Signature redacted

Department of Mechanical Engineering, MIT Sloan School of Management
May 12, 2017

Certified by _____ Signature redacted

Stanley Gershwin, Thesis Supervisor
Senior Research Scientist of Mechanical Engineering

Certified by _____ Signature redacted

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

Accepted by _____ Signature redacted

Rohan Abeyaratne PhD.
Chair, Mechanical Engineering Department Committee on Graduate Students

Accepted by _____ Signature redacted

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

This page is intentionally left blank.

Optimal Staffing Recommendation for Inbound Operations

by

Carla Li-Carrillo

SUBMITTED TO THE DEPARTMENT OF MECHANICAL ENGINEERING AND THE MIT SLOAN SCHOOL OF MANAGEMENT ON MAY 12, 2017 IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREES OF

MASTER OF SCIENCE IN MECHANICAL ENGINEERING
AND
MASTER OF BUSINESS ADMINISTRATION

Abstract

Amazon inbound operations are staffed following a 'staffing-to-charge' model in which labor is planned to match the incoming volume capacity required by the weekly Sales & Operations Planning (S&OP) forecast. Staffing-to-charge is a lean model of staffing that attempts to maximize labor utilization by minimizing the possibility of a labor surplus or deficit. However, due to inaccuracies in the S&OP freight forecast, poor visibility into incoming inventory, and last minute staffing changes, it is often the case that labor capacity is not adequately aligned with the actual unit receipts. This leads to additional labor costs and network inefficiencies.

This project explored the current staffing policies and current system constraints such as forecast accuracy, backlog management, and hiring schedules to understand the scope of the problem. From these findings, an alternate method for staffing, known as 'Level loading,' was proposed. Level loading consists of staffing to a known and consistent headcount every day of the week with the intent to reduce staffing costs and labor capacity variability. Level loading was found to improve the efficiency of inbound operations, leading to considerable costs savings for the distribution center. The project also created an optimization model that allows Fulfillment Center managers to plan the transition from their current shifts to level loading; Amazon's Production Planning Team will implement this model by mid-2017. To fully achieve the benefits from level loading, the system requires a change in the planning of incoming freight. In particular, the incoming freight should be scheduled and planned according to a known labor capacity, as set by the level loading policy. This change to freight planning is currently being investigated.

The study found that delayed restocking of the network is a costly inefficiency, similar in magnitude to the cost from excess labor capacity. To mitigate this, a labor plan that allows for greater capacity is necessary. The cost savings of more effective inbound operations offsets the additional labor costs of such a plan. The findings of this study are based on an Amazon warehouse, but a staffing model with greater labor capacity can be applied to inbound operations at any distribution center.

Thesis Supervisor: Stanley Gershwin
Title: Senior Research Scientist, Department of Mechanical Engineering,
MIT

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

This page is intentionally left blank.

Acknowledgements

This thesis would not have been possible without the help and support of many people throughout this process.

I would first like to thank my internship champion, Brian Donato, for the amazing opportunity to work at Amazon. My Amazon experience would never have been as fun and exciting had it not been for your help and encouragement along every step of the way. I would also like to thank Brian Urkiel and Diego Mendez de La Luz for being amazing company hosts and always making sure that I had all the resources I needed.

I would like to thank my supervisor, Julian Robinson, for the opportunity to visit and explore different areas of Amazon's intricate supply chain. Thank you for letting me visit whichever office or site I was curious about, regardless of how remote the location. I would also like to extend a big thank you to my team in Amazon Operations Finance: Pierre Dunkel, Tanmay Ramaiya, and Aditya Prathipathi, for always helping me get all the answers to my questions.

Additionally, I would like to thank Gabriella Mnyshenko and Brayden Billbe for being my first team at Amazon and my patient teachers as I tried to understand the labor planning processes. Special thanks to Niranjan Venkataramani for the many hours spent discussing S&OP forecasts with me. A big thank you to Sarah Holmes, Julian Martinez, Apoorva Prasad, Cameron Nelson, and the rest of the Network Operations Center for never failing to answer even the weirdest of hypothetical scenarios. Thank you to Matt Conlon and Zahir Papar for always being available to shred my whitepapers apart and engage in serious discussion about vendor management. Thank you to my fellow Ruby 6th floor interns: Khatia Chitashvili, Freddy Revah, and Baris Sevinc for being my daily brainstorm and teatime crew. Thank you to the MIT Sloan Amazon intern cohort for making my summer in Seattle a truly memorable experience. Finally, thank you to the many other incredible people at Amazon that I had the opportunity to work alongside.

I am grateful to the LGO alumni who form the LGO community in Seattle for their advice and guidance throughout the internship process. I would like to extend a special thank you to Naomi Arnold, Alberto Luna, and Guillermo Pamanés for not only being my role

models, but also my 'secuaces' in the RGB crew, making Seattle a true home away from home.

I am very grateful for the opportunity to call Professors Stephen Graves and Stanley Gershwin my thesis advisors. Professor Graves, thank you for all the time spent providing me guidance. Professor Gershwin, thank you for your unparalleled attention to detail. I would also like to acknowledge the Leaders for Global Operations Program for its support of this work.

I would like to thank my LGO classmates, my fellow commiserates in this experience, my support system whenever this endeavor seemed too overwhelming. A special thank you to Jordan Hoffmann, Sailashri Parthasarathy, and Zachary Stauber for the endless group banter that made it feel like I never left Cambridge in the first place.

Last, but certainly not least, I would like to thank my sister Ana and my parents; I owe all my accomplishments to you. Thank you for always believing in me and supporting all my crazy endeavors.

Table of Contents

Abstract 3
Acknowledgements 6
Table of Contents 8
List of Figures 10
List of Tables 11
Glossary 12
1. Introduction 13
1.1. Company Background 13
1.2. Problem Statement 14
1.3. Project Approach and Goals 16
2. Labor Planning at Amazon 18
2.1. Constraints in Labor Planning 18
2.2. Forecast Dependency 20
2.3. Headcount Calculations 21
2.4. Levers for Adjusting Labor Capacity 22
2.5. Inefficiencies of Labor Mismatches 24
3. Literature Review 25
3.1. Labor Staffing Practices for Supply and Demand Matching 25
3.1.1. Chase Demand 25
3.1.2. Level Capacity 26
3.2. Forecast Accuracy Metrics 26
3.3. Process Design and Execution 29
4. Ideal State Analysis 30
4.1. IXD Operations 32
4.2. IXD Constraints 34
5. Current State Analysis 36
5.1. Vendor Lead Time Variability 36
5.2. Inbound Forecasting 36
5.3. Backlog Management 39
5.4. Predicted vs. Actual Labor Capacity 41

5.5. Diminished Associate Experience 43
6. Backlog and its Consequences 45
6.1. Background into Backlog 45
6.2. Selecting a Representative Sample 46
6.3. Cost Calculation for Different Backlog Thresholds 48
6.4. Opportunity Cost of Missed Orders 50
7. Level loading Feasibility Study 54
7.1. Scenario 1: Current State 54
7.2. Scenario 2: Maximum Required Daily Capacity 57
7.2.1. 100% Daily Required Capacity 57
7.2.2. 65% Daily Required Capacity 59
7.3. Scenario 3: Average Required Daily Capacity 61
8. Implementation 66
9. Conclusion and Recommendations 69
9.1. Optimal Staffing Freight Mix 69
9.2. Automated Scheduling and Buying by Optimal Freight Mix 71
Bibliography 72

List of Figures

Figure 1. Process Flow Chart for Item Receiving at IXD Sites 15
Figure 2. Day-of-the-Week Curve for Incoming Units 20
Figure 3. Headcount and Forecasted Units per Day of the Week 21
Figure 4. Amazon Flywheel 30
Figure 5. Ideal Timeline for Inbound Operations 31
Figure 6. Ideal Timelines for Inbound Operations with IXD Implementation 34
Figure 7. WMAPE for Select FCs, 03/13 - 04/03 38
Figure 8. NACF Network WMAPE, 03/13 - 04/03 38
Figure 9. Percent Units processed to forecast 39
Figure 10. Headcount vs. Backlog 40
Figure 11. Units Arriving and Backlog at the IXD 41
Figure 12. Forecasted vs. Actual Work Hours 42
Figure 13. Planned vs. Actual Hires 44
Figure 14. Histogram of Top Selling 80,000 SKUs as a Percentage of total Retail Sales 47
Figure 15. Distribution of Item Types Within Missed Sales due to Backlog 51
Figure 16. Backlog Management given Staffing-to-Charge Model 55
Figure 17. Staffing-to-charge vs. Level loaded processed units 57
Figure 18. Backlog Management given Level loading to Maximum Required Daily Capacity per Week 58
Figure 19. Backlog Management given Level loading to 65% of the Maximum Required Daily Capacity per Week 60
Figure 20. Staffing-to-Charge vs. 7-day Average Level-Loaded Processing Capacity- .62
Figure 21. Backlog Management given Level loading to the 7-day Average Forecasted Capacity Required 63
Figure 22.Backlog Management given Level loading to 118% 7-day Average Forecasted Capacity Required 64
Figure 23. Headcount Changes as a Result of Level loading Implementation 68

List of Tables

Table 1. Simplified Shift Structures at Select Amazon IXD Site 19
Table 2. Weekly Processing Averages 41
Table 3. Service Levels for Different Backlog Thresholds 49
Table 4. Percentage of SKUs that face Stock Outs for Different Backlog Thresholds ... 49
Table 5. Population of SKUs within/ under the desired Service Level according to instances of stock outs. 52
Table 6. Cost of Backlog Calculations 53
Table 7. Costs of Backlog per Additional Day at Specific Site 53
Table 8. Arrived Units vs. Received Units for the week of 5/29/2016 55
Table 9. Cost Summary for Current Scenario 57
Table 10. Forecasted Required Capacity, Week of 4/17/2016 57
Table 11. Financial Summary of Different Level loading Scenarios Involving Maximum Daily Capacity Required 61
Table 12. Level loading based on 7-day average forecasted capacity, Weeks of 4/17 – 6/19 62
Table 13. Financial Summary of Different Level loading Scenarios Involving Variations of the 7-Day Average Forecasted Capacity 63
Table 14. Optimal Freight Mix Percentages 70

Glossary

- **FC:** Fulfillment center; the Amazon denomination for a distribution center.
- **IXD:** Inbound cross-dock site within the Amazon fulfillment network, meant to serve as a hub for vendor freight, within the greater hub-and-spoke system. This type of site only receives and reroutes freight, it does not serve as a storage facility.
- **Destination FC:** An FC that serves as a warehouse for inventory storage. For the purposes of this investigation, any FC that receives vendor freight from an IXD site will be denoted as a Destination FC.
- **NACF:** North America Customer Fulfillment; the network of fulfillment centers that serves customer orders within North America.
- **Associates:** Hourly workers at an FC.
- **Inbound:** Inventory items coming from vendors into the FC and the NACF network.
- **Received units:** Units that are processed by the site and thus considered to have entered into the Amazon network.
- **Arrived units:** Units that arrive at the site on a particular day. These units can either be received by the site or become a part of the backlog.
- **Assigned units:** The number of units that the site is expected to receive based on the forecast.
- **VLT:** Vendor lead time, the time between Amazon placing an order and the vendor delivering the order at the FC.
- **PO:** Purchase Order.
- **OT:** Overtime work hours required of FC associates.
- **VTO:** Voluntary time-off; the flipside of OT used for periods of very low volume to avoid idle workers.
- **COGS:** Cost of goods sold.
- **Freight mix:** The nature of the size and weight of the units arriving at the FC.

1. Introduction

1.1. Company Background

Amazon strives to be Earth's most customer-centric company where people can find and discover virtually anything they want to buy online. By giving customers more of what they want - low prices, vast selection, and convenience - Amazon continues to grow and evolve as a world-class e-commerce platform.

In order to fulfill customer orders within the time frame promised to the customer, Amazon has built a robust network of 85+ Fulfillment Centers (FCs), which compose the North America Customer Fulfillment network (NACF). Historically, vendors had a 3-day lead-time, starting from receiving a Purchase Order (PO) from Amazon, until shipping units to the FC where Amazon needed them stored. This system was problematic as it produced variability in vendor lead times (VLTs) from vendors shipping items to different locations nationwide.

To mitigate this problem, Amazon introduced inbound cross-dock (IXD) sites to simulate the hub-and-spoke system. In the hub-and-spoke model, material flow enters the system at critical hubs. Material is then consolidated into larger flows that move into and out of strategically placed spoke nodes. This consolidation of flow allows the operator to benefit from economies of scale by maximizing both the utilization of resources at every critical node and the truck utilization between hubs and spokes.[1]

Inbound operations at Amazon encompass any incoming vendor freight that enters the Amazon fulfillment network to be stored in FC, ready for customer orders. In this system, IXD sites are the critically located nodes; IXDs are located at strategic geographic locations and designed to receive vendor freight from nearby vendors. Items are then sorted in terms of their destination FC and shipped to their respective sites in the network in order to achieve the desired inventory placement. The IXD network allowed Amazon to have greater ownership of the supply chain of incoming products across the nation. Consolidation of items by destination FC enables greater truck utilization within the network and internalizing transportation between FCs provides Amazon with greater influence in the travel time between nodes. This sped up the

process of receiving and restocking vendor freight as destination FCs no longer had to wait for vendor to ship their items around the country; rather shipments are shipped under a time frame that is now within Amazon's control. Currently, IXD sites serve only certain regions, but the plan is to eventually switch to hub-and-spoke for the entire network.

As of December 2016, Amazon had six IXD sites to serve its network. The typical IXD site receives and injects around 1 million units daily into the network and employs around 2400 hourly associates, with 800 associates per shift. The key challenge for IXD sites is to ensure there is always a balance between available labor and incoming freight. Site leadership must make sure that there is enough labor to unload, receive, and ship the items that come to the IXD site each day in order to ensure timely injection of items into the NACF network. Alternatively, because labor is the main variable cost of the site, site leadership attempts to ensure that labor is always occupied in order to maintain low labor costs per unit processed.

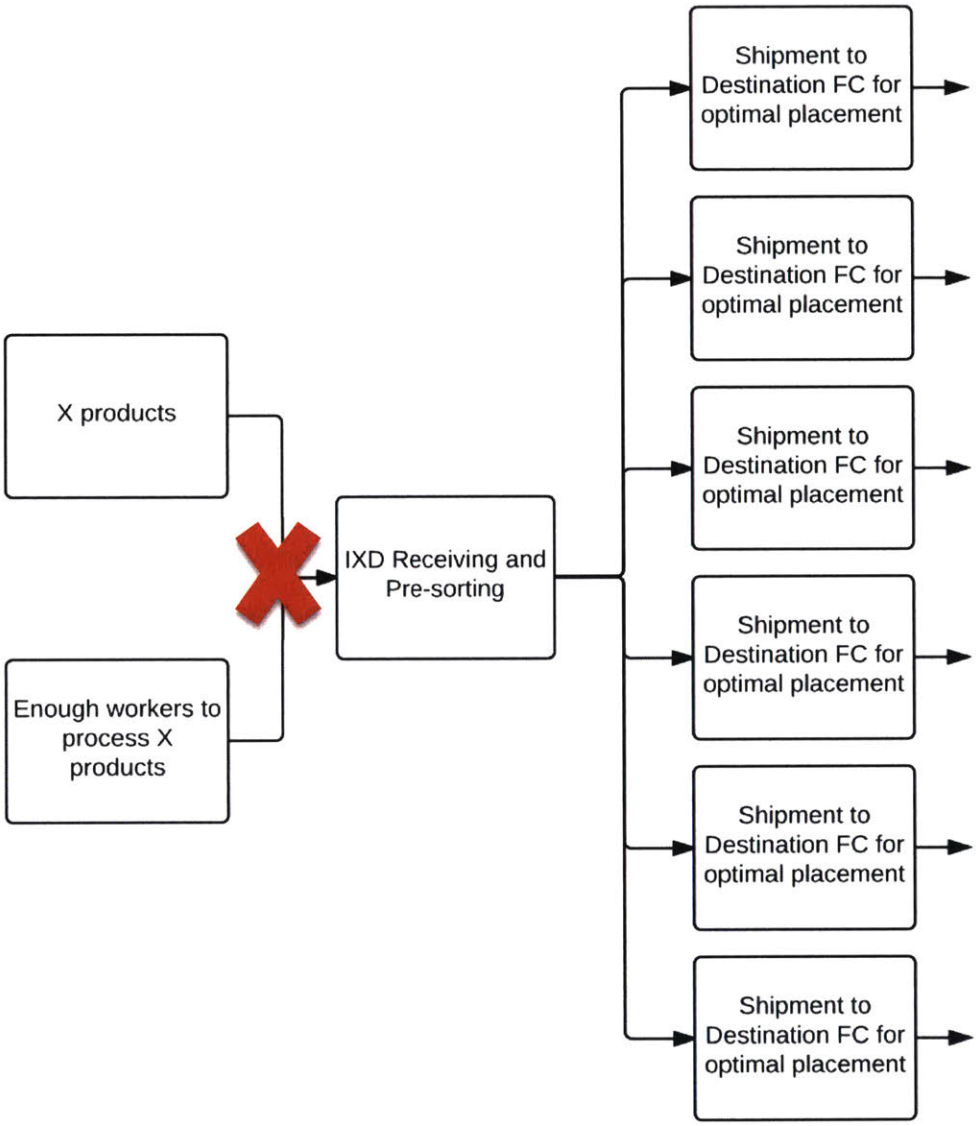
As the sortable network continues the transition into a hub-and-spoke system, with inbound volume entering the network through IXD sites exclusively, it is critical to improve the capacity and rate of inbound operations.[1]

1.2. Problem Statement

While the IXD network was designed to ensure timely injection of vendor freight into the NACF network, increasing instances of backlogs at IXD sites prevent this from occurring smoothly. This problem arises when IXD sites receive a number of units that is mismatched with their labor capacity (Figure 1).

[1] For a detailed explanation of IXD implementation, refer to Section 5.

Figure 1. Process Flow Chart for Item Receiving at IXD Sites



Currently, inbound operations are staffed following a 'staffing-to-charge' model in which labor is planned to match the inbound capacity required by the weekly freight forecast. Staffing-to-charge is a lean model of staffing that attempts to maximize labor utilization by minimizing the possibility of a labor surplus or deficit.

However, due to inaccuracies in the freight forecast, poor visibility into incoming inventory, and last minute staffing changes, it is often the case that labor capacity is not

adequately aligned with the actual receipts. This leads to additional variable costs in the form of overtime and not enough associates taking voluntary time off when requested.

Labor mismatches further lead to additional costs in logistics due to additional network inefficiencies. Such inefficiencies include the cost of suboptimal placement of inventory that leads to orders being shipped from FCs that might be further away from the customer, the costs of unplanned shipment upgrades to meet the promised delivery times, and finally the cost of a missed customer promise.

While these additional costs are necessary to ensure the network is replenished under the current system, a more robust solution is necessary. As Amazon transitions towards 100% of volume entering the network through IXDs, it is essential to adopt a staffing model that allows the network to operate smoothly (or efficiently). Most importantly, it is critical to control the flow of products from IXDs into the network at an optimal processing rate for destination FCs. Finally; it is critical to provide a solution that is conducive for the sustainable growth of the NACF network.

Since labor is the main variable cost of every FC, we develop a new way to set staffing levels so as to address these network inefficiencies.

1.3. Project Approach and Goals

This project was a study of the current state of Amazon labor planning, specifically inbound operations, in order to identify areas of opportunity. For purposes of scope, this investigation has focused on the inbound operations of one IXD site in particular. The specific site was chosen as a result of its proximity, receive volume, and vendors' willingness to cooperate in pilot studies.

The goal of the project was to meet the following objectives:

1. **Understand the current state of Amazon labor planning and freight allocation practices in inbound operations.**

In order to make an improvement, it was important to understand how the current scenario deviates from an ideal scenario. This required an understanding of the current labor planning practices.

Because work schedules require advanced notice, a thorough investigation into labor planning and forecasting was essential. More importantly, part of this exploration included a thorough understanding of what the constraints are when hiring and scheduling associates.

2. **Identify the root causes for labor mismatches and the consequences of this to the NACF network.**

A thorough understanding of the process for labor planning revealed that the root cause for labor mismatches was a systemic problem that involved additional organizations within Amazon, other than the labor-planning group. Thus potential causes such as freight forecasting, labor practices, FC utilization, and vendor relations were explored.

Additionally, the consequences of labor mismatching were explored. Besides the cost of excess labor, incoming inventory backlog was found to be an immediate consequence of labor mismatches. Thus, the phenomenon of backlog and the consequences of backlog to the entire network were investigated.

3. **Develop new system for staffing**

Different scenario analyses were performed based on the knowledge gained from labor planning, freight forecasting, and backlog implications. The scenario analyses explore the monetary costs of staffing and backlog implications for processing freight with different labor capacities.

The hypothesis for the different scenarios was that while seemingly more cost intensive, excess labor capacity actually leads to a net positive opportunity cost.

4. **Recommend a new staffing model that is more cost-effective and conducive to sustainable growth.**

Based on the results of the scenario analyses, it was possible to suggest a better staffing model. From this recommendation, a few implementation strategies were suggested, as well as potential projects that could further improve the system in the future.

2. Labor Planning at Amazon

Staffing requires a delicate balance between determining the number of associates necessary and timing this perfectly to account for hiring and training. Labor planning is an important element of capacity planning as labor is the main variable cost for an FC.

2.1. Constraints in Labor Planning

There are three main constraints to labor planning:

- **Hiring Lead Times**

Once a decision to hire an associate is made, there is on average a one-week period where the associate must be recruited. The length of the recruiting period is highly dependent on the labor market of the FC location.

Once hired, associates undergo a one-week training on the operations and different functions of the FC. Because there is a learning curve to the new functions, associates are not considered "locked" into the headcount until they have undergone 10 days of work. Therefore, there is a 3-week lead-time between a hire request and the associate being considered in the count of total work hours.

- **Shift Structures**

The FC is open 24 hours a day, 7 days a week, and associates are scheduled to work the same 10 hours, four days a week. When an associate is hired, he or she is placed in a shift that is set to work a certain combination of four days of the week (E.g.: Monday through Thursday OR Monday, Wednesday, Friday, and Saturday). This shift lasts ten hours (excluding one hour for breaks and lunchtime) and can take place either during the day, typically 6am to 5pm, or at night, typically 6pm to 5 am (The FC has two hours of downtime between shifts, during which handovers take place). Therefore, associates work a total of four 10-hour workdays, for a total of 40 hours a week. An additional fifth day is assigned to every shift in case of overtime, and this fifth day is completed as a 10-hour shift during the respective day or night shift. Table 1 shows an example of the weekly shift structures at one of Amazon's IXD sites: the letter D or N denotes whether this is a day or night shift; days shaded in black stand for days when the associate is not

required to work. The subsequent letter or number denotes the particular combination of four days of the week. In this example there are 19 shifts, each of which has an assigned crew of associates. The size of the crews for a shift range from 10 to 100 associates.

Table 1. Simplified Shift Structures at Select Amazon IXD Site

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
D8	■	■				■	
N8	■	■				■	
DA					■	■	■
NA					■	■	■
DB	■	■	■				
NB	■	■	■				
DC	■			■			■
NC	■			■			■
DF	■					■	■
NF	■					■	■
DH			■	■	■		
NH			■	■	■		
NJ				■		■	■
DK	■	■					■
NK	■	■					■
DL				■	■	■	
NL				■	■	■	
DN		■	■	■			
NN		■	■	■			

To provide stability for associates, these four days do not change week-over-week, and if overtime is ever assigned, it is always assigned on the same day of the week. This creates the constraint of very specific headcounts for specific days. In the event that 10 associates work on Tuesdays, and 5 work on Wednesdays; the headcount cannot be changed to 7 next Tuesday and 8 next Wednesday. The only instance in which an associate can change shifts is if he or she makes a request to transfer to a different shift.

- **Term of Contract**

Amazon hires temporary associates for 3-month periods, and full-time associates under at-will employment, meaning that either the employer or the employee can end the relationship at any time. Amazon cannot simply downsize the number of hired associates at any moment depending on need. Therefore, one of the key factors in labor planning is to ensure that before any hiring decision is made, that the shift in question is

truly in need of extra capacity, and that this will continue to be the case for at least the next 3 months.

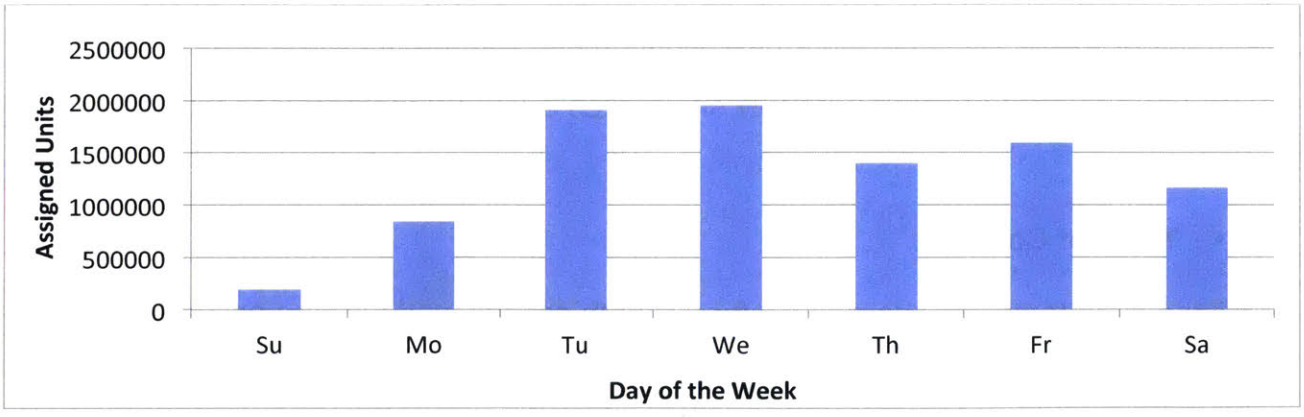
2.2. Forecast Dependency

Amazon's current staffing strategy determines the number of associates necessary at the site based on the forecasted number of items that the site is to receive day by day, for the upcoming week.

Amazon uses a combination of historical demand data and purchasing forecasts in an optimization model that allocates different units of volume to be sent to different locations across the NACF network. The optimization model produces a weekly total volume for each FC and IXD to receive, but this is not granular enough to determine daily headcounts.

To break down the volume into daily values, the total volume is broken down according to the "day-of-the-week" curve. The day-of-the-week curve is an estimate of volume allocations per day based on a 4-week average of historical data. The day-of-the-week curve looks similar for most sites, with the majority of the freight arriving between Wednesday and Friday, and lower volume scheduled to arrive from the weekend into Tuesday. Figure 2 illustrates this trend in arrivals for the studied site.

Figure 2. Day-of-the-Week Curve for Incoming Units



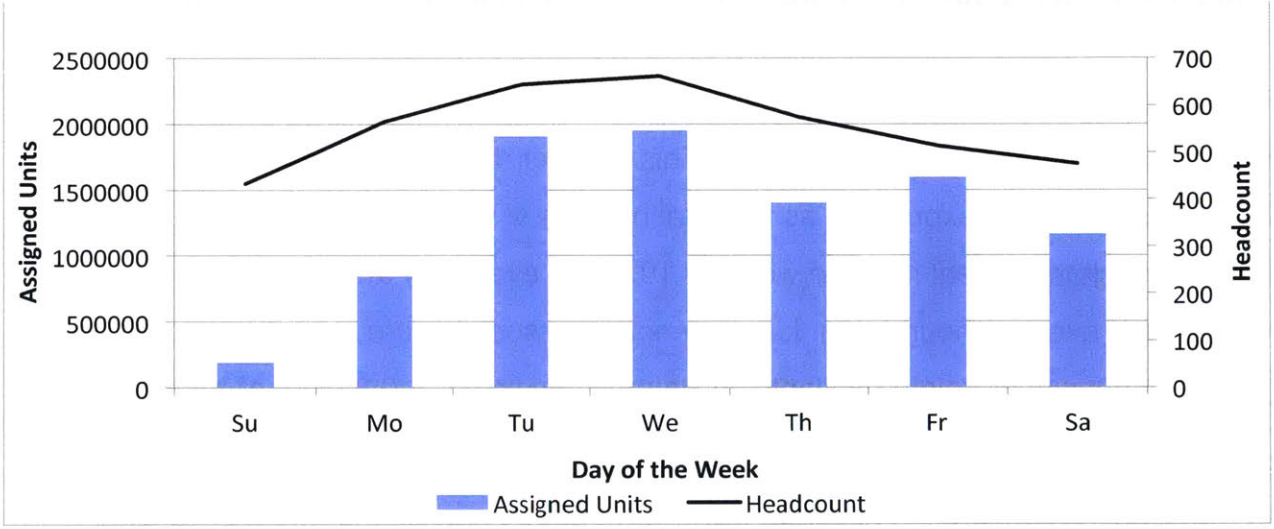
The forecast encompasses more than 10 upcoming weeks in order to foresee any hiring trends in the horizon. However, in practice, the forecast is only reliably accurate for the upcoming 3-6 weeks. For this reason, and because staffing changes require recruiting

and training for several weeks in advance, it is re-released every week and all further planning is adjusted on a weekly basis. For more detailed analysis of inaccuracies in the weekly forecast, refer to Section 5.2 Inbound Forecasting.

2.3. Headcount Calculations

Currently, Amazon uses a reactionary model of staffing known as "staffing-to-charge". Figure 3 shows the headcount relative to the number of assigned units for the staff in question. The contrast shows the staffing-to-charge strategy and how it aims to staff exactly the number of associates required to process the forecasted number of units, having capacity come as close as possible to the necessary number.

Figure 3. Headcount and Forecasted Units per Day of the Week



Because labor is the main variable cost for any FC, staffing-to-charge is a desirable model for staffing if the forecast is accurate; ideally, it minimizes any mismatch between labor and units by staffing to exactly the necessary headcount, thus maximizing labor utilization. The main downside of staffing-to-charge is that it leaves very room for error in the event of a unit surplus, and often leads to procurement of OT hours.

Because staffing-to-charge leaves very little buffer for error, FCs make assumptions regarding associates in order to ensure that the required number of associates is present every day. Some of these assumptions are:

- **Rate:** The hourly rate of units processed is based on historical averages of how fast the average-performing associate can process a certain number of units per hour. This rate is highly dependent upon not only the volume of units available to process, but also the nature of the units. Smaller units, such as books and DVD's, will be faster to process than larger, heavier units that require special handling such as furniture and TVs.
- **Attrition:** The projected decrease in the number of hired associates. This might be due to ending of temporary employment, or end of the employment-at-will contract by either party.
- **Attendance:** The number of associates present at the FC after any absences reported that day.

These assumptions are used to calculate the number of associates required to process the forecasted number of units arriving. This number determines the headcount required for that day which is expressed as the total number of work hours available given that every associate present is set to work for 10 hours every day and that they presumably work at this average hourly rate. Ideally, enough associates are scheduled so as to process the units of freight available, while maintaining a 0.5-0.8 days of backlog.[II]

2.4. Levers for Adjusting Labor Capacity

In the event of a mismatch between the current labor capacity (the number of work hours available to process a certain number of units) and the volume requirements dictated by the forecast, given the headcount is locked a week prior to execution, the FC can introduce last-minute levers to adjust the existing labor capacity:

- **Labor sharing**

Modeled after the Toyota Production System, FCs have direct and indirect functions. Direct functions are functions in which associates work directly with the flow of products into and out of the FC. Indirect functions are functions in which associates play a support role for the direct functions to run smoothly and continuously.

[II] The choice of backlog is further explained in Section 6.1 Background into Backlog.

FCs sometimes cross-train associates on direct and indirect functions. This is done to create flexibility in terms of capacity so that in the event of a labor mismatch, associates can be rearranged between direct and indirect functions. This way, the FC can alter the available labor capacity without the need to change the number of associates hired by the FC. For instance, an associate who normally works as a water spider (function based on the Toyota Production System function of the same name[2]) can be reassigned for a day to help with inbound if there is a labor shortage in inbound.

- **Overtime (OT)**

Associates are scheduled to work 10 hours, 4-days a week. However, in the event that labor capacity is insufficient associates can choose to work extra hours and earn higher wages for those hours past 10 hours that day or 40 hours that week. This usually occurs when the incoming backlog is greater than 1 day.[III]

If the need for overtime hours is too great to fill with volunteers, the FC can institute mandatory overtime, requiring an entire shift of associates to work an extra 10-hour day, also at the higher wage.

- **Voluntary time-off (VTO)**

Conversely, during periods of excess capacity, associates can choose to stop working, forfeiting the wages they would have earned during the hours they are not working. It is a form of unpaid time-off. The downside of VTO is that the FC can offer VTO, but associates can choose not to take it, in which case the FC is still forced to employ and pay for the extra labor capacity.

- **Freight Redirects**

In the event that OT is not sufficient to keep the backlog within manageable delay, the FC can escalate the issue. An escalation causes any vendor freight units currently en route to the FC to be redirected to a different IXD FC. The decision of where to reroute is not only based on proximity, but most importantly on available labor capacity at the receiving FC.

[III] For further detail on backlog management, refer to Section 6. Backlog and its Consequences.

This lever is an inconvenience not only to the receiving FC, which must have an unexpected additional labor capacity to process these units, but also to the network as a whole. These manual redirects undo the placement allocations previously determined by the forecasting model. Now shipments from vendors have to traverse a longer distance, increasing the VLT, cost of transportation to the IXD FC, and cost of transportation to any subsequent destination FCs.

2.5. Inefficiencies of Labor Mismatches

The objective in labor planning according to staffing-to-charge is to achieve the right balance of hourly associates to avoid fluctuations in the headcount on a daily basis. This is because any labor mismatch between number of units and labor capacity leads to unwanted labor costs.

Excess labor capacity translates into not enough work available and, therefore, idle workers. If the FC is unable to find more work for associates, or is unsuccessful in convincing associates to take VTO, the cost per unit processed increases, as less units are processed per associate hours paid.

Excess units translate into a growing backlog and the need for more labor capacity. Since more hiring is not an immediate solution, OT is the only lever available to avoid a growing backlog and delays of inventory being received into the network. Associates earn time-and-a-half hourly wages during overtime, leading to higher processing costs for units processed during overtime, and therefore a cost inefficiency. This cost is not inconsequential: during the first half of 2016 Amazon spent an additional $0.7 million on inbound overtime alone for one IXD site alone. Assuming all sites are operating similarly, this amounts to a total cost on the order of $6 million for inbound operations for all sites in the IXD network.

3. Literature Review

In order to provide an improved staffing recommendation, prior art was investigated for concepts relevant to this project. This included previous research on different practices for labor staffing, forecast accuracy measurements, and process design and execution for customer-centric businesses.

3.1. Labor Staffing Practices for Supply and Demand Matching

Balancing supply and demand when it comes to services is a challenge that is constantly faced by the service industry. While an Amazon warehouse could technically be considered a manufacturing plant due to its distance from direct consumer-facing functions, its performance dependency on labor does make its supply and demand matching challenges similar to those of the service industry.

There exist two main strategies for staffing when attempting to match demand and supply in terms of labor: Chase-demand and Level Capacity.[3]

3.1.1. Chase Demand

The strategy of chase demand consists of adjusting labor capacity to fit demand as closely as possible. Chase demand is the preferred strategy when the tasks at hand have been reduced to very simple actions that can be quickly mastered.[4]

Sasser, who introduced the notion of these strategies, describes chase demand, as the preferred strategy for jobs where workers require a low level of skills, work for a low pay, and the level of training per employee is relatively low. As a result, the chase strategy results in higher employee turnover rate, which can ultimately lead to higher training costs and higher defect rates. Thus, chase demand often ends in being a more costly strategy than anticipated given the low wages of employees.[3]

This document denotes chase demand as “staffing-to-charge”.

3.1.2. Level Capacity

The strategy of level capacity consists of maintaining supply capacity at a constant level, regardless of what the demand patterns may be. Level capacity is said to be preferable when demand is visible before the time of execution, which allows for a certain level of forecasting.[4]

Sasser describes level capacity as a longer-term strategy, requiring a longer labor capacity forecast. This is because the level capacity strategy assumes longer training time for employees, and therefore longer lead times to recruit them. However, the level capacity strategy is thought of as being more cost effective because it requires a lower level of supervision, and has positive externalities such as employee loyalty and low defect rates.[3]

This document denotes Level Capacity as "level loading".

3.2. Forecast Accuracy Metrics

Given the dependency of labor planning on the incoming units forecast, it was imperative to determine the accuracy of the weekly and daily forecast for inbound operations. There exist many different methods to benchmark and determine the accuracy of a forecast. Davydenko et al. suggest that the choice of an appropriate error measurement is almost as crucial the choice of forecasting method since the forecast method will be evaluated based on the error measurement chosen.[5]

Some of these methods were explored for suitability to assess the accuracy of the Amazon Forecast, as they are useful for comparing several series on the similar scales.[6]

- **Mean Absolute Percentage Error (MAPE)**

MAPE measures the average of the absolute value of the difference between the actual value and the forecasted value. The formula is as follows:

$$MAPE = \frac{1}{N}\sum \frac{|y_i - f_i|}{y_i}$$

where N is the number of data points, y_i is the actual value observed for time point i, and f_i is the forecasted value predicted for time point i.[7] Lewis et. Al suggest that the criteria for the MAPE is as follows:

- <10% - great accuracy
- 10-20% - good accuracy
- 20-50% - sufficient accuracy
- >50% - insufficient accuracy[8]

However, MAPE is often criticized due to the high percentage errors that can arise from relatively low values and near-infinite MAPEs due to near-zero values. Nonetheless, MAPE still remains one of the most popular measures of forecast accuracy among business forecasters due to its intuitive interpretation.[7]

- **Mean Absolute Error (MAE)**

MAE measures the average of the absolute value of the difference between the actual value and the forecasted value. It is a scale-dependent measurement, where the scale of the data affects the magnitude of the error.[7] The formula is as follows:

$$MAE = \frac{1}{N}\sum|y_i - f_i|$$

The MAE is deemed as unfit for series with intermittent trends,[6] and this intermittency is difficult to define.

Because of labor constraints, incoming volume patterns tend to be quite stable. However, because Amazon data is of large magnitudes due to the number of SKUs in question (Section 6.1), this scale-dependent measurement might not be the most appropriate.

- **Mean Absolute Scaled Error (MASE)**

The MASE uses the MAE as a benchmark, but attempts to improve error measurement with cases of intermittent demand. The formula is as follows:[9]

$$MASE = \frac{\sum|y_i - f_i|}{\sum|y_i - y_{i-1}|}$$

where the numerator is the MAE for forecast error and the denominator is the MAE using the actual value for the previous period instead of the forecast.

The MASE is a scale independent measurement, useful for measuring forecasts on different scales.[6] However, the MASE is criticized for overrating performance and being vulnerable to outliers.[5]

While the general weekly forecast is constant for any Amazon FC, subsequent application of the "day-of-the-week" curve (Section 2.2) does create outliers when studying the forecast of a day-to-day basis. Thus, this method might not be the most appropriate for the purposes of this investigation.

- **Mean Squared Error (MSE)**

MSE measures the average of the difference between the actual value and the forecasted value as a mean of their squares.[10] The formula is as follows:

$$MSE = \frac{1}{N}\sum(y_i - f_i)^2$$

Because MSE is sensitive to scales, it is prone to high forecast errors and therefore might not be the most suitable method.[10]

- **Weighted Mean Absolute Percentage Error (WMAPE)**

The WMAPE is a modification on the MAPE that attempts to minimize the effects of outliers. The WMAPE accounts the weighted contribution of each input to the total forecast error.[11] The formula is as follows:

$$WMAPE = \frac{\Sigma|y_i - f_i|}{\Sigma y_i}$$

Since the MAPE is still one of the most popular methods of forecast accuracy measurements, chosen for its ease of interpretation, the WMAPE seemed like a good choice for forecast accuracy measurements. It is applicable in that Amazon purchasing is not intermittent, the scale of the data does not affect it, and it is weighted for different percentages of SKUs within the freight in question.

3.3. Process Design and Execution

As a customer-centric business, Amazon strives to make sure that all activities within the business are targeted towards completing that customer promise. Operations research suggests many methods of implementation in order to achieve this in every area of the business.

A review of The Process Enterprise by M. Hammer suggested that in order to create customer value across all areas of the business, all business processes must work toward a common goal.[12] This is particularly true in supply chain.

Part of achieving this common goal is designing the processes in such a way that instructions are specific, allowing for repeatable and predictable process outcomes.[12] In the case of Amazon inbound operations; the common goal is to have the item in stock as soon as possible. However, the manner in which this occurs, such as correctly quantifying the labor capacity, is not standardized and therefore costly inefficiencies arise often within the network.

Another important component of process design is the creation of management systems, which work to support the processes under execution. Hammer argues that budgeting, planning, and financial systems should be modeled and aligned around the process they serve, rather than the inverse.[12]

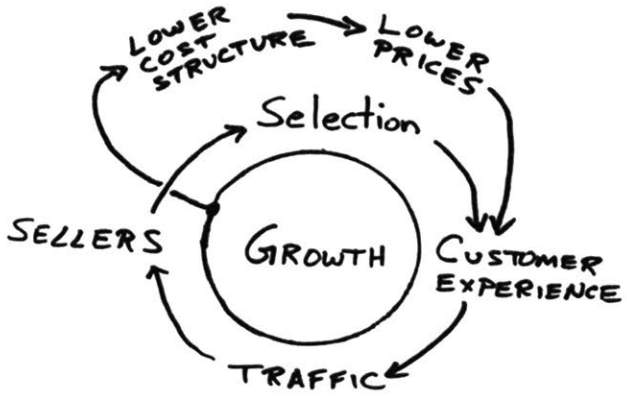
Amazon labor planning is in clear violation of this principle. Amazon's labor capacity, a key component to the execution of inbound operations, is modeled around purchasing planning. Moreover, decisions of headcount and freight allocation are mainly driven by financial incentives, and the desire to cut costs around the network.

Therefore, there exists the argument to change the priority of labor planning in the inbound operations process in order to create major executional improvements.

4. Ideal State Analysis

The Amazon flywheel consists of a greater selection of products, which leads to greater consumer purchases, in turn leading to greater online traffic that serves to increase the selection, thus creating a reinforcing loop (Figure 4).

Figure 4. Amazon Flywheel



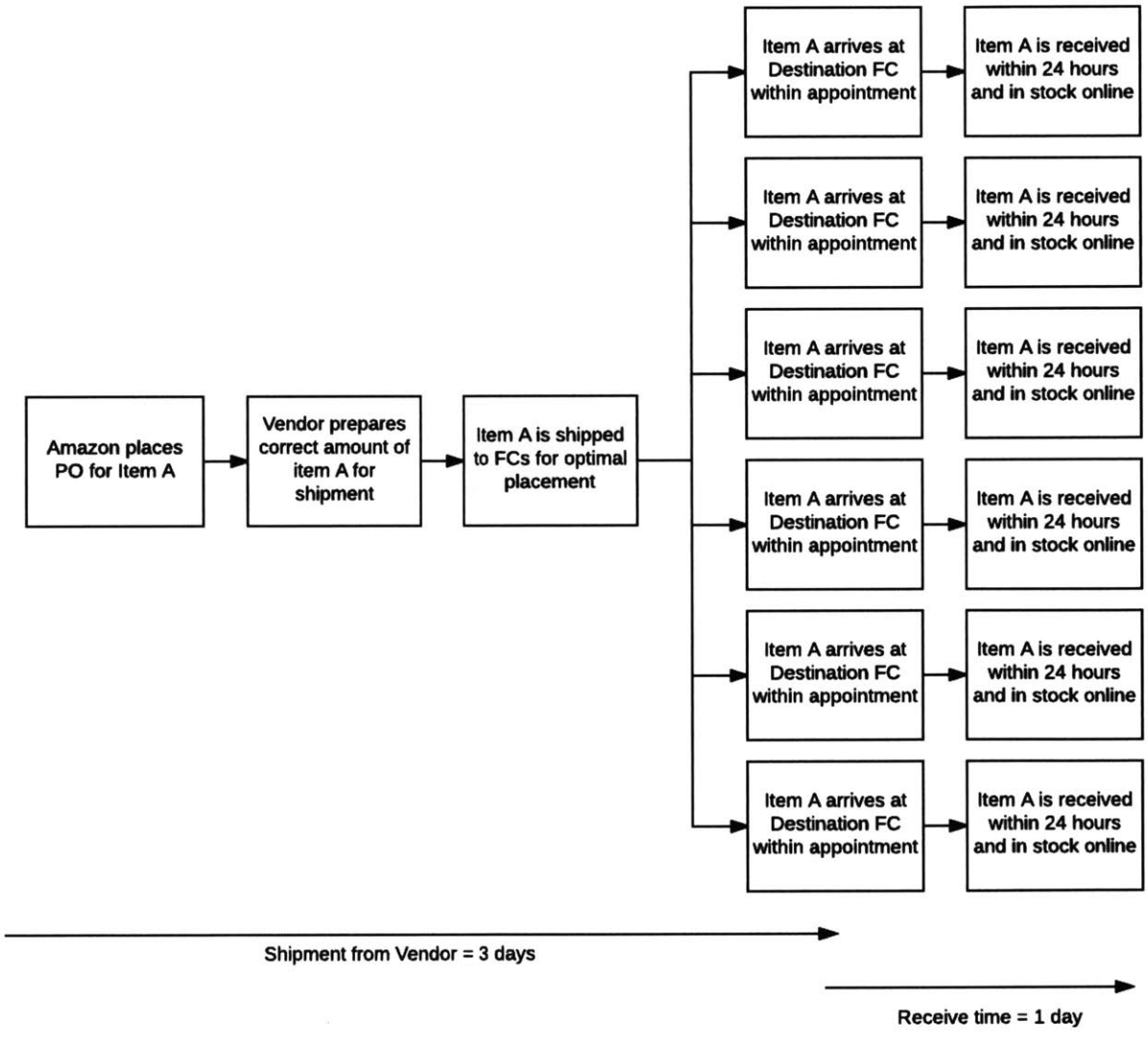
Keeping up the variety of products online requires timely inflow of items into the network inventory. Because items only become part of inventory, and therefore available for sale online, once they are received into the destination FC, streamlined inbound operations are crucial to maintain the Amazon flywheel. The ultimate goal of inbound operations is to have VLTs be as short as possible.

In an ideal state, the following timeline occurs when Amazon places an order for replenishment of item A (Figure 5):

- Amazon places order, known as placing a PO.
- The vendor ensures sufficient stock of Item A, and prepares Item A for shipment.
- Item A is shipped to the multiple destination FCs, chosen for planned placement within the network in terms of cost, and arrives within 3 days of PO placement.
- All destination FCs have backlogs of no greater than 0.5-0.8 days of backlog, and thus item A is received within its prescheduled appointment.

- Because the item is shipped in a timely manner and in the quantity specified in the PO, the incoming freight forecast is accurate in terms of predicting the arriving freight per day at the FC.
- As a result of accurate forecasting, all destination FCs have the right labor capacity to process the number of assigned units forecasted for the day, including Item A.
- Item A is received by all destination FCs, resulting in desired placement within the network, and is available for sale on the Amazon.com website within 24 hours of arrival at all sites.

Figure 5. Ideal Timeline for Inbound Operations



Of course, the ideal state is not possible. Three main reasons why VLTs are not strictly limited to four days are:

- Prep time: In order to provide the greatest selection, Amazon sells all kinds of products, from books to Jacuzzis. Given the drastically different nature of these products, lead times will have different lengths depending on the vendor and the nature of the product type. The mix of product Amazon receives is so vast and the staging and shipping process times so different, that it is challenging for all vendors to get freight prepped within the same time window. The prep time to palletize books will be drastically different from the prep time to palletize lawn mowers, which in turn will be drastically different from the prep time to palletize an order of axles and windshield wipers.
- Travel distance: In the Ideal scenario, vendors are able to send the product to any FC throughout the network in 3 days. However, this is not realistic. Amazon engages in business with a variety of vendors, from very established, to more amateur sellers. Thus, not all vendors have distribution centers and logistics that enable them to place their items anywhere in the network within 3 days of the PO.
- Backlog: Backlogs at FCs are not always within the manageable range of 0.5-0.8 days. One of the reasons for this is the fact that vendor freight is not always easy to scan and receive. Since items come from a variety of vendors, items must be carefully inspected before receiving, which leads to rate of items received per hour to decrease.

4.1. IXD Operations

As part of improving VLTs, Amazon implemented IXD sites. IXDs serve to avoid the vendor from having to ship items to multiple destinations in the network. Instead items are shipped to an IXD site, located within the same geographic region as the vendor. This way, Amazon ensures that the vendor can ship to a nearby location within the ideal timeline of 3 days.

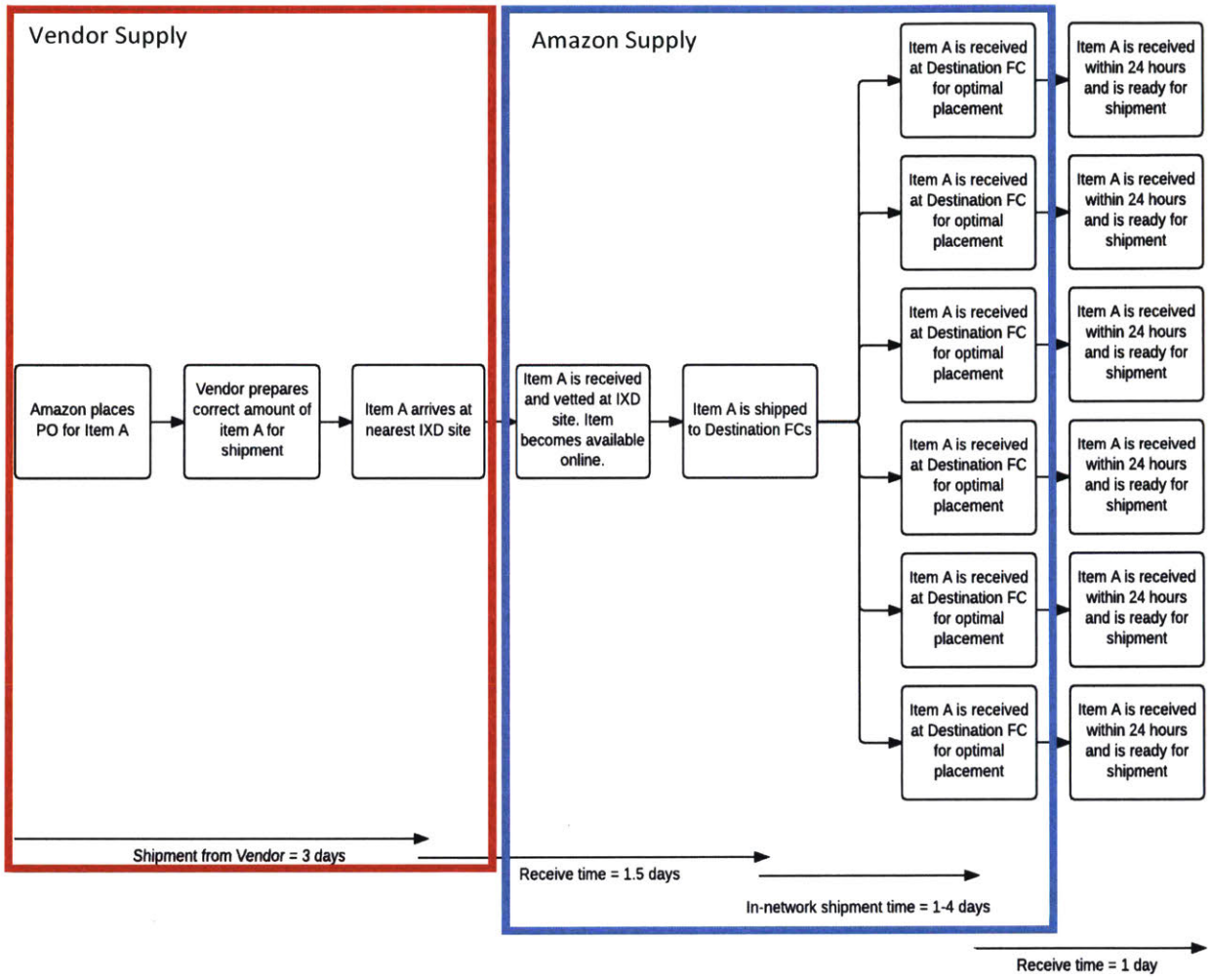
Items arriving at the IXD are inspected and vetted to determine that the items match the order placed. Vetted items then become available for sale online, and in an ideal scenario still three days after the PO is placed. Items are then placed in a truck and Amazon logistics transports the items to the multiple destination FCs to be placed in storage. Any subsequent transportation within the network is handled by Amazon, which ensures that Amazon is in control of the time frame to ship items across the nation, as opposed to having to depend on the vendor logistics as in the previous system. This maximizes Amazon truck utilization, and provides Amazon with economies of scale to lower transportation costs within the network as items moving within the network consist of both vendor freight but also customer orders.

The implementation of IXDs is meant to turn Amazon inbound operations into a hub-and-spoke network, where a network of IXDs act as the regionally-distributed hubs. Because Amazon owns the logistics within the network, travel times and freight arrivals can be more accurately monitored and predicted. IXD vetting allows for better visibility into what items are being sent to which FCs and when these items will arrive, and it provides opportunity for cost reductions as transportation is now sourced in-house.

Thus, the ideal VLT timeline changes to the following (Figure 6):

- Vendors now ship item A to the nearest IXD. Item A arrives at the IXD within its prescheduled appointment.
- The IXD has a 0.5-0.8 day backlog, allowing item A to be received within 24 hours of arrival. Item A is vetted and scanned, and thus immediately available online.
- Item A is then placed en route to the destination FC after 36 hours of arrival at the IXD.
- Transportation to the destination FC will vary, depending on the distance travelled (1-4 days).
- The destination FC also has 0.5-0.8 days of backlog, allowing item A to be received within 24 hours of arrival.

Figure 6. Ideal Timelines for Inbound Operations with IXD Implementation



4.2. IXD Constraints

IXD sites do not function as storage facilities, and so items entering the IXD do so in continuous flow through the building. Thus, some of the constraints for IXD sites are:

- Labor Capacity

The headcount available for the day, and assigned to inbound operations, is the upper bound constraint in terms of how much freight can be received. The labor capacity can vary due to absences, labor sharing, or wrong assumptions in terms of hourly rate of units processed.

If the labor capacity is mismatched with the volume of units, management can pull the different headcount adjustment levers in order to match capacity. If this is not possible, the issue is escalated to the freight redirect team.

- Dock Doors

Every IXD site has a set number of dock doors. This presents an upper bound constraint in terms of how much inventory can be received at any given time.

The freight redirect team is in charge of scheduling truck appointments and making sure that there is always dock capacity for any arriving trucks.

- Yard Space

The available yard space for incoming freight trucks to park and unload at an IXD site is a major site constraint. The yard can be occupied by trailers of items delivered by the driver, waiting to be received by the site. In cases in which the driver cannot leave the trailer, queuing trucks waiting to unload at the dock also occupy the yard.

Utilization of the yard is one of the main reasons for escalations in terms of overtime of freight redirects.

The optimization model for forecasting takes into account the yard space available at an IXD site.

5. Current State Analysis

The current state of inbound operations was analyzed for discrepancies with the ideal state scenario. All analysis was based on the IXD site chosen for this investigation. Inefficiencies that lead to this discrepancy are explained in further detail.

5.1. Vendor Lead Time Variability

As discussed in Section 4, in order to provide the greatest selection, Amazon sells all kinds of products, from books to Jacuzzis. Given the drastically different nature of these products, lead times will have different lengths depending on the vendor and the nature of the product type.

To engage in business with all types of vendors and maintain the maximum selection of products available, Amazon does not strictly enforce delivery dates from vendors. If the vendor fails to comply with the requested date, they are fined with a certain percentage of cost of goods sold (COGS) – the COGS penalty will differ by product group. Every year, Amazon receives around $30 MM in vendor non-compliance fees, accounting for the variability in VLT. In order to maintain the wide selection of items in the retail business and keep the flywheel growing, Amazon continues to do business with vendors who are periodically late.

Adding to the variable VLTs is the fact that, given the different scales of POs, Amazon is unable to provide a demand forecast to all vendors. The orders placed from vendors are highly variable in terms of product mix and quantity ordered. As a result, vendors (with the exception of high-confidence, established vendors) are generally unable or unwilling to carry the safety stock necessary to fulfill Amazon's changing orders. This results in historical VLT data that is highly variable, which makes it challenging to predict how many items will be received in reality.

5.2. Inbound Forecasting

Given that VLTs are variable and historical data is unreliable, the freight allocation optimization model produces a forecast for what is to arrive in the network in terms of incoming freight, and allocates it to the different FCs. Amazon measures its forecast accuracy using a series of metrics. One of those metrics is the Percent Actuals to Plan,

which measures the percentage difference between the actual receipts and the forecast from one week prior to execution, as a percentage of this forecasted value.

$$Percent\ Actuals\ to\ S\&OP\ Plan = \left|\frac{Forecasted\ number\ of\ weekly\ arriving\ units - Actual\ number\ of\ weekly\ arriving\ units}{Forecasted\ number\ of\ weekly\ arriving\ units.}\right|$$

Because sites cannot adjust labor upon short notice, the forecast aims to make the Percent Actuals to plan as stable as possible. The success of the forecast is then determined upon whether or not actual capacity from week to week changed by more than ±10%.

While Percent Actuals to plan is one way of measuring forecast accuracy, it only looks at the weekly bulk value. It does not give a representative picture of how the forecast actually performed on a daily basis. To measure this, a standard forecast accuracy metric was used: the Weighted Mean Absolute Percentage Error (WMAPE). The WMAPE measures the percentage error of the forecast with respect to the actual values as follows:

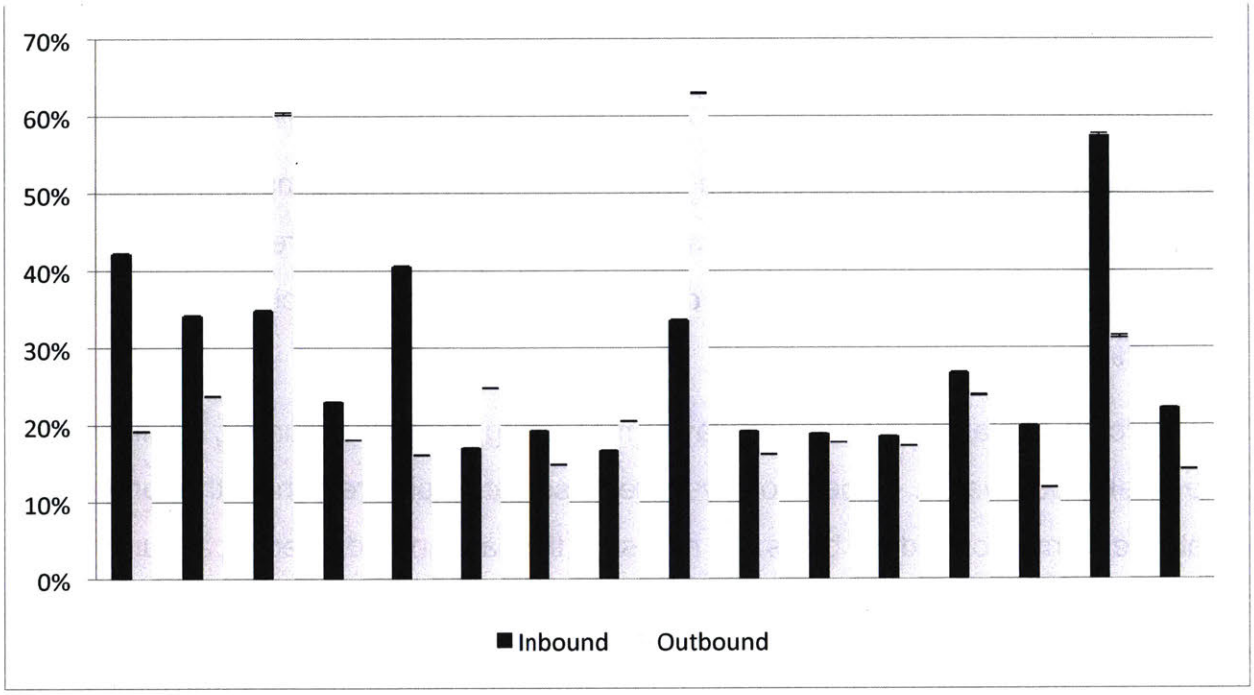
$$WMAPE = \left|\frac{\sum_{i=1}^{n}(Actual\ number\ of\ daily\ arriving\ units_{\ i} - Forecasted\ number\ of\ daily\ arriving\ units_i)}{\sum_{i=1}^{n}(Actual\ number\ of\ daily\ arriving\ units_i)}\right|$$

where n is number of days in the observed period of time. The period observed was the three-week period of March 13th to April 3rd, 2016, chosen for being a period with no holidays or sales, and therefore having an average demand pattern. The WMAPE was calculated for individual sites within the Amazon network (not exclusively IXDs), as well as for the network as a whole. The forecasts used were the same as for the previous method: the weekly forecasts from one week prior to execution.

The WMAPE for a few selected sites shows that when compared to actual values, inbound forecast error is consistently greater than 10% (Figure 7).[IV] Moreover, in general, inbound forecast errors seem to be much greater in comparison to outbound forecast errors. This is surprising, considering that inbound is a forecast for orders that buying has already planned and that are coming from known POs.

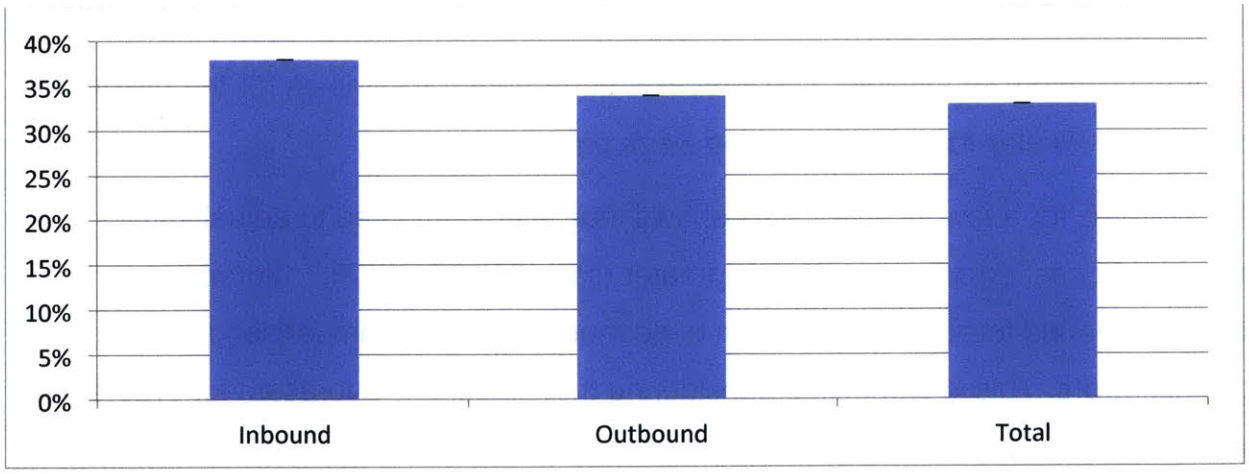
[IV] Inbound forecasts accounts for units arriving at the site, denominated as arriving units. Thus, this accounts both for units received and processed by the site that day as well as units added to the backlog that day.

Figure 7. WMAPE for Select FCs, 03/13 - 04/03



The WMAPE was then studied for the network as a whole. The WMAPE was calculated for every FC in the network, and then this error was averaged to compute the network WMAPE. A similar effect of forecast error was seen for the network as a whole, with inbound forecast error being off by more than 35% (Figure 8).

Figure 8. NACF Network WMAPE, 03/13 - 04/03



This large forecast error might be occurring due to the fact that this forecast was computed daily, and Amazon utilizes the Day-of-the-week curve for its daily forecast

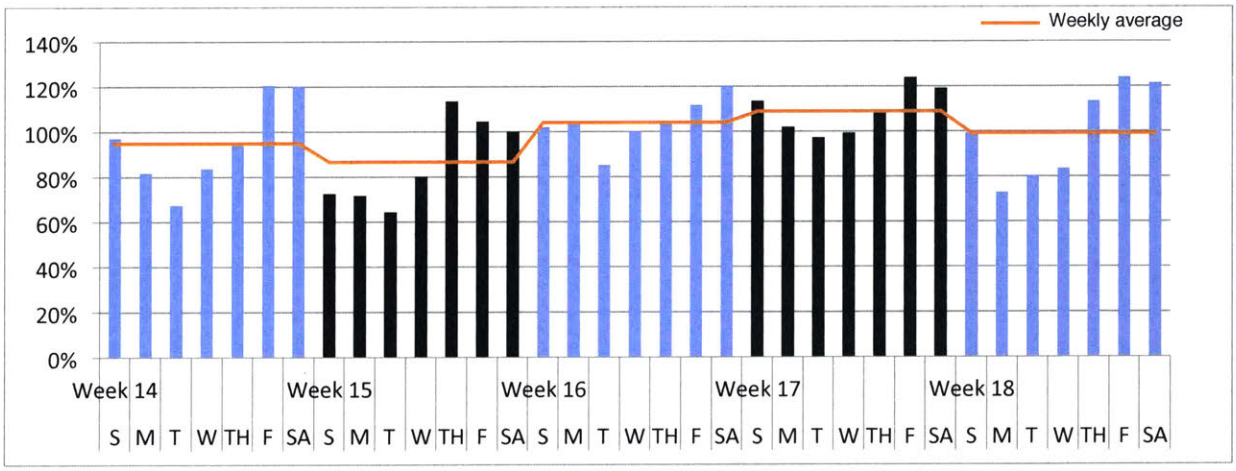
(Section 2.2). The day-of-the week curve is based on historical averages of vendor behavior. But given that VLTs are highly variable depending on the vendor, the day-of-the-week curve is not very reliable when predicting vendor behavior. Nonetheless, because historical data is currently the closest approximation to future vendor behavior, the forecasting model still produces a forecast according to the historical arrivals.

5.3. Backlog Management

Currently, staffing-to-charge follows the day-of-the-week curve for inbound arrivals. To understand how actual performance compares to the forecast, the actual number of units received was taken as a percentage of the forecasted volume of arriving units for a four-week period for the IXD in question. The weeks studied (weeks 14 through 18 of the year) were chosen for being weeks with regular demand. Because the period studied contained no major holidays, it was assumed to have no major abnormalities in the form of spikes in demand.

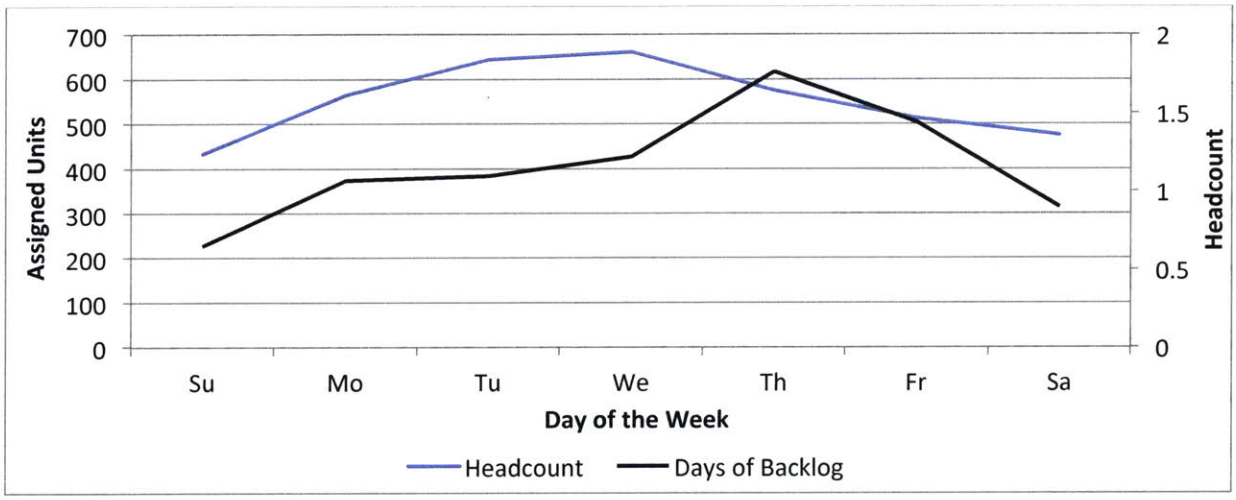
The study shows that a weekly cyclical pattern takes place in which lower than forecasted processing of assigned units, units that the site is expected to process according to the forecast, takes place from Sunday through Wednesday, and a greater than forecasted processing of assigned units takes place from Wednesday through Saturday (Figure 9). This is occurring because due to the following of the day-of-the-week curve, units are scheduled to arrive mostly from Wednesday through Friday. As a result, the FC tries to process as many units as possible.

Figure 9. Percent Units processed to forecast



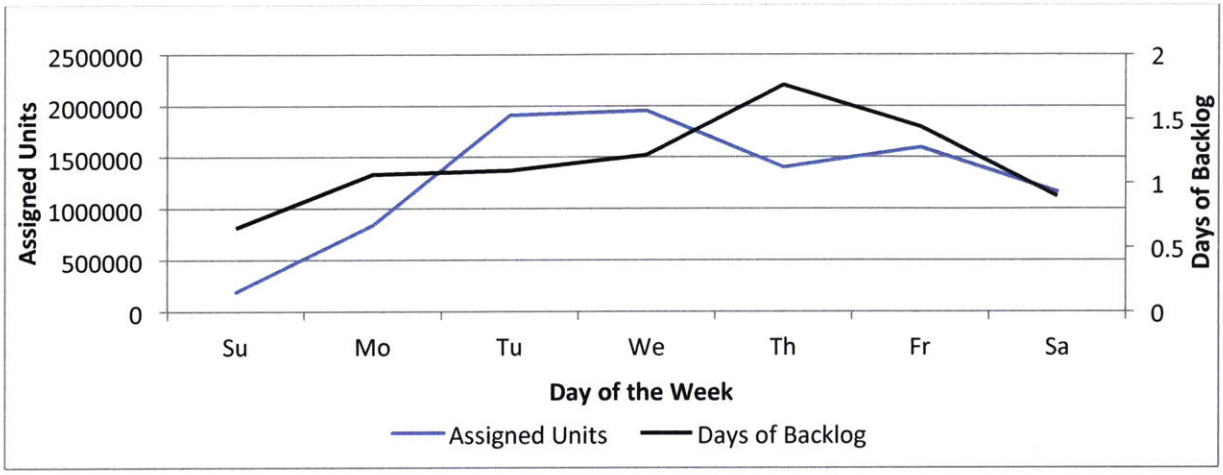
The reason for the discrepancy between actual and forecasted processed volumes is due to backlog management. Because staffing matches the forecasted day-of-the-week curve, on days on which forecasted volume is greatest (usually between Tuesday and Thursday), the FC processes as much freight arriving as possible. Simultaneously, the FC tries to build a backlog to provide work for the associates coming on lower volume days. On less frequented days, associates work through the backlog first and then process the new freight. This can be seen in Figure 10, which shows the 4-week average curve shapes of backlog vs. headcount.

Figure 10. Headcount vs. Backlog



Thus, on the weekends, associates are instructed to process the backlog before the units that arrive that day. To enable this, sites are assigned fewer units from Sunday through Wednesday and so fewer arrivals are seen during these days (Figure 11). By staffing to the day-of-the-week curve, inbound arrivals and processing result in a self-fulfilling prophecy.

Figure 11. Units Arriving and Backlog at the IXD



However, regardless on this handling of the backlog, the weekly percentage of units processed was found to be consistently within the range of forecasted units arriving at the site. When the weekly units processed was taken as a percentage of the forecasted arrivals, processing rates were found to be ± 13% of assigned units for the weeks studied (Table 2). This indicated that if the arriving units were evenly spread across the week, there exists a potential to process them in a timely manner given the labor available, without as much backlog management being necessary.

Table 2. Weekly Processing Averages

	Week 14	Week 15	Week 16	Week 17	Week 18
Processed Units as a Percentage of Arriving Units	95%	87%	104%	109%	99%

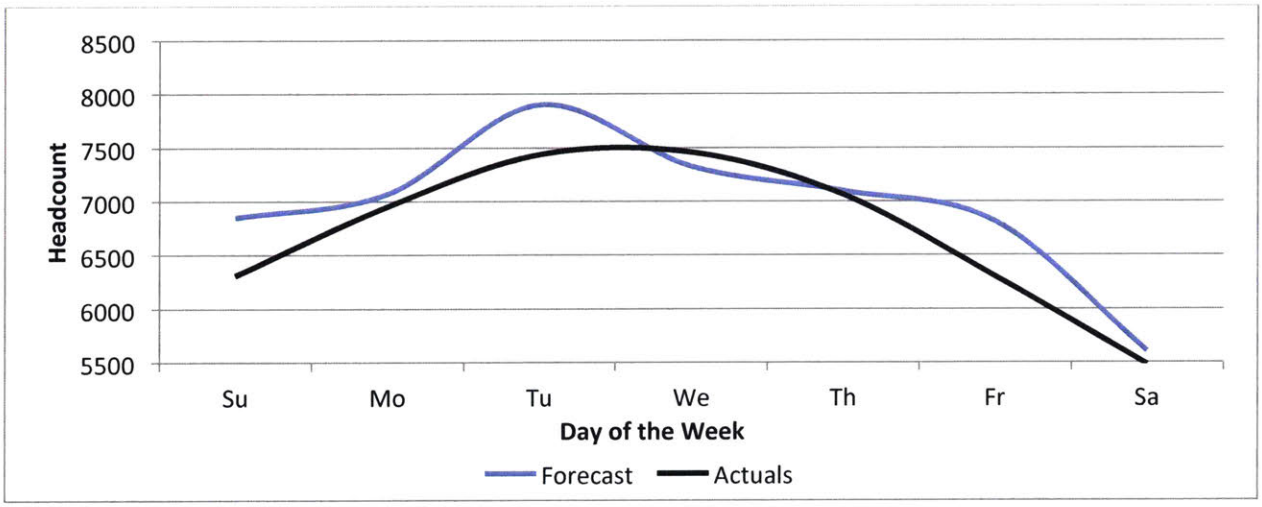
5.4. Predicted vs. Actual Labor Capacity

As previously mentioned, staffing-to-charge requires assumptions to be made in order to guarantee that the required labor capacity will be available everyday. While factors such as rate, attendance, and attrition are taken into consideration, human behavior can only ever be approximated.

Figure 12 shows a 4-week average of forecasted vs. actual show hours which shows that although the forecast makes predictions for attendance and attrition, the actual labor hours materialized for the day are still significantly different from forecasted. This

can be due the fact that associate performance can differ from the projected hourly rate. As mentioned previously, items in Amazon fulfillment centers vary greatly in nature, from books to jacuzzis. While the forecast could be correct in terms of the number of items arriving, such as 100 items, 100 books which are small and easy to carry will be processed at the site much faster than 100 jacuzzis which require multiple associates to load and unload. This variability in the nature of the units of freight is known as the 'freight mix'. Freight mix is one of the main causes of variations in labor capacity.

Figure 12. Forecasted vs. Actual Work Hours



Increased variability in labor capacity also occurs on a shift-by-shift basis. For instance a growing backlog on outbound operations could require labor sharing between inbound and outbound, which changes the headcount available for inbound receiving throughout the day.

Because the possibility of labor capacity changing drastically on a same day basis exists, Amazon logistics is setup to ensure that the FC only receives the volume it can process. There is a designated team in charge of redirecting if freight when the capacity of an FC is exceeded. This is a manual process, done on a daily basis, in order to guarantee timely insertion of units into the network. While it ensures that labor capacity is not exceeded, it creates network inefficiencies such as freight redirects and additional transportation costs as a result.

5.5. Diminished Associate Experience

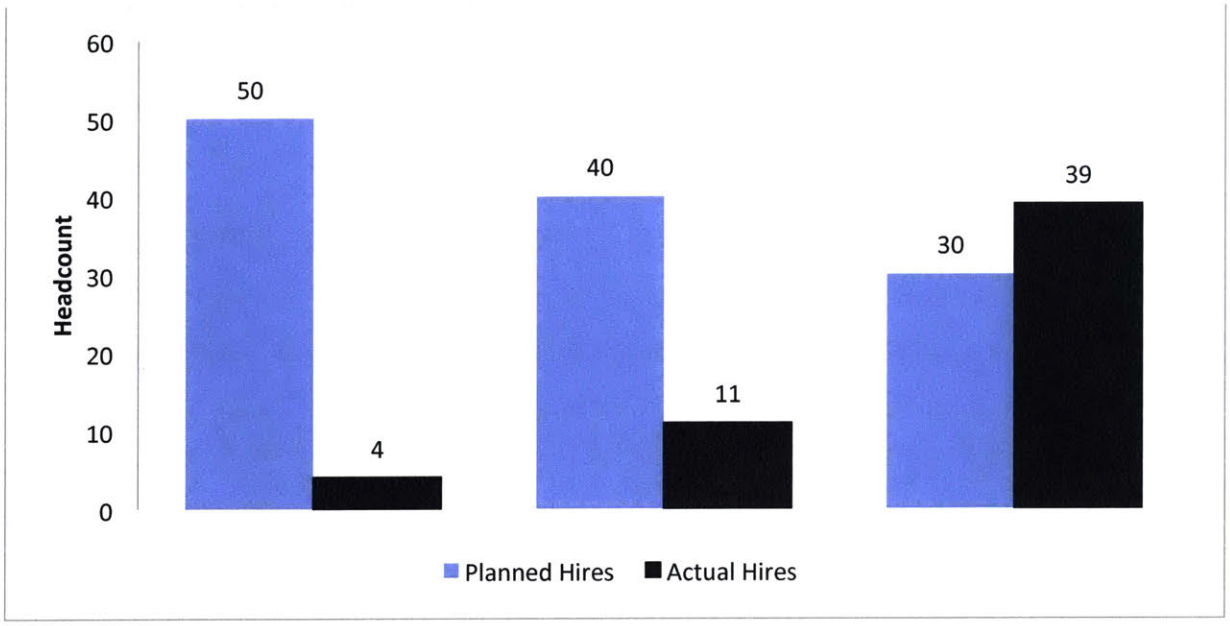
Inefficiencies in terms of labor planning lead to a detrimental effect for the associate experience. When working for Amazon, associates agree to a 40-hour workweek. Inconsistencies in the labor plan can have a material impact on the lives of associates.

For instance, cases in which the assigned units come in too light leads to lack of work and idle time. Associates are demotivated by idleness, in which case they have to option to VTO. Because VTO is a form of unpaid time off, too many consecutive weeks of VTO leads to a considerable decrease in income for associates.

Conversely, when the forecast comes in heavy, staffing lean implies that extra capacity is needed to bridge the gap between freight and labor capacity. When excess capacity is necessary on short notice, this takes the form of overtime. 50-hour workweeks can be overbearing for associates if they happen too frequently. This irregularity in shifts leads to a detrimental associate experience, potentially decreasing attendance and increasing attrition.

Given that certain FCs operate in tough labor markets where finding new associates is challenging, employee retention has become crucial for FCs. Certain labor markets have become tough for hiring, as the population is small and there is little untapped talent. Figure 13 shows that a week requiring hiring in preparation for Prime day, FCs in tough labor markets were only able to procure under 30% of the necessary hires, whereas in generous labor markets sites are able to hire over the requirement, to hedge against attrition.

Figure 13. Planned vs. Actual Hires



A consistently diminished associate experience eventually leads to higher levels of attrition. Thus, a stable staffing model that improves associate experience and decreases attrition is of critical importance.

6. Backlog and its Consequences

Per Section 5.3, inbound backlog is used as a lever for labor mismatch management. It is an important metric for site leadership and network capacity, and yet is understood very superficially. This section aims to understand backlog better in order to provide a more robust staffing recommendation.

6.1. Background into Backlog

To ensure that there is always work available for associates, the site maintains a certain level of backlog at any given time. The backlog consists of items that have arrived at the site and are waiting to be received. Backlog also indicates the efficacy of a site in handling the volume of units it was assigned. The number of days backlog is determined by the following formula:

$$Days\ of\ backlog = \frac{Number\ of\ units\ waiting\ to\ be\ received}{Average\ daily\ number\ of\ units\ to\ be\ received\ for\ the\ next\ 7\ days}$$

Additionally, backlog serves as a crucial metric to determine which site has extra capacity to handle additional freight and affects costs of labor and freight transportation between sites. However, while it is understood that backlog leads to increased costs due to delayed receipts, the exact magnitude of this cost is not clearly understood. In order to make any recommendation in terms of labor capacity, it was necessary to understand what is the cost of backlog and how changes in labor capacity can affect this cost. Only with a concrete cost of backlog can a true backlog threshold for the NACF sortable network be established, allowing the implementation of a relevant staffing recommendation.

Currently, NACF aims for a 0.5-0.8 day inbound backlog level at all times, to be used as a buffer in the event of short-term mismatches between arrivals and the planned labor capacity. Instances where the backlog exceeds 1 day lead to escalations in the form of overtime, as increased levels of backlog can lead to delayed restocking of inventory and downstream costs when fulfilling customer orders. If overtime is not sufficient to process the incoming freight, the escalation leads to freight redirects.

However, upon further investigation, it was found that the choice of 1-day backlog as a threshold for escalations was seemingly arbitrary. While the system worked with this escalation threshold, no employee at any level of management could truly explain the reason behind the numeric value of the threshold.

6.2. Selecting a Representative Sample

The calculation of a backlog cost required analysis into the different SKUs in Amazon's inventory that could be sitting in the backlog at any given time. Because backlog consists of items waiting to be received, items in the backlog run the risk of leading to a stock out for their respective SKU. The larger the backlog, the more inventory that is waiting to be received instead of being stored in the warehouse. As operations continue and customers continue to place orders, SKUs with items in the backlog run the risk of depleting warehouse inventory as backlog items wait to be received.

Because Amazon sells over 10 million SKUs, it is not possible to understand the network inventory level of all SKUs given the time constraints for this project. Therefore, it was necessary to select an appropriate sample size that would allow a generalization to be made regarding the cost of the backlog in the network.

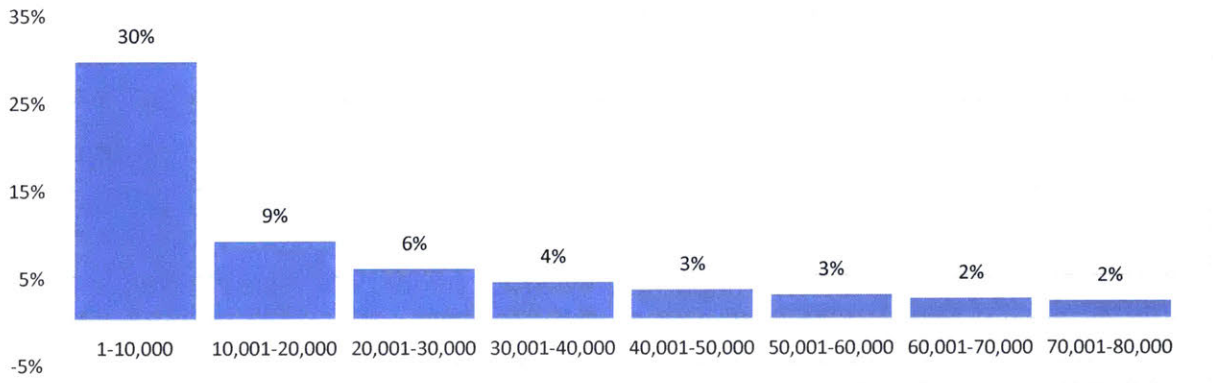
In trying to determine a sample of SKUs representative of network inventory, a few exceptions were made:[13]

- No digital SKUs: Digital SKUs consist on items that Amazon sells for consumption through devices, such as movies, e-books, audiobooks, etc. Digital SKUs are not part of traditional supply chain; they were excluded from investigation into network level inventory.
- No virtual or service SKUs: Virtual and service SKUs consist of Amazon services sold on the website, such as Amazon Music or Amazon Instant Video. These were excluded, as they are not delivered by Amazon fulfillment.
- No specialty, promotional, or seasonal SKUs: This category includes items that are sold mostly on certain holidays, such as Prime Day and Black Friday, as part of Amazon deals. These SKUs were excluded from the investigation for not being representative of regular demand patterns for Amazon.

All remaining SKUs were subsequently ranked in order of decreasing demand. To obtain a representative sample while keeping the data set within a manageable size, demand data was sampled from the first week of every month for the past 12 months (November 2015 through November 2016). The goal was to produce a list of the top selling SKUs for the year for the network as a whole, albeit the exceptions discussed previously.

Further investigation into the sales volume of this list of SKUs as a percentage of the total outbound sales volume (containing all possible SKUs) found that the top 10,000 selling SKUs account for 30% of total network sales throughout the year. For every subsequent range of 10,000 SKUs, the percentage of total sortable network sales for that range decreases. 10,000 SKUs ranges ranked past 80,000 SKUs were found to have an impact of under 2% per range on the total network sales and were therefore considered to have too small of an impact as a range (Figure 14).

Figure 14. Histogram of Top Selling 80,000 SKUs as a Percentage of total Retail Sales



Given the fact that these first ranges of SKUs are the top selling SKUs year-round and SKUs with smaller volume sales are likely to have fewer occurrences of freight redirects and escalations, it was assumed that 50% of total sortable network sales is a reasonable volume of sales to dictate the backlog threshold (~50,000 SKUs). Therefore, for the purpose of this project, the top selling 50,000 sortable SKUs were used to understand the network wide level of backlog.

Of the top selling 50,000 SKUs, a large portion of SKUs had been redacted, for protection of privileged information. As a result, there was data available for 38,059 SKUs, which accounted for about 40% of total network sales; these were studied for this project.

6.3. Cost Calculation for Different Backlog Thresholds

Having narrowed down a list of SKUs representative of over 40% of all sortable network sales, the general days of inventory available in the network at any given time was to be determined. The time period between September 1st and November 15th 2016 (76 days) was studied, based on most recent, comprehensive data available.

For each day and for each SKU, I computed the days of inventory available to the system. Total network inventory of a certain SKU was calculated as the sum of all items of that SKU either stored inside any FC in the network, traveling within internal network routes, en route from a vendor, eligible for manufacturing on-demand, and units of that item that can readily be ordered from a high confidence vendor. Therefore, given all of the sources of inventory replenishment considered, if an item has inventory of zero, it was assumed that there is truly no possibility of the item arriving in the warehouse on that same business day. In order to determine the right threshold for backlog, it was assumed that as long as a SKU is in stock anywhere in the network, it is considered to have an inventory level greater than zero.

A case study was done to determine how many SKUs might be affected if the backlog threshold were increased. This investigates how the level of inventory would be affected if the site no longer reacted once the backlog reached 1 day, but rather only escalating items once the backlog reached 2, 3, or more days. Increasing the backlog threshold implies that for every additional day of backlog, there are SKUs that are waiting in the backlog and not being replenished in the warehouses. Thus, there is the possibility of a continuation of stock out, if the SKU was already depleted, or new possibilities of stock outs for every other remaining SKU.

Assuming that a stock out occurs whenever there is less than one day's worth of average daily demand, the current service level was calculated for the 38,059 SKUs during the 76-day period using the following formula:

$$Service\ level = \left(1 - \frac{Number\ of\ instances\ of\ stock\ out\ within\ 76\ day - period}{38{,}059\ SKUs \times 76\ days}\right) \times 100\%$$

The service level for the period studied was found to be 95% for the current scenario. This was then calculated to see the effects on service level if the backlog threshold were increased to 2,3,4, and 5 day's worth of the average daily demand for all SKUs. Table 3 shows the changes in service level as the days of backlog increase.

Table 3. Service Levels for Different Backlog Thresholds

Days worth of Average Daily Demand	1	2	3	4	5
Service Level	95%	94%	93%	91%	90%

However, the service level calculation of all SKUs for the 76-day period is not granular enough to inform on how many SKUs actually face stock outs, since in actuality only certain SKUs face stock outs at different thresholds while other SKUs that would not face any stock out regardless of the days of threshold. Therefore, a more granular approach was taken in which the percentage population of SKUs that experience stock outs was measured for the different thresholds of backlog (Table 4).

Table 4. Percentage of SKUs that face Stock Outs for Different Backlog Thresholds

Backlog Threshold (Days)	1	2	3	4	5
Percentage of SKUs with at least one stock out	63%	65%	67%	69%	71%

The data shows that for all subsequent backlog thresholds, the population of SKUs with instances of insufficient stock increases by 2% for every additional day added to the backlog threshold. This leads to change from 63% of SKUs with insufficient stock for a backlog threshold of one day's worth of average daily demand, to 71% of SKUs with insufficient stock for a backlog threshold of five day's worth of average daily demand.

These minimal 2% increases in instances of stock out for every increasing day of backlog suggest that the size of the backlog threshold is not as relevant as the occurrence of any instance of stock out. The inventory data used takes into account any possible form of inventory replenishment, including items stored in an FC, items on trucks for any internal transfer, and items en route from the vendor. These instances of stock under one day of backlog would seem to highlight a problem in inbound receiving, not insufficient buying.

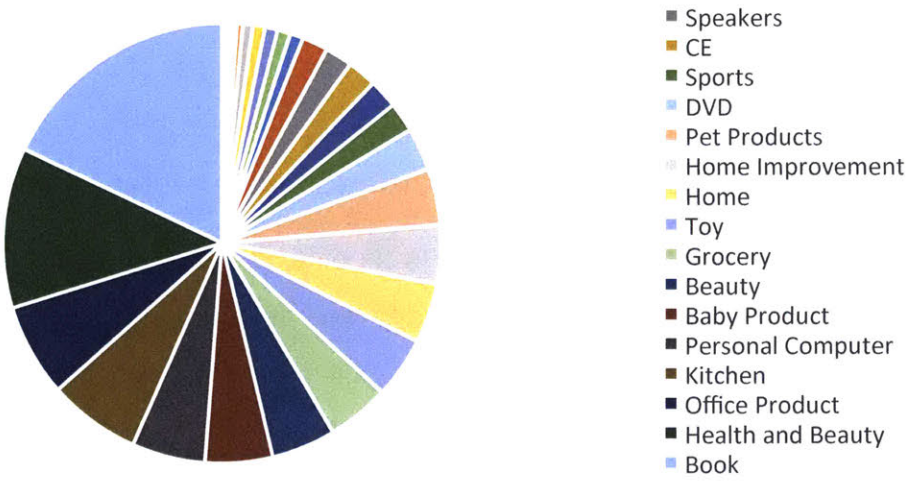
Therefore, the key problem with backlog inefficiencies appears not to be choosing a level of backlog that could provide the greatest level of stock but rather the opportunity cost of not having sufficient stock at any given moment for a particular SKU.

6.4. Opportunity Cost of Missed Orders

Since backlog is the main metric for IXD performance and a key lever for labor management, it is important to quantify the effect of backlog. Since backlog threshold did not appear to be relevant in determining the instances of missing stock; an alternate approach of determining the cost of missing stock was used in order to determine the true cost of backlog.

To calculate the cost of backlog, an assumption was made that the cost of backlog was equivalent to the cost of not being able to sell an item due to it being available but waiting in the backlog. Understanding the true cost of what these missed orders could be required an understanding of the magnitude of cost for every order missed. Thus, the population of SKUs who would be most affected by an increase in the backlog threshold was studied. The mix of SKUs can be seen in Figure 15.

Figure 15. Distribution of Item Types Within Missed Sales due to Backlog



There are many possible costs associated with an item being available but waiting in the backlog. For the purposes of this investigation, the cost of backlog was simplified to the cost of a lost retail order, simplified to the retail price of the item. This cost does not include any additional supply chain costs that could arise from not having inventory on-hand such as the cost of unplanned upgrades of shipments, the cost of transfer of items within the network, the cost of sending the item from an FC that is geographically further away but where the item is available for shipping, and ultimately the cost of a lost sale due to the item not being readily available. It also does not include the cost of a negative customer experience, which could deteriorate the consumer's loyalty to Amazon in the future. Because of lack of conclusive data on this subject, the calculation of this cost was not pursued further.

Thus, the cost of a missed order was simplified to the lost retail sales value of the item in the backlog for every additional day of backlog. This means that for a site with 1 day's worth of backlog, there is a certain population of missed items that results in a certain cost. For the same site, with a two-day backlog, there is an additional population of missed items that results in a new cost. This line of thought assumes that for every additional day of backlog, every order missed results in either a missed customer promise or a late order where the item is refunded and returned, both resulting in lost revenue.

Because every SKU has different demand frequency and order sizes, an average of the number of items in a missed order was determined for every SKU in the sample population, during the 76-day period. To determine the cost of these missed orders, the cost of every SKU order missed was quantified as the weighted average of the list price for all SKUs with a missed instance and their respective average order size.

Given the previous analysis on backlog thresholds, it was determined that items with 5 or more stock outs were outside of the 1-day service level of 95%: 4 instances of stock outs in 76 days is 5%, therefore anything past 4 instances was considered an undesirable service level for that SKU. A more granular look at the population of SKUs with stock outs suggests that the increase in days of backlog led to the increase of items with five or more instances of missing stock (Table 5).

Table 5. Population of SKUs within/ under the desired Service Level according to instances of stock outs.

Backlog Threshold (Days)	**1**	**2**	**3**	**4**	**5**
Percentage of SKUs within 95% service level (1-4 stock outs)	48%	46%	44%	42%	40%
Percentage of SKUs under 95% service level (5+ stock outs)	15%	19%	22%	27%	31%

This population of items under the 95% service level was determined to be the cost of backlog as these are the items that will become depleted completely out of stock if the site is unable to process its backlog within the one-day threshold. Therefore, the cost of missed orders was calculated as a contrast against the current scenario of 1-day backlog threshold and only for items under 95% service level given the previous analysis.

Table 6 shows the cost of backlog for the different threshold scenarios:

Table 6. Cost of Backlog Calculations

Minimum days of stock	1	2	3	4	5
Percentage Number of SKUs under 95% service level	15%	19%	22%	27%	31%
Number of Additional Orders missed per month	-	22,250	38,937	66,750	89,000
Average Price Per Order	-	$27.67	$ 28.35	$ 29.70	$ 29.93
Additional Cost of Missed Orders to NACF per month	-	$ 615,654	$ 1,103,872	$ 1,982,464	$ 2,663,755

Having established the cost of different levels of backlog for the network, the cost of backlog for the FC in question was calculated. The previous backlog calculations reflected the level of inventory in the entire network, resulting in the cost of backlog for the network as a whole. Throughout 2016, the IXD network has been responsible for receiving 68% of vendor freight. At the time of this investigation, the site in question was receiving 20% of the IXD vendor freight, amounting to 14% of total network new vendor freight. Therefore, this 14% was applied to the total network cost of backlog, and Table 7 shows the cost of additional missed orders if only considering the site in question:

Table 7. Costs of Backlog per Additional Day at Specific Site

Minimum Days of Stock	1	2	3	4	5
Additional Cost of Missed orders per month	-	$83,729	$150,127	$269,615	$362,271

This cost calculation allows the benchmarking of the cost of backlog for every additional day of backlog that the site accumulates. This will serve to contrast different labor implementation models, as different implementations of available labor capacity will affect the level of backlog at the site, affecting not only the cost of labor, but also the cost of backlog.

7. Level loading Feasibility Study

Given the historical difference of ± 13% between the forecasted units arriving and the actual units received (Section 5.3), a hypothesis was formed around how to improve the scheduling of associate headcount. This model, denominated 'level loading', proposes staffing to the two shifts to a cumulative same headcount of associates every day.

Level loading is advantageous as it leads to a known labor capacity every day, and is thus less sensitive to unprecedented increases in received unit volumes. However, the main disadvantage for level loading is that in creating this extra capacity for increased volume, the site becomes more vulnerable to stages where there is volume deficit, leading to increased instances of VTO. As discussed previously, this leads to a diminished associate experience, which can lead to more detrimental effects for the network in the long run, as well as higher labor costs if VTO is not successfully procured.

A series of scenarios were analyzed to determine the cost difference between staffing-to-charge and level loading. The scenarios were done using data for the specific FC throughout a 70-day period, between the weeks of 4/17/2016 and 6/12/2016. This particular period was chosen for its regular season demand and replenishment volumes, falling between Easter and Mother's day to avoid any seasonality effects. For all scenarios, the actual number of units arriving at the site was used as the actual daily charge for the period studied. For cost calculations, an hourly wage rate of $12.95 was used and associates were assumed to earn time-and-a-half during overtime.

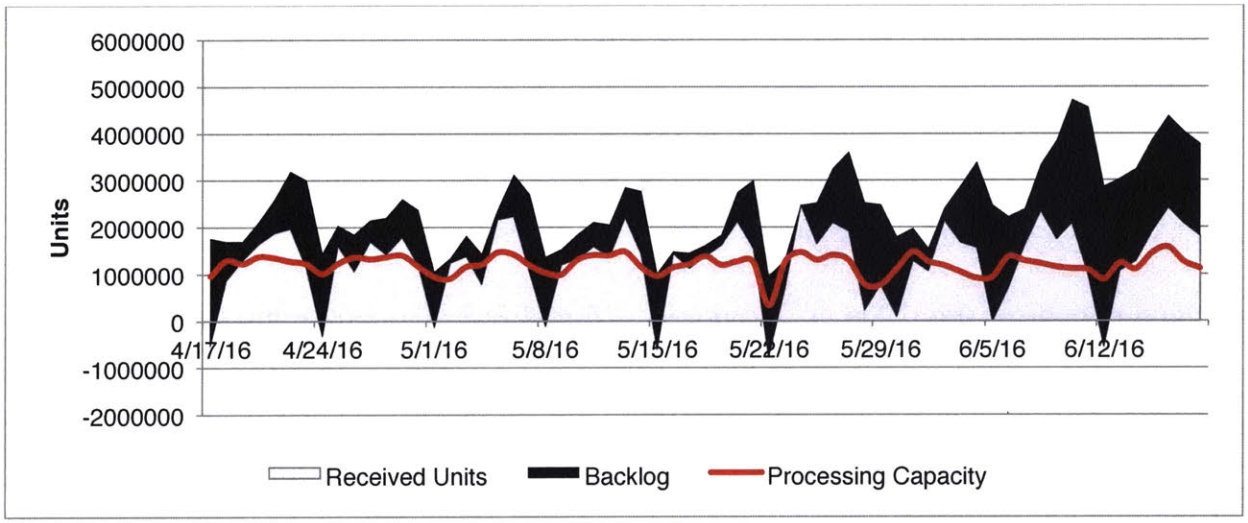
7.1. Scenario 1: Current State

This control scenario analyzed what occurred in actuality at the FC in question, given the Staffing-to-charge model. Figure 16 shows how the FC has attempted to mimic the trend in incoming freight with how many units are processed every day, while still maintaining a manageable backlog. For this study, backlog is defined as:

$$Backlog = (Previous\ day\ units\ in\ backlog + arriving\ units\ by\ EOD) - Units\ received\ by\ EOD$$

The current state scenario depicts the behavior observed during the period studied. It uses the actual number of arriving units as an input and depicts the labor capacity and backlog as they were observed. The average backlog was about 1 day, with a maximum backlog of 2.90 days on 6/9/2016.

Figure 16. Backlog Management given Staffing-to-Charge Model



However, given forecast variability, the number of processed units is not always the required amount to maintain the 0.5-0.8-backlog goal. This is particularly apparent during the week of 5/29/2016. Table 8 shows how this increase in the backlog from desired levels is due to mismatches between the headcount that the FC staffed given a forecasted volume of assigned units, and the actual volume of units of incoming freight, leading to a high discrepancy between the assigned units and the received units.

Table 8. Arrived Units vs. Received Units for the week of 5/29/2016

	5/29/2016	5/30/2016	5/31/2016	6/1/2016	6/2/2016	6/3/2016	6/4/2016
Previous day Backlog	1,714,613	1,708,482	685,345	490,860	271,107	1,210,752	1,829,055
Arrived Units	773,375	94,568	1,292,956	1,061,699	2,136,503	1,675,936	1,561,379
Received Units	779,506	1,117,705	1,487,441	1,281,452	1,196,858	1,057,633	922,550
End-of-day backlog	1,708,482	685,345	490,860	271,107	1,210,752	1,829,055	2,467,884
Backlog (in days)	1.52	0.60	0.41	0.23	1.03	1.56	2.08

It is also apparent that the current system requires the FC to be relieved from arriving freight on Sundays (the first day of every week). It is the lack of arriving freight on

Sunday (or sometimes negative freight[V]) that allows the site to maintain the backlog generally under 1.5 days. This is apparent when comparing the performance with that of the week of 5/29, where the FC receives freight on a daily basis, including 779,506 units on Sunday. It is apparent that given arriving freight on Sunday, the day when the FC usually processed the units in backlog, the site no longer has available capacity to process new arriving units. Since the site continues to staff according to the forecast, the backlog starts to accumulate, to the point where it reaches close to 4 million units. When the backlog reaches such unmanageable levels, any future freight is redirected to a different site, generating downstream problems such as transportation costs and suboptimal placement of inventory.

By staffing to charge, the FC is able to receive 75,424,257 units under the regular shift. However, as part of the backlog management, the site offered associates OT on a daily basis. At an assumed hourly rate of $12.95, this leads to a total overtime cost of $159,070 and an additional 2,576,239 units processed during overtime. This leads to a grand total of 78,000,496 processed units for a total processing cost of $4,688,305. Therefore, to increase efficiency, the new staffing model should have less overtime hours.

Moreover, with cost of backlog calculations, there is an added cost of backlog to the network. At the current scenario, the FC has 20 days of backlog greater than one day, and 7 days of backlog greater than 2 days. Using the cost calculations from Table 7, the total cost of backlog for the period studied was $6,149,686.

Table 9 shows the total cost of the current scenario, including labor costs, overtime costs, and the cost of backlog inefficiencies.

[V] The forecasting group uses negative arrivals when the site has asked for freight reallocation to somewhere else in the network. This means that not only were there no arrivals to the site, but units that had already arrived at the site were transferred to a different site.

Table 9. Cost Summary for Current Scenario

Regular labor cost	$4,529,236
OT labor cost	$159,070
Cost of Backlog	$6,813,663
TOTAL COSTS	$11,501,968

7.2. Scenario 2: Maximum Required Daily Capacity

7.2.1. 100% Daily Required Capacity

One hypothesis for implementing level loading was to staff to the highest daily volume of units projected for the week. For instance, Table 10 shows the forecasted arrivals for the week of 4/17; for a level-loading strategy, the headcount for the week would have been chosen as the number of associates required to process 1,975,998 units, the highest required labor capacity for the week.

Table 10. Forecasted Required Capacity, Week of 4/17/2016

	4/17/2016	4/18/2016	4/19/2016	4/20/2016	4/21/2016	4/22/2016	4/23/2016
Arrivals	-527957	881503	1305668	1649369	1875799	1975998	1120900

This method for choosing the level-loaded headcount was done for every subsequent week, resulting in the following processing capacity, in comparison to the current scenario (Figure 17).

Figure 17. Staffing-to-charge vs. Level loaded processed units

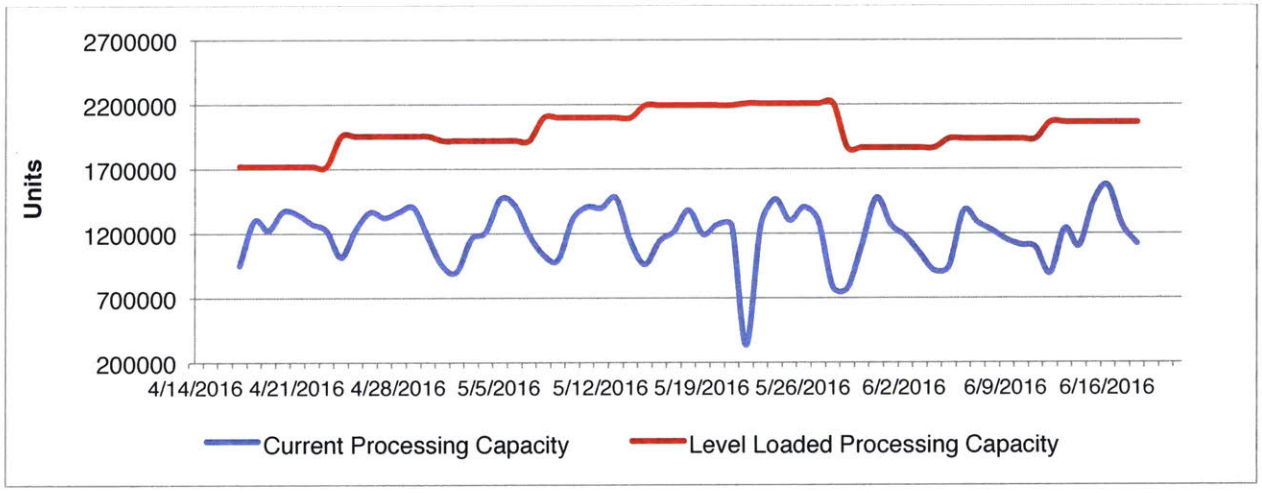
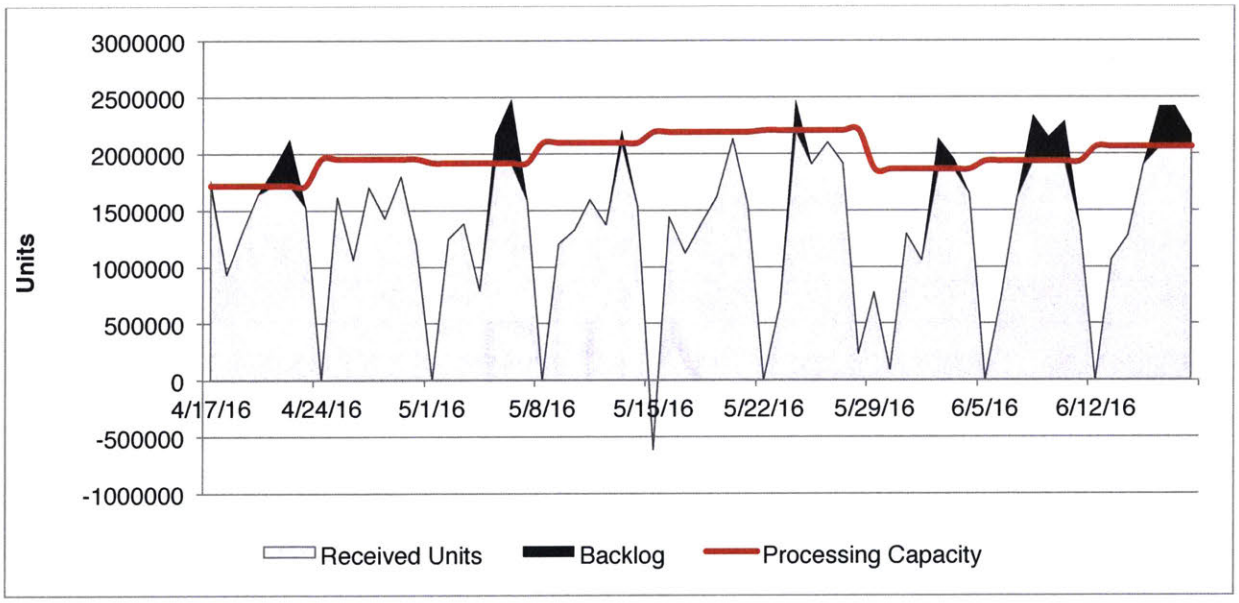


Figure 18 shows that level loading to the maximum required capacity every week would allow the FC to operate with almost no backlog. There is still a minimal level of backlog

because the labor capacity is still based on the week-ahead forecast of arrivals, which occasionally understates the level of units that arrive at the site in actuality. This scenario allows the FC to process 83,545,930 units within the same time period, about 5.5 million units more. This is due to the fact that level loading increases the FC's processing capacity from 78,000,496 units in the current scenario to a potential 129,277,721 units. This 49% increase in processing capacity would occur at the same cost of $0.06 per unit for the current scenario, leaving great room for opportunity in increasing daily arrivals and processing more freight on a daily basis.

Figure 18. Backlog Management given Level loading to Maximum Required Daily Capacity per Week



This model leads to a substantially higher cost for processed units of $7,184,421 when compared to the $4,529,236 staffing cost for the current state. However, in spite of a 53% increase in labor costs, the avoidance of any backlog costs makes this scenario plausible in terms of cost, as it avoids the additional costs of backlog and overtime labor.

While this is desirable in terms on units being processed into the network, it exemplifies the scenario that FCs try to avoid, in which there is excess capacity daily with idle associates. This is seen in the 21 instances of days in which there is not enough freight for the headcount of workers, requiring the site to request associates to VTO. As

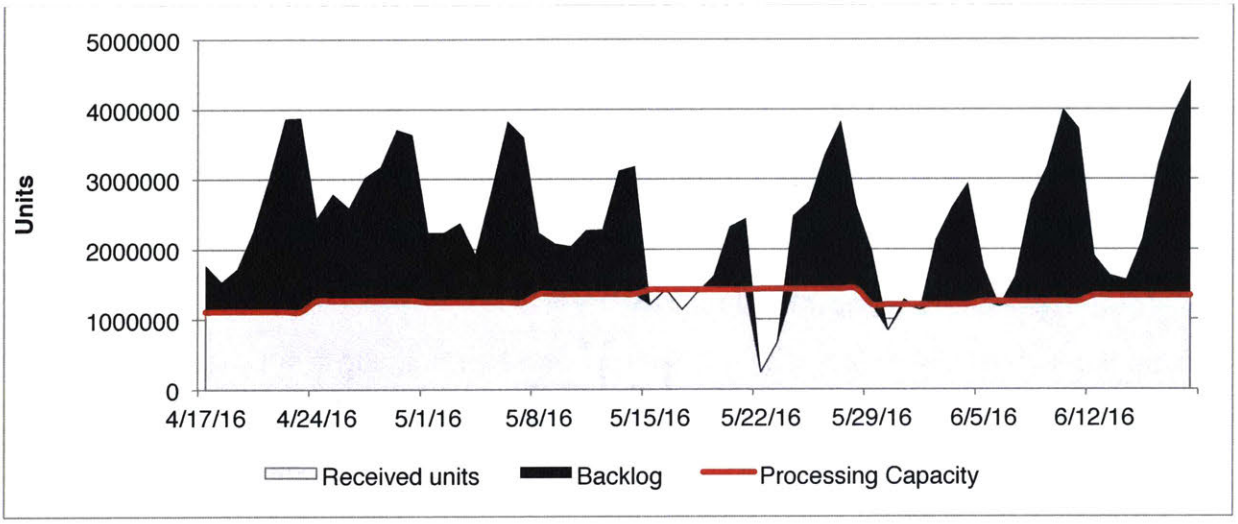
discussed in Section 2.4, VTO is a risky strategy in that associates can choose not to take it, forcing the site to pay for excess capacity. Therefore, these instances of VTO are highly undesirable for the FC in terms of unnecessary variable cost, as well as a diminished associate experience. Although more costly, the current scenario with a certain level of backlog is preferable in order to avoid idle labor. Thus, this solution was deemed as not feasible.

7.2.2. 65% Daily Required Capacity

A second level loaded scenario was further explored by choosing a sub-optimal percentage of the maximum capacity. Like in the previous scenario, the goal of level loading is to avoid overtime by staffing to a set labor capacity. Therefore, in this scenario, although backlog levels come to be high, the approach was not to institute overtime but rely on the extra capacity on days with low volume to process the backlog.

This was attempted at different percentages of the required capacity until a financially comparable solution without as much idle labor was found. This was found at 65% of the maximum capacity for the week; that is, for each week we set the daily processing capacity to be 65% of the maximum week-ahead forecast of daily arrivals. Figure 19 shows how a 65% maximum capacity would still allow the FC to process 89,811,127 units. Without requiring any overtime hours, this scenario has the capability to process more units than the current scenario, for a total labor cost of $4,669,874.

Figure 19. Backlog Management given Level loading to 65% of the Maximum Required Daily Capacity per Week



The 65% scenario leads to 21 days of backlog greater than 1 day, and 4 days of backlog greater than 2 days, for a total backlog inefficiency cost of $5,897,035. Thus, the total cost for this scenario is $10,566,909. However, while financially viable, this new solution had four more instances of backlog greater than 2 days than the 100% capacity scenario (Figure 19). Because the backlog calculation did not take into account additional supply chain costs or the cost of a negative customer experience, the instances of backlog greater than 2 days were considered highly undesirable. Additionally, this scenario was undesirable as it still produced four days in which there is not enough freight to justify the level-loaded headcount, requiring the site to employ VTO, if possible.

Table 11 shows a summary of how each scenario compares. The scenario, which processes 100% of the maximum required daily capacity results in much reduced backlog, but at an added expense of 53% of the current cost. Conversely, the scenario of 65% maximum required daily capacity requires a similar cost, but does not make any improvements to the backlog. Both level loaded scenarios present an improvement in the number of units processed in a timely manner, but the improvement is either too risky in terms of associate experience or not significant enough to convince Amazon to adopt the model for staffing sites across the network.

Table 11. Financial Summary of Different Level loading Scenarios Involving Maximum Daily Capacity Required

	Current	100%	65%
Average Backlog	0.99	0.03	0.91
Maximum Backlog	2.90	0.27	2.30
Days of Backlog >1 day	20	0	21
Days of Backlog >2 days	7	0	4
Days of VTO	0	21	4
Average Cost per Unit ($)	0.06	0.06	0.06
Days of OT	70	0	0
OT Average Cost per Unit ($)	0.09	0	0

	Current	100%	64%
Processed Units	78,000,496	83,545,930	78,688,969
Regular Time Cost of Processing ($)	4,529,236	7,184,421	4,669,874
OT Cost of Processing ($)	159,070	0	0
Cost of Backlog ($)	6,813,663	0	5,897,035
Total Cost ($)	11,501,968	7,184,421	10,566,909

7.3. Scenario 3: Average Required Daily Capacity

A second approach to implementing level loading was taken. The forecast used for the assigned units was a forecast made three weeks ahead of time. Table 12 shows how level loading would be implemented: the 7-day average of daily assigned units is taken for the upcoming week and this is determined to be the required daily labor capacity. For instance, in week 3/27, the forecasted daily average for week 4/17 is 1,177,000 units; this policy would then set the level-loaded headcount in week 4/17 to have a capability to process 1,177,000 units.

Table 12. Level loading based on 7-day average forecasted capacity, Weeks of 4/17 – 6/19

Forecast Date	Week Forecasted	Forecasted Assigned Units	7-Day Average
3/27/2016	4/17/2016	8,236,833	1,176,690
4/3/2016	4/24/2016	8,251,751	1,178,822
4/10/2016	5/1/2016	8,455,377	1,207,911
4/17/2016	5/8/2016	8,565,620	1,223,660
4/24/2016	5/15/2016	9,056,196	1,293,742
5/1/2016	5/22/2016	9,316,621	1,330,946
5/8/2016	5/29/2016	8,159,341	1,165,620
5/15/2016	6/5/2016	9,188,815	1,312,688
5/22/2016	6/12/2016	9,403,479	1,343,354
5/29/2016	6/19/2016	9,376,710	1,339,530

Figure 20 shows the comparison between current processing capacity and the new level-loaded processing capacity.

Figure 20. Staffing-to-Charge vs. 7-day Average Level-Loaded Processing Capacity-

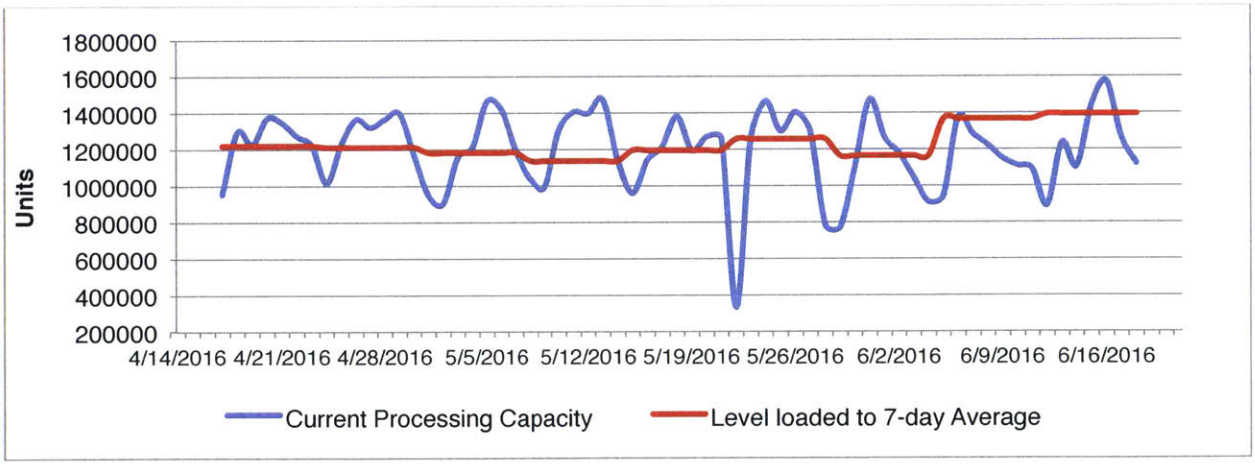
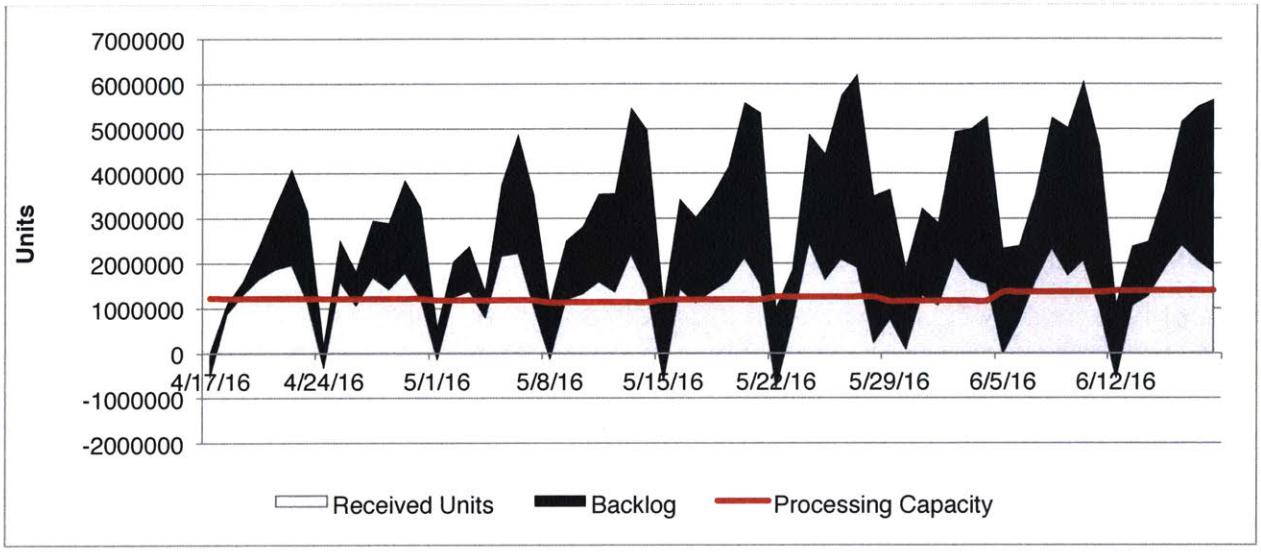


Figure 21 shows how this level loading method allows the backlog to be fully processed week-over-week; the backlog tends to grow to levels greater than 5 million units, which is around 3 days. This scenario leads to 68,157,538 units processed, which is a decrease of 12% in processed units when compared to the current scenario. Therefore, it is not a desirable solution.

Figure 21. Backlog Management given Level loading to the 7-day Average Forecasted Capacity Required



Since the 7-day average capacity did not provide sufficient capacity to process units, five other scenarios were studied. These included 100%, 110%, 115%, 118%, 120%, and 125% of the 7-day average capacity. Table 13 shows the results found for each scenario:

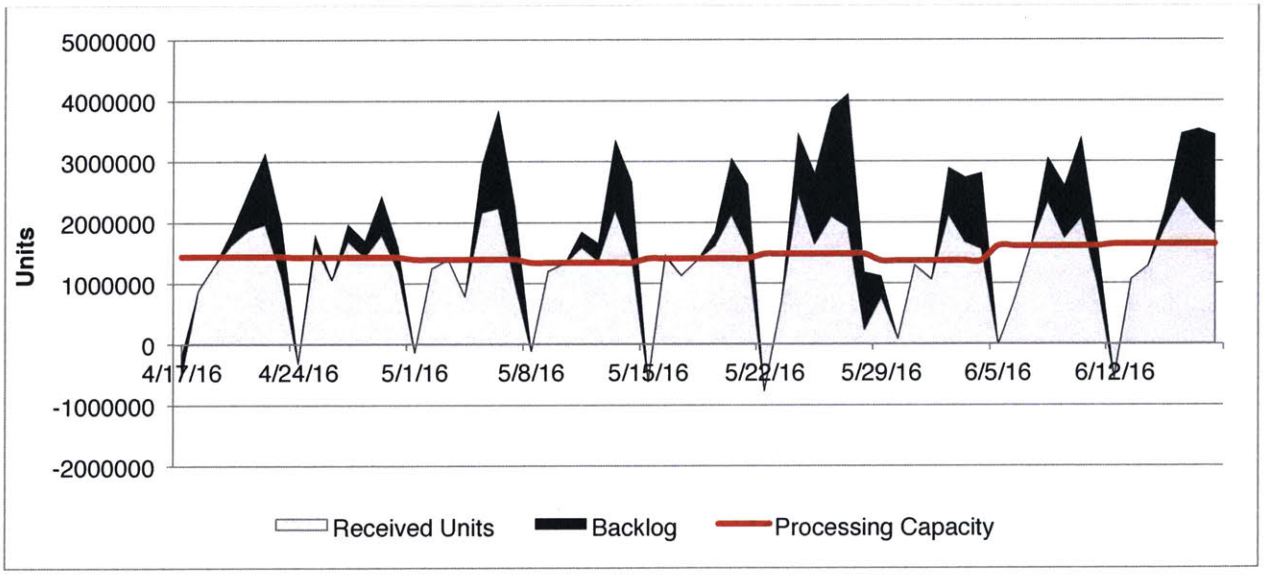
Table 13. Financial Summary of Different Level loading Scenarios Involving Variations of the 7-Day Average Forecasted Capacity

		Current	125%	120%	118%	115%	110%	100%
	Average Backlog	0.90	0.25	0.32	0.35	0.40	0.50	1.67
	Max Backlog	2.29	1.25	1.48	1.58	1.74	2.01	3.65
	Days of VTO	0	11	7	6	6	5	4
	Average Cost per Unit	$0.06	$0.06	$0.06	$0.06	$0.06	$0.06	$0.06
	Days of OT	70	0	0	0	0	0	0
	OT Average Cost per Unit	$0.09	0	0	0	0	0	0
COSTS								
Labor	Regular	$4,529,236	$5,595,145	$5,371,339	$5,281,817	$5,147,533	$4,923,728	$4,476,116
	OT	$159,070	0	0	0	0	0	0
	TOTAL	$4,688,305	$5,595,145	$5,371,339	$5,281,817	$5,147,533	$4,923,728	$4,476,116
Backlog	Days of Backlog >1	20	3	3	3	6	13	28
	Days of Backlog >2	7	0	0	0	0	1	21
	Cost of Backlog	$6,813,663	$627,967	$627,967	$627,967	$1,255,934	$3,096,507	$13,742,672
Total Cost		$11,501,968	$6,223,112	$5,999,306	$5,909,784	$6,403,468	$8,020,235	$18,218,788
Cost Savings		-	$5,278,856	$5,502,662	$5,592,184	$5,098,500	$3,481,733	$(6,716,820)

The results show that all level loading scenarios greater than 100% capacity lead to a greater volume of processed units on an overall basis with VCPU would remain unchanged. This increase in capacity is also reflected in the increase in instances of VTO being necessary, as all scenarios have an occurrence of 4 days or more where associates would run out of work and the FC would require VTO implementation. This shows that extra capacity given a level loaded model would come at an extra cost in the form of how much VTO is required and possible, and the success at which this VTO is procured. Nonetheless, all of these scenarios lead to much lower instances of VTO (compared to 21 days of VTO in the 100% 7-day average capacity scenario) while still leading to considerable cost savings of more than $3 million over the period studied.

Based on this study, level loading to 118% of the daily average of three-week ahead arrival forecast was chosen as the most cost effective option. With this 18% added buffer to capacity, the FC in question would be able to process an additional 2,133,149 units over the entire period studied, a 3% increase in processed units given the arriving units over the period studied. However, the site would actually be able to increase processing capacity from 78,000,496 to 13,978,696 units. This would be possible for an overall processing cost of $5,909,784, which is considerably more cost-effective than the current scenario. The specifics of this scenario can be seen in Figure 22.

Figure 22.Backlog Management given Level loading to 118% 7-day Average Forecasted Capacity Required



The extra capacity created would result in a maximum backlog of 1.58 days. In the level loaded model backlogs are dealt with at a faster pace, because there is always excess capacity available over the course of a week. Conversely, in the current model, the reason the backlog does not build sooner is because the site is avoiding new freight allocations on lowly staffed days. The backlog builds up the moment the site is assigned freight on a lowly staffed day (around 5/29). A level loaded staffing model would allow the site to receive freight every day. Although outside the scope of this study, with a controlled backlog, this scenario would not require as many redirects or overtime as was required in actuality due to backlogs exceeding 2.5 days.

Overall, a level loading staffing model performs better than the staffing-to-charge model in terms of cost, associate experience, and processed units for the current vendor freight allocation schedule which follows the day-of-the-week curve and dynamically adjusts the freight allocations based on backlog and labor schedules. With level loaded staffing, we expect that there can be additional benefits, as this will allow better planning of the vendor freight allocations and increase processing capacity, reducing the amount of safety stock necessary in the network.

.

8. Implementation

The main challenge with implementing level loading in an existing FC is the constraint of shift structures (Section 2.1). Associates in an existing shift can only be moved to a different shift if they are temporary associates, or if they themselves request to be switched to a different shift. For the site in question, it was found that about 6% of associates request a shift-change every month.

Therefore, an optimization model was developed to determine how long the implementation of level loading would take at the site in question. The model took into account the existing shift structures for that particular FC, and the number of temporary and full-time associates. The model operated under three main assumptions:

- Assumption 1: All temporary associates can be moved to any existing shift that is convenient for level loading in the first month.
- Assumption 2: Associate shift change requests are 6% every month consistently.
- Assumption 3: All shift change requests get approved and are assigned any existing shift that is convenient for level loading.

Thus, in the first month, all temporary associates plus 6% of full-time associates can be moved to different shifts. Every subsequent month, only 6% of associates are being moved to different shifts. The model was constructed as follows, and implemented on a month-to-month basis:

Inputs:

y_i - The number of associates assigned to shift i at the start of the current month.

z_i - The number of associates assigned to shift i that can be switched, as of start of month. This would be all of the temps, as well as 6% of the permanent associates.

$m_{ij} = 0,1$ - Indicator of whether or not shift i works on day j, where j is the day of the week starting with Sunday, $j = \{1,2,3,4,5,6,7\}$.

μ - The daily target for level loading; namely the total number of associates, multiplied by 4 days/week divided by 7 days/week.

Decision variables:

x_{ik} - The number of associates who are switched from shift i to shift k

X_j - The number of associates scheduled to work on day j.

Objective Function: To minimize the variance between the total headcount of associates for every day of the week.

$$\min \frac{\Sigma_j^7 (X_j - \mu)^2}{7}$$

Constraints:

- The number of associates switched from each shift should sum up to the number that are available for being switched.

$$s.t.$$
$$\sum_k x_{ik} = z_i, \forall i$$

- The number of associates working on each day must equal the number of associates scheduled to work that day.

$$\sum_i m_{ij} \left(y_i - z_i + \sum_k x_{ki} \right) = X_j, \forall j$$

- The number of associates who switch must be non-negative.

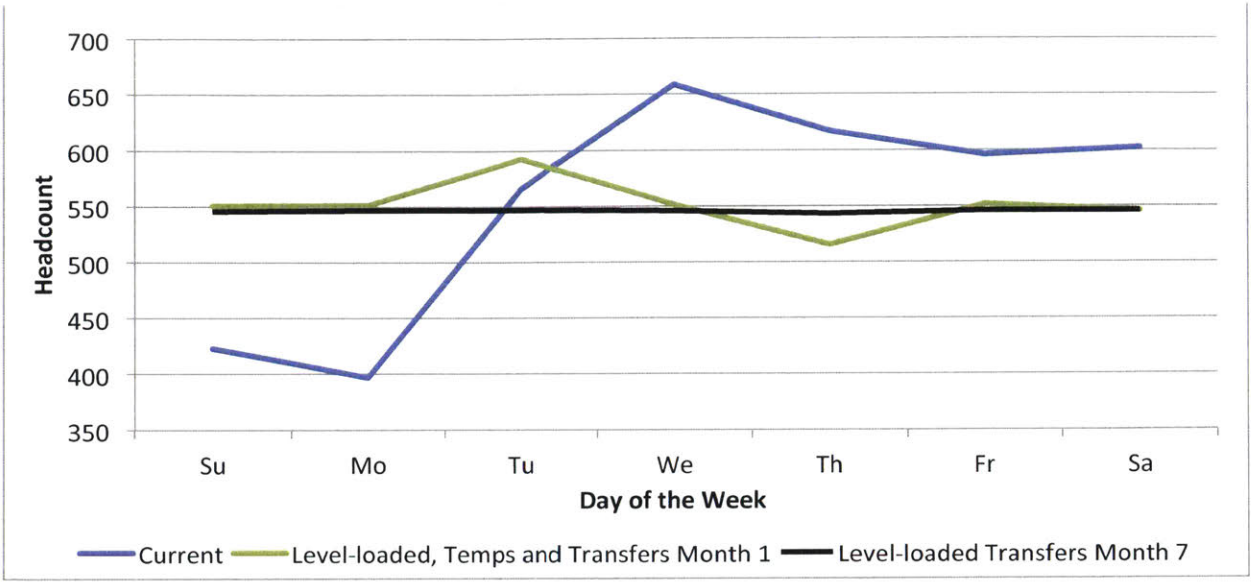
$$x_{ik} \geq 0, \forall i, k$$

- The number of associates scheduled to work on a certain day must be non-negative.

$$X_j \geq 0, \forall j$$

This optimization model was run for the FC in question, one month at a time. It was found that given the current shift structures and number of full-time and temporary associates hired on October 2016, after the first month, the daily headcount can be very much level loaded with only a higher than average headcount on Tuesday and a lower than average headcount on Thursday. Subsequently after that, it requires six more months of full time associates requesting transfers at a rate of 6% every month for the FC in question to completely level load its daily headcount to 550 associates everyday (Figure 23).

Figure 23. Headcount Changes as a Result of Level loading Implementation



This model could be further applied to any FC taking as input their own shift structures, number of temporary and full-time associates, and the percentage of shift request changes.

9. Conclusion and Recommendations

Level loading has proved to be a more cost-effective staffing model, even under the current freight allocation model. The chosen labor capacity of 118% the average forecasted daily required capacity led to cost savings of $5,592,184 over the 76-day period studied. The cost savings are possible because the site is no longer at the mercy of incorrect assumptions and a commonly inaccurate forecast, and thus ensures that labor is always available to process incoming freight. These costs savings apply to only the site in question, showing that there are greater cost opportunities available if applied to the remaining IXDs in the NACF network.

Moreover, level loading is a more desirable model of staffing because it minimizes the instances of overtime, bringing more stability to the associate experience. While instances of VTO increase, these occur under the current freight allocation model. However, by implementing level loading, Amazon can have freight scheduled and planned according to a known labor capacity, avoiding mismatches between labor and processed units, and thus avoiding the cost inefficiencies that arise as a result.

Additionally, level loading increases the overall processing capacity for the site. With an increased capacity, the site is able to process more units under less time. This increase in efficiency of processed units should lead to Amazon requiring lower levels of safety stock. With lower safety stock, the need for warehouse space is reduced. As Amazon continues to grow, the reduced need for safety stock could further reduce future costs in terms of future warehouse infrastructure.

By making daily freight allocation dependent on labor capacity, there are further improvements that Amazon can make in order to maximize labor utilization, optimize the injection of freight into the NACF network, and ultimately improve the customer experience.

9.1. Optimal Staffing Freight Mix

As mentioned in section 5.4, sites have major staffing challenges on a day-to-day basis because the nature of the units arriving changes too drastically day after day. This is

largely due to the current method of making a rate assumption for any freight mix arriving. As IXDs become the sites of inbound processing in the network, the inflow according to freight mix will become more important.

A more robust staffing model would not only level-load, but also determine a more accurate hourly rate and thus headcount based on the freight mix to be received. This should reduce the overall amount of inaccuracies in terms of labor capacity due to rates in the FC and the amount of OT/VTO used, as sites will be optimally staffed to process the type of units they will receive.

By knowing the type of items received, IXDs could guarantee a constant output of volume on a daily basis taking the form of a constant flow of the right mix of product into the network, outputting the same mix of products at a similar rate.

Based on the freight mix that the NACF network receives in a month, the optimal freight mix required from each IXD can be determined. An investigation into what this freight mix would be was done for FCs that currently receive volume from IXDs. The investigation determined what was received at every destination FC over a four-week average, and inferred the ratios of products that the network as a whole currently needed. From this, the optimal freight output for the site in question under the current state was determined by looking at the freight mix of destination FCs and determining what percentage of this freight came from the site in question. Table 14 shows the optimal freight mix for the site in question, under the current state of freight allocations.

Table 14. Optimal Freight Mix Percentages

	Small	Medium	Large	Heavy/Bulky	Case	LP	Pallet	Prep	TOTAL
%	50%	31%	0.4%	0%	0.2%	6%	11%	1.4%	100%

To make this staffing solution robust for the long term, freight mix will also consider product line proportions changing according to alterations in customer demand (E.g.: demand for physical books decreasing in the future).

9.2. Automated Scheduling and Buying by Optimal Freight Mix

Under the current system, buying takes place on a weekly schedule, where Purchase Orders (POs) are placed on the same day of the week. Simultaneous high volume arrivals and somewhat unpredictable VLTs lead to the current backlog management system of 0.5-0.8 days of backlog. As proven by this investigation, backlogs can lead to product being available to customers later than needed. If this is a recurring trend, product unavailability due to backlog leads to greater levels of safety stock in the future. Greater levels of safety stock require greater storage capacity, leading to the creation of additional FCs, which imply a major capital investment for Amazon.

However, by having a pre-determined labor capacity, freight could be ordered and set to arrive at a time when it can be processed and in the correct proportions, and buying would be able to place POs on a schedule that optimizes VLT per vendor. This would be possible through an optimization model that had information on every product regarding the unique VLT for the product and the frequency of inventory restocking. Instead of placing all orders on the same day, the model would be able to optimize PO placement according to which items are required by the network according to optimal freight mix taking into account the unique VLT.

Bibliography

[1] An, Yu, Yu Zhang, and Bo Zeng. "The Reliable Hub-and-Spoke Design Problem: Models and Algorithms." *Transportation Research Part B: Methodological* 77 (2015): 103–122. *ScienceDirect*. Web.

[2] Reis, L., et al. "APPLICATION OF LEAN APPROACHES AND TECHNIQUES IN AN AUTOMOTIVE COMPANY." *Romanian Review Precision Mechanics, Optics & Mechatronics* 50 (2016): 112.

[3] Sasser, W. Earl. "Match Supply and Demand in Service Industries." *Harvard Business Review* 54.6 (1976): 133–140. Print.

[4] Olhager, Jan, and Pontus Johansson. "Linking Long-Term Capacity Management for Manufacturing and Service Operations." *Journal of Engineering and Technology Management* 29.1 (2012): 22–33. *ScienceDirect*. Web. Creating Competitive Edge in Operations and Service Management through Technology and Innovation.

[5] Davydenko, Andrey, and Robert Fildes. "Measuring Forecasting Accuracy: The Case of Judgmental Adjustments to SKU-Level Demand Forecasts." *International Journal of Forecasting* 29.3 (2013): 510–522. *ScienceDirect*. Web.

[6] Prestwich, Steven, et al. "Mean-Based Error Measures For Intermittent Demand Forecasting." *International Journal Of Production Research* 52.22 (2014): 6782-6791. *Business Source Complete*. Web. 18 Feb. 2017.

[7] Kim, Sungil, and Heeyoung Kim. "A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts." *International Journal of Forecasting* 32.3 (2016): 669–679. *ScienceDirect*. Web.

[8] Lewis, C. D. *Industrial And Business Forecasting Methods : A Practical Guide To Exponential Smoothing And Curve Fitting*. n.p.: London ; Boston : Butterworth Scientific, 1982., 1982. *MIT Barton Catalog*. Web. 18 Feb. 2017.

[9] Franses, Philip Hans. "A Note on the Mean Absolute Scaled Error." *International Journal of Forecasting* 32.1 (2016): 20–22. *ScienceDirect*. Web.

[10] Kolassa, Stephan. "Evaluating Predictive Count Data Distributions in Retail Sales Forecasting." *International Journal of Forecasting* 32.3 (2016): 788–803. *ScienceDirect*. Web.

[11] Chase, Charles W. "Innovations in Business Forecasting." *The Journal of Business Forecasting* 33.1 (2014): 29. Print.

[12] "The Process Enterprise: An Executive Perspective (PDF Download Available)." *ResearchGate* n. pag. *www.researchgate.net*. Web. 22 Feb. 2017.

[13] Cui, Liying. *10k ASIN Project*. Rep. N.p.: Amazon, 2016. Print.