Dynamic Order Allocation for Make-To-Order Manufacturing Networks: An Industrial Case Study of Optimization Under Uncertainty

by

Gareth Pierce Williams

B.S., University of California, Berkeley (2006)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
AUG 0 1 2011
LIBRARIES
ARCHIVES

© Massachusetts Institute of Technology 2011. All rights reserved.

Author . . . .
Sloan School of Management
May 20, 2011

Certified by . . . .
Jérémie Gallien
Associate Professor of Management Science and Operations
London Business School
Thesis Supervisor

Accepted by . . . .
Patrick Jaillet
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science
Codirector, Operations Research Center

Dynamic Order Allocation for Make-To-Order Manufacturing Networks: An Industrial Case Study of Optimization Under Uncertainty

by

Gareth Pierce Williams

Submitted to the Sloan School of Management
on May 20, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Planning and controlling production in a large make-to-order manufacturing network poses complex and costly operational problems. As customers continually submit customized orders, a centralized decision-maker must quickly allocate each order to production facilities with limited but flexible labor, production capacity, and parts availability. In collaboration with a major desktop manufacturing firm, we study these relatively unexplored problems, the firm's solutions to it, and alternate approaches based on mathematical optimization.

We develop and analyze three distinct models for these problems which incorporate the firm's data, testing, and feedback, emphasizing realism and usability. The problem is cast as a Dynamic Program with a detailed model of demand uncertainty. Decisions include planning production over time, from a few hours to a quarter year, and determining the appropriate amount of labor at each factory. The objective is to minimize shipping and labor costs while providing superb customer service by producing orders on-time. Because the stochastic Dynamic Program is too difficult to solve directly, we propose deterministic, rolling-horizon, Mixed Integer Linear Programs, including one that uses recently developed affinely-adjustable Robust Optimization techniques, that can be solved in a few minutes. Simulations and a perfect hindsight upper bound show that they can be near-optimal. Consistent results indicate that these solutions offer several hundred thousand dollars in daily cost saving opportunities by accounting for future demand and repeatedly re-balancing factory loads via re-allocating orders, improving capacity utilization, and improving on-time delivery.

Thesis Supervisor: Jérémie Gallien
Title: Associate Professor of Management Science and Operations
London Business School

Acknowledgments

Foremost, I would like to thank my advisor, Jérémie Gallien, who made this work possible by introducing me to the problem, providing contacts and professional knowledge, and easing responsibility into my hands. I greatly appreciate his kind, just and wise guidance, which continued throughout my doctoral work.

Steve Graves and Cindy Barnhart, the other members of my thesis committee, imparted crucial feedback. John Foreman, who worked on a related project, aided me in many ways. Several employees at the firm motivating this work, especially Juan Correa, Spencer Wheelwright, and Jerry Becker, were instrumental in the acquisition of data and model development in this thesis.

This firm, the Singapore-MIT Alliance, research assistantships from Jérémie, and teaching assistantships funded my graduate eduction. It was an honor and a pleasure to teach with Jérémie, Vivek Farias, Arnie Barnett, Dick Larson, Cynthia Rudin, and Jim Orlin. Cindy Barnhart, Dimitris Bertsimas, and Patrick Jaillet, as ORC codirectors, all provided superb advice throughout my doctoral candidacy. The ORC administrators, Laura Rose, Paulette Mosley, and Andrew Carvalho, went above and beyond helping.

My colleagues at the Operations Research Center, many of whom contributed ideas to this thesis and are too many to list, were a great source of camaraderie, moral support, and intellectual stimulation. My officemates at the Operations Research Center, Alex Rikun, Ruben Lobel, and Jonathan Kluberg, made working at the office a pleasure. I give special thanks to Cristian Figueroa, Philipp Keller, Eric Zarybnisky, Michael Frankovich, Zachary Leung, Matthieu Monsch, Nikos Trichakis, David Goldberg, Diana Michalek Pfeil, Jason Acimovic, Dan Iancu, Adrian Becker, Matthew Fontana, and Vishal Gupta. My companions, Ying Zhu and Deborah Witkin, were always there for me.

Lastly, I would not be who I am without the loving and everlasting support of my family.

Thank you all!

Note on Confidential Information

So as to protect the firm's proprietary and confidential information, much of the data presented in this thesis has been changed to prevent access by competitors. Although specific values of many parameters vary from their true historical value, the relative values and qualitative results that we present still represent reality well.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis addresses production planning and control problems encountered by make-to-order manufacturers that have multiple production facilities. As rapid advances in technology improve the availability of information, global supply-chain controls are being developed to improve market responsiveness to shifting consumer demands. In make-to-order network manufacturing, firms must use automated controls to quickly and efficiently allocate thousands of custom orders to multiple manufacturing facilities.

This work is motivated by and performed in collaboration with a particular, large, make-to-order desktop computer manufacturing company, hereafter referred to as "the firm," that needed such controls. The firm is a $61B annual revenue corporation, based in Austin, Texas, United States, that designs, manufactures, sells and supports computer-related products. In North America, the firm manufactures hundreds of thousands of consumer and corporate desktop computers each week. Rather than selling via typical retail channels, the firm developed an innovative make-to-order and direct factory-to-customer shipping business model, which many other companies have now adopted. This business model provides great value to customers by tailoring products to their desires and reduces inventory requirements by assembling finished goods Just-In-Time. However, this business model makes outsourcing final assembly of products difficult and increases direct labor and shipping costs. In North America, the firm produces desktops from multiple factory locations to improve delivery lead times, reduce shipping costs, and mitigate the risk of losing all manufacturing capability. Uncertainty regarding the quantity, timing, and geographic destination of future orders for desktops complicates the decision of how to allocate orders to factories and adjust production capacity to match demand. Because computer manufacturing is one of the most rapidly growing and competitive industries with products that are becoming more difficult to differentiate, cost-advantages are critical to gaining a competitive advantage. The firm's flexible, complex, and cutting-edge supply chain presents an excellent opportunity to employ optimization-based solution techniques in an industrial setting.

The first major contribution of this thesis is exposition of this industrial problem which has received little attention in the literature. Chapter 2 presents the business problem faced by the firm and its innovative and evolving supply chain configuration, illustrating a problem with several sets of industrial data that requires further research and setting the context for the remainder of the thesis. The fundamental question posed in this problem and answered by this thesis is "Which desktops should be built, when and where?" Intricacies of the problem, its associated challenges, and the firm's approaches to solving it are discussed in detail. Chapter 3 reviews the relevant academic literature, including production planning and control in make-to-order and network settings, optimization-based solution techniques, similar industrial studies, and other work related to the firm.

Chapters 4, 5, and 6 contain the second major contribution of this work: three distinct models of the problem which incorporate the firm's actual data, testing, and feedback, discussions of challenges to optimization modeling in practice, and actionable solutions and insights. Analysis of these models demonstrates credible and realistic cost savings opportunities from optimization-based solutions. In all of these models, decisions include producing various desktops at factories over time and determining the appropriate amount of production capacity. The objective is to minimize the sum of several relevant supply-chain costs, including factory-to-customer shipping costs, the cost of labor that is used directly to produce the desktops, and the cost of poor customer service. Emphasis is placed on model realism and usability.

Several important questions are addressed in each of these chapters. How can we model the firm's problem mathematically? What choices are appropriate when balancing model tractability and realism? What challenges can be encountered when using mathematical optimization techniques to solve an industrial problem in practice? Compared to the firm's historical decisions, can mathematical optimization leverage the firm's manufacturing network to reduce relevant supply chain costs and improve customer service? If so, by how much and by making what decisions? Should the firm be producing orders in different factories or at different times? What is the appropriate amount of capacity to have at each factory? Can customer service be improved by satisfying more orders on time? How confident can we be in our answers to these questions?

In Chapter 4, the problem is formulated mathematically as a Dynamic Programming problem and demand is analyzed and modeled in great detail. A simulation study of various solution policies shows that a rolling-horizon, certainty-equivalent, linear-programming policy performs near-optimally, improving upon the firm's historical policy by several hundred thousand dollars per day.

In Chapter 5, the same problem is addressed deterministically by a large Mixed Integer Linear Program (MILP) with many details that were necessary for implementation at the firm but too intricate or unintuitive for the in-depth analysis of Chapter 4, exposing many issues and insights that arise when using optimization in practice. Analysis of the solution to the MILP indicates similar potential cost savings of several hundred thousand dollars per day relative to the firm's actual decisions.

Chapter 6 studies the same problem but plans for production over the next quarter-year rather than the next few days. Another MILP is developed, incorporating the cost to ship parts from suppliers to factories and decisions regarding how much staff to hire at each factory. Because demand data was limited and solutions suggested drastic reductions in labor levels, Robust Optimization techniques are introduced, including discussions of how to model uncertainty appropriately and techniques to maintain tractability. Results demonstrate that, even under extreme levels of protection against uncertainty, optimization-based solutions can provide similar cost savings of several hundred thousand dollars per day.

We coalesce these results and insights in Chapter 7, our conclusion. Many firms now face difficult decisions in a make-to-order network manufacturing environment. This thesis presents a thorough and grounded discussion of one such industrial problem. Realistic and tractable modeling choices, which often go without much discussion, in addition to substantial financial impact, are necessary for optimization-based solutions to be used in practice. Consistent and data-driven results show that these controls can provide significant cost savings by dynamically allocating orders among production facilities to continually re-balance factory loads. The insights gained from thoroughly studying this firm's problems are readily applicable to other firms facing similar production planning and control problems in make-to-order manufacturing networks.

Chapter 2

Case Study: Geo-manufacturing

This chapter describes the challenging production planning and control problems that were faced by the firm between 2006 and 2010 and are addressed in the remainder of this thesis. In §2.1 we describe the relevant history of the firm's supply chain, providing the problem framework. The problems are stated in §2.2. Further context is provided in the summary of several interviews in §2.3 which illustrate the firm's production capacity and labor force limitations at each factory. The solution that the firm was already using and provides the historical baseline for our study is described in §2.4.

2.1 The Supply Chain

The firm develops, assembles, sells and supports computers as well as related products and services. The firm is well known for its brand-name products and supply-chain innovation, shipping more than 110,000 computers every day to customers in over 180 countries. According to its website, in the third quarter of its 2011 fiscal year (ending October 29, 2010), the firm had a revenue of $15.4 billion, an operating income of $1.02 billion, a net income of $822 million, and earnings per share of $0.42.

The firm's unique and ground-breaking supply chain began in 1984 when it was founded in Austin, Texas, USA based on the idea that selling computers directly to the final customers would enable the best satisfaction of customer needs. Bypassing the wholesalers and retailers, which are common in other computer distribution channels, allowed the firm to let customers configure orders to their own specifications and conferred greater control over its supply chain. Whereas most personal computer vendors must forecast demand and build-to-stock, the firm's direct-sales and build-to-order business model allow it to have excellent performance in inventory turnover, overhead, cash conversion, and return on investment. Although the firm relies on outside suppliers and contract manufacturers to provide many components of its products, it performs the majority of final assembly for desktops itself. Instead of owning its own parts inventory, suppliers own and manage parts in Supplier Logistics Centers (SLCs) near each of the firms factories; every computer the firm builds has already been sold before the firm owns the parts, a new and enviable business model for the computer industry. By having such close relationships with both customers and suppliers, the firm had an immense amount of information, allowing them to quickly respond to customer demand. However, the direct-sales model requires quick responsiveness in manufacturing capability and the information technology to support swift order-fulfillment. Hence, the firm must maintain excess production capacity to deal with demand volatility and hedge against significantly higher outbound shipping costs, striking the correct balance of production at each factory over time.

Orders are configured and placed in-person, by phone, or via the firm's website. Material Requirements Planning (MRP) software, combined with supervision from the firm's Operations Center and factory managers, determines when and where desktops will be assembled; these decisions are the crux of what we study. Every two hours, supplies are then requested from nearby vendors for orders that the MRP system determines should be built in the next few hours; vendors have two hours to deliver those parts from the SLC. The factory then puts the parts for each desktop into kits, assembles the hardware, loads software, and tests basic functionality of these computers, using a substantial amount of human labor. The computers are then automatically packed in boxes that are later shipped from the firm's factories directly to consumer's doorsteps via third party logistics providers. This thesis focuses on the decision of when and where each desktop computer should be assembled.

Name	Location	Start	End
TX	Austin, TX, USA	1984	2008
TN	Nashville, TN, USA	1999	2009
NC	Winston-Salem, NC, USA	2005	2010
JM	San Jeronimo and Juarez, Mexico	2009	Present

Table 2.1: The name, location, first year of production (start), and final year of production (end) of relevant production facilities.

The firm's North American manufacturing network has evolved significantly. In 1984, the firm was founded in Austin, Texas (TX), where all manufacturing took place until 1996. In 1999, in order to increase production capacity, reduce the cost and lead-time of shipping directly to customers, and to reduce the risk of a disaster destroying all of its production capability, the firm opened a manufacturing facility in Nashville, Tennessee (TN). In 2005, it opened a third United States manufacturing facility in Winston-Salem, North Carolina (NC), where it received an incentive package "worth $240 million over 20 years from local and state governments" [Lad09] in exchange for meeting minimum employment targets. As demand for computers shifted toward notebooks, as the firm began selling via retail channels, and as investors pressured the firm to cut costs, starting in 2008, the firm began to terminate desktop production in its United States manufacturing facilities [Sch08]. The firm ended the manufacturing of new non-server desktops in Texas in 2008, in Tennessee in 2009, and in North Carolina in 2010. In 2009, it began outsourcing North-American desktop production to another firm with factories located in Mexico. Table 2.1 details the firm's North American manufacturing facilities, giving the names we refer to them by throughout this thesis, their geographic location, and the years they began and ended production.

This thesis focuses on desktop computer assembly in North American markets between September 2006 and April 2009, before the firm began retail distribution in North America, when it used a primarily build-to-order business model. At the time, it had two major desktop Lines of Business (or product categories), which we refer to as consumer desktops and corporate desktops. The consumer oriented desktop line focused on value, reliability, and modularity. The corporate desktop line, focused on longevity, reliability, and serviceability. In the time of this study, the firm assembled nearly 150,000 consumer desktops and 100,000 corporate desktops for customers across North America each week in two or three of its North American factories. The firm's North American customer base was distributed across the continent. The international nature of shipping to Mexico and Canada limited the production for most non-U.S. based customers to TX and TN, respectively. However, orders from across the United States were typically eligible to be built in almost any factory.

Having multiple factories capable of serving the same customer base with a Make-to-Order business model, enables much more dynamic production decisions than typical. An identical order made one day later or from a few miles away can easily be built in a different factory. However, the immense number of possible options and the complex dynamics of the system make such production decisions difficult. As shown by the following work, Operations Research techniques can help maintain efficient operations in such a dynamic production environment. As discussed in §3.2, although the work outlined below analyzes a supply-chain that no longer exists, many companies, often in other industries, have similar supply-chain configurations and should find this study useful. The Operations Center faced the difficult problem of determining which orders should be produced in which factories and at what times. This thesis addresses that problem.

2.2 Problem Scope and Definition

We began working with the firm's North-American Operations Center, responsible for centralized supply-chain coordination in North America, in 2006. The Operations Center was responsible for assigning demand for various desktops with varying due-dates, parts requirements, and shipping destinations across the continent, to the three active manufacturing facilities, which have various supply and manufacturing capacities. Although the Operations Center had developed heuristics to handle these tasks, it was unsure of their efficacy. The fundamental question answered by this thesis is "Which desktops should be built, when and where?"

At the time, the firm made decisions regarding this at three levels or scopes. At a strategic level, with a horizon of about three to twenty years, the firm's senior management decided to open or close assembly facilities in different locations, as described in §2.1; we do not address this problem. At what we call the *planning* level, the Operations Center planned production, staffing, and parts-sourcing targets for each factory for a quarter-year or more. At the *execution* level, which concerns day-to-day operations, looking at most two weeks into the future, factory managers and the Operations Center determined when and where each order is fulfilled and how long hired labor would be needed on the factory floor. This thesis focuses on the planning and execution problems, assuming that the factory locations are fixed but that production and labor-capacity decisions must be determined.

2.2.1 Planning

In order to inform parts supply decisions and factory staffing decisions, the firm plans its production for the next quarter (or occasionally year). Forecasts[1]for that quarter's sales volume, for each major Line of Business, were distributed among each week of the quarter based on historical percentages. The Operations Center, being responsible for production decisions, assigns this forecasted demand to different factories. Other groups within the firm then procure parts from vendors based on these production targets and factory managers hire sufficient labor to meet these production plans. Although the Operations Center does not make labor and parts sourcing decisions directly, it does consider the implications of its production decisions on other parts of the supply chain. In the planning problem, the major decisions that the firm plans for are 1) the volume of demand for various products that each factory will serve in each week, along with the associated production and backlog levels, and 2) the amount of labor necessary to serve that demand.

In the production of a desktop, parts components are assembled into into final

[1] We did not investigate alternatives to the firm's forecasting method, as it incorporates much beyond the scope of our project, including strategic marketing decisions and executive desires. However, we do analyze forecast data available to the Operations Center.

products. The firm shares forecasts of future demand with its parts suppliers who then manufacture and ship to the firm what they think will be a sufficient supply of parts. Although the suppliers own and manage the parts until just a few hours before assembly, the firm makes routing decisions regarding which purchased parts should be delivered to each factory about a month in advance of their arrival to the United States. Parts arrive from mostly Asian suppliers in Long-Beach, California, and are shipped via truck or train to each of the Supplier Logistics Centers near the firm's manufacturing facilities. Foreman [For08] addresses the problem of routing these parts for the same firm and its many complexities in great detail. The cost to ship various parts to different factories largely depends on the number of parts that fit on a shipping pallet, the mode of transit used, and the distance to the factory. Because parts routing decisions are made by another department within the firm and heavily depend on information that becomes available after production plans have been made, the Operations Center considers the implications of its production decisions on parts routing by using the average cost to ship parts to each factory, which is referred to as the *inbound shipping cost.*

Customers can choose how quickly they would like their order to be fulfilled from a set of limited options (e.g. 2, 3, 5, or 7 days) which, along with the Operations Center's choice of manufacturing location, determines the due date by which those orders must be produced. Failing to produce an order by its due date is considered poor customer service and can incur significant costs to the firm. The costs include order cancellations, contacting or being contacted by customers, concession of other valuable goods to appease customers, reduced likelihood of future purchases from the firm, and expedited third-party shipping. Dhalla [Dha08] analyzes these costs in great detail.

In a Make-to-Order business model, production for orders can only occur after customers configure those orders. The firm must produce these orders in the few days between when the order is made and when it is due. To do so, it must schedule sufficient capacity to assemble the desktops. Production capacity is limited by two expensive resources: 1) the *physical* layout and machinery of the factories and 2) the amount of *labor* available to operate the factory. The physical layout includes space for workers to assemble desktops and store work-in-progress (WIP) inventory on the factory floor and space to keep not-yet-shipped but finished goods. Machinery includes equipment that burns software onto hard-drives, tests machine functionality, boxes desktops, labels boxes, and sorts and shrink-wraps them for shipping. Purchasing additional machinery or changing the factory layout was beyond the scope of the Operations Center's decision making. However, the Operations Center did determine machinery utilization by assigning orders to each facility and thereby influence the amount of labor available to operate the machinery.

Producing desktops at the firm's factories requires a significant amount of direct labor to gather the correct components (called "kitting") and assemble them, which scales in proportion to production volumes. Labor varies in several ways that affect production capacity. The number of workers and quality of workers determines how quickly parts are gathered and assembled and therefore the rate that desktops can be assembled. Production in any period is limited by this rate multiplied by the amount of time that these workers operate the factory. Because limited space for WIP is available, as desired in a Make-to-Order environment, a steady flow of material must be supplied to the machinery; hence, another limit on production is the rate that machinery can process desktops multiplied by the amount of time that workers operate the factory. As such, the number and quality of workers as well as the amount of time they work are critical capacity decisions.

Factories employ both permanent and temporary laborers. Permanent hires tend to stay at the firm for at least a few months if not many years and can take weeks to recruit. Temporary laborers are available within a few days notice and may be hired for just one day or for a few weeks but are often less skilled at production tasks. The firm limits the number of temporary workers to be less than some fraction of permanent workers at each factory to both ensure quality and to be able to ensure enough people have appropriate training for each task. Although permanent labor tends to be more expensive and less flexible in quantity, their expertise is necessary for quality.

The workers who perform this direct labor are assigned to work-teams at each factory that operate shifts of varying lengths and frequencies. Typically, a work-team will be scheduled to eight-hour shifts on five days of each week, or ten-hour shifts four days per week, or twelve-hour shifts three days per week. These planned shift lengths and frequencies are often referred to as nominal or straight-time hours. Although the firm plans this shift structure, factory mangers often deviate from it and ask work-teams to either work longer shifts or go home early. Usually, all workers in a work-team will end their shift at the same time. If a work-team works less than their nominal number of hours in a pay period, the firm almost always pays the workers for all of the nominal hours anyways, making it a sunk cost. However, if a work-team works more than their nominal number of hours in a pay period, the firm pays an additional cost for each overtime hour. For instance, if a work-team has five eight-hour shifts per week, it will have eighty straight-time hours per pay period. If the work-team works less than eighty hours, it is still payed for eighty straight-time hours. If it works for eighty-five hours in those two weeks, it is payed for eighty straight-time hours and five overtime hours. Each shift has a minimum and maximum length which limits the amount of WIP they can kit, assemble and pass downstream to more automated machinery.

After a desktop has been assembled, loaded with software, tested, and boxed, it is shipped directly to the customer via a third party logistics provider. Although the customer may pay a shipping-fee to the firm at the time of ordering, the firm pays the third party logistics provider different prices based on both the manufacturing location and the customer's shipping address, which we typically call the destination, in addition to factors such as speed of delivery and the size or weight of each box. The choice of where each order is produced and what destination it is shipped to is a significant factor in how much the firm pays to third party logistics providers, which we call the *outbound shipping cost*. Because the firm ships desktops from its factories to customers in small quantities, preventing economies of scale, the outbound shipping cost can become relatively expensive and important in determining where an order should be produced.

The planning problem faced by the firm's Operations Center is determining how to allocate demand for a quarter-year's worth of desktops to the firm's North American production facilities. It should consider inbound shipping costs, outbound shipping costs, the cost of direct labor, limitations of the labor force, constraints on production capacity, and possibly uncertainty in demand and how demand differs from its forecast. Balancing all of these factors simultaneously can be difficult and mismanagement can cost several hundreds of thousands of dollars per day.

2.2.2 Execution

All of the issues encountered in planning problem, other than changing the size of the labor force, apply to the execution problem as well. Nonetheless, as the execution problem considers day-to-day operations, it contains many more fine details that must be considered. Some orders can only be satisfied by particular factories for a variety of reasons, including parts availability, labor expertise, customer requests, and legal issues. Scheduling the labor force become much more complex. Many more details are known about orders that have been configured and should be incorporated into the decision-making process.

In the execution scope, daily decisions are made regarding which orders are built in each factory and the how long each work-team operates the factory at its predetermined staffing level. Each day, from a backlog of available-to-build (ATB) orders from the recent past, often zero to four days worth of production, the Operations Center must decide whether to leave orders where they were assigned by the default plan or 'move' them from one factory to another. Further complicating this, more orders will be made and become due in the near future. The rate at which each work-team can produce desktops has already been determined by past staffing decisions, but the length of time they spend producing desktops has yet to be finalized. With projections for future sales and schedules for labor availability over the next two weeks, decisions must be made for what is to be done today. Factories then execute those decisions by producing the associated desktops.

Orders that customers have already configured, which are said to be in ATB, specify the quantity and type of desktops, necessary parts, due-date, shipping destination, and which factories can produce them. Little is known about orders that have yet to be configured other than forecasts of the total sales volume each day. As in the planning problem, orders must be fulfilled between the time that they are configured and the time that they are due; if an order is not fulfilled until after its due-date, customer service deteriorates. Because the volume, parts requirements, due-dates, shipping destinations, and eligibility of future orders is uncertain, making production decisions can be difficult. Large volumes can either force factories operate longer and at greater expense than planned or delay orders past their due dates. Producing too much too early can leave factories starved for work when sales volumes are low, wasting valuable resources such as labor that will be paid for anyway. Nonetheless, factories can pool some of their capacity by compensating for imbalances between them.

The Operations Center must decide which orders to satisfy immediately and which orders should be delayed until a later date or moved to other factories. Because bulky, heavy desktops are expensive to ship directly to customers, the choice should include considerations for outbound shipping costs. These decisions also alter the length of each work-team's shift which can cause non-trivial scheduling complications. The duration of most shifts is limited to an interval of time. In order for a shift to be extended beyond its nominal length, advanced notice must be given to the work-team one to two days in advance, depending on the day of the week. Similarly, work-teams can be called in for new shifts or added to existing ones. If a shift is extended long enough, it may overlap with another shift, which has different ramifications for physical bottlenecks and labor productivity; machines and space are still limited to the same production rate, but desktops can be kitted and assembled almost twice as fast. The number of hours worked so far in each pay period is tracked and can be used to predict the cost of overtime. Concerns about fairness in balancing the workload of each factory must be considered. Not only must the lengths of the current days' shifts be adjusted, but estimates of future shift lengths are important information for managing the workforce.

The Operations Center must assign orders to be built at locations that have parts available. Alternatively, parts can be transferred between factories to match demand within a few days through multiple modes of transit, including a regularly scheduled shuttle and trucking services that are available on-demand. Checking the availability of parts can be difficult. Although suppliers frequently update the firm about what parts are available in the Supplier Logistics Center and what deliveries can be expected over the next few weeks, variability in the time to delivery, substitution of parts for each other, and data inaccuracies complicate matters. Because the parts have already been routed to each factory's Supplier Logistics Center, inbound shipping costs are no longer relevant. However, part shortages occasionally occur at individual factories and throughout the network and can cause many orders to not be satisfied on-time which is costly.

The execution problem faced by the firm's Operations Center is determining which factories should satisfy each order, if at all, in the next twenty-four hours, accounting for how this affects production over the next two weeks. It should consider outbound shipping costs, the length of each work-team's shift, its impact on capacity and direct labor costs, and parts availability. Because demand and parts supply vary from planned values, the firm must repeatedly adjust its production tactics. Accounting for all of these factors simultaneously can be difficult but is critical to the firm's ability to operate and can be worth several hundreds thousands of dollars per day.

2.3 Understanding Factory Production Capacity

Given the importance of each factory's physical layout, machinery, and labor force, in 2006 we interviewed employees at each factory to understand that factory's production capacity. In some cases, multiple interviews and email correspondence were necessary to confirm the accuracy of the following details. Although many of the particular numbers described below changed throughout the course of this study, the constraints described continued to be exemplary of the limitations on production at the firm's factories and are referred to throughout this thesis.

TX Production Capacity

In [Fel06], Jennifer Felch, one of the managers at TX, the main desktop production facility in Texas, described the limitations on production capacity at TX. TX has six "kit lines" that gather parts into kits and assemble desktop hardware. Each kit line can produce up to 250 consumer desktops per hour or 300 corporate desktops per hour, or any mix of the two at those rates. However, only two of these six kit lines can assemble consumer desktops because the parts for consumer desktops are stored in only two kitting areas for this more customized Line of Business. These are physical constraints of the factory layout.

According to [Fel06], the labor shift structure also constrains productivity at TX. The default schedule has two work-teams operate two shifts per day that last eight hours per shift and have about 6.25 productive hours per shift, for five days each week. The most this schedule could be extended to is two shifts per day that last eleven hours per shift and have nine productive hours per shift, for seven days per week; this cannot be maintained indefinitely but can be done when facing extraordinarily high demand. Furthermore, most scheduled shifts must last at least six hours. Each work-team has up to eighty straight-time hours per two weeks and is paid overtime wages for any time spent on the factory floor in excess of eighty hours in a two-week pay-period.

The maximum weekly production output of TX, based on physical bottlenecks and the shift structure, is depicted in Figure 2-1. The inner region of the figure indicates production mixes of consumer and corporate desktops that are feasible without extending shifts beyond the default schedule; the outer region is possible by extending shifts. The number and skill of available workers for these shifts can further constrain production; the firm tracked this by estimating the average Units-per-Hour (UPH) production rate for each work-team; combined with the length of a shift, UPH provides a reliable estimate of the number of desktops that a work-team can produce during its shift.

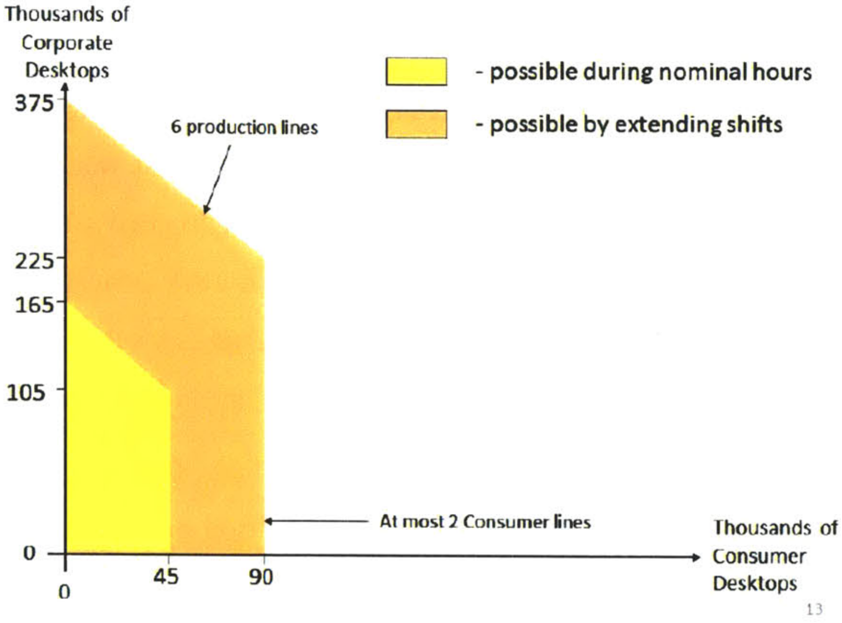


Figure 2-1: Maximum weekly production output of TX.

TN Production Capacity

In [Dol06, DH06], Eric Dolak, an employee at TN, and Michael Hoag, an Operations Center employee, describe limitations on production at TN, the firm's factory in Nashville, Tennessee. According to engineering specifications, TN has seven production lines that can produce 400 units per hour, yielding a total 2800 UPH, independent of product mix; however, if six or seven production lines assemble only consumer desktops, capacity drops to 2150 or 2250 UPH, respectively. Although each production line only builds one Line of Business at a time, over the course of a week or even shift, production can be smoothed to achieve any mix of products.

In addition to the number of production lines, boxing assembled desktops in preparation for shipping is a major physical bottleneck at TN. Large corporate orders that must be shipped together can consume most of the storage space in the the Automated Storage and Retrieval System (ASRS). At most 200 boxes can be work-in-progress (WIP) inventory before the two typically corporate desktop production lines must shut down; this 200 desktop build-up of WIP can be cleared when work-teams pause for a break every four hours. Given the rate that WIP builds for each Line of Business at TN, we can compute how the ASRS constrains the production mix.

The labor structure at TN includes two work-teams that work five eight-hour (7.3 productive hours) shifts per week and one work-team that works four ten-hour (9.3 productive hours) shifts per week. Shifts can be extended to two work-teams on four ten-hour (9.3 productive hours) shifts and two work-teams on three twelve-hour (10.75 productive hours) shifts. Minimum shift lengths varied by shift. The default shift length also determined the number of hours until overtime began.

The implications of the number of production lines, the ASRS bottleneck, and the shift structure on TN's maximum weekly production output are depicted in Figure 2-2. Labor availability can further restrict this.

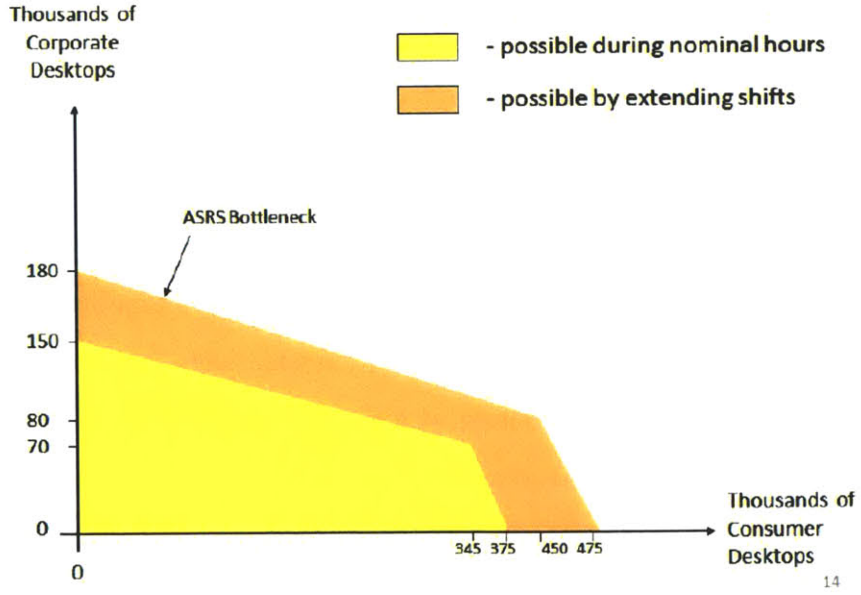


Figure 2-2: Maximum weekly production output of TN.

NC Production Capacity

Rebecca Fearing and Sean Holly, in [FH06], describe NC, located in Winston-Salem, North Carolina, as being the newest and most flexible North-American production facility. Because only 900 orders can be labeled per-hour, boxing is the biggest bottleneck at NC, limiting it to 900 UPH if producing solely consumer desktops and 1096 UPH if only producing corporate desktops, whose orders tend to contain multiple desktops. Over the course of this study, the capacity of NC increased to almost 1400 UPH. By this point, the production rate at NC was independent of the product mix.

NC has three work-teams, one working four ten-hour (8.75 productive hours), one working five eight-hour shifts (6.75 productive hours), and one working three twelve-hour (10.75 productive hours) shifts on weekends, each week. The first shift can be extended by one hour each day and the second can be extended by four hours each day. Minimum shift lengths varied by shift. The default shift length also determined the number of hours until overtime began. Figure 2-3 depicts the maximum weekly production output at NC if it is fully staffed.

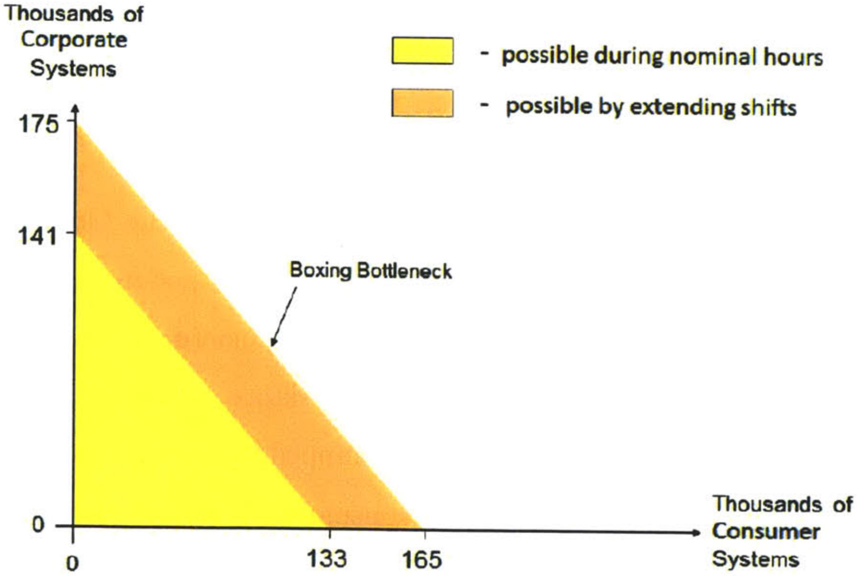


Figure 2-3: Maximum weekly production output of NC.

By 2008, when we developed and implemented a model to solve the problem at the execution scope, the NC factory had added "lean lines" in addition to the already existing conventional production lines. These lean lines focused on building specific high-volume products efficiently and had two additional work-teams with their own staff structure. We sometimes refer to these lean lines as "NCLL." In terms of production and labor, NCLL can be treated as a separate factory from NC. In most other ways, such as parts-availability and shipping logistics, NCLL can be treated as a part of NC.

2.4 The Operations Center's Geo-manufacturing Strategy

As the firm's supply chain evolved, the Operations Center developed solutions to the problems described in §2.2. In this section, we present how the firm handled these problems between 2005 and 2008, qualitatively. §2.4.1 covers the planning problem of §2.2.1 and §2.4.2 covers the execution problem of §2.2.2. Quantitative analysis of the firm's solutions is provided in Chapters 4 and 5 for the execution problem and Chapter 6 for the planning problem.

2.4.1 Geographic Manufacturing Plans

At the time we began working with the firm in 2006, the Operations Center employed a tactic called "geographic manufacturing," "geo-manufacturing," or "geo-man" which focuses on the geography of its manufacturing and customer network. The geo-manufacturing strategy allocated desktops to factories based on the geographic destination that the order will be shipped to, focusing on reducing the cost of shipping finished goods directly to the consumer. A map of the United States is split into thirteen geographic regions, with finer granularity for more central regions. This map, which the firm refers to as the "geoman map", can be seen in Figure 2-4 and changed rarely. Every fiscal quarter, the Operations Center partitioned or allocated those thirteen demand regions among the three factories. When a region is allocated to a factory, most orders from that destination will be, by default, produced in that factory. Exceptions were mostly orders that must be satisfied at particular factories. In Figure 2-4, a typical allocation, the one used by the firm in Fall 2006, is illustrated by the bold lines along with the associated percent of total demand assigned to each factory. The fundamental thought underlying the choice to split the map into western, central, and eastern segments is that this will minimize the cost of shipping to each destination while balancing the load on each factory. Because factory capacities vary over time (e.g. NC's productivity grew over its first year in 2005 as more production lines became operational), the proportion of total orders assigned to each factory was occasionally adjusted. Nonetheless, the firm usually made the same allocation decisions; the assignment of regions to factories in Figure 2-4 is representative of the Operations Center's plan for most quarters from 2006 to 2008.

Even though labor force and parts sourcing plans were based on the allocation decisions and hence production plans made by the Operations Center, the cost of direct labor and parts routing were not at the forefront of generating the geoman map. The total amount of volume given to each factory in the geoman map is chosen to balance factory loads by allocating a total quarterly sales volume that is proportional to that factory's physical production capacity. This does incorporate expected changes in capacity, such as NC bringing more assembly lines on-line. Between 2006 and 2008, each factory had between 27% and 40% of the firm's total North American manufacturing capacity, making the map split nearly evenly among the three active factories. Because demand was allocated to each factory in these proportions, each factory's labor force was also chosen to be of similar proportions. Production capacity based on only the permanent labor force working their nominal schedule ranged from 89% to 98% of total demand in quarters we observed. Demand in excess of this capacity would be met by a combination of temporary labor and overtime. As physical capacity and the allocation of regions rarely changed, the permanent labor force at each factory could remain relatively stable between quarters, helping maintain factory employees' morale and making planning for other operations easier. Because the firm

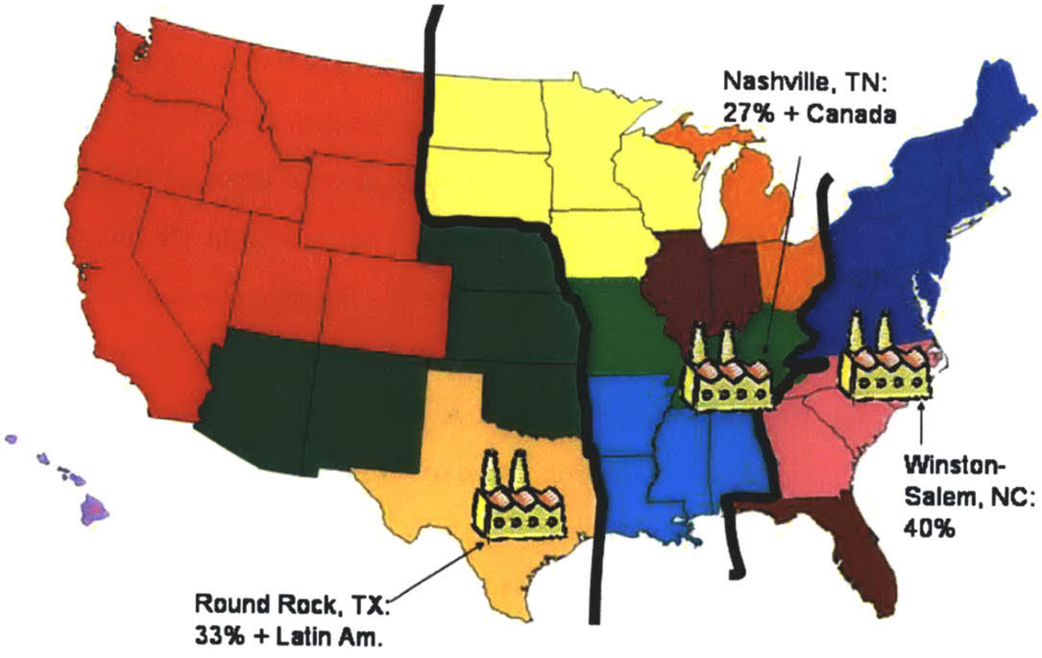


Figure 2-4: The firm's geoman, or geographic manufacturing, map partitions the United States into thirteen geographic regions (Alaska not shown), represented by different colors. Regions are assigned to factories, indicated by the bold lines; by default, factories produce orders that are to be shipped to the geographic regions that they are assigned. For each factory, its location, the percent of U.S. demand assigned to it, and additional international destinations are given in text. This particular allocation was used in Fall 2006 and was typical for most quarters from 2006 to 2008.

had a policy of producing all orders made within a fiscal quarter by the end of it, temporary labor was typically hired midway into each quarter and overtime was used heavily to satisfy demand at the end of the quarter. This change in the labor force is illustrated in much more detail in §6.1 where we model it mathematically. Although capacity played a large role in the allocation scheme, the cost of direct labor did not. Similarly, parts were routed to factories based on these production plans but were not incorporated into the choice of how much volume each factory received.

The firm's Material Resource Planning (MRP) system uses "default download rules," which, as the name suggests, are rules that determine the factory that will download (or be given) an order by default, i.e. without manual intervention. The

MRP system takes any new order and checks several logical conditions to determine which factory will be given instructions to produce that particular desktop. The regional assignments of the geoman map are the major deciding factor in the default download rules. Other factors include distinguishing orders with special requests, extremely early due dates, or rare parts. Although the geoman map does not distinguish between products, the download rules will only assign orders to factories that have the expertise to build a particular product family. For example, the factories in Mexico that handled outsourced orders beginning in 2009 could not produce orders that needed next-day delivery or orders that contained more one distinct configuration of computers. Once an order is assigned to a factory by these default download rules, the order will be produced there unless an employee makes a conscious choice to move the order to another factory.

2.4.2 Geo-move Execution

The firm's Operations Center continually evaluates the performance of its manufacturing network. As each new order arrives, it is automatically assigned a factory without regard for the current state of each factory or the availability of supplies. The firm must respond to variations in demand to maintain cost-effective operations. If total demand varies enough over a few days, it can induce excessive or insufficient labor capacity network-wide. Surges or lulls in demand from several geographic regions assigned to the same factory can cause imbalances between factory loads. Sometimes the default plan leads one factory to be starved for work while another would need to extend its shifts or even be unable to satisfy all of its orders. Similarly, imbalances in demand can cause unforeseen part shortages which also have costly consequences. The Operations Center responds to this by moving orders between factories and adjusting the length of shifts to produce as many desktops as possible.

The Operations Center re-allocates groups of orders to balance factory loads by using a mix of spreadsheets, heuristics, intuition, politics, and experience. Recall that the number of desktops available to be built is called ATB. Several employees track the ATB to capacity ratio, which is the number of desktops in backlog divided by the factory's daily production rate over the next few days, for each factory in several spreadsheets. Management prefers that this ratio be between zero and three days and at most five days worth of production. If the ATB to capacity ratio of two factories is "too" different, where the amount of difference is somewhat subjective but typically about one day worth of production, the Operations Center moves orders from the factory with high ATB to the one with low ATB. As a rule, order moves were made if otherwise one factory would be required to extend its shifts while another did not have enough ATB to last through its minimum shift lengths. Similarly, if a factory could not satisfy all of its orders within a day even with overtime while another factory did not need to extend its shifts, order moves would be made. Fairness and employee morale often played a role in order move decisions; factory managers occasionally called the Operations Center to inquire why they had so many or so few orders and whether orders could be moved so as to match the nominal shift structure more closely. Severe part shortages, indicated by spreadsheets that display the top five to fifty parts that are short, that are not network-wide in which multiple orders will become overdue if no action is taken are also reason to move orders between factories. Occasionally order moves were made when customers made special requests that their orders be delivered more quickly than they had originally stated. Extraneous information, such as large corporate orders that will become ATB soon, undocumented part shortages, or short notice of unexpected factory downtime, which are not obvious from the automatically generated data in the spreadsheets, are also used to inform whether order moves should be made.

Once the Operations Center has decided to move orders between factories, an employee queries the database of available orders to find a group of orders that have as many of the following attributes as possible. The orders' destinations should be in a region assigned to the factory with high ATB and are adjacent to (share a border with) regions assigned to the factory with low ATB; this ensures that the outbound cost increases by a relatively small amount. The number of desktops contained in the orders selected should be approximately half the difference in ATB between the two factories or at least enough to consider the two factories balanced; this was often several thousand orders. The orders should be past due, due today, or due tomorrow, which helps reduce the number of late orders because the factory they are moved to should be able to produce them within the day. The orders must be eligible to be built in the factory to which they will be transferred and required parts must be available; this is most commonly done by selecting one or two high-volume product families that can be built at any factory and almost always have parts available. Although orders with these attributes are preferable, a suitable set is often difficult to find. The person moving orders relies on expertise, intuition, and trial-and-error to find a suitable set of regions, due-dates, and product families. Once a set of orders is settled on and the ATB to capacity ratios are considered close enough, the move is executed by uploading the list of orders and their new factory assignments to the firm's MRP software.

Independent of whether any order moves have been made, factory managers will adjust the length of their work-team's shift lengths within limits to produce as many desktops as possible within a day. Delaying production now in anticipation of future excess capacity, especially at other factories, is rarely a consideration. Orders that are due in the next two days or already late are given highest priority. For a given allocation of orders to factories, this minimizes the number of late orders. Sometimes judgment calls must be made during the first shift of a day as to whether the first shift should be extended before they know how much ATB they will have for the second shift of the day. Occasionally, shifts will not be extended if the Operations Center foresees insufficient ATB in the near future. Most of these decisions are made by experienced employees inspecting time-series data of expected sales, ATB, and production. Generally, shift lengths are adjusted to be just long enough to produce all desktops in ATB.

Given the complexities and importance of this problem, the Operations Center was interested in understanding the efficacy of its current solution and how it could improve upon its current practices. In the Chapter 3, we review the literature that already addresses similar problems and practices. In the remaining chapters, our analysis of the firm's decisions and new solutions based on the literature suggest the firm did well at minimizing shipping costs, as planned, but also had potential to improve.

Chapter 3

Literature Review

In this chapter, we review the literature relevant to the problems defined in §2.2. Theoretical literature on production planning and control, which tends to focus on methodological issues such as characterizing and solving particular models, is reviewed in §3.1. The papers that are specific to Make-To-Order (MTO) manufacturing, which we cover in §3.1.1, tend to differ from those that deal with tactics for controlling multiple factories, which we discuss in §3.1.2. We present an industrial problem that belongs to both categories and is not fully addressed by either. Papers that discuss practical challenges and difficulties in optimization modeling for MTO manufacturing, such as balancing model realism and tractability or incorporating industrial testing, and have some similarities to our setting are covered in §3.2. Although these applied works tend to differ from ours in more ways than they are similar, they often contain relevant insights or illustrate other industries that could adopt our work. Other work done with the same firm that we study is discussed in §3.3.

3.1 Production Planning and Control

There is an extensive amount of relevant production planning and control literature, especially on basic concepts such as Lean Thinking, Just-in-Time inventory management, Linear Programming based models with inventory and backlog dynamics, forecasting and buffering against uncertain demand, and Make-to-Order business models. Missbauer and Uzsoy [MU11], Graves [Gra02], and Silver et al. [SPP+98] give broad introductions to and a plethora of references for the use optimization models for production planning and control problems in manufacturing of products with discrete parts. See Vidal and Goetschalckx [VG97] for a review supply chain models for production and distribution at a strategic level, emphasizing MILP formulations, case studies, and other modeling issues, providing a range of comparable work. Chen and Vairaktarakis [CV05] review more tactical and operational models for production and distribution; our work falls under the "general tactical production-distribution problems" class because a) it is tactical, b) it integrates inbound transportation, production, and outbound transportation, and c) it has a finite horizon with multiple time periods and dynamic demand. Sarmiento and Nagi [SN99] review work on the integration of production and transportation costs. Although many of models discussed in these reviews consider large networks and use MILP formulations, most allow for finished goods inventory and few address the stochastic issues inherent in MTO manufacturing. We first discuss what is known for MTO manufacturing and then return to the network setting.

3.1.1 Make-To-Order

Most of the literature on production planning and control focuses on Build-to-Stock (BTS) or Build-to-Forecast business models where production begins before customers order products. A common strategy in Make-to-Order (MTO) manufacturing is to use traditional BTS order-release mechanisms along with specialized order-acceptance, due-date setting, and contingency policies. Nonetheless, plenty of literature focuses solely on MTO production. Make-To-Order, Assemble-To-Order (ATO), Build-To-Order (BTO), and Configure-to-Order (CTO) are similar business models in which products are made, assembled, built or configured after the customer has placed an order; we use the more general and popular term Make-To-Order as our work is applicable to all of them, although the particular firm we collaborated with is considered to be CTO which is a subset of ATO which is a subset of MTO. Song and Zipkin [SZ03] survey the literature on dynamic ATO models, which tends to focus on the following topics: 1) order and due-date promising, often called Available-To-Promise (ATP), 2) scheduling or prioritizing already accepted orders, or 3) production, distribution, and inventory management for a given demand regime. Our work most closely falls under the third category, setting production levels and distributing orders through the network to customers, but consuming, not managing, parts inventories. The execution problem is also related to the second category, as individual orders must be scheduled for production.

Gunasekaran and Ngai [GN05] review Build-to-Order Supply Chain Management (BOSC), developing a framework for future work on the subject. They note that "there is a lack of adequate research on the design and control of BOSC. There is a need for further research on the implementation of BOSC... The trade-off between responsiveness and the cost of logistics needs further study... There are a noticeably limited number of research papers on BOSC from both academics and practitioners." They emphasize that BOSC must focus on optimizing logistics costs and delivering products to customers on time in the development of information systems. Contrary to lean manufacturing, BOSC requires quick production cycles, responsiveness to customers, and flexible rather than static schedules. Moreover, they note that "order-processing is time consuming and costly, multiple revisions of specifications are required, delivery dates are often not met, last-minute changes take up an increasing portion of resources, production plans are often inaccurate and over-ruled, and the more often this happens, the more profits decline." Although, plans must be made based on forecasts, when the orders that are available to be built differ enough from planned production schedules, recourse actions must be taken to control the situation. This thesis addresses these problems by developing information technology that dynamically accounts for all orders and minimizes the financial impact of such necessary but costly changes in production. Dynamic Programming, Linear Programming, simulation, multi-criteria optimization, and queuing models are typical solution techniques; our work employs all of these except queuing models. They further suggest that case studies be done on the implementation of BOSC in firms to develop insights. As "most companies are not yet prepared to completely disseminate the success behind their BOSC," this thesis presents and analyzes many aspects of one very successful firm's tactics, e.g. the geoman map, and improves upon them.

Lin and Shaw [LS98] show through simulation that synchronizing material and capacity availability in the order fulfillment process along with dynamic allocation of resources can be critical strategies. Iyer, Deshpande, and Wu [IDW03] show that postponing demand to handle potential surges in demand that would exceed production capacity can be an effective tactic in BOSC. Our MILP solutions re-iterate this in a realistic industrial context, as they match parts and capacity to demand and delay production until cost-efficient opportunities arise.

Before customers purchase from a MTO firm, their configuration options are typically filtered by an ATP system that must account for several of the same issues that MTO production planning and control do. Kaplan [Kap69] and Topkis [Top68] did early work on the issue of deciding whether to accept or reject orders for multiple inventory classes sold at set national prices with varying geographic transportation costs. Our problems only treat accepted orders but similarly must determine how to ration the available inventory of production capacity to geographically diverse orders. Chen, Zhao, and Ball [CZB01] formulate an Mixed Integer Linear Program for an ATP model that quotes due dates while dynamically reallocating parts and capacity when facing demand for a fixed set of orders; although our problem already has promised due dates, the parts and capacity adjustments are relevant. Moses et al. [MGGP04] study real-time promising of order due dates for BTO systems with dynamic order arrivals, accounting for 1) dynamic availability of resources, 2) individual orders with specific attributes, and 3) the backlog of previous commitments, similar to our execution problem; their scalable solutions perform well using an absolute lateness metric, similar to our linear lateness penalty, on similarly large problems of up to 100,000 orders with twenty resources. McNeil [McN05] gives a MIP for ATO systems that balances forecasted high-profit orders with delivery of already accepted low-profit orders, considering current commitments, delivery versus production expedites, and resource allocation, solving large problems but providing only sensitivity analysis. Hariharan and Zipkin [HZ95] show that the time between order acceptance and assembly, i.e. how long production is delayed, is equivalent to a reduction in manufacturing or parts-sourcing lead-time in Make-To-Stock systems, providing an interesting interpretation of production timing.

3.1.2 Network Manufacturing

Much of the work on network manufacturing focuses on Build-To-Stock models, where inventory can be held at various echelons in the network, and focuses on setting inventory levels rather than adjusting production capacity. Although our problem includes suppliers, manufacturing facilities, and customers, inventory is only held near the manufacturing facilities and is neither owned nor managed by the firm. Still, some of the methodological techniques are relevant.

Cohen and Lee [CL88] are widely referred to for models of production and distribution systems; their model includes raw materials, intermediate and final production facilities, distribution centers, warehouses, and customers, emphasizing the interaction of costs, service, and flexibility at each stage. Similar cost trade-offs are prevalent in our problem. Wu and Golbashi [WG04] study multiple factories making multiple products in a high-tech, capital-intense, short life-cycle context using multi-commodity flow and Lagrangian decomposition. Paquet, Martel, and Montreuil [PMM08] develop a MILP for multiple factories facing deterministic demand for multiple products, allowing for inter-facility transfer of parts and selecting the appropriate factory for each product based on local labor competencies. Dhaenens-Flipo and Finke [DFF01] study a multi-period, multi-product, multi-factory industrial problem with interrelated production and distribution costs, solving it quickly as a network flow problem with a few binary variables. In [DF00], Dhaenens-Flipo considers multiple facilities with high distribution costs, deciding geographically where to produce for scattered customers, akin to the geoman map. MILPs that are nearly network flows with Lagrangian relaxation techniques can model our problems well and can often be solved quickly.

Motivated by a large electronics manufacturer, Benjaafar, ElHafsi, and de Véricourt [BEdV04] develop a large and flexible (BTO or BTS) model of the problem of allocating stochastic demand for multiple products to multiple production facilities with varying capacities and inventory-handling costs, nearly modeling our problems. Some products can be restricted to particular factories and transportation costs can be included in production costs. Departing from our work, warehousing can be centralized or de-centralized, base-stock levels are decisions, multiple customer classes order the same products but have different demand rates and service costs, and the production rate of each facility is fixed. They give several managerial insights and a few characteristics of optimal solutions, some of which are "counter-intuitive," as is the case in our work. In [BLXE08], Benjaafar et al. allocate demand from multiple markets to multiple inventory locations, where production lead times depend on factory loads, and the goal is to minimize geographic transportation, inventory, and backordering costs. The non-linear problem is shown to be better solved by MILPs where each product is assigned to only one factory; in our problem, large groups of orders or market segments are assigned to single factories.

MILPs that are nearly network formulations with additional side-constraints are a common approach for problems similar to ours. Although these papers share many commonalities with our problems, they treat production capacity as fixed and demand is not satisfied in a brief interval of time, which is not the case in Make-To-Order manufacturing. Furthermore, they refrain from discussing the practical difficulties involved in using these models in practice.

3.1.3 Dynamic Programming Solution Techniques

Although the planning and execution problems faced by the firm are very interrelated, we solve them separately. The output solutions to the planning problem largely determine many inputs to the execution problem, making this system of control hierarchal. Sethi et al. [SHZZ02] survey the use of hierarchical control in stochastic, dynamic, manufacturing systems and show that the day-to-day fluctuations in demand, capacity, and other details need not be captured by longer-term planning models. Although it would be ideal to find global optimum when considering all of the firm's problems jointly or at least analyze the two models together, this is intractable because of the immense number of day-to-day details that would accumulate in a several-month horizon planning problem. Furthermore, other decisions are made and various uncertainties are realized between the scopes of the two problems; labor is staffed and parts are routed with the planning problem's solution as input but they do not always output the same staffing and parts quantities; the execution problem faces a different situation than dictated by the planning problem's output. However, if we did analyze the two problems together, we would expect similar cost savings potential, because optimal solutions to both problems cut costs in similar ways by matching production, parts, and capacity to demand while avoiding expensive shipping, labor, and order lateness. It is common in both the literature and in practice to solve and analyze these planning and control problems separately. For these reasons, we treat the problems separately but discuss how their results coincide.

Sethi [SS91] justifies the use of Rolling Horizon decision making theoretically by incorporating the cost of generating forecasts. As seen in Chand, Hsu and Sethi's survey [CHS02] on Rolling Horizon solutions to Operations Management problems, Dynamic Programs are often solved through rolling horizon heuristics and Linear Programs; notably, no Make-To-Order production systems were included. Bertsekas [Ber05b] indicates that rolling-horizon and certainty equivalence are common approaches to sub-optimal control for Dynamic Programming. Holt et al. [HMMS60] discuss the use of certainty equivalents, where replacing the stochastic elements in a problem with a particular deterministic ones leads to the same optimal expected value, in production planning problems. We use the more loose interpretation of certainty equivalent control, given by Bertsekas [Ber05a], where the uncertain quantity is replaced by a typical value that may not lead to optimal solutions but hopefully relatively good ones. Mayne et al. [MRRS00] surveys the theoretical results for and Qin and Badgwell [QB03] survey the industrial technology available for similar model predictive control problems. In our work, we formulate these rolling horizon, certainty equivalent, MILPs so that the industrial solver CPLEX can solve them in an appropriate amount of time, usually a few minutes.

3.2 Similar Implementations

Our work is relevant to more than just the desktop manufacturing industry; Make-to-Order networks arise in other industries, such as automotive, electronic retailing, custom design work, food catering, and grocery delivery, where many facilities can ship finished goods to geographically distributed customers and processing orders consumes scarce resources and takes time. The literature contains many reports of optimization models being implemented and used to control supply chains. We present those most relevant to ours and some of their insights that apply here.

In [JG95], Jordan and Graves consider the benefits of flexible manufacturing in the automotive industry and find that factories being able to cover some demand for products typically produced in another factory, by "chaining" the commonalities, improves profits when facing uncertain demand. They model demand as being Normally distributed but truncated at two standard deviations with a coefficient of variation of 40% and correlation coefficients within product groups of 30%. Our setting has a similar demand profile and can use the firm's full flexibility to have factories produce for destinations typically covered by other factories. For 3DayCar, Waller [Wal04] discusses automotive BTO demand forecasting, price management, and capacity planning, noting that "optimization technology is critical to build to order because it offers real-time constraint management and scenario planning." Ul-Haq and Naddem [UHN10] investigate BTO supply chain management strategies for the automotive industry with Volvo.

Klingman, Mote, and Phillips [KMP88] investigate a large, dynamic, multi-product production and distribution problem at a chemical products firm, decomposing it into a general network and a small, linear, Lagrangian, non-network component.

Zuo, Kuo and McRoberts [ZKM91] work with a corn seed producer and distributor, developing an MIP that allocates five major North American sales regions to different factories with inter-region product transportation and constraints on quality, work environment, minimum and maximum factory capacities, and market demand.

Wagner, Guralnik, and Phelps [WGP03] simulate dynamically distributed supply chains for sleeping bag and backpack manufacturing with capacity and raw material requirements. They show that demand variability in BOSC necessitates automated on-line coordination of production facilities rather than pre-computed solutions, often found through network flow formulations. In our case, solutions to the planning problem must be updated by solving the execution problem.

Ehrun and Tayur [ET03] optimize cost at a grocery retailer with highly variable demand and multiple distribution centers. In a pilot test, they reduced operating costs by 20.8% and boosted profit by 11.6%, while providing superior fulfillment to stores. Nonetheless, the model is mostly build-to-forecast and involves littler inter-factory interaction. Our work finds similar operational cost savings possible.

Biswas and Narahari [BN04] build a generic supply chain management algorithm and simulation based Object Oriented modeling language that supports BTO networks and Operations Research solution techniques. They provide an industrial case study of a petroleum gas supply chain.

Xu et al. [Xu05, XAG06] study e-tailer assignment of orders to fulfillment centers after accepting but before picking orders, similar to ATB orders in our problem, with considerations for inventory availability and transportation costs. As in our work, orders are re-assigned based on demand and supply information that arises over time in a rolling-horizon manner, accounting for the number and size of orders and correlation in demand, and heuristics are tested with industrial data.

3.3 Other Work on the Firm

Plenty of other work has been done with the same firm; much of it is relevant and informative when considering our problems.

In [Dha08], Dhalla quantifies the cost of parts supply shortages and product delivery delays for the firm's North American operations. The factory at which a shortage occurs, the type of products that are short, and the number of days that they will be late are the major determining factors in shortage costs. We use Dhalla's analysis of the cost of the firm delivering an order past it's due-date in our formulations of the execution problem.

In [Rey06], Reyner develops tools, metrics, processes, and organizational roles to improve routing of parts from suppliers in Asia to factories in North America. In [For08, FGA$^+$10], Foreman et al. extend these tools to include a MILP model, minimizing routing and shortage costs. The model was implemented, field-tested, and validated in a manner similar to the work in this thesis. Many of our information sources were shared. The planning problem determines the expected demand for each part at each facility, which their model would use as input; routing parts with their model would serve as input for the availability of parts in our execution problem.

Several others have investigated uncertainty in demand at this firm. Einhorn [Ein98] studies demand variability at the firm and the use of time-series forecasting of part-level demand along with the impact of hedging the firm's forecast upward. Our Robust Optimization approach to planning handles this in a similar but more systematic way. Hoffman [Hof09] addresses the impact of variation in CTO manufacturing, using the firm as its primary example, and suggests that companies understand 1) that variation and 2) how they provide value to their customers, while providing a framework to find the price of addressing variation. Gupte [Gup08] shows that the "dramatic changes" in the switch from BTO to BTS "have exposed some weaknesses in the firm's Build-to-Order supply chain including the demand forecast and capacity planning." The firm will need to increase order lead-times or decrease daily order variability in retail sales to manage with its current manufacturing capacity. Additionally, the firm will need to minimize the number of different suppliers it partners with in order to benefit from demand pooling and prevent a need for increased manufacturing capacity. Although our models do not consider changing the machinery or layout of factories, they do adjust the labor capacity to better utilize existing resources, even as the firm begins retail sales.

Vainio [Vai04] analyzes the firm's order fulfillment process and suggests methods to improve customer service and reduce logistics costs; by scheduling manufacturing and shipping based on time of day, air shipments can be sent by ground instead. Vainio considers only one factory at a time, using time-steps of two-hours and a horizon of one-day, to jointly analyze the manufacturing decisions of the firm's order release software with its distribution and routing software, making recommendations about how to merge the two. Vainio notes that the order-release software "does not factor order destination, scheduling time, and service level... into the scheduling algorithm;" our work does account for these at multiple factories simultaneously. Vainio's work is more granular than our model, focusing on only one factory, in smaller time-steps, with more production differentiation and shipping modes. It does address a similar issue of delaying less imminent orders so as to fulfill others more cost effectively. Stecke and Zhao [SZ07] consider a single factory in this firm's network and propose MILP models to integrate production and shipping to customers, re-scheduling the production so as to allow cheaper shipping transportation modes. This operational level of control is more granular in scope than our execution problem and would be used to release orders for production once our execution problem determines where they should be built.

To the best of our knowledge, this is the first work focusing on simultaneously allocating stochastic demand to factories while adjusting labor capacity in a multi-factory make-to-order setting. Exposing the problem's intricacies and the firm's solution to it in Chapter 2 is the first major contribution of our work. Solving similar problems using rolling-horizon, almost-network MILPs is not new. However, industrial data, testing, and feedback, along with discussions of the practical challenges in modeling and actionable managerial insights, are rare. Our second major contribution is providing these in a substantially different context, Make-To-Order network manufacturing.

Chapter 4

Execution Problem: Mathematical Formulation and Analysis

This chapter develops a mathematical model of the problems described in Chapter 2, using approaches commonly found in the literature described in Chapter 3, and evaluates the performance of various solution policies via a simulation study. It focuses on the execution problem's scope, especially with regard to quantitative value of parameters; most of the modeling still applies to the planning problem's scope.

We introduce our mathematical notation in §4.1 and then model the problem formally as a Dynamic Program in §4.2, discussing a few simplifying assumptions made in the modeling process in §4.3. The source of data for all of the model's parameters is examined in §4.4. We then detail the demand model in §4.5, analyzing the data and estimating parameters, making it as realistic as possible, introducing forecast error, and validating the demand model's correctness. We present solution policies for the Dynamic Program in §4.6 and develop the criteria through which we will evaluate them in §4.7. We then discuss the results of the simulation that evaluates these polices and discuss insights and conclusions in §4.8.

4.1 Mathematical Notation

We first develop some basic mathematical notation that will be useful throughout this chapter to describe the problem, detailed in Chapter 2, formally. Tables 4.1, 4.2, 4.3, and 4.4 summarize most of the notation from their respective subsections.

Lower case letters k, t, τ, l, d, ω, and c are indices; capital case versions of these letters tend to be either the set of all possible values for that index or the cardinality of that set. For all other letters, lower case letters refer to decision variables and capital letters refer to data parameters. When needed for clarity, bold symbols represent vectors.

4.1.1 Indices

Time is discretized into periods indexed by the letters k, t, and τ which take values in $\{1, \ldots, K\}$. In this chapter, each period is a day and the time horizon we study is typically one quarter year, making $K = 91$. The index k is typically reserved to represent the current day in the decision-making process, while τ will often refer to the day that demand becomes known and t is typically used for due-dates and the timing of future or past decisions. Subscript indices denote events that occur during time period t (or k) or at factory (facility) location $l \in \{1, \ldots, L\}$; occasionally we index the firm's three U.S. factories instead in $l \in \{TX, TN, NC\}$. Superscript indices refer to the demand $\phi_t^{\tau,d}$ (in number of desktop computers) that becomes known *at* time τ, must be shipped to destination $d \in D$, and is due *at* time t. D is the set of demand destinations, taken to be the set of U.S. states. The integer T represents the number of days that a policy forecasts (demand and decisions) into the future; data usually limits this to fourteen days. As we describe later, the only major source of uncertainty arises from the demand $\phi_t^{\tau,d}$; because it is later assumed exogenous from decisions made within the problem, we refer to each instance of uncertainty as the random instance $\omega \in \{1, \ldots, \Omega\}$ and use $\phi_t^{\tau,d}(\omega)$ as the demand on instance ω. The model for this uncertainty is detailed in §4.5. An additional index $c \in \{y, h, o, q\}$ denotes the category of decision or cost, based on the four types of decisions or costs,

Symbol	Domain	Description
k, t, τ	$\{0,\ldots,K\}$	time period or day in horizon
l	$\{1,\ldots,L\}$	factory (facility) location
d	D	U.S. states, shipping destinations for demand
ω	$\{1,\ldots,\Omega\}$	random instance of demand
c	$\{y,h,o,q\}$	category of costs or decisions

Table 4.1: Mathematical notation for indices.

Symbol	Description
$y^d_{t,l}$	production on day t for destination d at factory l
$h_{t,l}$	capacity on day t at factory l
$o_{t,l}$	overtime capacity on day t at factory l
q^d_t	desktops past-due (late) on day t for destination d

Table 4.2: Mathematical notation for decisions, all in units of number of desktops.

which are discussed next. Table 4.1 summarizes the notation for indices.

4.1.2 Decisions

Decision variables $y^d_{t,l}$ (in units of desktops) represent the number of desktop computers that are to be produced on day t at factory location l to be shipped to destination d. Decision variables $h_{t,l}$ are the amount of labor capacity (in desktops), representative of the labor force on the factory floor assembling desktops, on day t at factory location l; this quantity does include overtime. Auxiliary decision variables, used mostly for cost accounting in some policies, $o_{t,l}$ and q^d_t (both in desktops) respectively represent the amount of overtime capacity at location l in time period t and the number of desktops that are past due on day t for destination d. Table 4.1 summarizes the notation for decisions.

4.1.3 Costs

C^d_l is the shipping cost (in \$/desktop) from location l to destination d. $H_{t,l}$ is the non-overtime cost-per-unit-capacity (in \$/desktop) at location l during time period t and $O_{t,l}$ is the additional cost (in \$/desktop) of overtime capacity; given a total capacity

Symbol	Description
C_l^d	cost of shipping one desktop from factory l to destination d
$H_{t,l}$	cost per desktop of capacity on day t at factory l without overtime
$O_{t,l}$	additional overtime cost per desktop of capacity on day t at factory l
P	penalty cost per desktop past-due on each day for every destination

Table 4.3: Mathematical notation for cost parameters, all in units of U.S. Dollars ($) per desktop.

of $h_{t,l}$ which contains $o_{t,l}$ overtime capacity, the total labor cost is $H_{t,l}h_{t,l} + O_{t,l}o_{t,l}$. P is the scalar cost penalty (in $/desktop/day) for each time period that each desktop computer is past due. The objective is to minimize normal and overtime capacity costs, shipping costs, and late penalty costs. Table 4.3 summarizes the notation for cost data.

4.1.4 Data Parameters

In addition to $\phi_t^{\tau,d}$, we will use $\Phi_t^{\tau,d} = \sum_{t'=1}^{t} \phi_{t'}^{\tau,d}$ as demand that becomes known at time τ and is due *at or before* time t and $\overline{\Phi}_t^{\tau,d} = \sum_{\tau'=1}^{\tau} \sum_{t'=1}^{t} \phi_{t'}^{\tau',d}$ as demand that is known *by* time τ and is due *by* time t. Production for demand that arrives at time τ may not begin until it is known ($k \geq \tau$) but some policies use a cumulative forecast $\overline{F}_t^{\tau,d} = \overline{\Phi}_t^{k,d}(\omega) + \sum_{\tau>k} \hat{F}_t^{\tau,d}$ (in desktops) composed of prior demand plus a point forecast of future demand $\hat{F}_t^{\tau,d}$ (defined in §4.5.4) that will help coordinate current and future decisions. Total production at location l on day t cannot exceed the production capacity, $h_{t,l}$, which must be between a lower bound $\underline{H}_{t,l} \geq 0$ and an upper bound $\overline{H}_{t,l}$ (both in desktops). The firm's long-term planning shift structure included a nominal capacity $H_{k,l}^N$ (in desktops), which was the long-term plan or target for staffing and is used in defining the schedule for labor. $\hat{O}(t)$ is the first day of the "pay period" that contains day t. Overtime $o_{t,l}$ is the total capacity between $\hat{O}(t)$ and t in excess of the planned capacity $\hat{H}_{t,l}$ (in desktops, typically a very large number except for every fourteenth day when it is $\sum_{t=\hat{O}(k)}^{k} H_{t,l}^N$) for that time-frame; that is $o_{k,l} = [\sum_{t=\hat{O}(k)}^{k} h_{t,l} - \hat{H}_{k,l}]^+$. Similarly, the number of computers late on day k for destination d is $q_k^d = [\overline{\Phi}_k^{k,d} - \sum_l \sum_{t=1}^{k} y_{t,l}^d]^+$. Table 4.4 summarizes the notation for

Symbol	Description
$\phi_t^{\tau,d}$	demand for d arising at τ due at t
$\Phi_t^{\tau,d}$	demand for d arising at τ due by t
$\overline{\Phi}_t^{\tau,d}$	demand for d arising by τ due by t
$\overline{F}_t^{\tau,d}$	forecasted demand for d arising by τ due by t
$\underline{H}_{t,l}$	minimum capacity at l on t
$\overline{H}_{t,l}$	maximum capacity at l on t
$H_{t,l}^N$	nominal capacity at l on t
$\hat{H}_{t,l}$	overtime capacity threshold for pay-period beginning on $\hat{O}(t)$
$\hat{O}(t)$	first day of pay-period that contains day t

Table 4.4: Mathematical notation for data parameters, all in units of desktops other than the $\hat{O}(t)$ which is a day.

these additional parameters.

4.2 Dynamic Programming Formulation

The network manufacturing problem that we described in Chapter 2 can be stated as a stochastic dynamic program (DP) using the notation from §4.1, as follows. This formulation is the result of extensive study and reflects most important aspects of the problem described in §2.2.2. The definition of a DP requires an objective, a state evolution procedure, and a control space, which we now model.

The state at stage k is

$$x_k = \begin{cases} \overline{y}^d_{k-1,l} \triangleq \sum_{t=0}^{k-1} y^d_{t,l} & \forall l, d \\ \overline{h}_{k-1,l} \triangleq \sum_{\tau=\hat{O}(k-1)}^{k-1} h_{\tau,l} & \forall l \\ \Phi_t^{\tau,d}(\omega) & \forall \tau \leq k, \forall t, d \end{cases} \tag{4.1}$$

The state vector x_k contains cumulative production for each factory and destination $\overline{y}^d_{k-1,l}$, total capacity so far this pay-period $\overline{h}_{k-1,l}$, and a vector of all known demand $\Phi(\omega)$, which we do not collapse because it may contain information about future demand. Note that $\overline{\Phi}_K^{k,d}(\omega) - \sum_l \overline{y}^d_{k-1,l}$ is the number of outstanding orders for desktops from destination d on day k. The initial point x_0 is a vector of all zeros, since the system typically resets between horizons, as the backlog of orders is cleared at the end of every quarter.

The control at stage k is

$$u_k(x_k) = \begin{cases} y^d_{k,l} & \forall l, d \\ h_{k,l} & \forall l \end{cases} \tag{4.2}$$

which contains both the production decisions $\mathbf{y}$ and the capacity decisions $\mathbf{h}$ at each factory.

The noise at stage k is the demand

$$w_k = \Phi^{k+1,d}_t(\omega) \sim \mathbb{P}_k(\cdot|\, x_k)\, \forall t, d \tag{4.3}$$

where $\mathbb{P}_k(\cdot|x_k)$ will be specified in §4.5. Knowledge of the demand $\Phi^{\tau,d}_t$ is restricted to periods $\tau \leq k$; that is, past demand $\{\Phi^{\tau,d}_t(\omega) : \tau \leq k\}$ is data known by period k; future demand $\{\Phi^{\tau,d}_t : \tau > k\}$ is a stochastic quantity.

After decisions have been made in period k, the state evolves according to

$$x_{k+1} = \begin{cases} \overline{y}^d_{k-1,l} + y^d_{k,l} & \forall l, d \\ \mathbb{I}_{\hat{O}(k+1)\neq k+1}(\overline{h}_{k-1,l} + h_{k,l}) & \forall l \\ \Phi^{\tau,d}_t(\omega) & \forall \tau \leq k+1, \forall t, d \end{cases} \tag{4.4}$$

where $\mathbb{I}_a$ indicates whether event a is true (1) or not (0); the cumulative production $\overline{y}^d_{k,l}$ and cumulative capacity $\overline{h}_{k,l}$ are updated, capacity being reset if a new pay-period begins ($\hat{O}(k+1) = k+1$), and demand for the next day $\Phi^{k+1,d}_t(\omega)\, \forall t, d$ is observed.

The objective is to minimize over policies $\mathbf{y}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$ the relevant expected total supply-chain cost over the horizon:

$$\mathbb{E}_\Phi \sum_{k=1}^{K} \left[\sum_{l,d} C^d_l y^d_{k,l} + \sum_l (H_{k,l} h_{k,l} + O_{k,l}[\overline{h}_{k,l} - \hat{H}_{k,l}]^+) + P \sum_d [\overline{\Phi}^{k,d}_k(\omega) - \sum_l \overline{y}^d_{k,l}]^+ \right] \tag{4.5}$$

where $[a]^+ = \max\{0, a\}$. This captures the total shipping cost $\sum_{k=1}^K \sum_{l,d} C_l^d y_{k,l}^d$, the total non-overtime cost of capacity $\sum_{k=1}^K \sum_l H_{k,l} h_{k,l}$, the additional cost of overtime capacity $\sum_{k=1}^K \sum_l O_{k,l} o_{k,l}$, and the late penalty cost $\sum_{k=1}^K \sum_d P \cdot q_k^d$ for orders past-due. The terminal cost is zero, but unsatisfied demand is penalized by P per unit per day; the demand model in §4.5 will incorporate end-of-horizon effects.

The control constraints are

$$u_k \in U_k(x_k) \tag{4.6}$$

where

$$U_k(x_k) = \{u_k \text{ s.t.} \begin{cases} \underline{H}_{k,l} \le h_{k,l} \le \overline{H}_{k,l} & \forall l \\ \sum_d y_{k,l}^d \le h_{k,l} & \forall l \\ \sum_l (y_{k,l}^d + \overline{y}_{k-1,l}^d) \le \overline{\Phi}_K^{k,d} & \forall d \\ 0 \le y_{k,l}^d & \forall l, d. \end{cases} \} \tag{4.7}$$

The first line of constraints in (4.7) bounds the capacity $\mathbf{h}$ from below by the minimum capacity $\underline{\mathbf{H}}$ and from above by the maximum capacity $\overline{\mathbf{H}}$. The second set of constraints in (4.7) limits each factory's production $\sum_d y_{k,l}^d$ to be at most its capacity $h_{k,l}$. The third set of constraints enforces the Build-to-Order business model; it restricts cumulative production $\overline{y}_{k,l}^d$ by day k for each destination d to be only for orders that have already arrived, $\overline{\Phi}_K^{k,d}$ in total. Lastly, production, capacity, overtime capacity, and lateness for desktops are non-negative quantities.

The formal Dynamic Programming problem we wish to solve is: minimize (4.5) subject to (4.4) and (4.6). Once we have defined $\mathbf{\Phi}$, this is a well-defined dynamic programming problem.

4.3 Simplifications

The model developed in §4.2 has been simplified to make it both more tractable to analyze and more easy to understand. In doing so, we ignored a few practical difficulties which are addressed in the more detailed implementation models of Chapters

5 and 6. These issues can be categorized as follows.

- Product Differentiation, Availability and Transfer of Parts, and Geo-Eligibility
- Labor Shift Structure Details
- Managerial Constraints
- Solution Approaches

We address them in order.

4.3.1 Product Differentiation, Availability and Transfer of Parts, and Geo-Eligibility

In the formulation in §4.2, we treated all computer systems as identical. However, in practice, the firm builds computers to order, with particular parts being chosen by the customer. At the time, the firm produced two major categories of desktops, referred to as "Lines of Business"; these were the consumer desktop and the corporate desktop. Within each line of business, there are various product families which often shared many similar components. The largest source of distinction between product families (and also the two lines of business) are the parts components that combine to form them; these parts components, often referred to as "parts," include chassis, monitors, memory chips, processors, and video cards. Another distinguishing factor for product families is that some product families cannot be built at particular factories, sometimes referred to as geo-eligibility. If an order cannot be moved to another factory, it is called "non-geo-eligible." Unavailability of low-volume parts or unusually high labor-intensity in the production process for a particular product family are typical causes of non-geo-eligibility. Occasionally, individual orders with special delivery mechanisms, such as customer pick-up or international destinations, are also considered non-geo-eligible.

We choose to not model parts or geo-eligibility for multiple reasons. First, parts distribution is planned at a different level of scope; parts decisions are made by a separate planning team and need to be made about two weeks in advance (see [Dha08][For08] for more details) whereas the execution problem makes decisions less than a day in advance and the planning problem makes decisions at least a month in advance, limiting the possibility for dynamic interaction and hence usefulness. Secondly, data was largely unavailable to the Ops-Center regarding parts availability and hence little analysis could be done. Most importantly, when solving the problem for the firm in Chapters 5 and 6, we found that differentiating by parts or geo-eligibility rarely had a significant impact on the solution, other than that it was necessary for the firm's use of the model. If a particular desktop could not be built at a factory, a similar order assigned to another factory can often be swapped with it at little or no cost. The more important quantities to model correctly for analysis are the total volume of products being produced and how the factories were balanced based on total volume. Because adding these constraints would provide little of interest in analyzing the problem while adding significant complexities, we included neither parts nor geo-eligibility in the model.

Because we do not model parts and geo-eligibility, and these are the major differentiators for product families and lines of business, we also need not model multiple products and instead treat them as perfectly substitutable from a production planning standpoint. In the implementations in Chapters 5 and 6, we do account for lines of business or families and their respective parts.

4.3.2 Labor Shift Structure Details

The formulation in §4.2 includes capacity constraints on production and an overtime calculation for capacity in excess of certain thresholds. Because most of the factories had excess physical factory capacity, the major production capacity constraints stem from the rather expensive labor force that assembles the desktops. The underlying labor-force dynamics are much more complex and difficult to model appropriately. Dealing with capacity in units of desktops produced is much easier to analyze; hence, these constraints are formulated in units of desktops produced per time-period or over a collection of periods and are much simpler to state than the true underlying capacity constraints. However, in Chapters 5 and 6, the labor capacity constraints are modeled more explicitly. They are described in terms of units per factory-hour capacity and transformed by a units per labor-hour term to determine how many labor-hours are necessary for each factory hour, allowing costs and constraints to be written in units of labor hours.

In the planning problem of Chapter 6, a distinction is made between permanent labor and temporary labor. The permanent labor force is better trained and can produce desktops more efficiently, but cannot change in size as quickly, often being fixed for a quarter of a year. The temporary labor force can fluctuate in size from non-existent up to some fraction of the total workforce. Different employees can stay on the factory floor for varying amounts of time, so permanent and temporary labor can both stay for different amounts of overtime. Hence the decisions in the planning problem are "how much" labor to hire and "how long" that labor is on the factory floor; as such, the capacity decisions in the planning problem are actually the product of two decisions, making this problem much more difficult mathematically. For simplicity and because the model presented in this chapter is not analyzed at the planning problem's scope, we do not model it in this chapter.

In the execution problem of Chapter 5, each day is broken down into three work-shifts that have up to two work-teams staffed simultaneously. Overtime costs are tracked for each work-team individually. Each team's shift length can be extended beyond its nominal length in the labor schedule or be sent home early. Deciding to extend a shift beyond the nominal length requires at least one day's advance notice and sometimes two. Occasionally, two work-teams from different shifts may have overlapping time on the factory floor. During these periods, which may have up to double the planned amount of labor present, factory physical capacity can become the bottleneck. However, because we do not need decisions at this level of detail for analysis, we do not present these in this chapter.

These labor structure details are modeled appropriately in Chapters 5 and 6 where they are relevant.

4.3.3 Managerial Concerns

As with most mathematical formulations of real business problems, more is at stake than can be stated simply. The firm expressed corporate concerns other than financial impact and customer service which they would consider in the overall evaluation of a solution. The three major categories of concern were maintaining fair and balanced workloads at factories, producing stable solutions, and incorporating executive mandates that were made for reasons beyond the scope of our problem.

A major concern was maintaining fairness between factories and employee morale by having somewhat balanced workloads. For example, if one factory repeatedly required overtime while another sent its workers home early, the workers in either factory could perceive such production planning decisions as de-motivating or unfair, even if doing so is cost effective. The firm often asked for us to model constraints such as not deviating too far from a static schedule or that the ratio of orders assigned to planned factory capacity not deviate too much across factories. Most of these can easily be captured by linear constraints. In some cases these managerial requests could drastically alter a solution; by using sensitivity analysis or evaluating the cost of solutions with and without the constraint, we obtain useful estimates for the cost of such managerial policies and allow the firm to make the best decision. Because the parameters of such constraints or penalties were often determined by many widely-varying, subjective inputs and because these constraints usually do not create interesting changes in the solutions, we do not consider these types of constraints in our analysis.

A second concern was that the solutions should be relatively stable; in the absence of large or unexpected changes in the model's input, especially as small increments of time pass and expected orders become known, the output should be qualitatively the same. Creating such solutions would impose less work on various organizations within the firm, such as the production planning and parts routing teams. Furthermore, stable solutions would be more predictable and understandable. However, in such production planning problems with high substitutability (between due dates and destinations) and hence many near-optimal solutions, slight changes in the input, such as moving forward one day in the horizon, could cause mathematical optimization solutions to change substantially. An easy way to make these models have stable solutions is to add a small penalty to the objective for deviating from a previous solution. This is done in both implementations in Chapters 5 and 6, where we chose a penalty of \$0.01 per desktop whose production facility changes between solutions. However, from the perspective of analyzing the cost and aggregate decisions of various policies, such stability constraints add few insights and hence are not included in the above formulation.

A final concern was a major agreement that the firm's executives had made with a local government. In exchange for tax exemptions and other business incentives for locating a factory in Winston-Salem, North Carolina, the firm had agreed to maintain a minimum amount of permanent labor at that factory or pay expensive renumerations. This was equivalent to maintaining a minimum amount of production capacity at that factory. Such minimum production or capacity constraints are easy to include in the above formulation, but drastically alter the outcome in abnormal ways. In Chapter 6, the planning model implementation includes analysis of the cost of this constraint and solutions were provided both with and without it. Labor planning is beyond the scope of this formulation and therefore these constraints are not included in the present chapter.

4.3.4 Solution Approaches

Even though the model in §4.2 can be stated somewhat simply, it is difficult to solve both theoretically and in practice due to the well-known curse of dimensionality. This arises mostly from the uncertainty in the demand term Φ, which is indexed by arrival-date, due-date, and destination. A typical instance of this problem has ninety-one days and fifty destinations, as described in §4.4, making the state space several hundred-thousand dimensional. Additionally, the decision space includes decisions indexed by factories, destinations, and days within the decision horizon, making it several thousand dimensional. Because the state and decisions space are large, solving the Dynamic Program directly is computationally intractable. Even if we use a simplified demand model and collapse the state space to include only the cumulative demand realizations, production decisions, and capacity decisions to date, the state space would still have several thousand dimensions, leaving the problem intractable.

As discussed in §3.1.3, a common and simple approach to dealing with uncertainty in optimization is to replace the uncertain term by a deterministic one, hoping that the solution to the deterministic problem performs well for the problem with uncertainty. If the demand $\mathbf{\Phi}$ were deterministic, the formulation in §4.2 would be a Linear Programming problem, which can be solved easily. In §4.8, we show that this approach performs well and that more complex approaches are unnecessary; by solving this deterministic problem with a reasonable point-forecast for unknown demand, we find excellent solutions to the original problem, as shown by their proximity to a theoretical upper bound which can also be easily computed using Linear Programming. We compare several solution policies that treat demand deterministically, simplifying the problem to obtain solutions; we then analyze them stochastically. These policies are introduced in §4.6 and analyzed in §4.8.

4.4 Data Sources

Data from various sources was acquired by and in collaboration with the firm's North American Operations Center team and is used in the analysis of various solution policies to this dynamic program. Some of the sources, such as interviews with factory managers to determine factory production bottlenecks, are detailed in Chapter 2. Herein, we describe and comment on the data for the indices, decisions, and parameters described in Tables 4.1, 4.3, and 4.4, along with data necessary for developing the demand model in §4.5. New demand parameters, Γ and γ, the distributions of demand among due-dates and destinations, are introduced in order to construct other parameters, such as ϕ, that were not directly available. We first discuss the choice of indices in §4.4.1. We discuss the source of cost data in §4.4.2. The source of demand and labor parameters is discussed in §4.4.3. A brief summary of the source of each

Parameter Symbol	Description of Source
C_l^d	Contracts with third party logistics providers
$H_{t,l}$, $O_{t,l}$	Financial reports on labor and production
P	Data acquired by Dhalla [Dha08]
$\sum_{t,d} \phi_t^{\tau,d}$	Lookahead spreadsheets
Γ_k	ATB snapshots of currently known orders
γ_d	The firm's historical data for planning problem
$\phi_t^{\tau,d}$	See §4.5
$\overline{F}_t^{\tau,d}$	See §4.5.4
$\underline{H}_{t,l}$, $\overline{H}_{t,l}$, $H_{t,l}^N$, $\hat{H}_{t,l}$, $\hat{O}(t)$	Lookahead spreadsheets

Table 4.5: A brief description of the source of relevant data parameters.

data parameter is given in Table 4.5.

Most of the data collected comes from three distinct time periods, early 2008, late 2008, and early 2009, each separated by at least one month; we refer to these time periods as *data sets* 1, 2, and 3, respectively. Our analysis focuses on the time period of data set 1, February through April of 2008, which is a typical quarter for the firm; it avoids the "back-to-school" and winter break demand spikes. Spurious data was gathered from other time periods and helped in estimating parameters for these periods, but the demand patterns, factory capacities, and cost data are based on or tuned for February through April of 2008.

4.4.1 Indices

The horizon of $K = 91$ days was chosen because the firm has a policy of finishing production for all outstanding orders by the end of each fiscal quarter. The time granularity of using days, as opposed to shifts or weeks, matched the execution problem's scope of making daily decisions; furthermore, demand data was not available by shift nor were decisions made that frequently; given that we do not model shift structure in this chapter, this granularity is appropriate.

The three factories {TX, TN, NC} were the major production facilities during the time of this study. Each contained multiple production lines that had varying traits, but these played little role in the context of which orders were assigned to each factory; the differences that did affect production capacity are incorporated in the costs of capacity.

The firm's Operations Center planned order moves at a granularity of eight regions composed of U.S. states but executed its decisions by moving orders filtered by state; they already executed on but did not plan for the state level. Although data was available at a smaller granularity, such as zip-codes, U.S. states were fine enough to provide plenty of flexibility in moving orders between factories while being large enough to avoid moving orders for negligible returns in profits. Furthermore, given the size of the problem, using 3-zip or finer granularity would often introduce data storage, computation, and managerial oversight difficulties. Hence, we choose to use the U.S. states as the destinations in D because they are tractable for computation, sufficiently detailed for re-assigning orders with significant financial impact, and historically used by the firm.

4.4.2 Cost Data

The cost C_l^d to ship a desktop from each factory l to any destination d was acquired from the firm's early 2008 contracts with third party logistics providers to ship along those origin-destination pairs and weighting them by the relative volumes of various shipping priorities and desktop sizes/weights. An alternate set of shipping cost data, formed by averaging the empirical shipping costs, matched these quoted costs well and were the cost basis for the planning problem in Chapter 6. Because inbound routing (from vendor to factory) decisions was not within the scope of the execution problem, we do not include inbound shipping costs in this chapter.

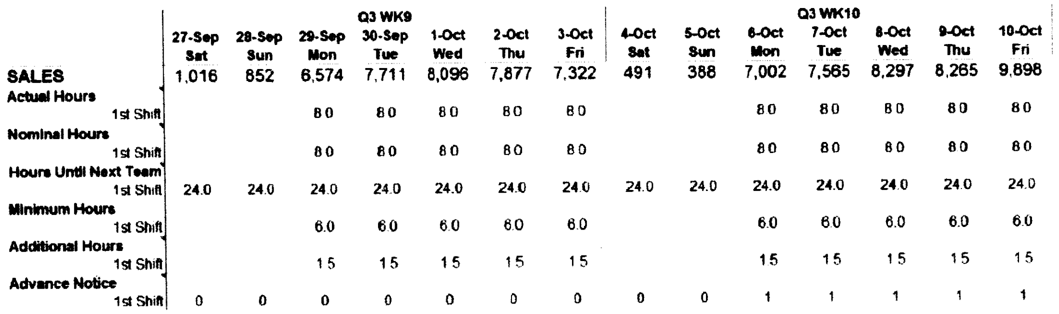
Data on the cost of labor was constructed by combining the firm's internal financial reports on the total cost of labor at each factory and historical production quantities. The cost of non-overtime capacity $H_{t,l}$ and the cost of overtime capacity $O_{t,l}$ were computed by totaling the non-overtime and overtime labor costs at each factory over many pay-periods and dividing it by the number of labor hours in that time-frame; averaging these yields the cost per labor hour at each factory. Dividing this by the units-per-labor-hour, another well-studied quantity that the firm tracked, yields the cost-per-unit of capacity at each factory, which ends up being the same for all time-periods t.

The penalty for late orders is based on the work of Dhalla [Dha08] in quantifying the cost of part shortages. Dhalla analyzed both the cost of lateness in existing orders (cancellations and consolations) as well as the cost of future customers not purchasing due to poor service. The data used in Dhalla's analysis indicates that after the first four days of being late, the cost for each day that a computer is late scales approximately linearly by P dollars per day that each computer is late. Those first four days are incorporated into the due date of order in the demand model, letting the DP of §4.2 treat the cost of each day that each desktop is late as scaling linearly with the penalty P. The model's sensitivity to this parameter is analyzed in §4.8.

4.4.3 Demand and Labor Data

The firm's North American Operations Center used Excel "Lookahead" spreadsheets to track both past and expected future desktop sales as well as the labor capacity assigned to each factory; they "look ahead" (into the future) up to at most two weeks. An example can be seen in Figure 4-1. A new spreadsheet was created each day and contained within it a tab for each factory. Within each factory's tab, each column represented a day and each row contained different data on sales and labor for those days. Columns from past days contained data on what had actually happened while future data were point estimates or forecasts of what they expected to happen. "Lookahead" spreadsheets from early 2008 through early 2009 were the largest source of dynamic input data for the model; they provided daily sales data, the firm's forecasts for future demand, and past decisions and future plans for labor capacity.

The most important row was daily sales; these numbers represented past sales along with forecasts for future sales. The word 'sales' is used interchangeably with 'demand' because decisions made within this model do not affect sales, outside of the customer service impact which is addressed by the order lateness penalty; thus



				Q3 WK9							Q3 WK10			
	27-Sep Sat	28-Sep Sun	29-Sep Mon	30-Sep Tue	1-Oct Wed	2-Oct Thu	3-Oct Fri	4-Oct Sat	5-Oct Sun	6-Oct Mon	7-Oct Tue	8-Oct Wed	9-Oct Thu	10-Oct Fri
SALES	1,016	852	6,574	7,711	8,096	7,877	7,322	491	388	7,002	7,565	8,297	8,265	9,898
Actual Hours														
1st Shift			8.0	8.0	8.0	8.0	8.0			8.0	8.0	8.0	8.0	8.0
Nominal Hours														
1st Shift			8.0	8.0	8.0	8.0	8.0			8.0	8.0	8.0	8.0	8.0
Hours Until Next Team														
1st Shift	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0
Minimum Hours														
1st Shift			6.0	6.0	6.0	6.0	6.0			6.0	6.0	6.0	6.0	6.0
Additional Hours														
1st Shift			1.5	1.5	1.5	1.5	1.5			1.5	1.5	1.5	1.5	1.5
Advance Notice														
1st Shift	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Figure 4-1: Example of the Lookahead spreadsheets used by the firm's Operations Center.

we have data with which to estimate demand. However, forecasts were rarely, if ever, updated, because they were generated by distributing total North American desktop sales targets amongst days, product lines, and destinations according to static proportions. Furthermore, data on what actually happened was only tabulated for each day, factory, and line of business; given that we do not model lines of business, our sales data and hence our demand data was indexed only by the time that it arose; that is, we have 310 data points (days τ) or empirical samples of $\sum_{t,d} \phi_t^{\tau,d}$ but none of $\phi_t^{\tau,d}$.

Nonetheless, the firm had estimates of the long-term distribution of demand among destinations,

$$\gamma_d = \frac{\mathbb{E}\sum_{t,\tau} \phi_t^{\tau,d}}{\mathbb{E}\sum_{t,\tau,d'} \phi_t^{\tau,d'}} \tag{4.8}$$

The firm historically used this to split the forecast among destinations when considering the planning problem. The firm's management had a strong belief that these fractions were stable and did not vary significantly over time.

Backlog or ATB "snapshots" of what orders were currently known but not yet produced were available for thirty days for which we had matching Lookahead spreadsheets. The total number of desktops in a snapshot from day k was $\sum_d [\overline{\Phi}_K^{k,d} - \sum_l \overline{y}_{k-1,l}^d]^+$. The snapshots contained almost all information that the firm had about any currently backlogged order, including the order-date, the due-date, the destination, which parts it required, and which factory it was currently assigned to. We could not reconstruct the total number of orders for each dimension of demand with this small data-set, mostly because the snapshots were not from consecutive days; even for those that were consecutive, Lookahead spreadsheets with larger sales quantities indicated that orders had become known and produced between the snapshots. However, we were able to use these ATB snapshots to estimate the fraction of orders that are due k days after arriving, namely

$$\Gamma_k = \frac{\mathbb{E} \sum_{d,\tau} \phi_{\tau+k}^{\tau,d}}{\mathbb{E} \sum_{t,\tau,d} \phi_t^{\tau,d}} \tag{4.9}$$

which is needed to estimate the demand distribution, as is done in §4.5.

The Lookahead spreadsheets also contained a large amount of data about the factory's capacity and more-so about the labor shift-structure. The most prominent figures were the "run-rate" or number of computers per hour that the scheduled work-team is expected to assemble and the planned (future) or actual (past) number of hours that each work-team is on the factory floor. There were two to five work-teams per factory, sometimes operating in different areas of the factory simultaneously. Although not available at first, for the purpose of model implementation in Chapter 5, we asked the Operations Center team to track the theoretical minimum and maximum length per shift, the timing between shifts, the effects of overlapping them, and other details. As mentioned in §4.3, we aggregated these shifts and work-teams into daily capacities that are directly usable by the model. We do this by multiplying these shift-lengths by the "run-rates" to obtain the minimum capacity $\underline{H}_{k,l}$, the maximum $\overline{H}_{k,l}$, and the scheduled (nominal) capacity $H_{k,l}^N$ for each work-team. We then add these together to obtain the daily capacities.

All of the data described above was reviewed and confirmed to be sufficiently accurate for implementation purposes by the firm's Operations Center members. The choices made in modeling and data acquisition reflect our best attempt to mirror the firm's operating environment.

4.5 Demand Structure

Herein we discuss the model of demand for the dynamic programming (DP) problem in §4.2. Because uncertainty in demand is largely what makes the problem difficult, we spend significant effort modeling it. The demand model must both realistically emulate the demand situation faced by the firm and be tractable enough for analyzing solutions to the DP problem. We are interested in the demand vector ϕ (the demand arising on day τ, due on day t, and destined for d) but only have data for $\sum_{t,d} \phi_t^{\tau,d}$ (the demand that arises on day τ), Γ_k, and γ_d. Constructing a realistic demand model from such limited data will require additional structure.

We begin in §4.5.1 by analyzing the available data and fitting distributions to the daily demand, wherein we find that a Log-Normal distribution that depends on the day of the week and the fiscal quarter fits best. We extrapolate this to fit a distribution to the demand vector ϕ, adding assumptions where necessary and estimating the distribution's parameters. This involves several conversions between $\sum_{t,d} \phi_t^{\tau,d}$ for which we have data and the more granular and data-less $\phi_t^{\tau,d}$, along with the logarithm of these quantities. In §4.5.2, we introduce correlation to make the model more realistic. In §4.5.3, we make a few additional adjustments, making the demand distribution discrete and accounting for the end-of-quarter "hockey-stick effect." In §4.5.4, we model error in the forecast F. Lastly, in §4.5.5, we validate that the approximations in this demand model generate demand that matches historical values. Many of the modeling techniques used in this section are common in practice; however, the correct choice and combination of these makes the modeling difficult and interesting. By the end of this complex modeling process, we have a simple method to generate random vectors ϕ from a distribution that reflects reality to the best of our ability, which will allow us to evaluate the performance of several policies in solving the stochastic DP of §4.2 via simulation.

The DP formulation is most easily stated in terms of $\overline{\Phi}_t^{\tau,d}$, the cumulative demand due (in number of desktop computers), which has become known by time τ, must be shipped to destination d, and is due by time t. To understand the nature of demand, it is easier to work with $\phi_t^{\tau,d}$, the demand that becomes known at time τ, is shipped to destination d, and is due at time t. We now model ϕ.

4.5.1 Data Analysis and Estimation

As mentioned in §4.4, in the firm's Lookahead spreadsheets, there are 310 data points (days τ) of the quantity $\sum_{t,d} \phi_t^{\tau,d}$ which are instantiations of the random quantity we call $\tilde{\phi}^\tau$, the daily demand. We also have the "splitting" parameters

$$\gamma_d = \frac{\mathbb{E}\sum_{t,\tau} \phi_t^{\tau,d}}{\mathbb{E}\sum_{t,\tau,d'} \phi_t^{\tau,d'}}$$

and

$$\Gamma_k = \frac{\mathbb{E}\sum_{d,\tau} \phi_{\tau+k}^{\tau,d}}{\mathbb{E}\sum_{t,\tau,d} \phi_t^{\tau,d}}.$$

which are the percent of total demand arising from each destination d and the percent of orders due within k days of arising, respectively, as described in §4.4. Note that $\sum_d \gamma_d = 1$ and $\sum_k \Gamma_k = 1$.

A key parameter for describing the distribution of demand is its mean, $\mu_t^{\tau,d} = \mathbb{E}\phi_t^{\tau,d}$. Because we have a good estimate of $\mathbb{E}\tilde{\phi}^\tau$, we would like any definition of $\mu_t^{\tau,d}$ to satisfy

$$\sum_{t,d} \mu_t^{\tau,d} = \mathbb{E}\tilde{\phi}^\tau. \tag{4.10}$$

We chose a simple, intuitive, and structural formula to define the mean of ϕ:

$$\mu_t^{\tau,d} = \mathbb{E}\phi_t^{\tau,d} := \gamma_d \Gamma_{t-\tau} \mathbb{E}\tilde{\phi}^\tau. \tag{4.11}$$

We can see that (4.11) satisfies (4.10) via

$$\sum_{t,d} \mu_t^{\tau,d} = \sum_{t,d} \mathbb{E}\phi_t^{\tau,d} = \mathbb{E}\tilde{\phi}^\tau \sum_d \gamma_d \sum_t \Gamma_{t-\tau} = \mathbb{E}\tilde{\phi}^\tau.$$

Although alternate definitions of $\mu_t^{\tau,d}$ were possible, this choice corresponds to assuming that daily demand (whose mean is $\mathbb{E}\tilde{\phi}$), the distribution of due dates (whose mean is Γ), and the distribution of demand among destinations (whose mean is γ) are independent; any other choice would require computing interaction effects between these parameters for which no data was available. Note that the mean of a distribution does not provide enough detail to evaluate the performance of policies in solving the DP; we need a model for the distribution of $\phi_t^{\tau,d}$ that incorporates more than its mean, such as its variance. Hence, we fit common probability distributions to the demand data.

We analyzed the fit of several distributions to the daily demand data, including Weibull, Log-Normal, Log-Logistic, Gamma, Exponential, and Uniform, with or without the intercept fixed to zero. Gamma, Weibull, Log-Logistic and Log-Normal distributions, all with zero intercept, fit well; other distributions did not. Both Log-Logistic and Log-Normal distributions passed (failed to reject) Anderson-Darling, Kolmogorov-Smirnov, and Chi-Squared tests at 10% significance (and lower) and had the best R^2 values. Log-Normal distributions are similar in shape to Log-Logistic distributions, but have lighter tails and more tractable covariance properties. Hence, we use a multi-variate Log-Normal (LN) fit for $\tilde{\boldsymbol{\phi}}$. Because Log-Normal distributions are appropriate for modeling demand and we have no way to estimate the distribution of $\boldsymbol{\phi}$, we also assume that $\boldsymbol{\phi}$ follows a multi-variate Log-Normal distribution. There is no known closed-form distribution for $\sum_{t,d} \phi_t^{\tau,d}$ when $\boldsymbol{\phi}$ is log-normally distributed. However, Fenton [Fen60] suggests a simple approximation that yields another log-normal; namely, match the first two moments of $\tilde{\phi}^\tau$ and $\sum_{t,d} \phi_t^{\tau,d}$. Although many papers since then (e.g. Mehta [MWMZ07]) have criticized this approach and provided alternatives, its simplicity and success in approximating most of the distribution (shown in both [Fen60] and [MWMZ07]) make it worthwhile.

In summary, we use the following fits:

$$\tilde{\phi}^\tau = \sum_{t,d} \phi_t^{\tau,d} \sim \mathrm{LN}(\hat{\mu}^\tau, \hat{\Sigma}^\tau) \sim e^{N(\hat{\mu}^\tau, \hat{\Sigma}^\tau)}$$

$$\boldsymbol{\phi} \sim \mathrm{LN}(\boldsymbol{\mu}', \boldsymbol{\Sigma}') \sim e^{\mathrm{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')}.$$

Because Log-Normal distributions are often easier to work with in the log-space, i.e.

Symbol	Distn.	Mean	Cov.	Description
$\phi_t^{\tau,d}$	LN	$\mu_t^{\tau,d}$	$\Sigma_{i,i'}$	r.v. for demand on τ due on t for d
$ln(\phi_t^{\tau,d})$	N	$\mu_t^{'\tau,d}$	$\Sigma'_{i,i'}$	logarithm of $\phi_t^{\tau,d}$
$\tilde{\phi}^{\tau}$	LN	$\mathbb{E}\tilde{\phi}^{\tau}$	$C_{\tau,\tau'}$	r.v. for total demand on τ; equals $\sum_{t,d}\phi_t^{\tau,d}$
$ln(\tilde{\phi}^{\tau})$	N	$\hat{\mu}^{\tau}$	$\hat{\Sigma}_{\tau\tau}$	logarithm of $\tilde{\phi}^{\tau}$

Table 4.6: Mathematical notation for demand parameters, in units of desktops. Distributions (Distn.) are either Normal (N) or Log-Normal (LN). The index i represents a triplet (t, τ, d), Cov. is the covariance, and r.v. means random variable.

$\ln(\tilde{\phi}^{\tau})$ or $\ln(\phi_t^{\tau,d})$, we convert between the Log-Normally distributed and Normally distributed counterparts frequently. $\hat{\mu}$ and $\hat{\Sigma}$ are the mean and covariance matrix of the Normally distributed logarithm of daily demand. We wish to derive μ' and Σ', the mean and variance of the Normal random variate $ln(\phi)$, in such a way to match the first two moments of $\sum_{t,d}\phi_t^{\tau,d}$ and $\tilde{\phi}^{\tau}$. To do so, we first estimate $\hat{\mu}$ and $\hat{\Sigma}$ and then use equation (4.11) along with some extra structure. A summary of the notation used for various demand parameters of interest is provided in Table 4.6.

The data was largely analyzed using linear regression on $\ln(\tilde{\phi}^{\tau})$ as $\tilde{\phi}^{\tau}$ consistently gave worse results, using the following independent variables: Day of the Week (indicator for each of the seven days of the week), Weekend (indicator for whether τ is a Saturday or Sunday), Weeknumber (in which the first Sunday in 2007 has a Weeknumber of 1 and Weeknumber increases by one for each subsequent week, the first week of 2008 having Weeknumber of 53 or reset to 1, and so forth), Demand k Days Ago for $k \in \{1, \ldots, 9\}$, and Data Set (as explained in §4.4 in $\{1, 2, 3\}$ or as indicators). Our major metric of regression power is the adjusted R-squared

$$\overline{R}^2 = 1 - \frac{K-1}{K-p-1}\frac{\sum_{\tau}(\hat{z}_{\tau} - \overline{z})^2}{\sum_{\tau}(z_{\tau} - \overline{z})^2}$$

where z_{τ} is the data, $\overline{z} = \frac{1}{K}\sum_{\tau} z_{\tau}$ is the mean of the data, $\hat{z}_{\tau}$ is the regression's prediction, p is the number of independent variables, and K is the number of samples.

We begin by estimating $\hat{\mu}$ using regression. While all of the independent variables had significant (under 5%) p-values, the coefficients for Weeknumber and Demand k Days Ago are too small to have non-negligible impact on the total demand value.

	Coefficients	P-value	Lower 95%	Upper 95%
Intercept	7.659	1.74E-115	7.259	8.058
DataSet1	0.347	4.23E-02	0.012	0.682
DataSet2	0.199	5.18E-02	-0.002	0.400
Monday	2.073	9.28E-114	1.963	2.183
Tuesday	2.274	1.73E-121	2.162	2.387
Wednesday	2.309	2.02E-126	2.200	2.418
Thursday	2.344	4.29E-80	2.169	2.518
Friday	2.635	7.94E-86	2.450	2.819
Saturday	0.272	4.93E-04	0.120	0.425
Weeknumber	0.002	6.03E-01	-0.005	0.008
1 Day Ago	0.000	6.00E-03	0.000	0.000
2 Days Ago	0.000	7.61E-02	0.000	0.000

Table 4.7: Regression of $\ln(\tilde{\phi}^\tau)$ (log desktops) for $\hat{\mu}$ with Weeknumber and Demand k days ago. $\overline{R}^2 = 0.941$.

	Coefficients	P-value	Lower 95%	Upper 95%
Intercept	8.038	2.45E-307	7.950	8.127
DataSet1	0.342	2.79E-20	0.274	0.409
DataSet2	0.184	9.52E-06	0.103	0.264
Monday	2.113	1.40E-116	2.003	2.223
Tuesday	2.313	1.71E-126	2.203	2.423
Wednesday	2.343	5.98E-128	2.234	2.453
Thursday	2.278	8.18E-125	2.168	2.388
Friday	2.361	5.01E-130	2.253	2.470
Saturday	0.092	1.01E-01	-0.018	0.202

Table 4.8: Regression of $\ln(\tilde{\phi}^\tau)$ (log desktops) for $\hat{\mu}$ used in demand generation. $\overline{R}^2 = 0.939$.

The results of regression with and without Weeknumber and Demand k Days Ago are listed in Table 4.7 and Table 4.8, respectively. We use the latter table as the structural model for $\mathrm{E}\tilde{\phi}^\tau$; that is, for each day τ, we can use the coefficients in Table 4.8 to obtain $\hat{\mu}^\tau$.

In order to estimate $\hat{\Sigma}$, we first note that our samples $\ln(\tilde{\phi}^\tau)$ are not independent and identically distributed; they are temporally dependent and interrelated. To gain tractability, we assume there is some constant coefficient of variation β_0 such that

$\hat{\Sigma}_{\tau,\tau} = (\beta_0 \hat{\mu}^\tau)^2$ which gives us normalized samples

$$L(\tilde{\phi}^\tau) := \frac{\ln(\tilde{\phi}^\tau) - \hat{\mu}^\tau}{\hat{\mu}^\tau} \sim \mathrm{N}(0, \beta_0^2).$$

To deal with the lack of independence, we assume that our 310 samples provide sufficient mixing that short-term inter-temporal correlations can be ignored; that is, we have enough samples to have a representative view of the sample space and can treat these observations as independent. We then estimate β_0 by $\hat{\beta}_0 = \sqrt{\mathrm{Var}(L(\tilde{\phi}))}$ which yields $\hat{\beta}_0 = 0.0277$ with a 95% confidence interval of $[0.025, 0.030]$. We then use this to generate the diagonal of $\hat{\Sigma}$ by $\hat{\Sigma}_{\tau,\tau} = (\hat{\beta}_0 \hat{\mu}^\tau)^2$. We refrain from developing the off-diagonal entries of $\hat{\Sigma}$ and instead construct the off-diagonal of Σ indirectly in subsection §4.5.2.

We can then convert $\ln \tilde{\phi}^\tau$to $\tilde{\phi}^\tau$ by the well known formulas for Log-Normal distributions

$$\mathbb{E}\tilde{\phi}^\tau = e^{\hat{\mu}_\tau + \hat{\Sigma}_{\tau\tau}/2} \ \forall \tau$$

$$C_{\tau,\tau'} = \mathrm{Cov}(\tilde{\phi}^\tau, \tilde{\phi}^{\tau'}) = (e^{\hat{\Sigma}_{\tau\tau'}} - 1) e^{\left(\hat{\mu}_\tau + \hat{\mu}_{\tau'} + \frac{\hat{\Sigma}_{\tau\tau} + \hat{\Sigma}_{\tau'\tau'}}{2}\right)} \ \forall \tau = \tau'.$$

We can now apply (4.11) to have $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{\phi}$.

We now develop the diagonal entries of $\boldsymbol{\Sigma}$, the covariance matrix of $\boldsymbol{\phi}$, by determining their relationship to the above quantities. Again, we make the assumption that there is a constant coefficient of variation β_τ such that $\Sigma_{i,i} = (\beta_\tau \mu_i)^2 \ \forall i \in \{(t, \tau, d)\}$. In order to gain further tractability, suppose that, on each day τ, orders from different destinations d and due dates t are independent; we will adjust for this in §4.5.5 by re-normalizing the final covariance matrix to match the empirical variability. Then we have

$$C_{\tau,\tau} = \sum_{t,d} \Sigma_{i,i} = \beta_\tau^2 \sum_{t,d} \mu_i^2$$

and therefore $\beta_\tau^2 = \frac{C_{\tau,\tau}}{\sum_{t,d} \mu_i^2}$ and hence we have the diagonal of $\boldsymbol{\Sigma}$.

Later, we will wish to generate random variates $\boldsymbol{\phi}$; to do so, we follow some suggestions outlined by Law [Law]. We generate a multivariate Gaussian $N = \mathrm{N}(\mu', \Sigma')$

and set $\phi_i = e^{N_i}\ \forall i$ where

$$\mu'_i = \ln\left(\frac{\mu_i^2}{\sqrt{\mu_i^2 + \Sigma_{ii}}}\right)$$

and

$$\Sigma'_{i,i'} = \ln\left(1 + \frac{\Sigma_{i,i'}}{|\mu_i \mu_{i'}|}\right)$$

with $\Sigma_{i,i'} = \Sigma'_{i,i'} = 0\, \forall i \neq i'$, i.e. that all demand components are independent. Using this, we also know that $N_i = \mu'_i + Z_i\sqrt{\Sigma'_{ii}}$ where $Z_i \sim \mathrm{N}(0,1)$ is a standard Normal random variable. Having a basic structure for the demand ϕ, we now address the independence assumption.

4.5.2 Correlation Structure

In this section, we incorporate dependency between components of the demand vector ϕ. Correlation in demand can arise from many factors, ranging from temporary market trends and fluctuations in the economy to corporate sales promotions. Kahn [Kah87] and Maccini and Zabel [MZ96], show that serial (temporal) correlation in demand for backlogged system causes greater production variance than demand variance. We also expect that, without dynamic re-balancing, geographic correlation in demand can cause imbalances in factory utilization. Fluctuations in the total demand for large subsets of the demand vector, i.e. positive correlations among those components, can cause some factories to have insufficient capacity to meet demand and induce order lateness. Correlating large geographically and temporally contiguous subsets of the demand vector allows us to analyze how the firm's supply chain production and capacity respond to large changes in demand under various solution policies. Incorporating correlation improves the realism of the model and can illustrate the supply chain's flexibility.

Attempting to construct the off-diagonal entries of Σ using inter-temporal and geographic correlations proved too dense for computation purposes. Although correlations in demand cannot be observed from the available data, we must model it in order to evaluate the performance of policies in solving the DP. The following simple technique introduces correlation in an understandable and controllable form. It separates the demand into the weighted sum of two vectors, one where all of the components are independent and the other with perfect correlation between large subsets of the vector; changing the weight parameter allows us to analyze demand situations from complete independence to perfectly correlated subsets.

We introduce correlation by decomposing ϕ_i into

$$\boldsymbol{\phi} = [(1-\alpha)\boldsymbol{u} + \alpha\boldsymbol{v}] \tag{4.12}$$

where $\alpha \in [0,1]$ and

$$u_i \sim v_i \sim \phi_i \sim \mathrm{LN}(\mu_i', \Sigma_{ii}') \ \forall i;$$

that is, u_i and v_i have the same distribution as ϕ_i . The components of $\boldsymbol{u}$ are independent while large subsets of the components of $\boldsymbol{v}$ are completely dependent, but each subset is independent from each other.

We divide the nation's geography into thirds (West, Central and East) and the due dates t into three "months," all having similar average demand. We then have nine subsets, denoted $g \in G$ and set

$$v_i := e^{\mu_i' + Z_i^v \sqrt{\Sigma_{ii}'}}$$

with

$$Z_i^v = Z_{i'}^v \ \forall (i,i') \in g \times g \subset G \times G$$

when generating v_i rather than sampling them randomly. The rather large size of these subsets is sufficient to drive imbalances across factories and throughout the horizon, allowing us study the impact of correlation in demand upon the performance of various policies, while being simple enough to easily generate, evaluate, analyze, and understand.

All components of $\boldsymbol{u}$ are generated independent of each other. Within $g \subset G$, all components of $\boldsymbol{v}$ have a correlation of one. Choosing $\alpha = 1$ makes all components of $\boldsymbol{\phi}$ within g have a correlation of one. Similarly, choosing $\alpha = 0$ makes all components of ϕ within g independent and hence have a correlation of zero.

Using the structure of equation (4.12) maintains $\mathbb{E}\phi = \mu$. However, for any $\alpha > 0$, we will have $\forall(i, i') \in g \times g$:

$$Var(\sum_{t,d} \phi_t^{\tau,d}) = \sum_{t,d} \Sigma_{i,i} + \sum_{t,d} \sum_{t \neq t', d \neq d'} \Sigma_{i,i'} > C_{\tau,\tau} = Var(\tilde{\phi}^{\tau})$$

since $\Sigma_{i,i'} = \rho_{i,i'} \sqrt{\Sigma_{i,i} \Sigma_{i',i'}} > 0$ as i and i' are positively correlated. Any demand generated by using this technique will induce greater variance than that observed empirically; this is corrected in §4.5.5. The value of α is chosen to ensure that variation in total demand matches historical quantities, as is explained in more detail in §4.5.5.

4.5.3 Additional Adjustments

Two additional adjustments must be made to the model. These are to 1) make demand discrete and 2) to incorporate an end-of-quarter phenomena that appears in production control in practice.

Randomized Rounding

Our demand generation process is continuous, but in practice, orders come in discrete numerical quantities. For some regions, such as early due dates in low population destinations, the average demand per day can be close to zero. In such cases, using a continuous approximation to demand would be misleading. To conservatively correct for this, we use randomized rounding to create discrete demand with the same mean and slightly higher variability than the continuous demand described previously.

We round a fractional demand $x.y$ to

$$z_{\tau} = \begin{cases} x & \text{w.p. } 1 - y \\ x + 1 & \text{w.p. } y \end{cases}$$

(w.p. stand for "with probability") which maintains $\mathbb{E}[x.y] = \mathbb{E}[z_{\tau}]$ but will have slightly higher variability.

Hockey-Stick Effect

The firm claimed that demand traditionally has followed a "hockey-stick" effect, wherein demand rises sharply at the end of the quarter as sales teams try to finish deals to meet sales quotas. The following t-tests, for the two populations with unknown means and variance on the residuals $\ln(\tilde{\phi}^{\tau}) - \hat{\mu}_{\tau}$, exemplifies how our data does not reflect this effect. The data points $\ln(\tilde{\phi}^{\tau}) - \hat{\mu}_{\tau}$ are split into two groups, those when τ represents a day in the last two weeks of a fiscal quarter and those in the first eleven weeks. The Null Hypothesis that $\mathbb{E}[\ln(\tilde{\phi}^{\tau}) - \hat{\mu}_{\tau}]|_{\tau\geq 12} = \mathbb{E}[\ln(\tilde{\phi}^{\tau}) - \hat{\mu}_{\tau}]|_{\tau\leq 11}$ has a p-value of 0.240 on weekdays and 0.315 on weekends and hence cannot be rejected. Alternatively, the Null Hypothesis that $\mathbb{E}[\ln(\tilde{\phi}^{\tau}) - \hat{\mu}_{\tau}]|_{\tau\geq 12} \geq \mathbb{E}[\ln(\tilde{\phi}^{\tau}) - \hat{\mu}_{\tau}]|_{\tau\leq 11}$ would be rejected at a 16% but not at a 10% significance level. Therefore, it is reasonable to proceed as if demand did not follow the "hockey-stick" effect.

However, the firm did have a policy of trying to "clear the ATB" by the end-of-quarter, i.e. all orders that arrive during a fiscal quarter were to be shipped by the last day of the quarter, day K, if possible, which does create such an effect on production. We illustrate this in §6.1 when describing the firm's production capacity for the planning problem. We emulate the firm's end-of-quarter "hockey-stick" effect by shifting demand that arises during $\{1, \ldots, K\}$ but is due after K back to day K by setting

$$\phi_K^{\tau,d} := \sum_{t\geq K} \phi_t^{\tau,d} \, \forall \tau, d.$$

Note that $\Gamma_k = 0 \; \forall k > 14$, so we have at most "two weeks" of extra orders due at the end of a thirteen week horizon.

This shift in end-of-quarter demand is also applied to the cumulative demand parameters when we compute them via

$$\Phi_t^{\tau,d} := \sum_{t'\leq t} \phi_{t'}^{\tau,d} \, \forall t, \tau, d$$

and

$$\overline{\Phi}_t^{\tau,d} := \sum_{t' \leq t, \tau' \leq \tau} \phi_{t'}^{\tau',d} \forall t, \tau, d.$$

This technique not only reflects reality, but it ensures that all demand arising within the Dynamic Program's horizon is due within the horizon. The penalty P for not producing such orders acts as a cost structure on the terminal state of the system.

4.5.4 Forecast Error

The demand model is largely concerned with modeling uncertainty in the dynamic programming problem and the results in §4.8 reflect the impact of demand uncertainty upon the performance of various policies. The policy that we recommended to the firm uses a point forecast of that demand F_i, which obviously differs from the actual demand ϕ_i due to this uncertainty; for this policy to be useful as a solution, it must be able to adapt to the difference between its forecast and the actual demand that arises. Mathematical programs are often extremely sensitive to their input data, which is a cause for concern that such solutions may not be useful when demand differs significantly from the forecast. We are interested in the following two questions: 1) How does this policy that depends on the forecast perform when the forecast differs significantly from the true demand? 2) Can improving forecasts for this policy or developing more complex stochastic techniques provide significantly better performance? In order to answer these questions and to evaluate this policy's performance overall, we must model the inaccuracy in this policy's forecast.

The scope of our project did not involve determining new forecasting techniques for the firm's operations team; instead, we use a forecast based on the data available from the firm. As described in §4.4, forecast data in the Lookahead spreadsheets was rarely updated because forecasts were made based on quarterly sales targets and distributed across the horizon; furthermore, this makes any error in the forecast, at more granular levels, perfectly (positively) correlated[1]. After cleaning the data, only fifteen data points of $F'_\tau = \sum_{t,d} F_t^{\tau,d}$ were available to understand forecast error. With such limited data, we could not use the firm's forecasts directly; instead, we construct a forecast by distorting the mean of the true demand distribution by an error term whose magnitude is estimated from empirical data, as follows.

It is common to assume that a forecast is unbiased, i.e. that

$$\mathbb{E}F_i = \mathbb{E}\phi_i \, \forall i, \tag{4.13}$$

as otherwise one could adjust the forecast by the difference of an estimate of $\mathbb{E}\phi_i$. However, it would be unrealistic to have $F_i = \mathbb{E}\phi_i$. Because the forecast is an attempt to estimate the sum of many somewhat independent choices to purchase desktops, we model forecast error as being Normally distributed. A reasonable choice for F_i is to perturb it about the mean via

$$F_i = \mu_i(1+\epsilon) \tag{4.14}$$

where $\epsilon \sim N(0, \sigma^2)$ is Normally distributed with zero mean and standard error σ. This makes F_i unbiased and have a constant (across i) coefficient of variation σ. Also, by using the same ϵ for each i, the perfect correlation in the empirical data is captured. We now determine the forecast's variability about the mean, σ.

Because most forecasts attempt to estimate the mean of the demand distribution, we compare our data on daily forecasts F'_T to the mean daily demand $\mu_T := \mathbb{E}\tilde{\phi}^T$. The relative forecast error $\frac{F'_T - \mu_T}{\mu_T}$ was calculated for these fifteen data points; they had a variance of $\sigma'^2 = 0.160$, which is the square of our point estimate for the coefficient of variation. Using our assumption of the Normality of ϵ_i, the estimator for variance in daily demand should be χ^2 distributed and have a 95% confidence interval of $[0.087, 0.383]$.

We must re-normalize this variance σ'^2 from daily values to σ^2, the co-efficient of

[1] The assumption that forecast components are perfectly correlated makes cost estimates about the recommended policy conservative, because positive correlations yield the highest variance in aggregate and higher variance forecasts contain the least amount of useful information.

variation for a triplet i. To do so, note that

$$\begin{aligned}
\sigma'^2 \mu_\tau^2 &= Var(F'_\tau) \\
&= Var(\sum_{i \ni \tau} F_i) \\
&= \sum_{i \ni \tau} \sum_{i' \ni \tau} Cov(F_i, F_{i'}) \\
&= \sum_{i \ni \tau} \sum_{i' \ni \tau} \rho_{i,i'} STD(F_i) STD(F_{i'}) \qquad (4.15) \\
&= \sum_{i \ni \tau} \sum_{i' \ni \tau} 1(\sigma\mu_i)(\sigma\mu_{i'}) \\
&= \sigma^2 \sum_{i \ni \tau} \sum_{i' \ni \tau} \mu_i \mu_{i'}.
\end{aligned}$$

The notation $i \ni \tau$ represents the indices i for demand that arises on τ, $\{i = (t, \tau', d) : \tau' = \tau\}$. The first and fourth equalities follow from our assumption of a common coefficient of variation or standard error in the forecasts, with $\rho_{i,i'} = 1$ coming from perfect correlation. Because we do not have enough data to calculate σ'^2 for each τ, we sum both sides of (4.15) over τ, and solve for σ to get

$$\sigma = \sigma' \sqrt{\frac{\sum_\tau \mu_\tau^2}{\sum_{i,i'} \mu_i \mu_{i'}}} = 4.79\% \qquad (4.16)$$

with a 95% confidence interval of [3.54%, 7.41%].

We use the standard error value $\sigma = 4.79\%$ in (4.14) to generate the Forecasts F_i. We then apply the hockey-stick effect of §4.5.3 via

$$F_K^{\tau,d} := \sum_{t \geq K} F_t^{\tau,d} \, \forall \tau, d. \qquad (4.17)$$

On day k of instance ω, when $\overline{\Phi}_t^{k,d}(\omega)$ is known $\forall t, d$, the forecast for $\tau \leq k$ is set to reflect past demand via

$$\overline{F}_t^{\tau,d} := \overline{\Phi}_t^{\tau,d}(\omega) \, \forall t, d, \tau \leq k. \qquad (4.18)$$

For future periods $\tau > k$, the cumulative forecast is

$$\overline{F}_t^{\tau,d} = \overline{\Phi}_t^{k,d}(\omega) + \sum_{t' \leq t} \sum_{\tau'=k+1}^{\tau} F_{t'}^{\tau',d} \; \forall t, d, \tau > k \tag{4.19}$$

where the first term is known demand from $\overline{\Phi}$ and the second term is the cumulative forecast for the next day onward.

4.5.5 Validation

Having made many assumptions to generate the demand vector ϕ, including modeling the sum of Log-Normal random variables as Log-Normal, incorporating correlation, and implementing the hockey-stick effect, one might be concerned that ϕ no longer reflects reality. To test this, we develop four metrics, based on available data, to validate whether the demand model is generating demand that is appropriate for evaluating the efficacy of various policies in solving the dynamic program. The relevant data to benchmark ourselves against is γ_d, Γ_k, $\hat{\mu}$, and $\hat{\beta}_0$.

The first two metrics are

$$M_1 = \frac{1}{2} \sum_d |\gamma_d - \frac{1}{|\Omega|} \sum_\omega \frac{\sum_{t,\tau} \phi_i(\omega)}{\sum_i \phi_i(\omega)}|$$

and

$$M_2 = \frac{1}{2} \sum_k |\Gamma_k - \frac{1}{|\Omega|} \sum_\omega \frac{\sum_{d,\tau} \phi_i(\omega)}{\sum_i \phi_i(\omega)}|$$

which give us the mismatch between the empirical and generated distributions of demand at an aggregate destination or due-date level.

Similarly,

$$M_3 = \frac{1}{K} \sum_\tau |\hat{\mu}^\tau - \frac{1}{|\Omega|} \sum_\omega ln(\sum_{t,d} \phi_i(\omega))|$$

yields the average mismatch between the empirical and generated mean daily log-demand. We will also use

$$\frac{e^{\hat{\mu}^\tau + M_3} - e^{\hat{\mu}^\tau}}{e^{\hat{\mu}^\tau}} = e^{M_3} - 1$$

as an estimate of the percentage difference between the generated and empirical mean daily demand.

Lastly, our best information on demand variability, $\hat{\beta}_0$, gives us

$$M_4 = \hat{\beta}_0 - \frac{1}{K|\Omega|}\sum_{\omega,\tau}\frac{ln(\sum_{t,d}\phi_i(\omega)) - \hat{\mu}^\tau}{\hat{\mu}^\tau},$$

the mismatch between the estimated co-efficient of variation and that produced by the model. Because the sum of Log-Normal random variables is not actually Log-Normal and because we introduce correlation in the model after computing its variance, this last metric reflects enormous error if we choose to set the parameter β_0 (coefficient of variation of daily demand used for demand generation) to the estimated coefficient of variation $\hat{\beta}_0$. To fix this, we choose β_0 so that the metric above is nearly zero; different values of β_0 will be necessary for various levels α, as shown in Table 4.9. As we introduce more correlation into demand, we reduce the coefficient of variation to maintain the same amount of total covariation. Except where noted otherwise, results are reported for $\alpha = 0$ and $\beta_0 = 0.0370$ for clarity of analysis and exposition.

Several thousand random variates ϕ were generated using the structural models detailed in this section. Table 4.9 presents how these simulated demand values compare to the empirical demand values, via the above metrics. The columns for M_1 and M_2 indicate that, for either choice of α, the demand model generates vectors ϕ whose distribution of due-dates and destinations differ from Γ and γ respectively by no more than 1% total. The values of M_3 indicate that the demand model generates vectors ϕ that differ from the empirical mean daily demand by at most 0.034% and 0.0196% for $\alpha = 0.00$ and $\alpha = 0.16$, respectively. Our choices of β_0, for a given α, make M_4 nearly zero, indicating that the demand model generates vectors ϕ that match the empirical daily log-demand variability. Overall, the data in Table 4.9 validates that, for these choices of α and β_0, our demand model generates demand vectors that match historical statistics.

α	β_0	M_1	M_2	M_3	M_4
0.00	0.0370	0.0089	0.0044	0.000335	0.0000
0.16	0.0277	0.0099	0.0029	0.000197	0.0001

Table 4.9: Demand model validation metrics for various values of β_0 and α. Note that $\hat{\beta}_0 = 0.0277$ and (α, β_0) is chosen to zero M_4.

4.6 Policies

In this section, we formalize the four policies whose performance is to be analyzed in-depth. Those polices are

1. Greedy (G)
2. Historical (H)
3. Rolling-Horizon Certainty-Equivalent Linear-Programming (LP)
4. Perfect Hindsight (PH)[2].

As described in detail in §2.4, when we first began this study, the firm used a heuristic policy, called "geographic manufacturing," which focuses on minimizing outbound shipping costs, to generate feasible solutions to the problem[3]. We call this policy Historical (H). It is derived from a much simpler policy, a set of "default download rules." This simpler policy essentially assigns each order to the factory that has the lowest shipping cost for that destination, making it a Greedy (G) policy. Using G as a baseline, the firm employed human oversight to re-balance factories by "moving" orders from factories with excessive workloads to ones with excess capacity; the historical policy (H) reflects this.

We developed and tested many alternative solutions policies. Many of these were improved upon and evolved into the solution that we eventually recommended to the firm: a Rolling-Horizon Certainty-Equivalent Linear-Programming (LP).

[2] Although the Perfect Hindsight policy is not a policy but more of a benchmark in the sense that its decisions depend on future information, we refer to it as a policy for notational convenience. Later, we show that its cost performance is a lower bound on the optimal policy's performance.

[3] During the course of our study, the firm began to use a second heuristic policy which focuses on minimizing the labor costs; it was also greedy algorithm with human oversight. This policy is analyzed in §5.3.2 when the firm's production facilities were in Winston-Salem, USA and Juarez, Mexico.

In order to benchmark the suboptimal (with respect to the optimum of the original dynamic programming (DP) problem) policies, we solve an anticipative Perfect Hindsight (PH) linear programming problem, which has perfect knowledge of demand throughout the horizon and satisfies $z_{PH} \leq z_{DP} \leq z_p$ for any non-anticipative policy p, where z denotes the cost of a policy. In general, this PH policy can perform arbitrarily better than any non-anticipative policy; however, in our case, some deterministic policies achieve values near the z_{PH}, indicating that this bound is nearly tight and that those policies perform well.

We now formalize these four policies.

4.6.1 Greedy Policy

Although rarely executed without alteration by the firm, the Greedy Policy is used as a default solution for the firm's production planning team and as a tool for the firm's management team to understand the impact of their own actions. Furthermore, since G represents the default plan, it serves as a baseline for comparison of the policies.

The Greedy policy is based on "default download rules," described in §2.4.1, that determined which factory an order would be built at based on various attributes of that order. By far, the most relevant rule was to allocate orders to factories based on a geographic manufacturing (or "geoman") map. Other order attributes that played a role in the default download rules include parts availability, technical labor expertise for particular product lines, and legal issues. However, the most prominently used rule for determining where orders were assigned was the geographic destination. At the time, the firm had a map of the United States divided into three approximately equal (with respect to demand contained) thirds, each corresponding to one of three factories, in Austin, TX; Nashville, TN; or Winston-Salem, NC; this is illustrated in Figure 2-4. Orders to be shipped to a particular third were by default allocated to the factory corresponding to that third.

This map was designed with outbound shipping costs in mind. The westernmost factory, TX, in Austin, Texas, was responsible mostly for orders destined for the U.S. West Coast. The easternmost factory, NC, in Winston Salem, North Carolina, was responsible for most orders destined for the East Coast. TN, the factory in Nashville, Tennessee, covered most orders from the middle third of the United States. Because each factory had nearly the same amount of production capacity, and demand was split approximately evenly among them based on geography, this heuristic does a great job of minimizing the total shipping cost, as shown in §6.6 for the planning problem which includes inbound shipping costs.

In order to model this policy mathematically, we need an ordering for the factories; according to the geo-manufacturing map, this is naturally $TX < TN < NC$, going from West to East, numbered 1, 2, 3. We also must choose an ordering for the destinations $d \in D$, to help us decide which states' production will be most cost effective to re-assign to another factory. Our choice was in increasing order of $C^d_{NC} - C^d_{TX}$, numbered $1, \ldots, |D|$, so the first items in the list are cheapest to ship from TX (mostly West Coast states) and later items on the list are cheaper to ship to from NC (mostly East coast states), while many destinations in the middle of the ordering will have similar shipping costs across factories but often be cheapest to ship to from TN. Alternative greedy cost structures are possible, but this choice best reflects the geo-manufacturing map that the firm used historically, as described in Chapter 2. In problems with more factories, where a single dimensional ordering is not obvious, a simple linear assignment problem could be solved.

Let D_l be the set of destinations whose production is assigned to factory l. Recall that γ_d is the historical (and mostly static) percent of all orders that arise from destination d. We wish for $\sum_{d \in D_l} \gamma_d \geq \gamma^* \, \forall l$, where $\gamma^* < 1/L$ (L is the number of factories) is a threshold for the minimum percentage of total demand that each factory should have. In our case of $L = 3$ we took $\gamma^* = 32\%$. Using the orderings above, we implement Algorithm 1 to assign factories to destinations. Following this procedure yields an assignment of destinations to factories that is very similar to the geo-man map that the firm historically used in its default download rules. Assigning destinations to factories is not enough to constitute a policy for the original dynamic programming problem; we need more detailed production decisions $\boldsymbol{y}$ and capacity decisions $\boldsymbol{h}$.

Algorithm 1 Algorithm for assigning destinations to factories for the Greedy policy's default download rules.

```
$d := 1$ and $D_l := \emptyset \forall l$
for $l = 1$ to $L$ do
  while $\sum_{d \in D_l} \gamma_d < \gamma^*$ do
    $D_l := D_l \cup d$
    $d := d + 1$
  end while
end for
```

Before we can specify the production decisions, we need to specify how the how the firm adjusted capacity. First, we briefly develop some notation, which may seem excessive at first, but will be necessary to define the Historical policy as well. Note that $\sum_d[\overline{\Phi}_K^{k,d} - \sum_l \overline{y}_{k-1,l}^d]^+$ is the number of outstanding orders at the beginning of day k and is often called the Available-To-Build backlog of orders. By default, the firm would allocate these outstanding orders among factories according to D_l, as given by Algorithm 1, yielding

$$ATB_{k,l} = \sum_{d \in D_l} [\overline{\Phi}_K^{k,d} - \sum_l \overline{y}_{k-1,l}^d]^+, \tag{4.20}$$

the Available-To-Build at factory l at the beginning of day k.

The amount of capacity used by the firm usually depended on a few states that are defined by $ATB_{k,l}$, the minimum capacity $\underline{H}_{k,l}$, the nominal capacity $H_{k,l}^N$, and the maximum capacity $\overline{H}_{k,l}$. Note that $\underline{H}_{k,l} \leq H_{k,l}^N \leq \overline{H}_{k,l}$. We now define a quantity A that indicates which regime a factory's Available-to-Build (ATB) to Capacity ratio lies within. Let $A_l \in \{1, \ldots, 4\}$ for factory l (always for day k) be defined by the following

$$A_l = \begin{cases} 1 & \text{if } ATB_{k,l} < \underline{H}_{k,l} \\ 2 & \text{if } \underline{H}_{k,l} \leq ATB_{k,l} < H_{k,l}^N \\ 3 & \text{if } H_{k,l}^N \leq ATB_{k,l} < \overline{H}_{k,l} \\ 4 & \text{if } \overline{H}_{k,l} \leq ATB_{k,l}. \end{cases} \tag{4.21}$$

Then the capacity $h_{k,l}$ on day k at factory l is set to match $ATB_{k,l}$ as much as possible via the following:

$$h_{t,l} := \begin{cases} \underline{H}_{k,l} & \text{if } A_l = 1 \\ \overline{H}_{k,l} & \text{if } A_l = 4 \\ ATB_{k,l} & \text{otherwise.} \end{cases} \tag{4.22}$$

Now that we have D_l and $h_{k,l}$, we can now determine the production decisions $y_{k,l}$ by executing Algorithm 2.

Algorithm 2 Algorithm for determining production decisions $y^d_{k,l}$ given D_l and $h_{k,l}$ for the Greedy and Historical policies.

```
$\overline{y}^d_{k,l} := \overline{y}^d_{k-1,l}\ \forall l, d$ where $\overline{y}^d_{0,l} = 0\ \forall l, d$.
for $l' = 1$ to $L$ do
  for $t = k$ to $K$ do
    for $d' \in D_{l'}$ do
      $\nu^{d'}_{k,l'} := \frac{[\overline{\Phi}^{k,d'}_t - \sum_l \overline{y}^{d'}_{k,l}]^+}{\sum_{d\in D_{l'}}[\overline{\Phi}^{k,d}_t - \sum_l \overline{y}^d_{k,l}]^+}$
      $v^{d'}_{k,l'} := \min\{\sum_{d\in D_{l'}} [\overline{\Phi}^{k,d}_t - \sum_l \overline{y}^d_{k,l}]^+, U_{k,l'}h_{k,l'} - \sum_d(\overline{y}^d_{k,l'} - \overline{y}^d_{k-1,l'})\}$
      $\overline{y}^{d'}_{k,l'} := \overline{y}^{d'}_{k,l'} + \nu^{d'}_{k,l} \cdot v^{d'}_{k,l}.$          (4.23)
    end for
  end for
end for
```

In Algorithm 2, $\nu^d_{k,l}$ is the fraction of ATB at factory l on period k due on period t that is destined for d and $v^d_{k,l}$ is the lesser of 1) the total ATB at factory l on period k due on period t and 2) the remaining capacity at factory l on period k after producing orders due before period t. This ensures that highest priority is given to orders with the earliest due date. Note that for each due date t, we will either produce all currently known orders assigned to l due by t or use the entire remaining capacity at l on period k. In the case of using the entire capacity, production is split proportionally to the amount of demand arising from each destination due on t.

We are now ready to compute the policy for all demand instances over the entire horizon, by executing Algorithm 3.

Algorithm 3 Algorithm for computing the Greedy Policy.

```
for $\omega \in \{1,\dots,\Omega\}$ do
  for $k \in \{1,\dots,K\}$, do
    run Algorithm 1 to compute $D_l \forall l$
    compute $ATB_{k,l}$ by (4.20), $A_l$ by (4.21), and $h_{k,l}$ by (4.22) $\forall l$
    run Algorithm 2 to compute $\overline{y}^d_{k,l} \forall l, d$
  end for
end for
```

The Greedy Policy does not allocate destinations to factories exactly according to the geoman map in Figure 2-4 because the outbound cost structure and the firm's manufacturing capacity changed between 2006, when the geoman map was generated, and 2008, when these policies were analyzed. To ensure that G does mirror the firm's decision to focus on outbound shipping cost, we compared the shipping cost-per-box of the two solutions in the following manner. Let $X^p_{l,d} = 1$ if policy p allocates orders destined for d to factory l and $X^p_{l,d} = 0$ otherwise, where $p = G$ is the Greedy policy and $p = M$ is the geoman map. Then $\sum_d \gamma_d C^d_l (X^G_{l,d} - X^M_{l,d}) = \$0.02$ is the difference in expected shipping cost-per-box; given that the expected shipping-cost-per-box is nearly \$10, the shipping costs of these policies differ by 0.2%, which is insignificant compared to the results we find in §4.8. Hence, we are confident that the allocation strategy in G represents the firm's geoman map and default download rules well.

4.6.2 Historical Policy

The Greedy Policy uses the firm's default download rules to split demand "evenly" among factories throughout the quarter. However, from day to day, imbalances in the ATB to Capacity ratio between factories may occur. In the policy historically employed by the firm, which we call the Historical policy (H), the firm's production planning team attempted to account for these imbalances by "moving" orders (almost equivalently destinations) between factories, on a daily basis. In order to maintain "fairness" between factories, $ATB_{k,l}$ was moved to ensure that $A_l = A_{l'}$ $\forall l, l'$ if possible.

We now mathematically adjust the default Greedy solution to incorporate the reallocation of orders among factories which were made manually by the firm, yielding the Historical policy that the firm actually used. Our modeling here captures what the human oversight at the firm executed on a regular basis. The ordered set of destinations $\{1, \ldots, |D|\}$ is partitioned into three components by two breakpoints; all destinations less than and including the first breakpoint are assigned to TX; all destinations greater than the first breakpoint and less than and including the second breakpoint are assigned to TN; all destinations greater than the second breakpoint are assigned to NC. The historical policy attempts to equalize the components of A by shifting these two breakpoints along $\{1, \ldots, |D|\}$. Let d^l be the greatest (last) destination in the ordered set that is assigned to factory l, as in

$$d^l := \max\{d \in D_l\}. \tag{4.24}$$

These destinations d_l for $l \in \{1, \ldots, L-1\}$ are the breakpoints.

To determine the Historical Policy for all instances and time periods in the horizon, run Algorithm 4. It begins by initializing $D_l \, \forall l$ to the same map as in G via Algorithm 1. The only major change from G is that H iteratively updates D_l until $A_l = A_l' \, \forall(l, l')$, computing the same capacity quantities as in G repeatedly. It does so by moving a destination from factory l to factory l' if $A_l > A_{l'}$, through the updates to D_l and $D_{l'}$ in the two 'if' statements, effectively shifting some breakpoint d_l up or down by one. Once all values of A_l are as similar as possible, it finishes by using Algorithm 2 to compute the production decisions $y_{k,l}$.

4.6.3 Linear Programming Policy

After many iterations of developing alternate solution techniques for the Dynamic Programming problem with the firm, we recommended the following solution, which we call the Rolling-Horizon, Certainty-Equivalent, Linear-Programming Policy, or

[4] Due to the discrete nature of the assignments, A may not stabilize after several iterations; if so, exiting the while loop after $|D||L|$ iterations will leave the ATB to capacity ratios nearly equal even if there exists (l, l') such that $A_l \neq A_{l'}$. This reflects the firm's historical actions well.

Algorithm 4 Algorithm for computing the Historical Policy.

```
for $\omega \in \{1, \dots, \Omega\}$ do
  for $k \in \{1, \dots, K\}$, do
    run Algorithm 1 to compute $D_l \forall l$
    while $A_l \neq A_{l+1}$ for some $l$ do
      compute $d^{l'}$ by (4.24), $ATB_{k,l'}$ by (4.20), $A_{l'}$ by (4.21), and $h_{k,l}$ by (4.22) $\forall l'$
      if $A_l > A_{l+1}$ then
        $D_l := D_l \setminus d^l$ and $D_{l+1} = D_{l+1} \cup d^l$
      end if
      if $A_l < A_{l+1}$ then
        $D_l := D_l \cup d^l + 1$ and $D_{l+1} = D_{l+1} \setminus d^l + 1$
      end if
    end while[4]
    compute $ATB_{k,l'}$ by (4.20), $A_{l'}$ by (4.21), and $h_{k,l'}$ by (4.22), $\forall l'$
    run Algorithm 2 to compute $y_{k,l}$
  end for
end for
```

Linear Programming (LP) policy for brevity. Relevant literature on similar techniques is provided in §3.1.3.

On each day k in some instance ω, when information has been gathered on the previous day's activities, the past day's orders have been collected, and forecasts for future days have been updated by other organizations within the firm, the LP policy solves Linear Program (4.25) with decision variables ψ, η, θ, and q. The solution is to then implement the decisions $h_{k,l} = \eta_{k,l}$ and $y^d_{k,l} = \psi^d_{k,l}$ immediately (on day k) but leave the decision of $h_{t,l}$ and $y_{t,l}$ with $t > k$ for later days, when more information will be available.

$$\begin{array}{rll} \textit{minimize}: & & \sum_{t,l,d} C_l^d \psi_{t,l}^d + \sum_{t,l} (H_{t,l}\eta_{t,l} + O_{t,l}\theta_{t,l}) + P q_t^{\tau,d} \\ \textit{Demand}: & \forall \tau, d & \overline{F}_\tau^{\tau,d} - q_\tau^d \le \sum_l \sum_{t=1}^{\tau} \psi_{t,l}^d \le \overline{F}_{k+T}^{\tau,d} \\ \textit{Capacity}: & \forall t, l & \sum_d \psi_{t,l}^d \le \eta_{t,l} \\ \textit{Labor}: & \forall t, l & \underline{H}_{t,l} \le \eta_{t,l} \le \overline{H}_{t,l} \\ \textit{Overtime}: & \forall t, l & \theta_{t,l} \ge \sum_{\tau=\hat{O}(t)}^{t} \eta_{\tau,l} - \hat{H}_{t,l} \\ \textit{NN}: & & \psi,\ \eta,\ \theta,\ q \ge 0 \\ \textit{Past}: & \forall t < k, l, d & \psi_{t,l}^d = y_{t,l}^d,\ \eta_{t,l} = h_{t,l} \end{array} \tag{4.25}$$

Linear Program (4.25) is exactly the same as the original Dynamic Programming problem, except that the uncertain term $\overline{\mathbf{\Phi}}$ is replaced by a deterministic term $\overline{\mathbf{F}}$. This can be seen by noting a few transformations. First, note that for $\tau = k$ we have

$$q_k^d \ge \overline{F}_k^{k,d} - \sum_l \sum_{t=1}^{k} \psi_{t,l}^d \tag{4.26}$$

along with $\overline{F}_k^{k,d} = \overline{\Phi}_k^{k,d}(\omega)$ and $\sum_l \sum_{t=1}^k \psi_{t,l}^d = \sum_l \overline{y}_{k,l}^d$. At optimality, either (4.26) or $q_k^d \ge 0$ holds with equality because $P > 0$ and we have no other constraints on q_k^d. Substituting these yields

$$q_k^d = [\overline{\Phi}_k^{k,d}(\omega) - \sum_l \overline{y}_{k,l}^d]^+ \quad \forall k, d \tag{4.27}$$

from the original problem. Similarly, for $t = k$, either

$$\theta_{k,l} \ge \sum_{\tau=\hat{O}(k)}^{k} \eta_{\tau,l} - \hat{H}_{k,l} \tag{4.28}$$

or $\theta_{t,l} \geq 0$ holds with equality at optimality because $O_{t,l} > 0$, yielding

$$\theta_{k,l} = [\overline{h}_{k,l} - \hat{H}_{k,l}]^+ \quad \forall k, d \tag{4.29}$$

because $\sum_{\tau=\hat{O}(k)}^{k} \eta_{\tau,l} = \overline{h}_{k,l}$ since $\hat{O}(k)$ is the last time period that reset $\overline{h}_{t,l}$ to zero. It is clear now that the objectives are equivalent. The remaining constraints are drawn directly from the Dynamic Programming control space (4.7), except that $\sum_l \sum_{t=1}^{\tau} \psi_{t,l}^d \leq \overline{F}_{k+T}^{\tau,d}$ holds for more than just $\tau = k$ and may be tighter because $\overline{F}_{k+T}^{k,d} \leq \overline{\Phi}_K^{k,d}$ for each decision stage k. By replacing the uncertain term $\overline{\Phi}$ with the deterministic forecast $\overline{\mathbf{F}}$, the difficult Dynamic Programming problem becomes a Linear Program, which is tractable. However, the solution given by this sequence linear programs will be sub-optimal for the Dynamic Programming problem. In §4.8, we show that the optimality gap is relatively small.

As opposed to the previous policies, the LP policy does use a forecast, which is detailed in §4.5.4. It is common practice to assume that a forecast is unbiased, i.e. $\mathbb{E}\overline{\mathbf{F}} = \mathbb{E}\overline{\Phi}$; an optimization problem where the uncertain term is replaced by its mean is known as the Certainty Equivalent problem, giving our policy that part of its name. Note that $\overline{F}_t^{\tau,d} = \overline{\Phi}_t^{\tau,d}(\omega)\, \forall t, d, \tau \leq k$, is known demand while $\overline{F}_t^{\tau,d} = \overline{\Phi}_t^{k,d}(\omega) + \sum_{t' \leq t} \sum_{\tau'=k+1}^{\tau} F_{t'}^{\tau',d}\, \forall t, d, \tau > k$ is a combination of known and forecast demand. Furthermore, we assume that the indices τ and t obey $\tau \leq k + T$ and $t \leq k + T$ for some lookahead value T; for the firm and for our analysis here, T is taken to be fourteen days. This limits the number of periods into the future for which the policy has a forecast and therefore is able to plan for, limiting the policy's "horizon;" because this updates daily or on "a rolling basis," it has the name "Rolling Horizon."

In §4.3.2 we noted that our model ignores the managerial rule that extending shift hours beyond the nominal value requires advanced notice. For realism regarding the analysis of this policy and to ensure that the decisions it suggests could be used by the firm, we often constrained the model to disallow extending shifts without advanced notice. To model it, we do the following: if $\eta_{k+1,l} < \overline{H}_{k+1,l}$ on day k, $\eta_{k+1,l}$ replaces

$\overline{H}_{k+1,l}$ when generating day $k + 1$'s policy, because day k's plan indicated that no more than that $\eta_{k+1,l}$ capacity would be necessary. In all other policies, advanced notice is always given, since it is optimal for the DP, even though it may degrade factory morale. Imposing this condition on the LP Policy can only increase its total cost and make statements about LP's efficacy conservative. The results in §4.8 do include advanced notice.

To compute the LP policy for all demand instances over the entire horizon, execute Algorithm 5.

Algorithm 5 Algorithm for computing the Rolling-Horizon, Certainty-Equivalent, Linear-Programming (LP) Policy.

```
for $\omega \in \{1, \ldots, \Omega\}$ do
  for $k \in \{1, \ldots, K\}$, do
    solve the linear program (4.25)
    set $h_{k,l} = \eta_{k,l}$ and $y^d_{k,l} = \psi^d_{k,l}$ $\forall l, d$
  end for
end for
```

4.6.4 Perfect Hindsight Policy

In order to evaluate the performance of various policies as solutions to the Dynamic Programming problem, which is too difficult to solve to optimality, we develop a lower bound on the optimal cost. We do this by solving a single linear program per instance ω that makes decisions for the entire horizon while knowing all values of uncertainty, something that cannot be done in practice. It generates the same decisions that one would if one were at the end of the horizon, knowing all of the uncertainty, and made the best possible decisions; this gives it the name "Perfect Hindsight" (PH) and makes it perform better than any policy.

For each $\omega \in \{1, \ldots, \Omega\}$, PH solves the same linear program that LP solves K times, but with $k := K$ and no "Past" constraints. In particular, it solves the Linear Program 4.30. For the same reasons as in LP, this is exactly the same problem as the original Dynamic Programming problem but with the previously unknown term $\overline{\Phi}$ replaced by its eventual value $\overline{\Phi}(\omega)$.

$$\begin{array}{rll} \textit{minimize}: & & \sum_{t,l,d} C_l^d \psi_{t,l}^d + \sum_{t,l} (H_{t,l}\eta_{t,l} + O_{t,l}\theta_{t,l}) + Pq_t^{\tau,d} \\ \textit{Demand}: & \forall \tau, d & \overline{F}_{\tau}^{\tau,d} - q_{\tau}^d \leq \sum_{l} \sum_{t=1}^{\tau} \psi_{t,l}^d \leq \overline{F}_{K+T}^{\tau,d} \\ \textit{Capacity}: & \forall t, l & \sum_{d} \psi_{t,l}^d \leq \eta_{t,l} \\ \textit{Labor}: & \forall t, l & \underline{H}_{t,l} \leq \eta_{t,l} \leq \overline{H}_{t,l} \\ \textit{Overtime}: & \forall t, l & \theta_{t,l} \geq \sum_{\tau=\hat{O}(t)}^{t} \eta_{\tau,l} - \hat{H}_{t,l} \\ NN: & & \psi,\, \eta,\, \theta,\, q \geq 0 \end{array} \tag{4.30}$$

To compute the PH policy for all demand instances over the entire horizon, execute Algorithm 6.

Algorithm 6 Algorithm for computing the Perfect Hindsight (PH) Policy.

```
for $\omega \in \{1, \ldots, \Omega\}$ do
    solve the linear program (4.30)
    set $h_{k,l} = \eta_{k,l}$ and $y_{k,l}^d = \psi_{k,l}^d \; \forall k, l, d$
end for
```

Although the PH policy cannot be implemented, it provides a lower bound on the best possible cost. If any policy performs close enough to this cost of the PH policy, that policy cannot be far from optimal.

4.7 Evaluation Criteria and Simulation Structure

For many academics, improving upon the mathematically stated objective is often the most important criteria. In a business context, this criteria is improving profitability, often referred to as the bottom line. In our case, the objective is to minimize the total relevant supply-chain cost including lateness penalties[5]. Although this is not the full supply chain cost of delivering desktops to consumers, it does account for

[5] The lateness penalty incorporates the cost of poor customer service on future sales; otherwise, revenue should not be affected by the decisions made within the scope of this problem.

the relevant supply chain costs that can be significantly affected within the decision making scope. Profitability or financial impact is measured by (4.5).

Note that the term within $\mathbb{E}$ in (4.5) is stochastic due to the constraints on the lateness quantities $\mathbf{q}$ which involve the uncertain demand $\overline{\mathbf{\Phi}}$. We must incorporate this uncertainty into our analysis in order for statements about costs to be well-defined, to mitigate risks and increase confidence in using various solutions, and to garner insights into how various policies should and do behave. As commonly done in Economics and Operations Research literature, we focus most of our attention on the expected cost of a policy. However, we also look at the distribution of costs, including extreme quantiles and the stochastic dominance of various policies over each other. Because the demand model has a high-dimensional uncertainty space, analytical and even numerical integration prove too difficult. As a proxy for integration, it is common to use simulation wherein random points are drawn from the uncertainty distribution, the policy is evaluated at those points, and the cost is averaged over them. The point estimate derived from such simulation is an unbiased estimator for the expected cost that would be obtained from direct integration. In order to determine the profitability of the policies described in §4.6, we perform such a simulation study using the models and data described throughout this chapter.

For each demand instance $\omega \in \{1,\ldots,\Omega\}$ (which determines the values of $\overline{\Phi}(\omega)$ throughout the horizon) and for each period k in the horizon $1,\ldots,K$, a policy maps the state x_k to capacity decisions $h_{k,l}$ and production decisions $y^d_{k,l}$. Note that $o_{k,l} \triangleq [\overline{h}_{k,l}(\omega) - \hat{H}_{k,l}]^+$ and $q^d_k \triangleq [\overline{\Phi}^{k,d}_k(\omega) - \sum_l \overline{y}^d_{k,l}(\omega)]^+$ are auxiliary decisions; their minimal values are fixed once $\mathbf{h}$ and $\mathbf{y}$ are fixed. To evaluate the cost of a policy, for each $\omega \in \{1,\ldots,\Omega\}$, and for each $k \in \{1,\ldots,K\}$, we generate $y^d_{k,l}(\omega)$ and and $h_{k,l}(\omega)$, $\forall l, d$, which are the decisions implemented on day k given the state x_k.

We then evaluate the cost of the policy on day k of instance ω. The four relevant cost categories, indexed by $c \in \{y,q,h,o\}$, are the shipping costs $z^y_{k,l,d} := C^d_l y^d_{k,l}(\omega)$, the direct capacity costs $z^h_{k,l} := H_{k,l}h_{k,l}(\omega)$, the additional cost of overtime capacity $z^o_{k,l} := O_{k,l}[\overline{h}_{k,l}(\omega) - \hat{H}_{k,l}]^+$, and order lateness $z^q_{k,d} := P[\overline{\Phi}^{k,d}_k(\omega) - \sum_l \overline{y}^d_{k,l}(\omega)]^+ = Pq^d_k$.

We are interested in comparing both costs and decisions along many dimensions.

Table 4.10 lists the major dimensions, our notational index, the domain of values for that dimension, and how we aggregate across that dimension when computing summary statistics. It indicates that we are interested in averaging (or standard deviation of or quantiles of) costs and decisions across demand instances ω, adding costs (but not decisions) across categories c (shipping y, lateness q, capacity h, and overtime o), and adding costs and decisions across time periods k[6], factories l, and destinations d, always aggregating over instances last. Often we will compare across a few dimensions while aggregating across the rest.

Dimension Name	Index	Domain	Summary Aggregate
Policy	p	G, H, LP, PH	Not Applicable (N\A)
Instance	ω	$1,\ldots,\Omega$	Average, Quantiles, Std. Dev.
Category	c	y, q, h, o	Add costs, N\A for decisions
Period	k	$1,\ldots,K$	Add (occasionally Average)
Factory	l	TX, TN, NC	Add
Destination	d	U.S. States	Add

Table 4.10: Dimensions to aggregate results by for evaluation.

With these summary statistic aggregation techniques in mind, we introduce a useful summary statistic notation. Let '*' be some set of index values and let u_* and z_* be the decisions and costs respectively that have index '*' and are aggregated along all other dimensions. For example, if $* =$ "LP, y, TX", then $u_{LP,y,TX} = u_*$ is the LP policy's ($p = LP$) production ($c = y$) at the TX factory ($l = TX$); we will add all of the LP Policy's production at TX across destination and periods in the horizon and then average these values across instances.

In addition to summarizing decisions and the total cost, another important metric is the "cost-per-box" (CPB). Let $\mathbf{e}$ be a vector of all ones of appropriate size and recall that $\phi(\omega)$ is the vector of demands from all arrival-dates, due-dates, and destinations for instance ω, making $\mathbf{e}\cdot\phi(\omega)$ the total demand on instance ω. Note that $z_\omega/(\mathbf{e}\cdot\phi(\omega))$ is the total cost divided by the total demand in that instance. We average this quantity across instances to get the CPB. CPB differs from the total cost divided by average demand $z/(\mathbf{e}\cdot\mathbb{E}\phi)$; normalizing by demand before averaging across instances

[6] Occasionally we may also be interested in averaging across time periods k to get the the cost per period, but this is a constant factor $1/K$ times the cost added across the horizon.

ω makes instances with higher demand (and thus higher cost) more comparable to other instances, leading to higher confidence in cost estimates. We use the same notation CPB$_*$ to indicate that we aggregate the cost-per-box across all dimensions other than those fixed by '$*$.'

We are most often interested in comparing two policies. In doing so, it is often worth noting the aggregate decisions $u_{p,c}$, cost z_p, and CPB$_p$ of each policy p, the difference between these aggregate values (e.g. $(z_{PH} - z_{LP})$), and the relative difference (e.g. $(z_{LP} - z_{PH})/z_{PH}$). Furthermore, we will often discuss these at a more granular level, by not aggregating across one or more dimension. For example, one quantity of interest is the percent of total demand for destination d that is produced at factory l by policy p, given by $u_{y,p,l,d}/u_{y,p,d}$; this quantity will let us understand which destinations are assigned to which factories and with further granularity on ω which destinations change factories between instances. Other quantities of interest include cumulative production $\sum_{t\leq t'} u_{y,t}$, cumulative capacity $\sum_{t\leq t'} u_{h,t}$, cumulative demand due $\sum_{d\in\mathbf{D}} \overline{\Phi}_{t'}^{t',d}$ over time, and how much overtime $u_{o,t,l}$ is used when and where.

We are now prepared to discuss the results of the simulation study on policy profitability.

4.8 Simulation Results and Insights

Now that we have formally defined our notation, the problem, the demand structure, the policies, and our evaluation criteria, we instantiate the data described in §4.4, generate the demand vector according to the model in §4.5, and compute the corresponding policy decisions and costs as described in §4.6 and §4.7. Results are presented for the case of independent demand components where $\alpha = 0$ and $\beta_0 = 0.0370$ unless otherwise noted.

As a high level summary of the results, Table 4.11 gives the mean, standard deviation (across instances ω), and 95% confidence intervals for the cost-per-box of each policy, CPB$_p$. The data in this table suggest that we can conclude with high confidence that

$$\mathrm{CPB}_{PH} < \mathrm{CPB}_{LP} < \mathrm{CPB}_{H} < \mathrm{CPB}_{G} \tag{4.31}$$

in expectation, as the confidence intervals are nearly non-overlapping. The Historical policy nearly halves the cost of the Greedy policy and the recommended LP policy nearly halves the cost of the Historical policy, achieving nearly the same cost as the Perfect Hindsight policy. Furthermore, the standard deviation of the costs has the same ordering; the more costly policies have a higher variance in their costs while the LP policy repeatedly performs near-optimally.

Metric\Policy	G	H	LP	PH
Mean	72.81	32.87	18.76	17.64
Standard Deviation	38.87	19.57	2.12	1.25
95% Lower Bound	59.74	26.29	18.04	17.22
95% Upper Bound	85.88	39.45	19.46	18.06

Table 4.11: 95% confidence intervals on total Cost-Per-Box by policy in dollars per desktop.

Figure (4-2) depicts this more explicitly by plotting the relative optimality gap

$$\frac{\mathrm{CPB}_{p,\omega} - \mathrm{CPB}_{PH,\omega}}{\mathrm{CPB}_{PH,\omega}} \tag{4.32}$$

against ω in increasing order of $\mathrm{CPB}_{PH,\omega}$ for each policy. The variability in $\mathrm{CPB}_{G,\omega}$ and $\mathrm{CPB}_{H,\omega}$ is readily apparent, even among scenarios in which $\mathrm{CPB}_{PH,\omega}$ is relatively low, while $\mathrm{CPB}_{LP,\omega}$ is consistently near the lower bound of $\mathrm{CPB}_{PH,\omega}$. Furthermore, G never outperforms H and H outperforms LP only once, in 34 instances, according to this metric.

With an average total demand of nearly three million desktops, the total cost is approximately three million desktops times the cost-per-box, or at least fifty million dollars. Table 4.12 reiterates the cost ordering of (4.31) but in terms of total relevant supply chain costs z_p:

$$z_{PH} < z_{LP} < z_H < z_G.$$

Note that although $z_{p,\omega} e \cdot \phi(\omega) = \mathrm{CPB}_{p,\omega}$, it *not* true that $z_p \mathbf{E}\phi(\omega) = \mathrm{CPB}_{\omega}$; Table

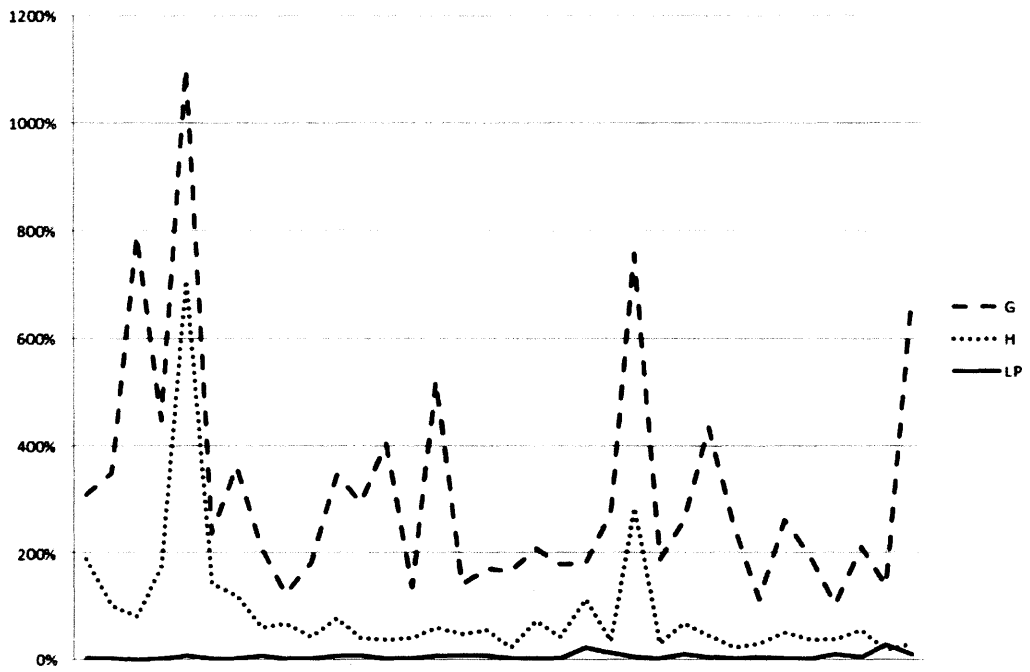


Figure 4-2: Relative Optimality Gap (4.32) of each policy on each instance ω in increasing order of $\text{CPB}_{PH,\omega}$.

4.12 emphasizes instances ω with larger total demand $e \cdot \phi(\omega)$ more than Table 4.11. Nonetheless, we see similar results with or without normalizing by the total demand. The Operations Center's current practice for solving the execution problem, which is represented by the difference in costs between G and H, saves the firm $126M in this fiscal quarter; clearly the Operations Center's plays a key role in cost management. The most striking result is that the recommended LP policy offers $47M in quarterly cost savings or $516K per day, a significant cost savings opportunity.

Metric\Policy	G	H	LP	PH
Mean	232	106	59	56
Standard Deviation	130	68	7	3
95% Lower Bound	189	83	57	55
95% Upper Bound	276	129	62	57

Table 4.12: 95% confidence intervals on the total quarterly cost z_p by policy in millions of dollars.

Table 4.13 details the cost-per-box of each policy broken down by category, $\text{CPB}_{p,c}$.

Category\Policy	G	H	LP	PH
Shipping CPB	10.42	10.51	10.97	11.03
Late CPB	56.06	16.28	1.09	0.20
Capacity CPB	6.19	6.05	6.46	6.17
Overtime CPB	0.14	0.02	0.24	0.23
Total CPB	72.81	32.87	18.76	17.64

Table 4.13: Cost-Per-Box by policy and cost category in dollars per desktop.

Contrary to CPB_p in (4.31), the ordering for the shipping costs $\mathrm{CPB}_{p,y}$ is reversed:

$$\mathrm{CPB}_{PH,y} < \mathrm{CPB}_{LP,y} < \mathrm{CPB}_{H,y} < \mathrm{CPB}_{G,y}. \tag{4.33}$$

The Greedy policy proves to be a good heuristic for minimizing the shipping cost, which without considering customer service makes up more than 62% of relevant costs for all policies. The other policies pay more to ship desktops from other facilities that have capacity available to serve demand more promptly. The PH solution spends the most on shipping in order to better match capacity with demand over time.

The Historical policy spends the least on capacity and overtime, producing as early as possible from all factories with available capacity. The Greedy policy also spends less on capacity than optimal, producing orders as soon as possible at each factory. The LP policy uses the most capacity because it occasionally delays production when capacity is available with the incorrect expectation that a less expensive factory will be available in the near future.

Lateness is the largest differentiator between policies, ranging from 1% of the PH total cost to 77% of the Greedy policy's cost. Even though H produces orders as early as possible, if all factories are beyond their maximum capacity, H does not re-prioritize orders; if one factory's orders become late while the other factories are busy on not-yet-due orders, the sets D_l do not change. In G, factories do not aid each other even if $\mathrm{ATB}_{k,l}$ is less than maximum capacity, further exacerbating order lateness. On the other hand, LP dynamically re-balances orders every day, accounting for where orders are currently late. This reflects the advantage of using more dynamic policies to more frequently and more comprehensively re-evaluate and re-balance factory loads.

We are interested in how correlation in demand affects the performance of policies. As described in §4.5.2, we introduce correlation by increasing α in the demand model. As described in §4.5.5, we must also reduce the coefficient of variation for log-daily demand, β_0 to maintain the same total covariation. Table 4.14 contains $\mathrm{CPB}_{p,c}$, the cost-per-box of each policy broken down by category, for $\alpha = 16\%$ and $\beta_0 = 0.0277$. When demand is heavily correlated, as in this scenario, the relative efficacy of each policy is similar to the uncorrelated results and the explanations above still hold. However, average cost of each policy is generally greater with more correlation but the same variance in total demand, due to regional and temporal spikes and lulls in demand. Furthermore, the variability, measured by the standard deviation of $\mathrm{CPB}_{p,\omega}$, is 28% to 565% higher.

The lateness cost of G more than doubles, on average, after introducing correlation because it cannot move production from factories whose destinations have demand spikes to those with lulls. The costs of LP and PH rise slightly as well because using the network's flexibility in capacity does have slightly higher shipping and capacity costs. The total cost for H actually decreases because less total variation helps with smoothing demand for capacity across the whole system; H addresses the month-long regional demand spikes by repeatedly bailing out any factory facing consistently high demand; having all factories at maximum capacity with one being far behind is less likely to happen with lower total variance and hence H has a lower lateness cost.

Category\Policy	G	H	LP	PH
Shipping CPB	10.33	10.64	11.15	11.24
Late CPB	122.18	15.14	1.23	0.18
Capacity CPB	6.46	6.37	6.72	6.49
Overtime CPB	0.13	0.03	0.28	0.26
Total CPB	139.10	32.18	19.39	17.88
Standard Deviation	219.53	25.14	4.30	2.70

Table 4.14: Cost-Per-Box by policy and cost category in dollars per desktop when demand is correlated via $\alpha = 16\%$.

Given that these results are largely driven by lateness, we analyze the model's sensitivity to the lateness penalty P. The solutions for policies G and H do not depend on the value of P. On each day, once orders are assigned to factories, these policies either satisfy all orders or use all available production, making the lateness quantity $u_{p,q}$ (number of desktop-days late) constant as P varies. For these policies, changing the value of P only changes the lateness cost $z_{p,q} = Pu_{p,q}$. Policies LP and PH do depend on the value of P. Figure 4-3 depicts the lateness quantity $u_{q,p,\omega}$ for several values of P on a particular instance ω that displayed significant order lateness, for policies H, LP, and PH. Policy G had a lateness quantity of 744,341 which is not displayed. For values of P less than about 16 ($/desktop/day), the shipping and capacity costs outweigh the lateness penalty, leaving no incentive for LP or PH to produce desktops on time, causing a sharp rise in the lateness quantity, especially at the end of the horizon when there are few days for the penalty to accrue. Although changes in the lateness quantity as P varies tend to occur near the end of the horizon, in some cases, the lateness quantity changes mid-horizon, indicating that lateness can influence decisions throughout the horizon. For $P \geq 30$, the lateness quantity is fairly stable, having one slight change between $P = 60$ and $P = 70$; for $P \geq 30$, customer service (avoiding order lateness) is the dominant term of the objective (4.5), forcing the model to produce desktops by their due-dates as much as possible. Sensitivity to changes in P for other demand instances ω displayed a similar pattern, becoming stable when $P \geq 30$, even though the magnitude of the lateness quantity varies by instance. Because we used the empirical value $P = 109.30$, which was estimated from the work of Dhalla [Dha08] and lies well within a relatively flat interval, we are confident that this section's results would yield similar conclusions for most reasonable values of P.

In §4.5.4, we modeled forecast error so as to understand the performance of LP. Recall that the forecast F was generated by $F = \mu(1+\epsilon)$ where μ is the mean demand and ϵ is a zero-mean scalar Normal random variable with standard error $\sigma = 4.79\%$. The error ϵ is applied to the entire demand vector, making this an ϵ error for the total demand during the horizon. The scatter plot in Figure 4-4 shows the relative optimality gap (4.32) of LP for several demand instances (generated independently of ϵ) that had different forecast errors ϵ. Fitting a line to this via linear regression

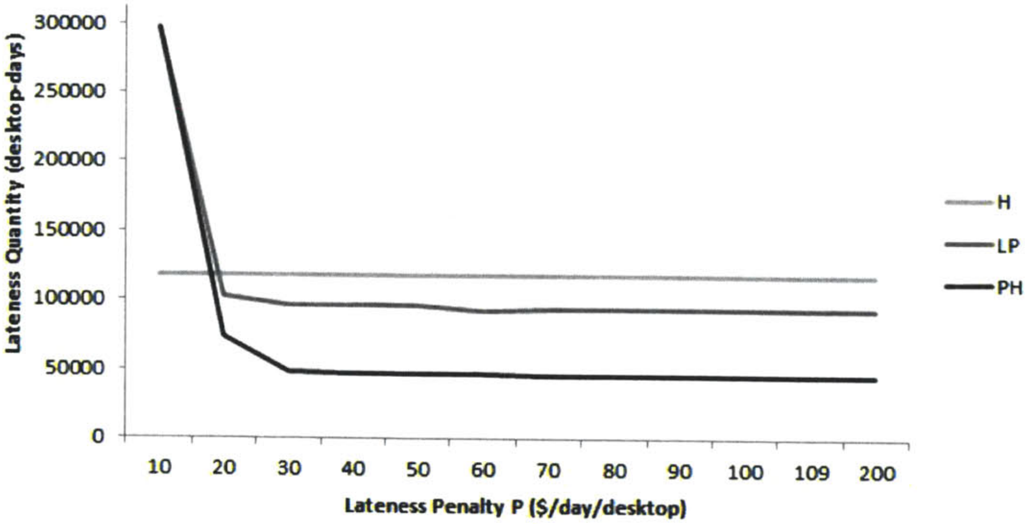


Figure 4-3: The lateness quantity $u_{q,p,\omega}$ for several values of the lateness penalty P on a particular instance ω that displayed significant order lateness, for policies H, LP, and PH.

indicates that the optimality gap is $(6.14 + 11.9\epsilon)\%$ with an R^2 value of 0.0095; for errors ϵ as large as 10%, this predicts a 1.19% decrease in optimality. Surprisingly, this also indicates that the optimality gap decreases as the forecast underestimates demand. The scatter plot for the case of correlated demand and the scatter plot for the absolute cost of LP $z_{LP,\omega}$ appear similar to Figure 4-4. With no apparent pattern and such a low R^2, these results suggest that forecast error has little impact on the performance of LP; even when the forecast differs significantly from true demand, the LP continues to perform near optimally, leaving little room for improvement by more complex forecasting or stochastic optimization techniques. This occurs because the already known demand in ATB is crucial to today's decisions whereas better information about future demand will become available before those orders need to be satisfied; the forecast only needs to be on the correct order of magnitude to guide today's decisions.

We would also like to understand how these policies differ in their decisions. Some of these differences can be inferred from the cost results, such as Table 4.13. LP spends more per desktop in shipping, capacity, and overtime capacity than G or H. Similarly, PH spends more per desktop in shipping and overtime than G or H and

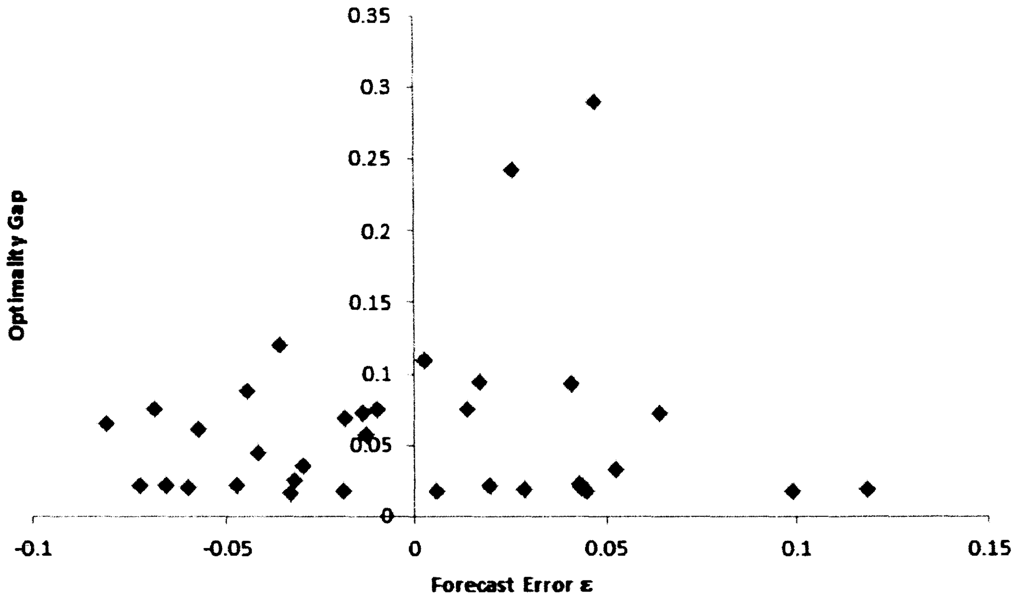


Figure 4-4: A scatter plot of the relative optimality gap (4.32) against the forecast error ϵ for the LP policy.

more on capacity than H. This is the price that LP and PH pay to reduce the number of desktops that are late. Beyond order lateness and the cost results, we analyze the distribution of production and capacity among factories, the number of destinations that each factory serves, and the use of overtime.

The distribution of production among the factories is given by $\frac{u_{y,p,l}}{u_{y,p}}$ for each policy in Table 4.15, where $u_{y,p,l}$ is the average number of desktops produced at a factory l under policy p. G splits demand close to evenly among the three factories. The other policies differ from this by moving production from TN mostly to NC. The distribution of capacity $\frac{u_{h,p,l}}{u_{h,p}}$ was similar to this; because capacity necessarily exceeds production, we show the excess capacity $\frac{u_{h,p,l}}{u_{y,p,l}} - 1$ for each l and the total excess capacity $\frac{u_{h,p}}{u_{y,p}} - 1$ in Table 4.16. It indicates that each policy used almost all capacity at TN while other factories are not fully utilized. With perfect hindsight, the total excess capacity was not be reduced beyond 1.1%. The historical policy H was more efficient with its use of capacity than LP because H produces any order in backlog if it has capacity available; the LP policy sometimes delays production when capacity is available if the LP forecast indicates cost savings can be acquired by producing the order at a later time at lower cost.

Factory \Policy	G	H	LP	PH
TX	34.1%	35.6%	32.7%	34.5%
TN	32.8%	26.4%	29.4%	28.3%
NC	33.1%	37.9%	37.9%	37.3%

Table 4.15: Distribution of production among factories, $\frac{u_{y,p,l}}{u_{y,p}}$, for each policy.

Factory \Policy	G	H	LP	PH
TX	9.9%	4.9%	9.5%	1.9%
TN	0.0%	1.8%	1.8%	0.0%
NC	12.6%	6.6%	7.3%	1.2%
Total	7.5%	4.7%	6.4%	1.1%

Table 4.16: The excess capacity $\frac{u_{h,p,l}}{u_{y,p,l}}-1$ at each factory l and the total excess capacity $\frac{u_{h,p}}{u_{y,p}}-1$ under each policy.

In order to understand how the manufacturing network is utilized by each policy, Table 4.17 presents the number of destinations for which each factory produces (at least one desktop or at least 1000 desktops); that is, it contains $D^1_{p,l} := |\{d \in D : u_{y,p,l,d} \geq 1\}|$ and $D^{1000}_{p,l} := |\{d \in D : u_{y,p,l,d} \geq 1000\}|$ for each (p, l) where $|\cdot|$ is the cardinality of a set. Note that $|D| = 51$. Policy G has the least number of destinations per factory; in fact, G assigns each destination to exactly one factory and never reallocates it. Policy H however, uses almost the full flexibility of the network, shipping to 50 of the 51 destinations from each factory. LP and PH strike a balance between these two extremes, using some of the network's ability to satisfy some destinations from multiple factory locations, incurring greater shipping costs but avoiding capacity discrepancies and order lateness. They usually avoid using all three factories for one destination, transferring imbalances between TX and NC through TN. For example, if TX cannot satisfy all demand from the western destinations it usually serves and NC has excess capacity, rather than having NC satisfy TX's excess orders, these policies will have TN satisfy some of TX's eastern-most destinations and NC satisfy some of TN's eastern-most destinations. This is illustrated by $D^{1000}_{p,l}$ being nearly two-thirds of $D^1_{p,l}$ for each factory under H, LP, and PH and TN serving the most destinations for all policies.

$D^1_{p,l}$	G	H	LP	PH	$D^{1000}_{p,l}$	G	H	LP	PH
TX	14	50	41	30	TX	14	34	24	24
TN	20	51	51	46	TN	20	37	35	32
NC	17	50	49	36	NC	17	36	29	29

Table 4.17: The number of destinations that have more than 1 or 1000 desktops produced for it by each factory for each policy.

The amount of overtime used in each factory $u_{o,l}$ is depicted in Table 4.18. One might expect that overtime is used more often at factories that have less excess production capacity. By comparing Table 4.18 to Table 4.16, we see that this is not the case, as TX had excess capacity in most policies but also used the most overtime while TN (excluding G) had relatively little overtime even though it had little excess capacity. This largely stems from the overtime cost structure; nominal labor levels, which vary by factory, were determined with the static geo-manufacturing map, which G is based on, in mind; labor in excess of the nominal amount for a pay period accrues as overtime. The nominal labor levels at NC were high enough that it need not incur much overtime while TX had planned for less production and required more overtime. TN had little overtime for H, LP, and PH because they moved production to the other factories. H needs significantly less overtime because it produces orders as soon as possible while LP and PH can delay production.

Factory \Policy	G	H	LP	PH
TX	14,957	7,953	76,886	75,663
TN	38,150	1,460	3748	672
NC	75	0	1523	980

Table 4.18: Average overtime capacity (in units of desktops) per quarter for each policy at each factory.

According to the evaluation criteria of minimizing (4.5), the human oversight that re-balances factories in the firm's Historical policy is a significant improvement over the static map used by the Greedy policy. The Rolling-Horizon Certainty-Equivalent Linear Program significantly improves upon the Historical policy, offering $47M per quarter in cost savings, coming very close to the lower bound for this model. The primary driver of these cost savings is using flexibility within the network to dynamically re-balance factories so as to best match demand patterns while accounting for the costs of capacity and shipping from each factory.

Chapter 5

Execution Problem: Implementation

5.1 Introduction

This chapter discusses the implementation of a mathematical optimization model for the firm's execution problem. See §2.2.2 for a detailed, qualitative description of this problem. Chapter 4 evaluates the performance of several policies in solving a simplified version of this problem. This chapter covers many implementation details that are too intricate or unintuitive for the in-depth stochastic analysis of Chapter 4, many of which are discussed in §4.3.

The purpose of the model presented in this chapter was foremost to solve the firm's problem. The model was also used to understand the source of potential cost savings by using mathematical optimization techniques. Furthermore, many practical issues, relevant to anyone interested in the use of similar solutions, arose in solving the problem. Additional insights were gained from the firm's employees using the model daily, giving valuable feedback as to what makes a model usable.

A large Mixed-Integer Linear Programming (MILP) formulation is developed in §5.2, where we discuss many issues encountered in modeling the problem and we present the formulation that was implemented at the firm in full detail. Results comparing this model's decisions with the firm's actual decisions in two different time periods and the insights gained through the implementation process are discussed in §5.3.

5.2 Model Development

We first describe the model qualitatively in §5.2.1. We then develop the mathematical formulation of the model, starting with some additional notation in §5.2.2. Constraints on production, parts, and labor are modeled in §5.2.3, §5.2.4, and §5.2.5, respectively, followed by factory bottlenecks in §5.2.6 and costs in §5.2.7. All of these modeling details, including many complex substitutions and extensions discussed herein, are combined and presented in the complete formulation in §5.2.8. Additional practical challenges beyond the scope of modeling and our solutions to them are given in §5.2.9.

5.2.1 Qualitative Description

We formulate a large MILP, with all necessary details, for use by the firm in solving the problem posed in §2.2.2. The model is deterministic; up to a few days worth of demand are already known from past demand and the firm's point forecast is used to help plan for production of future, unknown orders. The time horizon is up-to two weeks, often only ten days, limited by data availability. Although modeled in time increments of shifts, of which there are three per day, the model is intended for daily use to align itself with the firm's decision-making time-frame; any decisions for the first three shifts are intended to be implemented; later decisions are postponed until additional information is available and the model is re-run on a subsequent day.

The objective of the MILP is to minimize the sum of the following five quantities: 1) outbound shipping costs, 2) labor costs, 3) part shortages, 4) order lateness, and 5) deviation from prior allocation decisions. Multiplier penalties ensure the last three quantities are measured in dollars; varying these penalties offered flexibility in emphasizing various parts of this objective. Outbound shipping costs are the costs to ship desktops from factories to the customers (destinations) via third party logistics providers. Inbound shipping costs, which are the costs to ship desktop parts components from the firm's suppliers in Asia to the firm's factories in the United States, are ignored because their cost is sunk at this problem's scope; inbound parts routing decisions are made at least one month in advance while this problem only looks two weeks into the future; this problem treats parts availability as input that constrains production at each factory. Labor costs include only the variable cost of increasing available staff to produce additional desktops; this includes both their standard wages and overtime wages. A penalty is applied for each day that any one part component that the model suggests using when we expect it to be unavailable. Similarly, for each day that any desktop is due but not yet delivered, we apply another penalty. In order to avoid the solution changing for insubstantial gains, a very minor penalty is applied for re-assigning a desktop from one factory to another factory.

The primary decisions fall into four categories: 1) allocation of subsets of demand for desktops to factories, 2) production quantities over time for those desktops, 3) the shift-lengths for work-teams, determining how long factories operate, and 4) transfer of parts between factories. Often, multiple orders must be allocated (or assigned) to the same factory, making allocation a discrete decision. However, the timing of their production at a factory can be spread over multiple days, making them non-negative, continuous decisions; the discrete nature of producing "whole" desktops is ignored because most solutions return integral production decisions and because model parameters are likely inaccurate at such a small scale. Decisions for known orders (those already in ATB) are treated separately from those for future forecasted orders because we know the quantity, parts requirements, destinations, geo-eligibility and due-dates of known orders with certainty and can produce them immediately if capacity is available while we can only plan for forecasted orders. Shift lengths are also non-negative and continuous decisions, but involve other discrete decisions to overlap two consecutive shifts. Production and shift-length decisions are separated into those in overlapping and non-overlapping parts of shifts. Part transfers are also modeled continuously, but the number of pallets used to transfer those parts are non-negative integers.

Our formulation also includes many auxiliary decisions that follow from these primary decisions. Auxiliary decisions tracking the quantity of late desktops can be determined from desktop production. Similarly, auxiliary decisions for the quantities of parts available and parts short can be determined from production and parts transfers. Many costs are treated as auxiliary decisions that can be computed after making the primary decisions. All of these auxiliary decisions are piece-wise linear functions of the primary decisions; using multiple linear lower-bounding inequalities, along with their positive costs in the minimization objective, ensures these auxiliary decisions take the correct values (determined by the primary decisions) at optimality.

The major constraints in the formulation fall into five categories: 1) production, 2) labor, 3) parts, 4) factory bottlenecks, and 5) cost computations.

Production constraints includes allocating each subset of orders to a factory, producing desktops on-time or counting them as late, not producing desktops before orders for those desktops are configured by the customer, and restricting the eligibility of various desktops for production at different factories.

Most labor constraints model the labor-force staffing structure at the firm's factories. Production can occur within a factory nearly 24 hours per day, seven days per week; work-teams operate the factory in shifts, which are typically scheduled to be eight, ten, or twelve hours in length. However, one of the decisions within the scope of the model is to change this shift length. Shift lengths have upper and lower limits; the upper limit can increase if factory employees are given advanced notice that their shift may be extended. Shifts can overlap; i.e. two work-teams are operating in the same factory simultaneously. The overlapping portions of shifts must be tracked separately because some constraints apply to individual work-teams (e.g. minimum shift lengths) and some to the whole factory (e.g. physical production bottlenecks) during the interval of overlap. This also forces us to separate production and shift length decisions into those made during periods of overlap and non-overlap. Complicated overtime calculations involve each work-team's nominal schedule and number of hours worked so far this pay period. Most importantly, shift length decisions constrain production because each work-team produces computers at a particular rate.

Constraints on parts track parts inventory and shortages by accounting for the current inventory, expected future deliveries, the consumption of parts during planned production, and the transfer of parts between factories. It also includes computing the number of pallets required to execute such parts transfers.

Factory bottlenecks constrain production based on shift-lengths and the physical layout of factories.

Cost computations are mostly equalities used to simplify exposition of the objective.

A summary of the ways in which the model developed in this chapter differs from the model developed in Chapter 4 is given below.

- Demand is deterministic.
- Already known orders are treated separately from future demand.
- Groups of orders must be produced in the same factory, making part of the decision space discrete.
- Products are differentiated into Lines of Business and product families.
- Eligibility of an order for production at different factories is captured in several ways.
- Parts availability limits production and parts transfer decisions can be made.
- Time is measured in shifts, which can overlap, rather than days.
- Capacity is measured by length of work-teams' shifts rather than in desktops.
- Advanced notice must be given before capacity (shift lengths) can be increased.
- Overtime is modeled much more accurately, accounting for hours worked in pay-periods.
- Capacity is separated into labor productivity and physical bottlenecks.
- Deviation from previous solutions is penalized.

Details and justifications of these modeling choices are discussed in the following subsections. Although all of the solution policies in Chapter 4 treated demand as deterministic, we analyzed the model stochastically and found that the deterministic optimization-based policy performed near-optimally in the stochastic problem. By refraining from stochastic analysis in this chapter, many additional details can be incorporated to make the model useful in practice.

5.2.2 Notation

As much as possible, the notation in this chapter mirrors that in Chapter 4. However, the large number of additional details to make the model usable in practice requires additional notation. Here, we develop most of the notation as we describe each part of the formulation.

t, g, k, l, i, d, p, m, and w are indices and when capitalized represent the cardinality of the index set. When unspecified, assume that we consider all possible values of each index. Any other upper case letters denote input data parameters. Other lower-case letters represent decision variables, including x, $\hat{x}$, y, $\hat{y}$, y^o, $\hat{y}^o$, s, s^-, z, $\overline{z}$, h, h^o, o, $\overline{o}$, and c. Subscripts are reserved for indices. Superscripts are mostly used to add further notational depth for data or decisions.

We break the time index t into "shifts" and have exactly three shifts per day, possibly of zero length. The timing for parts transfers and assembly lead times are rounded to be measured in time units of shifts. The time horizon is denoted by T; the upcoming shift is $t = 1$; the last shift in the current labor pay-period is denoted $\underline{T}$; $t = 0$ represents the initial state. Occasionally we refer to labor hours and factory hours; a labor hour is one person working for one hour; a factory hour is all people in a factory working for one hour.

To use all available information, allocation and production variables are separated into two types: 1) decisions whose demand has already been realized (orders in ATB or known orders), which are represented by unmarked variables, e.g. "x," and 2) decisions based on forecasted demand, which are marked by a hat, e.g. "$\hat{x}$."

5.2.3 Production Constraints

Desktops can be differentiated by 1) the time t, measured in shifts, that it becomes known, is due, or is assembled, 2) the order's destination d, usually a U.S. State, such as "TX" or "MA," 3) the Line of Business (LoB) k it belongs to, such as "Consumer" or "Corporate," 4) the product family it belongs to, which itself is a part of a LoB, 5) the group, g, of orders they are associated with, and 6) the current factory location l to which the order has been assigned.

D is the number of destinations; we use the 50 U.S. states, Washington D.C. and "non-geo-eligible," the latter being explained in §4.3.1.

K is the number of LoBs we consider; we use "Consumer" and "Corporate." Each LoB is composed of several product families, a more granular category of products; we only use this attribute to differentiate between groups of orders and to limit production of some product families to particular factories; otherwise, the LoB attribute captures the relevant distinctions between desktop configurations in the data that was available. Known orders, which are Available-to-Build (in ATB), are grouped together and organized by index g of cardinality G; orders within a group have additional information associated with them, such as the number of parts required to build all desktops within the group, and must be produced in the same location.The groups are chosen to be tractable for both computation and implementation. Group sizes could theoretically range from individual orders, to all orders for a particular destination, to one LoB across many destinations with a particular due date. Because there are often several hundred thousand individual orders in ATB, using individual orders becomes intractable for optimization. Nonetheless, we would like to be able to make decisions for subsets of orders that share the same LoB, destinations, and due-dates. In our results, we chose to group orders by 1) by product family, 2) their due date (overdue, today, tomorrow, 2 days out, or extended), 3) their destination, and 4) the location at which they are currently scheduled to be built. This choice of groups allows the model to move orders at least at the granularity that the firm's Operations Center had done while keeping the number of such groups manageable.

Although we group orders in ATB by index g, future demand is tracked by LoB k and destination d, because additional information to distinguish them is not available. Where necessary, we convert groups g into LoBs and destinations using $\hat{B}$, a $K \times G$ matrix of data that encodes the proportion of desktops of each LoB k that are in group g. Furthermore, $\hat{E}$ is a $L \times K$ binary matrix of data that indicates whether (1) a LoB k can be produced (is eligible) at factory l or (0) it cannot.

Input data $F_{t,k,d}$ is the demand forecast (in number of desktops) input for period t, for LoB k, with destination d. It is obtained by adding sales across factories in the firm's lookahead spreadsheets, which are described in §4.4.

$y_{t,l,g}$ is the non-negative production decision (number of desktops) for known orders for group g at facility l in period t. Similarly, $\hat{y}_{t,l,k,d}$ is the non-negative production decision (number of desktops) for forecasted demand for desktops of LoB k with destination d in period t.

All orders in a particular group g (or for LoB k destined for d in the case of forecasted demand) must be moved together; this makes executing these decisions manageable in practice and ensures that individual orders for multiple desktops are kept together, a firm policy. We define binary allocation decision variable $x_{l,g} \in \{0,1\}$ to indicate whether (1) or not (0) group g will be produced at factory l. Similarly, $\hat{x}_{t,l,k,d}$ is the binary allocation decision during period t for LoB k with destination d that assigns all forecasted demand to facility l via $\hat{x}_{t,l,k,d} = 1$. Exactly one production facility must be chosen, so we require

$$\sum_l x_{l,g} = 1 \ \forall g$$

$$\sum_l \hat{x}_{t,l,k,d} = 1 \ \forall t, k, d.$$

In order to capture due dates for assembly of orders that are already known, let $Y_{t,l,g}$ be the number of desktops in group g that are due on shift t if g is to be produced at location l. We require that cumulative production exceed the cumulative number of desktops due via

$$\sum_{t'=1}^{t} y_{t',l,g} \geq x_{l,g} \sum_{t'=1}^{t} Y_{t',l,g} \;\; \forall t, l, g. \tag{5.1}$$

In order to enforce timely production of forecasted desktops, we use Γ_τ ($0 \leq \Gamma_\tau \leq 1$, non-decreasing in τ) which represents the percentage of forecasted demand that is *due to the customer* within τ shifts after the demand becomes known, as described in §4.5. $L_{l,k,d}$ is the input average manufacturing (assuming typical work-in-progress) and shipping time, both in units of shifts, from factory l to destination d for LoB k. Let $\hat{\Gamma}_{\tau,l,k,d} := \Gamma_{\ulcorner \tau + L_{l,k,d} \urcorner} \; \forall \tau, l, k, d$ be the percentage of forecasted demand that must be *produced* within τ shifts after demand is realized if desktops of LoB k destined for d are served by factory l. Then, in order to ensure that $\hat{\Gamma}_{t+\tau,l,k,d} F_{t,k,d}$ desktops are produced by $t + \tau$ for each $\tau \in \{0..t - T\}$, given that $F_{t,k,d}$ is served by l (which is decided by $\hat{x}_{t,l,k,d}$), we require that

$$\sum_{t'=1}^{t} \hat{y}_{t',l,k,d} \geq \sum_{t'=1}^{t} \hat{\Gamma}_{t-t',l,k,d} F_{t',k,d} \hat{x}_{t',l,k,d} \;\; \forall t, l, k, d, \tag{5.2}$$

which forces cumulative production up to t to be sufficient.

In order to ensure that desktops are built to order instead of built to stock (i.e. backlog is non-negative), we require that cumulative production be no more than cumulative demand. The constraint

$$\sum_{t'=1}^{t} \sum_{l} \hat{y}_{t',l,k,d} \leq \sum_{t'=1}^{t} F_{t',k,d} \;\; \forall t, k, d$$

does this but we choose to implement it more explicitly (i.e. more constraints but possibly stronger LP relaxations) as

$$\sum_{t'=1}^{t} \hat{y}_{t',l,k,d} \leq \sum_{t'=1}^{t} F_{t',k,d} \hat{x}_{t',l,k,d} \;\; \forall t, l, k, d \tag{5.3}$$

to make the MILP formulation stronger.

LoB k can only be built at factory l if it is eligible. We can express this by eliminating assignments based on the binary matrix $\hat{E}_{k,l}$ via the following constraints:

$$\hat{B}_{k,g} x_{l,g} \leq \hat{E}_{l,k} \ \forall l, k, g$$

$$\hat{x}_{t,l,k,d} \leq \hat{E}_{l,k} \ \forall t, l, k, d.$$

We also check whether an order's more granular product family can be built at a particular factory by $B_{i,g} x_{l,g} \leq E_{l,i}$ where B and E are similar to $\hat{B}$ and $\hat{E}$ but use this more granular product family, indexed here by i, instead of LoB which is indexed by k. Other than in defining groups g of orders currently in ATB, this is the only other place where we use the more granular product family instead of Line of Business (consumer or corporate).

In addition to some groups or LoBs, many individual orders are marked as having a "non-geo-eligible" destination, which we denote by destination 'n', indicating that they can only be produced in one particular factory. International orders fulfilled from North America, labor intensive orders, and orders with low-volume parts are often considered non-geo-eligible. To capture these individual orders within forecasted demand, we let $F^{n}_{t,l,k}$ be the forecast for the additional number of units of LoB k demanded in period t that must be satisfied by factory l. We then implement constraints similar to (5.2) and (5.3) to enforce due dates and build-to-order production for non-geo-eligible, forecasted demand via

$$\sum_{t'=1}^{t} (\hat{y}_{t',l,k,\text{`}n\text{'}} - \hat{\Gamma}_{t-t',l,k,\text{`}n\text{'}} F^{n}_{t',l,k}) \geq 0 \qquad \forall t, l, k \tag{5.4}$$

$$\sum_{t'=1}^{t} (\hat{y}_{t',l,k,\text{`}n\text{'}} - F^{n}_{t',l,k}) \leq 0 \qquad \forall t, l, k. \tag{5.5}$$

$$\tag{5.6}$$

Note that we cannot simply fix $x_{l,\text{`}n\text{'}}$ because multiple factories may have non-geo-eligible demand and must instead constrain non-geo-eligible production at each factory.

5.2.4 Parts Constraints

Parts components from suppliers that are assembled into a final desktop can be differentiated by 1) the particular type of part, p, 2) the location of the parts, l, and 3) the time period that they available, t. We only consider parts that are causing or are expected to cause significant shortages, as determined by the firm's logistics team, because the relevant data is collected only for those parts. P is the number of different parts considered. This usually includes at least some chassis and monitors and occasionally includes other significant items, such as memory and processors.

$S_{t,p,l}$ is the expected the number of parts of type p that will arrive at l at the beginning of period t. Data for this was collected from several supplier databases and lookahead spreadsheets that contained information on part shortages.

A is an input data matrix of size $G \times P$ where each entry $A_{g,p}$ is the known average number of parts of type p required to produce a desktop in group g. This is computed by totaling the number of type p parts required to produce all desktop in group g divided by the quantity of desktops in group g.

$\hat{A}$ is a $K \times P$ matrix of data where each entry $\hat{A}_{k,p}$ represents the expected number of parts of type p required to build a forecasted desktop of LoB k. Although $\hat{A}$ does not apply to known orders, the data for $\hat{A}$ was estimated by averaging the number of parts required among known orders for desktops in available ATB snapshots. Combined with the quantity of orders for each group, A provides the exact number of parts required for currently known orders which differs from the expected number of parts required for the same number of forecasted orders. $\hat{A}$ is estimated from a much larger set of orders because the currently required parts can significantly differ from the parts required in the near future. For example, if there is a shortage for a particular part, orders that require it will accumulate in backlog; the proportion of orders in backlog requiring that part will exceed the proportion of future demand requiring that part.

$z_{t,l,l',p}$ is the decision variable representing the number of parts of type p to transfer from factory l to factory l' starting in shift t. We assume that transfers from l to

l' take $L_{l,l'}$ shifts to reach their destination, and hence an item shipped during shift $t - L_{l',l}$ will arrive at the beginning of shift t.

$s_{t,l,p}$ is the auxiliary inventory stock level decision for part p at factory l at the end of shift t. $s_{0,l,p} := S_{0,l,p}$ is the current inventory of part p at facility l.

We model parts inventory by setting the inventory at the end of period t equal to the previous period's ending inventory, plus forecasted part arrivals, minus parts used in production, plus net transfers, via

$$\begin{aligned} s_{t,l,p} = s_{t-1,l,p} + S_{t,l,p} - \sum_{k,d} \hat{A}_{k,p}\hat{y}_{t,l,k,d} - \sum_{g} A_{g,p}y_{t,l,g} \\ + \sum_{l'}(z_{t-L_{l',l},l',l,p} - z_{t,l,l',p}) \ \forall t, l, p. \end{aligned}$$

Because data errors during live use often made $s \geq 0$ infeasible but solutions must be generated anyway for the model to be useful, we allow $s < 0$ and penalize its negative component by introducing a negative stock variable, $s^-_{t,l,p} \geq 0$ with $s_{t,l,p} + s^-_{t,l,p} \geq 0$. We then substitute recursively and use the following instead:

$$\begin{aligned} s^-_{t',l,p} + S_{o,l,p} + \sum_{t=1}^{t'}(S_{t,l,p} - \sum_{k,d} \hat{A}_{k,p}\hat{y}_{t,l,k,d} - \sum_{g} A_{g,p}y_{t,l,g} \\ + \sum_{l'}(z_{t-L_{l',l},l',l,p} - z_{t,l,l',p})) \geq 0 \ \forall t', l, p. \end{aligned} \tag{5.7}$$

This relaxation was also necessary because, during a stock-out for part p, if $\hat{A}_{k,p} = a > 0$ for some LoB k, the model believes that no products of LoB k can be produced; however, it would usually be the case that only a fraction a (for $a < 1$) of future orders for LoB k cannot be produced. This relaxation allows the model to continue generating production plans without directly modeling data errors and uncertainty in parts availability.

Although the Operations Center did not make transfer decisions, another closely-related department within the firm did. Foreman [For08] analyzes parts transfer and inbound supply routing decisions in much greater detail than we do here. The firm has many parts transfer mechanisms, many of which we have not modeled, such as semiweekly "Redball" shuttles that travel between factories on somewhat complicated schedules and full truckloads. Because the other department ships parts that we do not account for, we cannot determine whether a full truckload is cost-efficient. We capture the flexibility of the firm's parts transfer abilities with only one mode of transit, less-than-truckload shipping on pallets, which is always available and allows for the smallest volumes of cost-effective transfers. We assume there is a cost-per-pallet (calculated by cost-per-truck divided by pallets-per-truck) where the number of pallets used must be integral, because parts are shipped with only one type of part on each pallet. Let $\underline{z}_{t,l,l',p}$ be the integer number of pallets filled with part p leaving on shift t from l to l' . Let Z_p be the number of parts of type p that fit on a pallet. Then $z_{t,l,l',p} \leq Z_p \underline{z}_{t,l,l',p}$ where $\underline{z}_{t,l,l',p} \in \mathbb{N}$ and the cost is $\sum_{t,l,l',p} C^z_{l,l'} \underline{z}_{t,l,l',p}$.

5.2.5 Labor Constraints

Labor is a complex, prominent and expensive bottleneck in desktop production capacity.

Shift hours or shift length $h_{t,l}$ is a (rational) decision variable, with units in hours, that represents how long the active work-team during period t at factory l should be producing desktops.

Associated with each work-team is that team's average production rate, often referred to as Units-per-Hour (UPH), $U_{t,l}$ which is the number of desktops that work-team (t, l) is able to produce every hour. $U_{t,l}h_{t,l}$ is then the total number of desktops that can be assembled during shift t. In practice, the UPH for a production line varies for each desktop, often due to each desktop configuration having different parts and complexities. However, because the number of desktops that can be produced in each labor hour does not vary significantly based on LoB and this determines the cost of such labor, we do not need to determine how much labor works on each LoB. UPH can also vary significantly based on the expertise of the labor force. The planning problem considered in Chapter 6 decides how much permanent and temporary staff to hire. For the execution problem which we consider in this chapter, those staffing decisions are already fixed. Although temporary labor could be sent home earlier than permanent staff, allowing UPH to vary within a shift, we do not model it because this decision often has an insignificant impact on costs and capacities or it is unimplementable due to complex labor management issues.

Labor management considerations constrain the length of each shift. Each shift length $h_{t,l}$ satisfies

$$0 \leq \underline{H}_{t,l} \leq h_{t,l} \leq \overline{H}_{t,l} + N_{t,l} H^A_{t,l} \ \forall t, l$$

where $\underline{H}_{t,l}$ and $\overline{H}_{t,l}$ are the minimum and maximum number of hours that a shift can assemble desktops during shift t at plant l, respectively, when given no advance notice. These minimums and maximums are based on a variety of factors, including promises made by management to labor and timings of other operations within a facility. N is a binary input data matrix whose entries $N_{t,l}$ indicate whether advance notice has ($N_{t,l} = 1$) or has not ($N_{t,l} = 0$) been given for shift t at facility l, which can allow the shift length to be extended. If advance notice has been given, $H^A_{t,l}$ additional hours can be added to shift t. To avoid confusion, we call these "additional hours" instead of "overtime"; we use the name "overtime" to denote hours in excess of the nominal hours per pay period that induce a higher labor pay-rate. In every scenario, $N_{t,l} = 1 \, \forall t > 6$ because notice never needs to be given more than 48 hours (6 shifts) in advance. Similarly, for shifts that are part of the normal work schedule, $N_{t,l} = 1$ if $t > 3$. However, as advance notice is only given when the firm expects to need additional hours a few days in advance, $N_{t,l}$ is often zeroed for $t < 6$. In all other cases, $N_{t,l}$ actually depends on whether the firm has given advanced notice to that shift.

The labor constraint on production during shift t at factory l is then simply

$$\sum_{k,d} \hat{y}_{t,l,k,d} + \sum_{g} y_{t,l,g} \leq U_{t,l} h_{t,l} \ \forall t, l. \tag{5.8}$$

However, this simple model does not account for an important interaction effect between shifts. The ordering of the shifts may not be strict and there may be periods of (possibly complete) overlap between two work-teams operating simultaneously in the same factory. Overlaps between shifts tend to occur when one shift is extended beyond its nominal (planned) length by more than an hour or two. We must modify many previous decisions and constraints to account for this overlap between a shift t and a previous shift $t-1$ because some constraints will apply to both work-teams combined (e.g. factory bottlenecks) and some will apply to individual work-teams (e.g. shift length minimums) during the overlap. We denote decisions made during the overlap of shift t and $t+1$ by a superscript 'o'. The decision $h_{t,l}$ is actually composed of three segments: the overlap with the prior shift, denoted $h^o_{t-1,l}$; the non-overlapping time which we still denote $h_{t,l}$; and the overlap with the next shift, $h^o_{t,l}$. The shift length quantity $h_{t,l}$ in previous constraints is replaced by $h^o_{t-1,l} + h_{t,l} + h^o_{t,l}$. Furthermore, production decisions y and $\hat{y}$ are replaced by $y^o + y$ and $\hat{y}^o + \hat{y}$, respectively. All of these variables are non-negative.

We now discuss when and how constraints on production y and $\hat{y}$ account for overlapped shifts. In the due date constraints (5.1), (5.2), and (5.4), we allow the overlap with the next shift to help meet demand on-time. In the build-to-order constraints (5.3) and (5.5), which disallow production before demand arrives, we do not count production during the overlap with the next shift, as new demand information will have arrived and the next shift may need to produce for it. Similarly, we do not count parts consumption in (5.7) during overlap with the next shift, allowing the next shift to consume parts that arrive during it, even if it still overlaps with the current shift. Labor constraints on production (5.8) use the sum of both work-teams' UPH values during overlap. Physical capacity bottlenecks and constraints (5.14) and (5.15), which we describe later, retain the same capacity per hour; the production capacity during the overlap of two shifts is enforced separately from the capacity during non-overlapping time periods. The objective function includes decisions during both non-overlap and overlap. These redefined constraints appear in the final formulation in §5.2.8.

We must do the accounting to ensure that the model captures overlap if and only if the labor schedule indicates it will occur. Input data $H^S_{t,l}$ is the number of hours between the start of shift t and the start of shift $t+1$. Overlaps occur when shift t at factory l exceeds $H^S_{t,l}$. We create the binary decision $v_{t,l}$ which indicates whether (1) or not (0) to force $h^o_{t-1,l} + h_{t,l} = H^S_{t,l}$. The relevant constraints defining shift overlaps are:

$$h^o_{t-1,l} + h_{t,l} \leq H^S_{t,l} \qquad \forall t, l \qquad (5.9)$$

$$h^o_{t,l} \leq v_{t,l}(\overline{H}_{t,l} + N_{t,l}H^A_{t,l} - H^S_{t,l}) \qquad \forall t, l \qquad (5.10)$$

$$h^o_{t-1,l} + h_{t,l} \geq v_{t,l}H^S_{t,l} \qquad \forall t, l \qquad (5.11)$$

Inequality (5.9) limits the non-overlapping shift-length to be less than the time until the next shift begins, H^S. Inequality (5.10) defines the maximum length of any overlap: zero if there is no overlap and the remaining time for that shift in excess of H^S otherwise. Inequality (5.11) forces the non-overlapping shift-length to be exactly H^S if overlap does occur.

Cost accounting for labor is also quite complex to model accurately and requires many additional constraints. We first describe the relevant cost parameters. We then develop some notation using modular arithmetic to map work-teams and the pay-periods that they work to the time (shift) index t. We then define auxiliary decision variables for the cumulative overtime using inequalities and convert this to overtime for individual shifts by subtracting the difference between subsequent cumulative terms.

The cost per labor-hour of non-overtime ("straight-time") direct-labor wages C^s_l and the unit-per-labor-hour production rate $\hat{U}_l$ are empirical averages estimated from the firm's data, based on the current permanent to temporary laborer mix. In order to maintain $U_{t,l}$ units-per-hour of production with $\hat{U}_l$ units-per-labor-hour, the firm must be paying

$$\frac{U_{t,l}}{\hat{U}_l}C^s_l$$

dollars per factory hour when not in an overtime situation.

Overtime costs C^o_l dollars per labor-hour, also computed using an average of labor costs weighted by the permanent to temporary labor mix. Overtime is only paid when the total straight-time hours $H^N_{w,l}$ (input data given in hours) for work team w

at factory l has been exceeded.

$w \in \{1, 2, 3\}$ indexes the work teams; we assume that each factory only has one work-team per shift; work team w works every third shift. Here, 'mod' is the modulo operator and 'div' is the division operator that returns integers by rounding down. At time t the working shift is

$$w(t) = ((t + 2) \bmod 3) + 1.$$

We know the number of hours that work team w has already worked this week and denote it $H^0_{w,l}$. The last of shift of the current pay-period is $\underline{T} \leq T$. We let

$$\overline{w}_1(t) := (t - 1) \text{ div } 3 \ \ \forall t \leq \underline{T}$$

represent the number of shifts that work-team $w(t)$ performs before shift t for $t < \underline{T}$, i.e. in the current pay-period. We let

$$\overline{w}_2(t) = (t - \underline{T} - 1) \text{ div } 3 \ \ \forall t > \underline{T}$$

represent the number of shifts that work-team $w(t)$ performs before time t after time $\underline{T}$, i.e. in the pay-period after the current pay-period. Also, we let

$$\underline{w}(t) = t - 3\overline{w}_2(t)$$

represent the first shift that $w(t)$ works after $\overline{T}$.

Auxiliary decision variables $\overline{o}_{t,l} \geq 0$ represent the cumulative overtime hours for work-team $w(t)$ at facility l up to shift t. The following inequalities, at optimality, will cause the cumulative overtime to be the non-negative difference between the total shift lengths and the total straight-time hours for each work team $w(t)$ during the pay-period that contains t:

$$\bar{o}_{t,l} \geq \sum_{i=0}^{\overline{w}_1(t)} h_{w(t)+3i,l} + H^0_{w(t),l} - H^N{}_{w(t),l} \ \forall t \leq \underline{T}, l \tag{5.12}$$

$$\bar{o}_{t,l} \geq \sum_{i=0}^{\overline{w}_2(t)} h_{\underline{w}(t)+3i,l} - H^N{}_{w(t),l} \ \forall t > \underline{T}, l. \tag{5.13}$$

These inequalities also need the substitution $h_{t,l} = h^o_{t-1,l} + h_{t,l} + h^o_{t,l}$. By requiring that $o_{t,l} \geq 0 \ \forall t, l$ we ensure that all work teams get at least $H^N_{w,l}$ hours of work per pay-period, a policy used by the firm to be fair to its employees.

For cost-accounting purposes, the number of hours in each shift that are considered overtime is $o_{t,l}$ and is computed by one of the following four equations:

$$\begin{aligned}
o_{t,l} &= \bar{o}_{t,l} & \forall t &\leq min\{3, \underline{T}\}, l \\
o_{t,l} &= \bar{o}_{t,l} & \forall \underline{T} + 1 \leq t &\leq min\{\underline{T} + 3, T\} \\
o_{t,l} &= \bar{o}_{t,l} - \bar{o}_{t-3,l} & \forall 4 \leq t &\leq \underline{T}, l \\
o_{t,l} &= \bar{o}_{t,l} - \bar{o}_{t-3,l} & \forall \underline{T} + 4 \leq t &\leq T, l.
\end{aligned}$$

The first two equations deal with the first shift for each work-team in the current and next pay periods; that period's overtime is simply the cumulative overtime so far in the pay-period. The last two equations compute the current shift's overtime as the difference in cumulative overtime between t and the most recent shift for $w(t)$ in the same pay-period.

5.2.6 Factory Bottlenecks

Each factory has physical limitations or production bottlenecks that would constrain production even if they had unlimited staffing, as discussed in interviews with the firm's factory managers and illustrated in §2.3. These constraints are best described using the maximum achievable production rate for a given mix of different LoBs. Special consideration must be given when only one LoB k is considered.

Factory hours $h_{t,l}$ are multiplied by a productivity factor α_l to approximate the *effective* time that a factory is assembling desktops, accounting for down-time due to rest breaks, machine failures, and other unexpected events. The factor α_l is not used for modeling labor capacity because the labor capacity parameters $U_{t,l}$ are estimated from the empirical production achieved whereas the physical limitations are based on engineering specifications that assume constant utilization.

Input data parameter $R_{l,k}$ is number of desktops-per-effective-factory-hour that can be produced at factory l if it only produces LoB k. Input data parameter $Q_{l,k,i}$ defines, for each factory l and each constraint i, the k-intercept for the number of desktops-per-effective-factory-hour that can be produced at factory l. Here, the index i is used to differentiate between multiple constraints whose indices are otherwise the same; in Figure 2-2, we see that TN encounters different bottlenecks when producing mostly consumer desktops than it does when mostly producing corporate desktops; to model this, we need more than one mixed-production physical capacity constraint. By k-intercept, we mean that if constraint i were the only physical bottleneck and only produced LoB k is produced, $Q_{l,k,i}$ desktops-per-effective-factory-hour can be produced. Dividing the production for LoB k at factory l, $\sum_d \hat{y}_{t,l,k,d} + \sum_g \hat{B}_{k,g} y_{t,l,g}$, by $R_{l,k}$ or $Q_{l,k,i}$ yields the number of effective hours needed to produce that many desktops of LoB k if that constraint is the active bottleneck.

For each bottleneck, the number of effective hours needed to produce desktops for all LoBs cannot exceed the number of effective factory hours available. Mathematically, physical bottlenecks limit production by the following constraints:

$$\sum_d \hat{y}_{t,l,k,d} + \sum_g \hat{B}_{k,g} y_{t,l,g} \leq R_{l,k} h_{t,l} \alpha_l \qquad \forall t, k, l \qquad (5.14)$$

$$\sum_k \frac{\sum_d \hat{y}_{t,l,k,d} + \sum_g \hat{B}_{k,g} y_{t,l,g}}{Q_{l,k,i}} \leq h_{t,l} \alpha_l \qquad \forall t, l, i. \qquad (5.15)$$

These constraints are also enforced during periods of overlap, in the compete formulation in §5.2.8.

5.2.7 Costs

This section describes how we modeled many costs at the firm. For clarity, enforcing penalties, and efficient post-processing, many cost components are defined by additional auxiliary variables constrained by inequalities. Because we minimize total cost, these cost coefficients are positive, and the decisions are non-negative, these variables will equal the actual cost of the primary decisions at optimality.

The cost to ship a desktop of LoB k from factory l to destination d is $C^x_{l,k,d}$ (in units of \$ per desktop), which we call the outbound shipping cost. Normally, we could multiply $C^x_{l,k,d}$ by the production decisions to get the total cost of shipping produced desktops; however, because this model allows orders to be delayed past the time horizon, we instead calculate outbound shipping costs by using the allocation decisions and the number of desktops due, via

$$c^x_g = \sum_{t,l,k,d} x_{l,g} Y_{t,l,g} \hat{D}_{g,d} \hat{B}_{k,g} C^x_{l,k,d} \qquad \forall g \tag{5.16}$$

$$\hat{c}^x_{t,k,d} = \sum_{l} \hat{x}_{t,l,d,k} FF_l F_{t,k,d} C^x_{l,k,d} \qquad \forall t,k,d. \tag{5.17}$$

Inbound shipping cost, or the cost of sourcing parts from Asia to the Supply Logistics Centers near the firm's factories, is ignored because its decisions are beyond the scope of this model and are by this point sunk costs.

However, the cost of sourcing parts from within the firm's factories is within the scope of the model. Each pallet transferred from factory l to l' costs $C^z_{l,l'}$ (in units of \$/pallet). The cost of transferring parts is then

$$c^z_{t,l} = \sum_{p,l'} \underline{z}_{t,l',l,p} C^z_{l',l}.$$

Each hour of straight-time labor costs C^s_l (in units of \$/labor-hour); the cost of straight-time labor, which does not include any overtime-hours, for shift t at factory l is

$$c^s_{t,l} = C^s_l \frac{U_{t,l}}{\hat{U}_l} (h^o_{t-1,l} + h_{t,l} + h^o_{t,l} - o_{t,l}).$$

Here, straight-time hours are computed as the difference between the shift-length and the number of overtime hours and $C^s_l \frac{U_{t,l}}{\hat{U}_l}$ is the cost per factory-hour. Similarly, each hour of over-time labor costs C^o_l where $C^o_l > C^s_l$; the cost of over-time labor for shift t at factory l is

$$c^o_{t,l} = C^o_l \frac{U_{t,l}}{\hat{U}_l} o_{t,l}.$$

We use penalties to relax the constraints on cumulative production due, (5.1) and (5.2), and parts non-negativity constraints, which would often otherwise be infeasible, in order to have a solution that can at least be partially used. These infeasible instances arose because the firm occasionally encountered parts shortages or was unable to satisfy a few orders by their due-date, no matter the decisions they made, making the problem of satisfying orders on-time impossible. For instance, orders in ATB can be past-due before the first day of the horizon. Data errors in parts availability, such as vendors reporting negative parts inventory, and data errors in production eligibility, occasionally caused by incorrect manual data entry, also led to infeasible instances. Nonetheless, for the model to be useful, it must return solutions in these circumstances.

To do so, we let P^q (\$/unit/day) be the late penalty applied for each overdue desktop on each day and P^s (\$/part/day) be the stock penalty per part per day for a part that is used but is predicted to be unavailable. Because customer service is a high priority for the firm, these penalties were often set to values larger than their estimated cost, so that late orders and parts shortages are highly discouraged in the model. Typical values for these penalties were \$500/unit/day, whereas the estimated costs were well under \$200/unit/day. As seen in §4.8 when discussing Figure 4-3, the cost to the firm of a late order justifies producing as many orders on-time as possible. Similarly, part shortages induce order lateness if demand arises for that part and should be avoided as much as possible. Furthermore, the model's solutions are not sensitive to the value of these parameters in a wide range.

We let P^u (\$/unit) be the penalty associated with moving a desktop in ATB away from its original factory. This penalty is used to encourage only moving orders when significant savings are predicted, so as to help mitigate uncertainty for parts supply and labor planners, who have a longer planning horizon than this model. P^u was often given the value \$0.01/desktop. This helps significantly when multiple optimal solutions exist, which easily occurs because many decisions share the same cost parameters. Adding this penalty addresses the second managerial concern of §4.3.3, ensuring that significant changes in the model's solution only arise from significant differences in model input, making its solutions more stable between subsequent uses.

$q_{t,l,g}$ and $\hat{q}_{t,l,k,d}$ are auxiliary decision variables representing the quantity of desktops that are not fulfilled by their due date. We relax inequalities (5.1) and (5.2); $q_{t,l,g}$ and $\hat{q}_{t,l,g}$ represent the amount by which the due date constraints, for orders and forecasted orders respectively, are violated. Similarly, the parts inventory constraints in (5.7) can be violated but auxiliary decision $s^-_{t,l,p}$ tracks the amount that each constraint is violated and is penalized. The amount that an allocation strategy deviates from the original plan, u, is defined by

$$\begin{aligned} u_{l,g} &\geq x_{l,g} - O_{l,g} && \forall g, l \\ u_{l,g} &\geq O_{l,g} - x_{l,g} && \forall g, l \end{aligned}$$

where O is a binary matrix representing which groups in the ATB are currently assigned to which factories. We also use $O_{l,g}$ to ensure that non-geo-eligible groups of orders g remain allocated to their original factory.

The complete objective function is included below in the complete formulation.

5.2.8 Complete Formulation

We now present the complete Mixed Integer Linear Programming formulation that was used in the software prototype that the firm used in the live implementations of §5.3. This includes all of the substitutions suggested above but previously left out for clarity of exposition. It also includes many initial and terminal conditions. The results in this chapter are based on this model.

$$\text{minimize: } \sum_{g} c^{x}_{g} + \sum_{t,k,d} \hat{c}^{x}_{t,k,d} + \sum_{t,l} (c^{z}_{t,l} + c^{s}_{t,l} + c^{o}_{t,l})$$

$$+ P^{q}(\sum_{t,l,g} q_{t,l,g} + \sum_{t,l,k,d} \hat{q}_{t,l,k,d}) + P^{s} \sum_{t,l,p} s^{-}_{t,l,p} + \frac{1}{2} P^{u} \sum_{l,g} u_{l,g} \sum_{t} Y_{t,l,g}$$

The production constraints are:

$$\sum_{l} x_{l,g} = 1 \qquad \forall g$$

$$\sum_{l} x_{t,l,k,d} = 1 \qquad \forall t, d, k$$

$$\sum_{t'=1}^{t} (y_{t',l,g} + y^{\circ}_{t',l,g} - Y_{t,l,g} x_{l,g}) + q_{t,l,g} \geq 0 \qquad \forall t, l, g$$

$$\sum_{t'=1}^{t} (\hat{y}_{t',l,k,d} + \hat{y}^{\circ}_{t',l,k,d} - \hat{\Gamma}_{t-t',l,k,d} F_{t',k,d} \hat{x}_{t',l,k,d}) + \hat{q}_{t,l,k,d} \geq 0 \qquad \forall t, l, k, d$$

$$\sum_{t'=1}^{t} (y_{t',l,g} + y^{\circ}_{t'-1,l,g} - Y_{t,l,g} x_{l,g}) \leq 0 \qquad \forall g, l$$

$$\sum_{t'=1}^{t} (\hat{y}_{t',l,k,d} + \hat{y}^{\circ}_{t'-1,l,k,d} - F_{t',k,d} \hat{x}_{t',l,k,d}) \leq 0 \qquad \forall t, l, k, d$$

$$\sum_{t'=1}^{t} (\hat{y}_{t',l,k,\text{‘}n'} + \hat{y}^{\circ}_{t',l,k,\text{‘}n'} - \hat{\Gamma}_{t-t',l,k,'n'} F^{n}_{t',l,k}) \geq 0 \qquad \forall t, l, k$$

$$\sum_{t'=1}^{t} (\hat{y}_{t',l,k,\text{‘}n'} + \hat{y}^{\circ}_{t'-1,l,k,\text{‘}n'} - F^{n}_{t',l,k}) \leq 0 \qquad \forall t, l, k$$

$$y^{\circ}_{0,l,g} = 0 \qquad \forall l, g$$

$$\hat{y}^{\circ}_{0,l,k,d} = 0 \qquad \forall l, k, d$$

$$\hat{x}_{t,l,k,d} - \hat{E}_{l,k} \leq 0 \qquad \forall t, l, k, d$$

$$\hat{B}_{k,g} x_{l,g} \leq \hat{E}_{l,k} \qquad \forall l, k, g$$

$$B_{i,g} x_{l,g} \leq E_{l,i} \qquad \forall l, i, g$$

$$x_{l,g} - O_{l,g} \leq 0 \qquad \forall l, g : \hat{D}_{g,'n'} = 1.$$

The parts constraints are:

$$\begin{aligned}
s_{t,l,p} &= s_{t-1,l,p} + S_{t,l,p} - \sum_{k,d} \hat{A}_{k,p}(\hat{y}_{t,l,k,d} + \hat{y}^{\circ}_{t-1,l,k,d}) \\
&\quad - \sum_{g} A_{g,p}(y_{t,l,g} + y^{\circ}_{t-1,l,g}) + \sum_{l'}(z_{t-L_{l',l},l',l,p} - z_{t,l,l',p}) && \forall t, l, p \\
s^{-}_{t,l,p} + s_{t,l,p} &\geq 0 && \forall t, l, p \\
s_{0,l,p} &= S_{0,l,p} && \forall l, p \\
z_{0,l,l',p} &= 0 && \forall l, l', p \\
z_{t,l,l',p} &\leq Z_p \underline{z}_{t,l,l',p} && \forall t, l, l', p.
\end{aligned}$$

The labor constraints are:

$$\begin{aligned}
\underline{H}_{t,l} - (h^{\circ}_{t-1,l} + h_{t,l,} + h^{\circ}_{t,l}) \leq 0 &\qquad \forall t, l \\
(h^{\circ}_{t-1,l} + h_{t,l,} + h^{\circ}_{t,l}) - (\overline{H}_{t,l} + N_{t,l} H^{A}_{t,l}) \leq 0 &\qquad \forall t, l \\
h^{\circ}_{t,l} - v_{t,l}(\overline{H}_{t,l} + N_{t,l} H^{A}_{t,l} - H^{S}_{t,l}) \leq 0 &\qquad \forall t, l \\
h^{o}_{t-1,l} + h_{t,l} - H^{S}_{t,l} \leq 0 &\qquad \forall t, l \\
h^{o}_{t-1,l} + h^{n}_{t,l} - v_{t,l} H^{S}_{t,l} \geq 0 &\qquad \forall t, l \\
h^{\circ}_{0,l} = 0 &\qquad \forall l \\
h^{\circ}_{T,l} = 0 &\qquad \forall l \\
\overline{o}_{t,l} - \sum_{i=0}^{\overline{w}_1(t)} (h^{\circ}_{w(t)+3i-1} + h_{w(t)+3i,l} & \\
+ h^{\circ}_{w(t)+3i,l} + H^{0}_{w(t),l} - H^{N}_{w(t),l}) \geq 0 &\qquad \forall t \leq \underline{T}, l \\
\overline{o}_{t,l} - \sum_{i=0}^{\overline{w}_2(t)} (h^{\circ}_{\underline{w}(t)+3i-1} + h_{\underline{w}(t)+3i,l} & \\
+ h^{\circ}_{\underline{w}(t)+3i,l} - H^{N}_{w(t),l}) \geq 0 &\qquad \forall t > \underline{T}, l \\
o_{t,l} - \overline{o}_{t,l} = 0 &\qquad \forall t \leq min\{3, \underline{T}\}, l \\
o_{t,l} - \overline{o}_{t,l} = 0 &\qquad \forall \underline{T} + 1 \leq t \leq min\{\underline{T} + 3, T\} \\
o_{t,l} - (\overline{o}_{t,l} - \overline{o}_{t-3,l}) = 0 &\qquad \forall 4 \leq t \leq \underline{T}, l \\
o_{t,l} - (\overline{o}_{t,l} - \overline{o}_{t-3,l}) = 0 &\qquad \forall \underline{T} + 4 \leq t \leq T, l. \\
\sum_{g} y_{t,l,g} + \sum_{k,d} \hat{y}_{t,l,k,d} - U_{t,l} h_{t,l} \leq 0 &\qquad \forall t, l \\
\sum_{g} y^{\circ}_{t,l,g} + \sum_{k,d} \hat{y}^{\circ}_{t,l,k,d} - (U_{t,l} + U_{t+1,l}) h^{\circ}_{t,l} \leq 0 &\qquad \forall t, l.
\end{aligned}$$

The factory bottlenecks are:

$$\sum_{d} \hat{y}_{t,l,k,d} + \sum_{g} \hat{B}_{k,g} y_{t,l,g} \leq R_{l,k} h_{t,l} \alpha_l \qquad \forall t, k, l$$

$$\sum_{d} \hat{y}^o_{t,l,k,d} + \sum_{g} \hat{B}_{k,g} y^o_{t,l,g} \leq R_{l,k} h^o_{t,l} \alpha_l \qquad \forall t, k, l$$

$$\sum_{k} \frac{\sum_d \hat{y}_{t,l,k,d} + \sum_g \hat{B}_{k,g} y_{t,l,g}}{Q_{l,k}} \leq h_{t,l} \alpha_l \qquad \forall t, l$$

$$\sum_{k} \frac{\sum_d \hat{y}^o_{t,l,k,d} + \sum_g \hat{B}_{k,g} y^o_{t,l,g}}{Q_{l,k}} \leq h^o_{t,l} \alpha_l \qquad \forall t, l.$$

The constraints regarding costs are:

$$c^z_{t,l} = \sum_{p,l'} \underline{z}_{t,l',l,p} C^z_{l',l} \qquad \forall t, g$$

$$c^s_{t,l} = C^s_l \frac{U_{t,l}}{\hat{U}_l} (h^o_{t-1,l} + h_{t,l} + h^o_{t,l} - o_{t,l}) \qquad \forall t, g$$

$$c^o_{t,l} = C^o_l \frac{U_{t,l}}{\hat{U}_l} o_{t,l} \qquad \forall t, g$$

$$c^x_g = \sum_{t,l,k,d} x_{l,g} Y_{t,l,g} \hat{D}_{g,d} Fam_{g,k} C^{ob}_{l,k,d} \qquad \forall g$$

$$\hat{c}^x_{t,k,d} = \sum_{l} \hat{x}_{t,l,k,d} F_{t,k,d} C^{ob}_{l,k,d} \qquad \forall t, k, d$$

$$u_{l,g} \geq x_{l,g} - O_{l,g} \qquad \forall g, l$$

$$u_{l,g} \geq O_{l,g} - x_{l,g} \qquad \forall g, l.$$

The vectors of decision variables are:

x, $\hat{x}$, v binary

$\overline{z}$ non-negative integer

s unrestricted sign

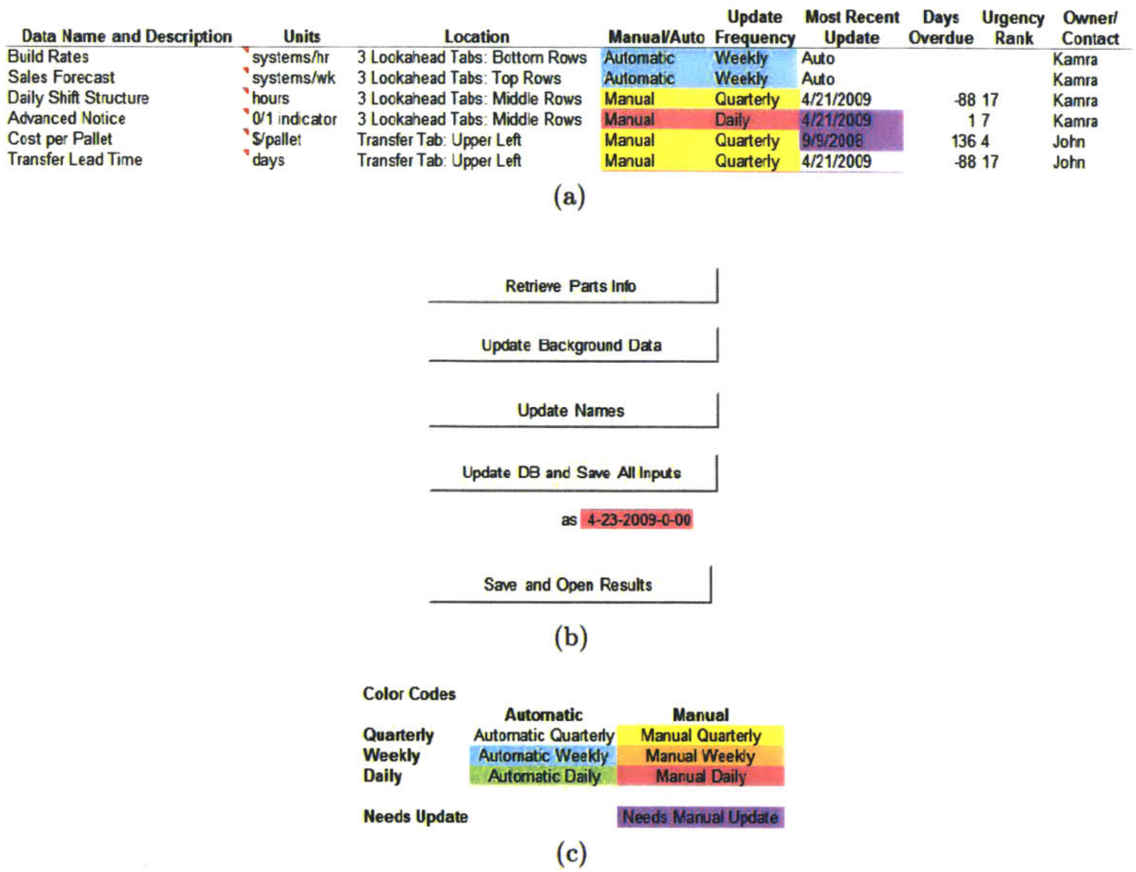
y, $\hat{y}$, y^o ,$\hat{y}^o$, q, $\hat{q}$, s^- , u, z, h, h^o, o, $\overline{o}, c$ non-negative.

5.2.9 Practical Challenges and Solutions

Many difficulties arose in making the mathematical model work in practice. In order to make the model usable for the firms employees, the model's user interface and data connections needed to be intuitive, visually informative, and somewhat fail-proof. A large amount of effort was necessary to make both input and output data accessible and accurate.

The largest area of difficulty was data acquisition. Data was maintained by different employees in different formats. A large number of data parameters or viable substitutes were not readily available. Acquiring some data, such as actual duration of shifts, required asking employees to take on the additional task of tracking such data. Other data, such as the factory capacities in §2.3, required interviewing employees at each factory. A large amount of data was automatically updated by extracting information from databases or spreadsheets that were already actively being maintained. However, a significant amount of effort is needed to maintain data and eliminate errors to make the model useful; we found the most useful techniques to be automating the process and visually alerting the user to errors.

In order to help the firm keep the model's data up-to-date, we developed a maintenance schedule, depicted in Figure 5-1, which the user would check before solving the problem to see if any data updates were necessary. Figure 5-1a includes the name or description of the data, the units of measurement, the location within the spreadsheet that the model reads that data from, whether updates were manual or automatically done, the frequency with which the data needed to be updated, the last day that it was updated, the number of days since the last update, a ranking of which data needs updating the most, and the person responsible for (owner) or to contact about the data. Conveniences such as the buttons in Figure 5-1b, some of which initiate macros to update data, save inputs, or indicate that all data was just updated, help keep the schedule and data accurate. Figure 5-1c shows the automated color coding of the schedule and most data throughout the spreadsheet which helps the model's user know what needs to be updated and at what frequency. Often new product families



Data Name and Description	Units	Location	Manual/Auto	Update Frequency	Most Recent Update	Days Overdue	Urgency Rank	Owner/ Contact
Build Rates	systems/hr	3 Lookahead Tabs: Bottom Rows	Automatic	Weekly	Auto			Kamra
Sales Forecast	systems/wk	3 Lookahead Tabs: Top Rows	Automatic	Weekly	Auto			Kamra
Daily Shift Structure	hours	3 Lookahead Tabs: Middle Rows	Manual	Quarterly	4/21/2009	-88	17	Kamra
Advanced Notice	0/1 indicator	3 Lookahead Tabs: Middle Rows	Manual	Daily	4/21/2009	1	7	Kamra
Cost per Pallet	$/pallet	Transfer Tab: Upper Left	Manual	Quarterly	9/9/2008	136	4	John
Transfer Lead Time	days	Transfer Tab: Upper Left	Manual	Quarterly	4/21/2009	-88	17	John

(a)

(b)

Color Codes

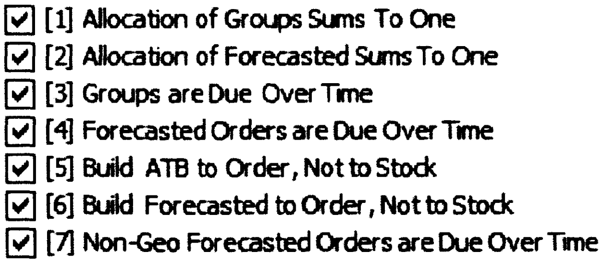
	Automatic	Manual
Quarterly	Automatic Quarterly	Manual Quarterly
Weekly	Automatic Weekly	Manual Weekly
Daily	Automatic Daily	Manual Daily
Needs Update		Needs Manual Update

(c)

Figure 5-1: The maintenance schedule, convenient buttons for updating, and color codes used to help employees keep data up-to-date.

were introduced (or had their names mistyped) and had no data on factory-eligibility available; to avoid this, unknown product families were automatically given a default set of factory-eligibility information and the name and number of desktops from this unknown product family were displayed prominently in model output to alert the user that this family was given default parameters. New or misspelled destinations were considered non-geo-eligible. Parts components often had many different aliases, often because the same part was made by multiple manufacturers; to account for this, frequent updates from the firm's MRP system were necessary. All of this made data maintenance simpler on the user.

Another difficult issue was dealing with infeasible model instances; usually this occurred due to incorrect input data. One useful technique was relaxing constraints and harshly penalizing violations of those constraints in the objective, as has already



☑ [1] Allocation of Groups Sums To One
☑ [2] Allocation of Forecasted Sums To One
☑ [3] Groups are Due Over Time
☑ [4] Forecasted Orders are Due Over Time
☑ [5] Build ATB to Order, Not to Stock
☑ [6] Build Forecasted to Order, Not to Stock
☑ [7] Non-Geo Forecasted Orders are Due Over Time

(a)

Figure 5-2: User interface to easily toggle sets of constraints on and off.

been done for due-date and part-shortage constraints in §5.2.7. Doing so allows the optimization engine to return infeasible solutions which often can either be implemented anyway, because many constraints are 'soft' or concern future actions, or can be used to alter the model's input in attempt to make the problem feasible by helping the user identify constraint conflicts. Another useful technique was adding a simple user interface for toggling large sets of constraints on or off, depicted in Figure 5-2. This interface was used to test solution sensitivity, to evaluate "what-if?" scenarios, and to narrow down the list of conflicting constraints. Additionally, automated data checking mitigated the occurrence of many common data errors that caused infeasibility. For example, if $\underline{H}_{t,l} > \overline{H}_{t,l}$ for some (t, l), i.e. the minimum shift length exceeds the maximum, those data cells were automatically highlighted in bright red, alerting the user to the data entry error.

The model often presented solutions that did not conform to the firm's normal thought process or procedures; in some cases, this is because it found unorthodox, excellent solutions; in other cases, the solution was unacceptable, often because several of the firm's internalized constraints had not been expressed because they were thought to be obvious. For example, having different allocation decisions x for each Line of Business had not been evaluated by the firm and was welcomed with enthusiasm. Additionally, the firm appreciated that the model sometimes delayed production in anticipation of a lull in future demand. However, it was unacceptable for the model to frequently choose shift lengths below their nominal values because of labor management issues. Often such solutions were eliminated by adding additional constraints that could be toggled on and off or by the user adjusting input data parameters manually. In order to identify such issues quickly, model output was post-processed and formatted to help the user understand the model's reasoning and quickly decide the solution's value to the firm. For example, summaries of model output include comparisons to the "planned" solution which would take effect if the firm did nothing. Making the model's solution understandable by providing meaningful summary statistics was crucial for its use in practice.

5.3 Results and Insights

The above model was implemented virtually, solving the model daily using live data, in two different time periods. Much of the data described in §4.4 was acquired during these two live implementations of this model and are the basis for the analysis of this model; additional data was necessary to capture individual orders, more detailed products, numerous exceptions, and additional constraints. We present results from Fall 2008 in §5.3.1 and Spring 2009 in §5.3.2. Between these two time periods, the firm discontinued desktop production at TX and TN and began outsourcing production to JM. We conclude with §5.3.3.

5.3.1 Results from Fall 2008

We analyzed the performance of this model from September 7th to October 14th 2008. This time period is representative of typical time periods at the firm, having no exceptional promotions or emergencies. At this time, the TX, TN, and NC factories were all still active. The NC Lean Lines (NCLL) had recently become operational and were incorporated into the model.

Over this period, on most weekdays, we collected and saved all relevant model parameters; this includes daily snapshots of ATB (the already known orders), the lookahead spreadsheets with demand, shift structure, and parts availability forecasts, most decisions the firm made such as order moves, parts transfers, and shift lengths, and up-to-date model input and output from a firm employee actively using the model.

The majority of our analysis was done for three consecutive days, September 29th, September 30th, and October 1st, which was the ninth week of the third fiscal quarter for the firm. We chose this period because critical data was missing on other days and this set of days had plenty of recorded Operations Center activity (e.g. order moves). Doing so allowed us to reconstruct most of the firm's decisions for the whole horizon, allowing for a more direct comparison than in other periods.

We tracked what the firm did and compared it to what the model would have done. We solve the linear programming formulation above twice with almost the same set of inputs, once to optimality and once with most of the firm's decisions fixed to their historical values (which we inferred from data collected on later dates). We knew the firm's default allocation scheme, the actual length of shifts as recorded in later lookahead files, and the availability of high-priority parts, in this time-frame. However, because we did not know the timing of the firm's production decisions, we had the optimization engine determine the firm's production decisions by using the same model but with some variables pre-determined; this makes predictions about the firm's order lateness and parts shortages lower than they may have actually been. We then evaluated the performance of the firm's policy by comparing it to the model optimum. With many of the firm's decisions fixed, the the firm's solution had a significantly smaller feasible space and hence guaranteed worse performance in every scenario.

Because the firm was not confident in some of the input parameters or constraints, we analyzed many different scenarios. To test the sensitivity of the model to surges in demand and to test the limits of the firm's supply chain flexibility, we evaluated the firm's solution and the model optimum with forecasted future demand set to 100%, 120%, and 150% of the empirical demand. As the lean lines at NC were just becoming operational, the firm was interested in analysis with and without the lean lines enabled. The model often suggested shorter shift lengths than the firm felt comfortable with; as a result, we analyzed three labor flexibility scenarios: inflexible (where all shifts must be at least their empirical length), some flexibility (where the sum of all shift lengths must exceed the empirical sum at each factory), and full flexibility (where only the shift-structure, as described in §5.2.5, constrained shift lengths). Lastly, each scenario starts on a specific date and uses the data available on that day. To summarize, a scenario consists of choosing one element from each of the following four sets:

Demand {100%, 120%, 150%}

NCLL {Enabled, Disabled}

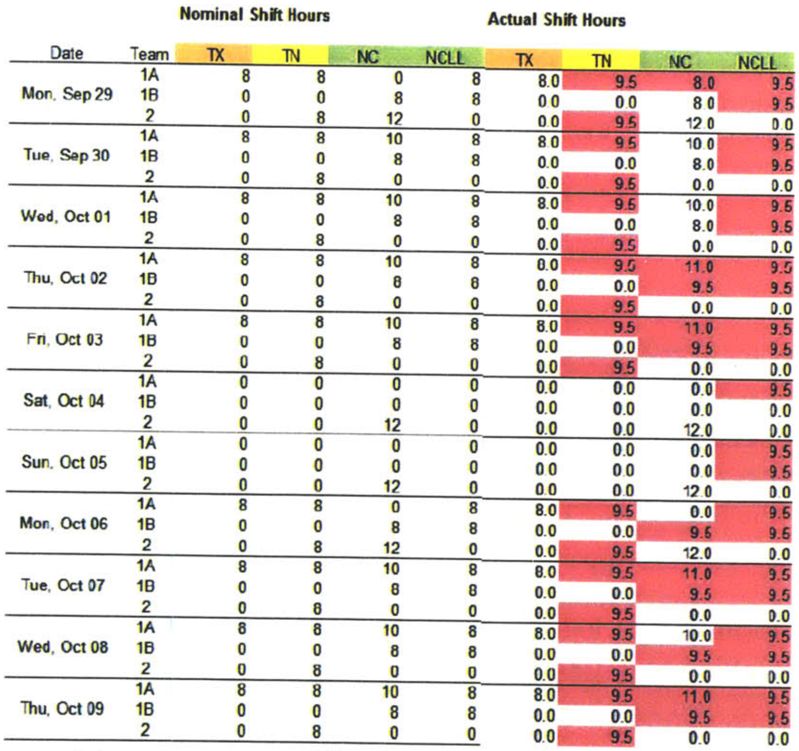
Labor {Inflexible, Some Flexibility, No Flexibility}

Start Date {9/7/2008,...,10/14/2008}.

The firm's Operations Center was also interested in the impact of its actions; to this end, we also evaluated a solution we call "plan," which uses default download rules, no order moves, default or nominal shift lengths, and no parts transfers. The solution called "actual" uses the firm's actual shift length, order allocation, and parts transfer decisions. Although the order moves that the firm made were known, we did not know which orders were produced when; to account for this, the timing (as opposed to location) of production is optimized in all three solutions, making estimates of the optimal solution's cost savings conservative. The model optimum, often denoted "opt" is the solution to the MILP in §5.2.8.

Comparing the firm's actual solution to the model's optimum in the scenario with the actual demand (100%), NCLL enabled, and inflexible labor is the best benchmark and will be our main focus. Understanding the scenarios with 150% demand instead gives interesting insights into how the solutions behave in more extreme situations.

Figures 5-3 and 5-4 display shift lengths for various solutions and illustrate the difference in labor flexibility scenarios, for scenarios that begin on September 29th 2008 with NCLL enabled and 100% of forecast. Similar displays were automatically generated as model output to help the user understand various solutions. Sub-figures 5-3a and 5-4a give the date and the work-team, which are equivalent to shifts, indexing time. Sub-figure 5-3b contains the nominal or default shift-length, corresponding to the "plan" solution, in hours for each shift and factory. Sub-figure 5-3c contains the "actual" shift-lengths for the firm's empirical solution. Similarly, sub-figures 5-4b, 5-4c, and 5-4d show the optimal shift lengths when labor is restricted to be inflexible, to have some flexibility, and to have full flexibility, respectively. Values are colored if they deviate from the nominal shift-length in sub-figure 5-3b by at least half an hour, pink if the shift is extended and light-blue if it is shortened. These figures illustrate that, for these scenarios, the firm extended almost every shift within the horizon; with some flexibility, the optimal solution extends other shifts; with full flexibility, the optimal solution shortens almost every shift. We discuss the impact of these decisions later in this section.

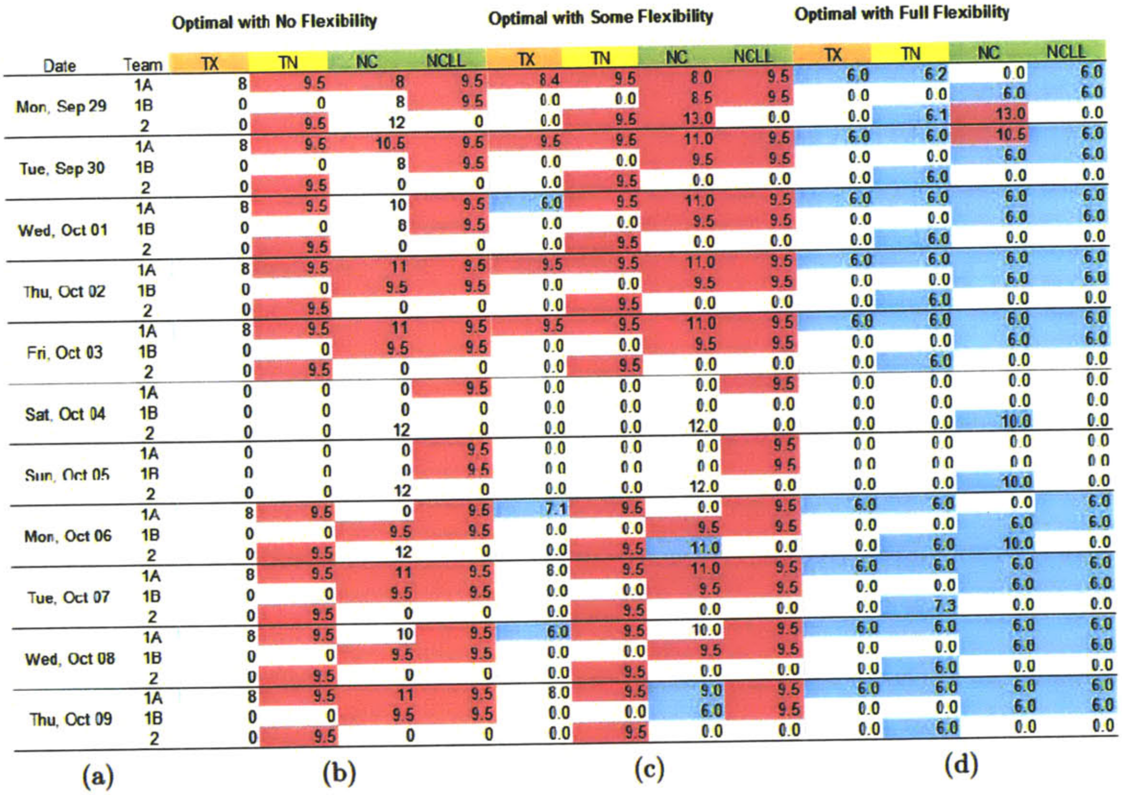


		Nominal Shift Hours				Actual Shift Hours			
Date	Team	TX	TN	NC	NCLL	TX	TN	NC	NCLL
Mon, Sep 29	1A	8	8	0	8	8.0	9.5	8.0	9.5
	1B	0	0	8	8	0.0	0.0	8.0	9.5
	2	0	8	12	0	0.0	9.5	12.0	0.0
Tue, Sep 30	1A	8	8	10	8	8.0	9.5	10.0	9.5
	1B	0	0	8	8	0.0	0.0	8.0	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0
Wed, Oct 01	1A	8	8	10	8	8.0	9.5	10.0	9.5
	1B	0	0	8	8	0.0	0.0	8.0	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0
Thu, Oct 02	1A	8	8	10	8	0.0	9.5	11.0	9.5
	1B	0	0	8	8	0.0	0.0	9.5	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0
Fri, Oct 03	1A	8	8	10	8	8.0	9.5	11.0	9.5
	1B	0	0	8	8	0.0	0.0	9.5	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0
Sat, Oct 04	1A	0	0	0	0	0.0	0.0	0.0	9.5
	1B	0	0	0	0	0.0	0.0	0.0	0.0
	2	0	0	12	0	0.0	0.0	12.0	0.0
Sun, Oct 05	1A	0	0	0	0	0.0	0.0	0.0	9.5
	1B	0	0	0	0	0.0	0.0	0.0	9.5
	2	0	0	12	0	0.0	0.0	12.0	0.0
Mon, Oct 06	1A	8	8	0	8	8.0	9.5	0.0	9.5
	1B	0	0	8	8	0.0	0.0	9.5	9.5
	2	0	8	12	0	0.0	9.5	12.0	0.0
Tue, Oct 07	1A	8	8	10	8	8.0	9.5	11.0	9.5
	1B	0	0	8	8	0.0	0.0	9.5	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0
Wed, Oct 08	1A	8	8	10	8	8.0	9.5	10.0	9.5
	1B	0	0	8	8	0.0	0.0	9.5	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0
Thu, Oct 09	1A	8	8	10	8	8.0	9.5	11.0	9.5
	1B	0	0	8	8	0.0	0.0	9.5	9.5
	2	0	8	0	0	0.0	9.5	0.0	0.0

(a) (b) (c)

Figure 5-3: Planned and actual shift lengths for each factory and shift on September 29th, 2008 with NCLL enabled and 100% of forecast.

Table 5.1 displays the average daily difference between actual and optimal labor costs, shipping costs, quantity of late desktops, and quantity of parts short. Quantities are given as daily averages; when unspecified, we have averaged across the other



		Optimal with No Flexibility				Optimal with Some Flexibility				Optimal with Full Flexibility			
Date	Team	TX	TN	NC	NCLL	TX	TN	NC	NCLL	TX	TN	NC	NCLL
Mon, Sep 29	1A	8	9.5	8	9.5	8.4	9.5	8.0	9.5	6.0	6.2	0.0	6.0
	1B	0	0	8	9.5	0.0	0.0	8.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	12	0	0.0	9.5	13.0	0.0	0.0	6.1	13.0	0.0
Tue, Sep 30	1A	8	9.5	10.5	9.5	9.6	9.5	11.0	9.5	6.0	6.0	10.5	6.0
	1B	0	0	8	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	6.0	0.0	0.0
Wed, Oct 01	1A	8	9.5	10	9.5	6.0	9.5	11.0	9.5	6.0	6.0	6.0	6.0
	1B	0	0	8	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	6.0	0.0	0.0
Thu, Oct 02	1A	8	9.5	11	9.5	9.5	9.5	11.0	9.5	6.0	6.0	6.0	6.0
	1B	0	0	9.5	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	6.0	0.0	0.0
Fri, Oct 03	1A	8	9.5	11	9.5	9.5	9.5	11.0	9.5	6.0	6.0	6.0	6.0
	1B	0	0	9.5	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	6.0	0.0	0.0
Sat, Oct 04	1A	0	0	0	9.5	0.0	0.0	0.0	9.5	0.0	0.0	0.0	0.0
	1B	0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2	0	0	12	0	0.0	0.0	12.0	0.0	0.0	0.0	10.0	0.0
Sun, Oct 05	1A	0	0	0	9.5	0.0	0.0	0.0	9.5	0.0	0.0	0.0	0.0
	1B	0	0	0	9.5	0.0	0.0	0.0	9.5	0.0	0.0	0.0	0.0
	2	0	0	12	0	0.0	0.0	12.0	0.0	0.0	0.0	10.0	0.0
Mon, Oct 06	1A	8	9.5	0	9.5	7.1	9.5	0.0	9.5	6.0	6.0	0.0	6.0
	1B	0	0	9.5	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	12	0	0.0	9.5	11.0	0.0	0.0	6.0	10.0	0.0
Tue, Oct 07	1A	8	9.5	11	9.5	8.0	9.5	11.0	9.5	6.0	6.0	6.0	6.0
	1B	0	0	9.5	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	7.3	0.0	0.0
Wed, Oct 08	1A	8	9.5	10	9.5	6.0	9.5	10.0	9.5	6.0	6.0	6.0	6.0
	1B	0	0	9.5	9.5	0.0	0.0	9.5	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	6.0	0.0	0.0
Thu, Oct 09	1A	8	9.5	11	9.5	8.0	9.5	9.0	9.5	6.0	6.0	6.0	6.0
	1B	0	0	9.5	9.5	0.0	0.0	6.0	9.5	0.0	0.0	6.0	6.0
	2	0	9.5	0	0	0.0	9.5	0.0	0.0	0.0	6.0	0.0	0.0

(a) (b) (c) (d)

Figure 5-4: Optimal shift lengths for each factory and shift on September 29th, 2008 with NCLL enabled and 100% of forecast, when labor is restricted to no flexibility (flex), some flex, and full flex.

Forecast	Labor Flex.	Labor Cost	Shipping Cost	Qty Late	Parts Short
100%	Full Flex.	$239,562	$3,592	-6	-11
100%	Some Flex.	$17	$4,778	0	15
100%	Inflexible	$(293)	$4,778	0	37
120%	Inflexible	$(373)	$4,851	422	36
150%	Inflexible	$(647)	$4,490	4913	45
150%	Full Flex.	$239,665	$(262)	4913	-8

Table 5.1: Average daily difference between actual and optimal quantities from September 7th to October 14th, 2008, with NCLL enabled for several scenarios with varying labor flexibility (Flex.) and forecasted demand.

unspecified aspects of scenarios. The firm did not transfer any parts that we tracked nor did the model suggest parts transfers in any scenario on any instance; hence we do not display transfer costs.

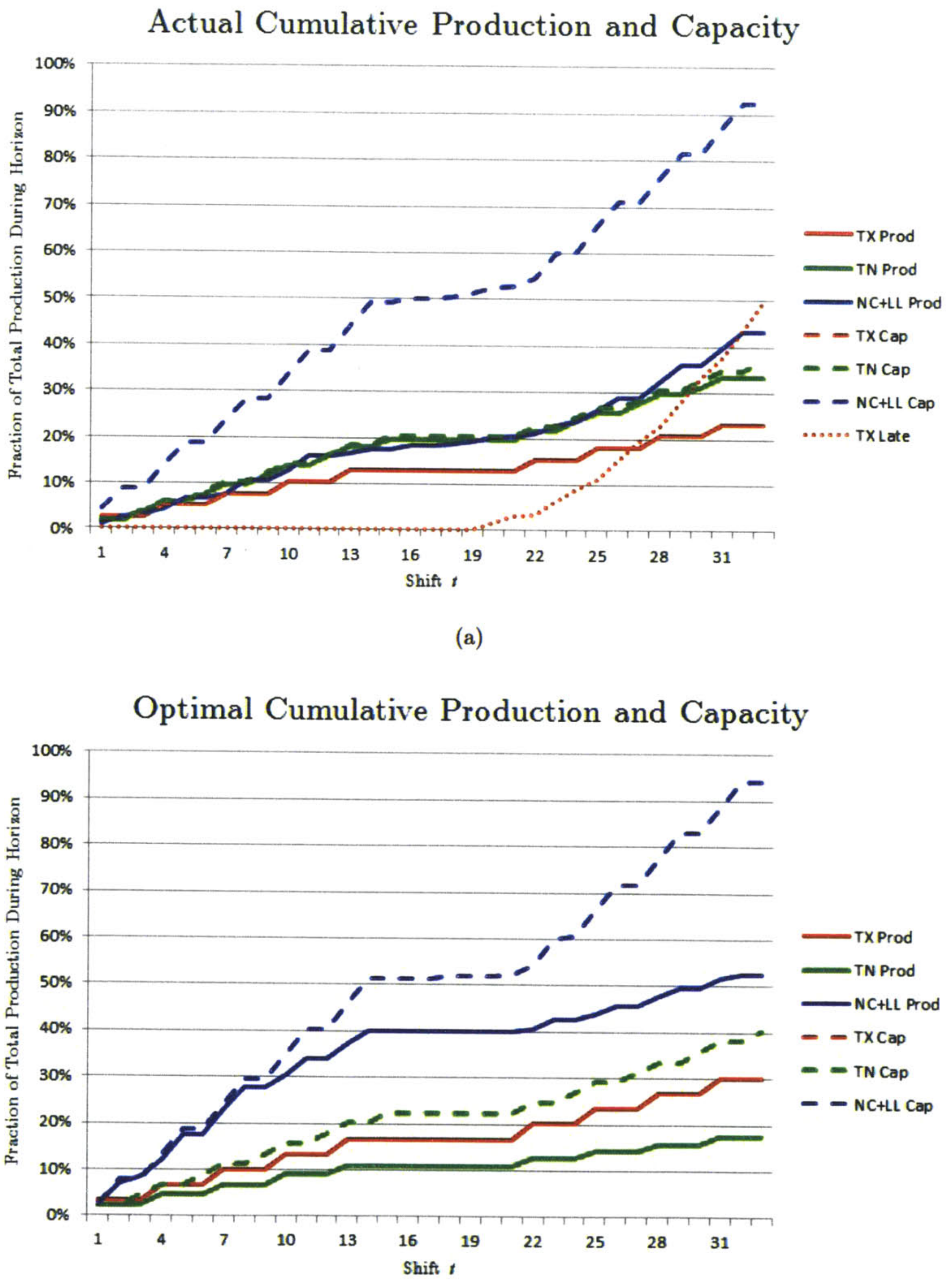
Not indicated in the table are baseline figures: typical labor costs are almost two million dollars; typical shipping costs are a few hundred-thousand dollars; the optimal solution almost always had zero late orders; every solution had several thousand parts short.

The most striking figure is labor cost. With full flexibility, the model suggests that several hundred thousand dollars in cost savings is possible; the firm was skeptical that such drastic changes in the labor force were possible, and hence suggested focusing on less flexible labor. As flexibility decreases, the labor costs for actual and optimal solutions become more similar and the optimal solution begins to focus on shipping cost savings.

Several thousands of dollars per day can be saved in shipping costs by simply producing desktops in different factories. This is in stark contrast to the results of §4.8 and §6.6 which indicate that the historical (and greedy) policies came close to minimizing the outbound shipping cost. Those policies are our best representation of what the firm planned to do; however, their actual actions differed at the time of execution, as seen in the results of this section. The firm moved orders from their default factories to re-balance factory loads; given that labor costs are significantly higher than shipping costs, this makes sense. However, the firm's manual choice of which orders to move to which factories left several thousand dollars per day in improvement potential; optimization techniques offer the opportunity to capture these savings by dynamically analyzing the network's cost structure in relation to factory loads.

When demand rises, costs rise marginally, cost savings potential increase almost proportionally to demand, and most metrics behave as they do in the 100% demand scenario. However, the number of late desktops rises swiftly for the "actual" solution because its production capacity is not adaptive; this does not indicate that the firm cannot respond to rising demand. Nonetheless, it does illustrate that an optimized solution can respond well to upward variations in demand. Figure 5-5 graphs the cumulative production and capacity (as a fraction of total production) at each factory over the horizon for the 150% demand scenario with full flexibility, which corresponds to the last row of Table 5.1. Both graphs indicate excess production capacity at NC, near the third and fourth day of the horizon. Sub-figure 5-5a shows further excess capacity at NC at the start of the horizon for the actual solution. It also shows the firm's actual production matching labor capacity at TX and TN throughout the horizon, with a sharp increase in late orders, which were assigned to TX, starting seven days (nineteen shifts) into the horizon. Even though NC is producing near its full capacity at this time, TX has much more urgent orders to satisfy. The optimal solution avoids this by moving many orders from TX and TN to NC to use NC's excess capacity early in the horizon, before orders become late. Responding early to forecasts indicating that a factory will not be able to produce all of its demand allows for more opportunities to reduce order lateness and mitigate costs. Even if the optimal solution did not respond to this imbalance before orders became past-due, it would immediate re-prioritize producing TX's late orders before not-yet-due orders at other factories. This indicates that the proportion of late orders at a factory, in addition to the ATB to capacity ratio, is a good indicator of imbalance in factory loads.

The most representative scenario in Table 5.1 is the third row, with 100% forecast and inflexible labor. The optimal solution spends several hundred dollars more each day on labor while saving several thousand dollars in shipping costs, while building


Actual Cumulative Production and Capacity

(a)

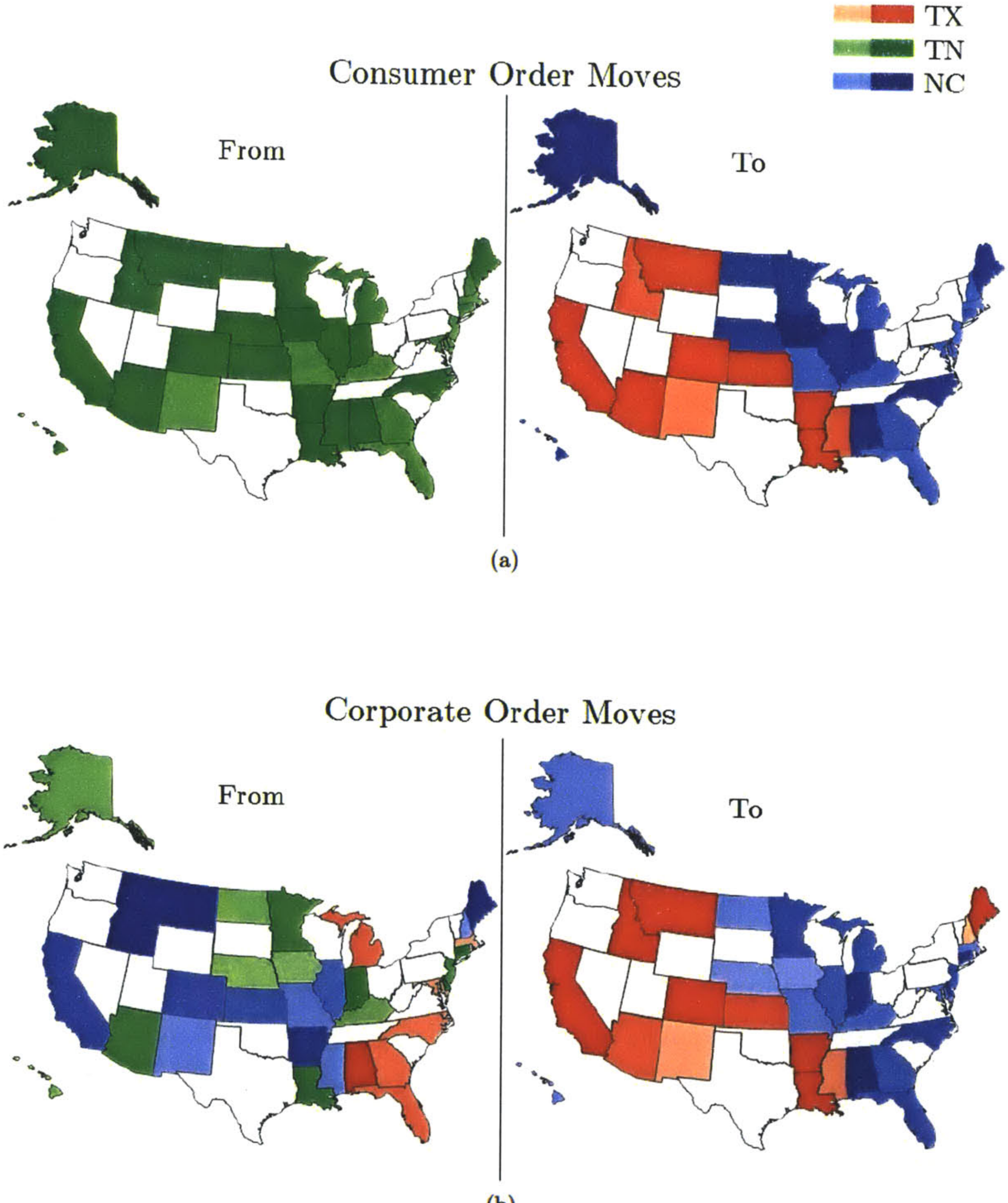
Optimal Cumulative Production and Capacity

(b)

Figure 5-5: Cumulative production and labor capacity (as a fraction of total production during the horizon) at each factory over the horizon starting on September 29th, 2008 with 150% demand and NCLL enabled, for the firm's actual solution (a) and the optimal solution with full labor flexibility (b).

orders on time and avoiding a few more part shortages. In developing the model for use, we repeatedly found that the model offered reductions in shipping costs, order lateness, or part shortages, while maintaining similar labor levels. Order lateness and part shortage reductions occur in a manner similar to that of Figure 5-5, where orders are moved to match production with capacity or part consumption with part availability.

Figure 5-6 illustrates the order moves suggested by the model. It suggests moving many orders for consumer desktops across the United States from TN to TX for more western states and to NC for more eastern states. For corporate desktops, which had significantly higher volume at the time, most order moves were made for orders that were previously assigned to TN or were assigned to a factory on the other side of the United States. Given that the objective function depends explicitly on these assignments, they are the major driver of several thousand dollars in daily shipping cost savings potential. Table 5.2 gives the total (over the horizon for the same scenario) outbound shipping cost, for orders that were in ATB at the start of the horizon, broken down by factory and line of business, depicting the source of cost savings just described. The firm typically moves orders in the "right direction" but the model suggests that additional order moves can significantly improve the objective. In this case, the model suggests moving a large portion of consumer desktops from TN to NC; the firm did this, but not to the extent that the model suggested. Furthermore, the model re-balances the distribution of corporate desktops to exploit idiosyncrasies in the outbound cost structure; shipping costs are not always proportional to the distance between factories and destinations; third-party logistics providers often set contract prices according to their infrastructure. For example, in Figure 5-6, the model suggests moving orders for corporate desktops destined for Maine and New Hampshire, the north-eastern most U.S. states, from NC to TX; at the same time, it suggests moving orders destined for Nebraska and North Dakota, states directly north of TX in Texas, to NC in North-Carolina. Orders that were moved in the past to mitigate imbalances in factory loads can be returned to their original cheaper location if factory loads even out. Similar savings were seen in all scenarios.



(a)

(b)

Figure 5-6: Optimal solution's suggested order moves for 100% demand, inflexible labor, NCLL Enabled, starting on September 29th, 2008. Each factory has a color; each state is colored to match the factory that had the most orders moved from or to it from that state. More intense colors indicate higher portions of that states desktops being moved.

Solution	Line of Business	TX	TN	NC	Total
Actual	Consumer	$3,885	$66,414	$26,796	$97,095
Actual	Corporate	$150,326	$64,899	$175,539	$390,765
Actual	Total	$154,211	$ 131,314	$ 202,335	$ 487,860
Optimal	Consumer	$ 31,308	$ 3,918	$ 58,946	$ 94,173
Optimal	Corporate	$ 155,566	$ 1,994	$ 197,623	$ 355,183
Optimal	Total	$ 186,875	$ 5,911	$ 256,570	$ 449,357

Table 5.2: Outbound shipping cost of already known orders for Actual and Optimal policies, by factory and line of business, for the horizon starting on September 29th, 2008.

5.3.2 Results from Spring 2009

From April 16th 2009 to May 6th 2009, we performed another study similar to the one in §5.3.1. During almost every week-day, the firm's data files were updated, our software prototype retrieved relevant input data, an Operations Center employee updated a few relevant parameters, the MILP was repeatedly solved under a few varying conditions, and model input, model output, and actual decisions enacted by the firm were saved. We then studied these results and analyzed the potential impact of the model.

This study was based on the full history of relevant, live, data from the firm. This includes daily ATB snapshots of all outstanding orders, lookahead forecasts with plans for the next weeks, all order moves made by the firm, the actual length of every shift at each factory, up-to-date outbound shipping costs and lead times, updated factory bottleneck and productivity data, and the prevailing shift structure (labor constraint data). At the time, part shortage lookahead files which contain data on parts availability were unavailable; however, the study in §5.3.1 does include analysis of parts.

During this time period, only the North Carolina (NC) and Juarez, Mexico (JM) factories were assembling desktop computers. The firm still had control over which computers were built in each factory, though it had little control over labor decisions at JM, instead paying proportional to the number of desktops produced. The results and insights gained by this new and different setting for implementation are still

	Quantity Late	Shipping Cost	Labor Cost
Historical	1579	$230K	$5006K
Optimal	183	$169K	$4849K
Hist - Opt	1396	$61K	$157K

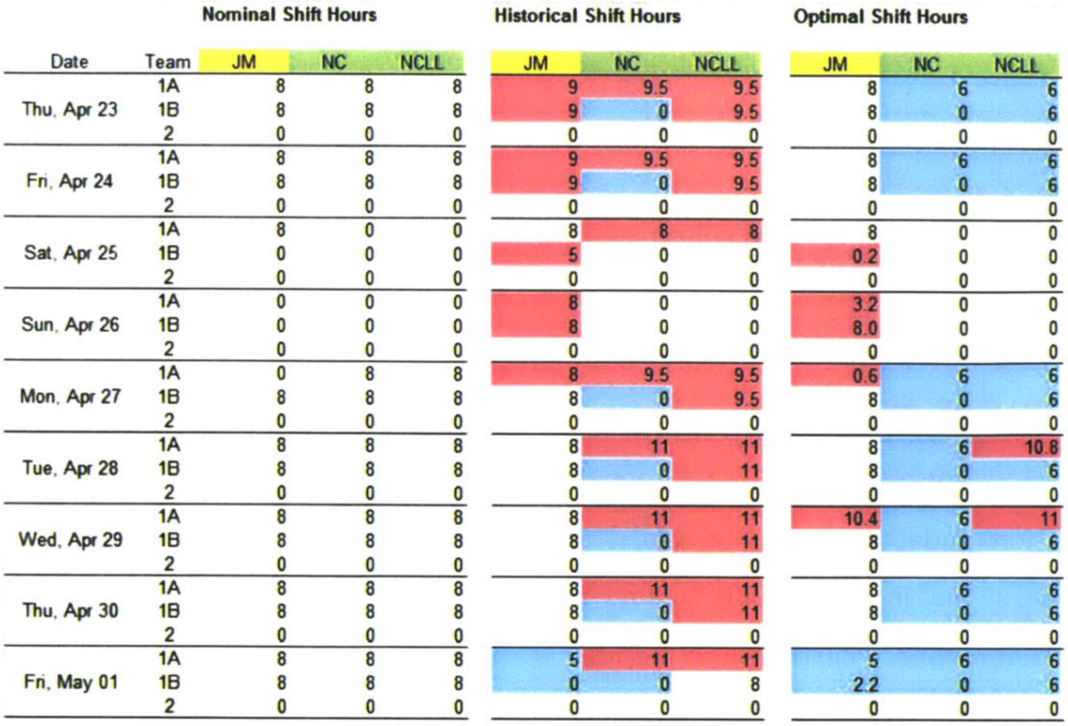
Table 5.3: Average daily quantity of late desktops, shipping cost, and labor cost for the historical and optimal solutions in the Fall of 2009.

relevant to many build-to-order network settings.

Other than having had discussions with us to develop the model, the employees responsible for order moves, parts transshipment, and labor scheduling enacted decisions independent of the model. The firm's decisions were not the same as the model's, making it difficult to find parameters to evaluate the model in a rolling horizon manner. Furthermore, because many orders are produced between the daily ATB snapshots, it is difficult to reconstruct the details of every order. However, the model's data was updated daily after having implemented the firm's decisions the previous day; using this data, we had knowledge of the volume of daily demand by Line of Business. We use this data to compare the model's optimal solution to what the firm actually did and the corresponding costs over a nine to fourteen day horizon.

A quantitative summary at the most aggregate level is given in Table 5.3. It gives the daily average quantity of late desktops, shipping cost, and labor cost for the firm's empirical solution and the model's optimal solution. Again, the firm did not transfer any parts that we tracked nor did the model suggest parts transfers in any scenario on any instance; hence we do not display transfer costs. Large sunk costs from previous shifts and minimum shift lengths that must be included in the labor cost calculations make labor costs appear disproportionately large. However, given that Dhalla [Dha08] suggests that each day a desktop is late costs about $100, labor cost savings potential tends to be about the same order of financial impact as cost savings or delivering more desktops on time. The model suggests up to a 75% reduction in outbound shipping costs worth $61,000 each day, about triple that in potential labor savings, and a significant reduction in the number of late orders.

Figure 5-7 illustrates the shift-lengths that the firm had planned (nominal hours),



		Nominal Shift Hours			Historical Shift Hours			Optimal Shift Hours		
Date	Team	JM	NC	NCLL	JM	NC	NCLL	JM	NC	NCLL
Thu, Apr 23	1A	8	8	8	9	9.5	9.5	8	6	6
	1B	8	8	8	9	0	9.5	8	0	6
	2	0	0	0	0	0	0	0	0	0
Fri, Apr 24	1A	8	8	8	9	9.5	9.5	8	6	6
	1B	8	8	8	9	0	9.5	8	0	6
	2	0	0	0	0	0	0	0	0	0
Sat, Apr 25	1A	8	0	0	8	8	8	8	0	0
	1B	0	0	0	5	0	0	0.2	0	0
	2	0	0	0	0	0	0	0	0	0
Sun, Apr 26	1A	0	0	0	8	0	0	3.2	0	0
	1B	0	0	0	8	0	0	8.0	0	0
	2	0	0	0	0	0	0	0	0	0
Mon, Apr 27	1A	0	8	8	8	9.5	9.5	0.6	6	6
	1B	8	8	8	8	0	9.5	8	0	6
	2	0	0	0	0	0	0	0	0	0
Tue, Apr 28	1A	8	8	8	8	11	11	8	6	10.8
	1B	8	8	8	8	0	11	8	0	6
	2	0	0	0	0	0	0	0	0	0
Wed, Apr 29	1A	8	8	8	8	11	11	10.4	6	11
	1B	8	8	8	8	0	11	8	0	6
	2	0	0	0	0	0	0	0	0	0
Thu, Apr 30	1A	8	8	8	8	11	11	8	6	6
	1B	8	8	8	8	0	11	8	0	6
	2	0	0	0	0	0	0	0	0	0
Fri, May 01	1A	8	8	8	5	11	11	5	6	6
	1B	8	8	8	0	0	8	2.2	0	6
	2	0	0	0	0	0	0	0	0	0

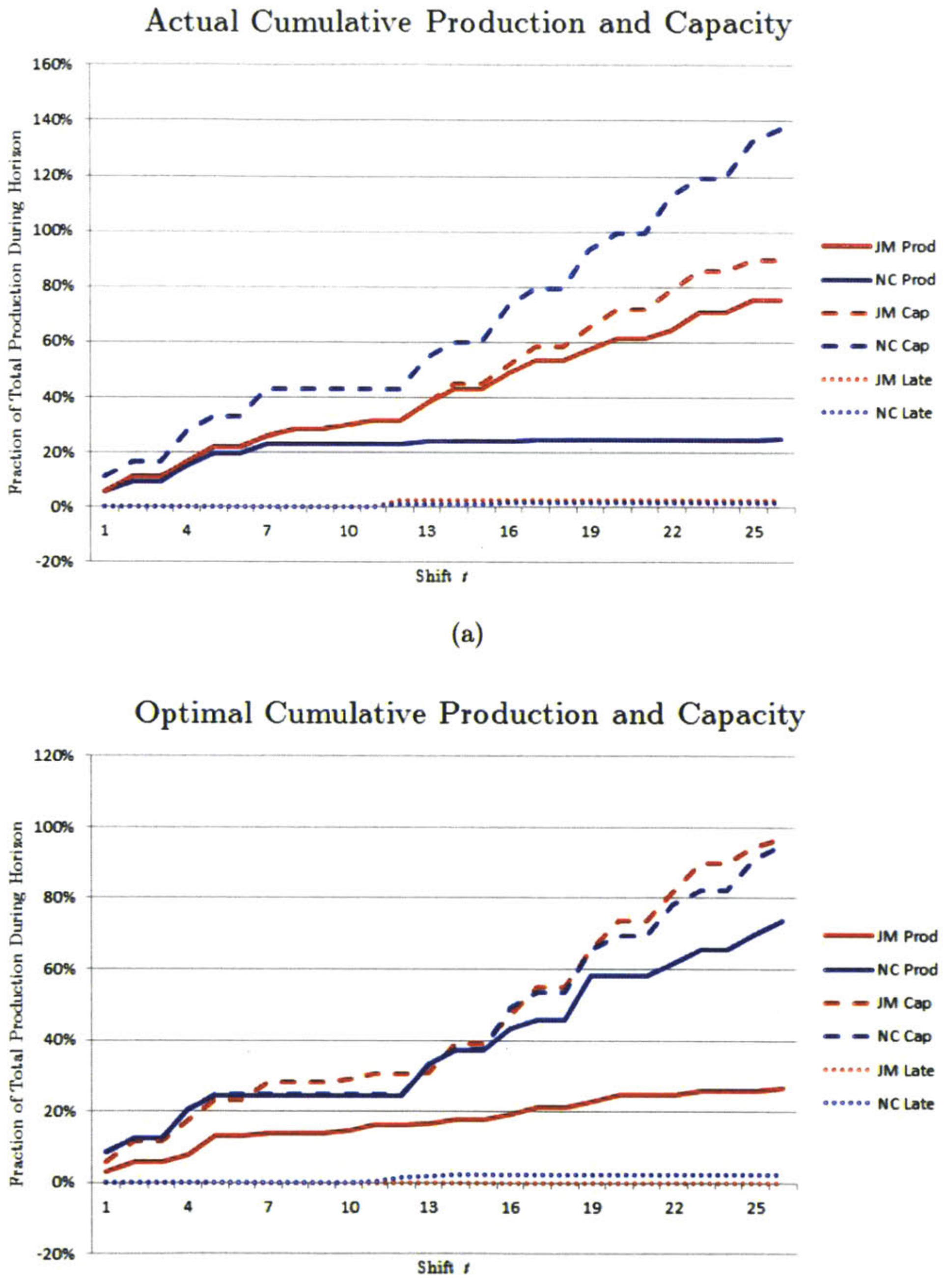
Figure 5-7: Planned, actual (historical), and optimal shift lengths for each factory and shift for the horizon starting on April 23rd, 2009.

the actual or historical shift lengths, and the optimal solutions' shift lengths. At this time, each of NC, NCLL, and JM were to have two eight-hour shifts on weekdays. Values are colored if they deviate from the nominal shift-length in sub-figure 5-3b by at least half an hour, pink if the shift is extended and light-blue if it is shortened. Due to the results in §5.3.1, the optimal solution here is for the scenario of some labor flexibility, where each work-team had to receive at least their total nominal (planned) hours over the horizon. Historically, the firm cut the second shift at NC but extended almost every other shift at NC or NCLL and extended some shifts at JM. The model optimum extended shifts early in the horizon to avoid order lateness and reduced the length of shifts near the end of the horizon. At other starting dates, we saw the optimal solution reverse this, conserving labor now when expecting a future lull in demand.

By re-arranging the timing and location of labor, the optimal solution matches capacity and production with demand. This is illustrated in Figure 5-8, which depicts the cumulative production and capacity (as a fraction of total production) at each factory over the horizon starting on April 23rd, 2009 for (a) the firm's historical solution and (b) the optimal solution. The firm stopped producing at NC once JM could handle all production. This reflects the firm's policy at the time to outsource as much as possible to JM, using NC to handle 1) demand that JM could not, 2) more complex products that required more skilled labor, and 3) non-geo-eligible orders, such as those for the U.S. government. They did so, in part, because the firm believed that the labor cost was less expensive at JM and the firm had plans to outsource all North American desktop production over the next few years; reducing reliance on NC at this time eased that transition. Our analysis suggested that the two factories had similar labor costs, which made shipping costs and delivery lead time relevant. Hence, the model moved many orders from JM to NC, as can be seen in Figures 5-8b and 5-9, where NC capacity is maximally utilized at the start of the horizon, to reduce the number of late orders from 3.9% to 2.3% of total production during the horizon and reduce costs by $343K per day. This is an example of a solution that did not meet unexpressed executive concerns. Nonetheless, it does provide insight into the cost of such executive decisions and illustrate the model's ability to adapt to new network configurations. The model moves orders to reduce outbound shipping costs and deliver orders on-time. It also re-arranges the timing and amount of labor to improve throughput.

5.3.3 Conclusions

In order to solve the execution problem §2.2.2, we implemented a more detailed and discrete version of the rolling-horizon certainty-equivalent linear program analyzed in Chapter 4. Individual orders and their already known details are distinguished from future forecasted orders. To model labor, we introduced overlapping shifts, overtime, and lean lines at NC. Parts transfers were included. Practical concerns largely regarding data acquisition, model usability, and infeasibility were addressed. After several iterations with the firm's Operations Center employees, the model was solved daily using live data in two time periods, providing insights into how the firm



Actual Cumulative Production and Capacity

(a)

Optimal Cumulative Production and Capacity

(b)

Figure 5-8: Cumulative production and capacity (as a fraction of total production) at each factory over the horizon starting on April 23rd, 2009.

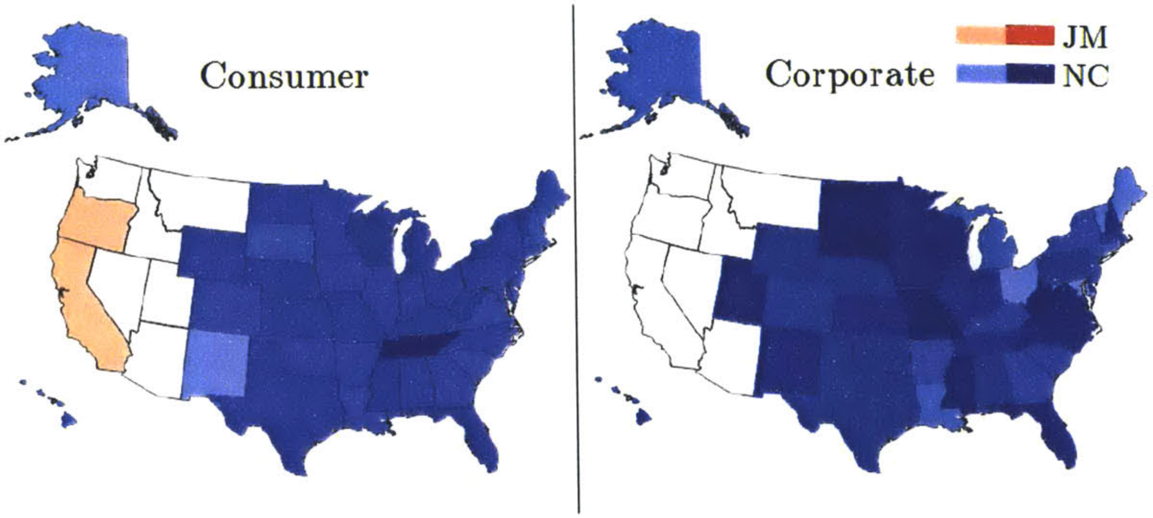


Figure 5-9: Optimal solution's suggested order moves when the horizon starts on April 23rd, 2009. Each factory has a color; each state is colored to match the factory that had the most orders moved to it from that state. More intense colors indicate higher portions of that state's desktops being moved.

did and should act.

The firm's static map allocates orders to factories based solely on the shipping destination. The Operations Center improves this by executing order moves based on destination, due-dates, and the amount of ATB relative to capacity. However, the order moves that the firm made left room for several thousand dollars in further daily cost savings.

The results above, such as the Inflexible 100% scenario in Table 5.1, suggest that significant improvements can be made by distinguishing between different products, even if only two categories (Consumer and Corporate) are used. Exploiting the different production capacities of each factory, as described in §2.3, allows for more production in the same amount of time. This is re-iterated in the results for the planning model in Chapter 6. Further shipping cost savings can be captured by accounting for idiosyncrasies in the shipping cost structure.

Although parts transfers were not observed in the solutions we observed in the instances we analyzed and a more detailed investigation of alternate parts deployment techniques are beyond the scope of this project, this work shows that the parts transfers we modeled were less cost effective than moving the orders to the parts or delaying production and adjusting capacity appropriately, as seen under Parts Short in Table 5.1.

The most significant source of cost savings potential was labor cost, as capacity exceeded demand in most scenarios. This is in part because desktop sales were declining as notebook computers became more popular. Both the model and the firm responded to this change in demand by adjusting the labor force, the firm by discontinuing its own production and the model by reducing shift lengths where possible. Although management was not comfortable with the drastic reductions in shift lengths that the model suggested, additional constraints on the minimum amount of labor can help strike an acceptable balance. Even with no labor flexibility, the model still provides significant cost savings.

The results also suggest that accounting for order lateness, not just capacity utilization, can significantly improve customer service in a network with multiple manufacturing facilities, especially if predicted in advance. Satisfying orders from a more expensive factory can provide future cost savings if the inexpensive factory becomes over-loaded.

As in Chapter 4, the model in this chapter shows cost saving potential of several hundred thousand dollars per day. Table 4.12 indicates the recommended solution outperforms the historical one by \$47M on average over 91 days, or \$516K per day. Table 5.1 shows \$243K in daily cost savings and Table 5.3 shows \$218K in daily cost savings, with further cost savings via reduction of order lateness and part shortages. Because two different models, using several different parameter estimation techniques and analyzing different periods of time, report similar estimates of potential financial impact, these results seem very consistent.

Using mathematical programming, we balance several competing objective simultaneously and at a much more detailed level than simple heuristics and human oversight. Not only does our model account for shipping costs like the firm's static map, but it also accounts for differences in labor costs and customer service. Repeated re-optimization allows one to incorporate new information much more quickly and easily adapt to changes in circumstances. Mathematical optimization can exploit peculiarities in both the shipping and labor cost structure as well as delivery times. The impact of delaying or expediting production and its impact on shift lengths and the timely satisfaction of customer orders is much more readily apparent. These criteria are much easier to evaluate and adapt to in a mathematical optimization environment, which revealed that re-scheduling labor and moving orders to match capacity has potential for several thousand dollars in daily cost savings while improving customer service by delivering more orders on time.

Chapter 6

Planning Problem: Formulation, Implementation, and Analysis

This chapter discusses mathematical optimization models for the firm's planning problem. In §2.2.1, we give a detailed, qualitative description of this problem. One purpose of these models was to understand the effectiveness of and improve the firm's default download rules that allocated orders to factories. Another purpose of developing them was to inform other decision-making groups within the firm, such as labor management and parts sourcing teams within the firm, of plans for production across the manufacturing network, helping them make their decisions. With a typical horizon of a quarter year and time discretized into weeks, the primary decisions are allocation of groups of orders to factories, production of the desktops in those orders, the amount of labor to hire, and how much time that labor should spend in the factory.

We begin by describing two models that were developed as planning software prototypes in collaboration with and for use by the firm. We first formulate a deterministic MILP based on the firm's point forecasts for demand, which we call the *nominal model*, in §6.1. Whereas the execution problem had plenty of data on daily demand for desktops, data on demand for the planning problem's larger temporal scope was scarce. To address concerns about the model's sensitivity to demand, which arose largely because the model suggested significantly less labor capacity was necessary to maintain adequate production and customer service levels, we introduce

Robust Optimization techniques in §6.2, formulating what we call the *robust model.* Challenges such as maintaining tractability and modeling uncertainty sets for Robust Optimization are addressed while modeling the problem. We extend both of these models to the case of multiple products in §6.3. We discuss details of implementing the MILP in §6.4. The methodology for validating model accuracy and evaluating solutions is presented in §6.5. Numerical results and managerial insights showing significant cost savings potential acquired from using the models are given in §6.6.

Input data will be denoted by upper-case letters while decisions will be denoted by lower-case letters, excepting t, d, l, k, and s, which are indices whose capital letters are the cardinality of the index set. We consider a planning horizon T, typically one quarter of a year, with time discretized into weeks which we index by t. The firm's three factories are indexed by l, the demand destinations are indexed by d, and the firm's Lines of Business are indexed by k. Index s is used for further differentiation of some sets of constraints or variables.

6.1 Nominal Formulation

In the planning model, the objective is to minimize the sum of inbound shipping costs, outbound shipping costs, and direct labor costs. The relevant decisions are allocating and producing demand at factories and planning permanent and temporary staffing levels, including planning for overtime. The major constraints are that demand must be satisfied by the end of the horizon in a build-to-order manner, labor restrictions, and the resulting capacity at each factory.

The firm has a forecast of future demand $F_{t,d}$ for each period t and destination d, which this model treats as being the true future demand. The plan allocates this forecasted demand among factories, represented by the decision $x_{t,d,l} \in \{0,1\}$ or its relaxation $x_{t,d,l} \in [0,1]$, with $\sum_l x_{t,d,l} = 1 \forall t, d$. The plan includes non-negative production decisions $y_{t,d,l}$ for each period t, destination d, and factory l. Because of the firm's build-to-order business model, cumulative production must not exceed the cumulative demand to date. Because the firm fulfills all outstanding orders by the end of each fiscal quarter, which is the typical planning horizon, demand is reset to zero at the start of the horizon and cumulative production must equal cumulative demand by the end of the horizon; this termination criteria forces the model to produce desktops promptly. An inbound shipping and outbound shipping cost of $C_{d,l}$ dollars is incurred for every unit shipped from factory l to destination d.

Staffing decisions regarding the amount of permanent labor p_l and temporary labor $r_{t,l}$ are made for each factory l, measured in units-per-factory-hour, i.e. the number of desktops that can be produced for each hour that the staff works in the factory. Permanent labor takes several weeks to recruit and train, making it difficult to change the permanent labor level during the horizon; however, because we are planning in advance, permanent labor levels can be adjusted once before the quarter begins. The temporary workforce is more flexible and can be adjusted with a few days notice, making it a weekly decision. For quality assurance purposes, the temporary labor force cannot exceed M_l percent of the permanent labor force. Both types of labor work S_l "straight" factory-hours every week (a fixed quantity determined by the factories staffing structure) and $o_{t,l}$ overtime factory-hours, a decision to extend the amount of time they work for additional pay. Each *labor-hour* provides approximately U_l (units-per-labor-hour) production capacity per period, regardless of labor type, but costs $C_l^{rs} \leq C_l^{ro} \leq C_l^{pS} \leq C_l^{oS}$ where the superscripts denote the type of labor, which is either permanent (p) or temporary (r), and either straight-time (S) or overtime (o). Furthermore, each factory produces at most A_l units-per-factory-hour due to physical bottlenecks within the factory, as described in §2.3.

Formulating the above, we have the following *nominal model*:

$$\text{minimize:} \sum_{t,d,l} C_{d,l} y_{t,d,l} + \sum_{t,l} \frac{1}{U_l} \left(\begin{bmatrix} S_l & o_{t,l} \end{bmatrix} \begin{bmatrix} C_l^{pS} & C_l^{rS} \\ C_l^{po} & C_l^{ro} \end{bmatrix} \begin{bmatrix} p_l \\ r_{t,l} \end{bmatrix} \right) \tag{6.1}$$

subject to:

$$\sum_{l} x_{t,d,l} = 1 \qquad \forall t, d \tag{6.2}$$

$$\sum_{\tau=1}^{t} y_{\tau,d,l} \leq \sum_{\tau=1}^{t} F_{\tau,d} x_{\tau,d,l} \qquad \forall t, d, l \tag{6.3}$$

$$\sum_{\tau=1}^{T} y_{\tau,d,l} = \sum_{\tau=1}^{T} F_{\tau,d} x_{\tau,d,l} \qquad \forall d, l \tag{6.4}$$

$$r_{t,l} \leq M_l p_l \qquad \forall t, l \tag{6.5}$$

$$\sum_{d} y_{t,d,l} \leq (p_l + r_{t,l})(S_l + o_{t,l}) \qquad \forall t, l \tag{6.6}$$

$$\sum_{d} y_{t,d,l} \leq A_l(S_l + o_{t,l}) \qquad \forall t, l \tag{6.7}$$

$$x,\ y,\ o,\ p,\ r \geq 0$$

The first term of the objective (6.1) accounts for shipping costs and the second term captures the four different types of labor costs. Equation (6.2) ensures that demand for each destination d and week t is allocated to a factory. Inequality (6.3) prevents the model from building desktops before customers demand them; however, the termination criteria, equality (6.4) ensures that (6.3) holds with equality for $t = T$ forcing all demand to be satisfied by the end of the horizon. Inequality (6.5) provides the quality assurance upper bound on the temporary labor force. Inequalities (6.6) and (6.7) define the production capacity in terms of labor productivity and physical bottlenecks, respectively. All of the decisions are non-negative.

At the beginning of this study, the firm's information technology and management allocated orders to factories in large geo-graphic groups, the smallest being U.S. states; as such the binary constraints $x \in \{0,1\}^{T|D|L}$ were necessary for implementation purposes. Binary constraints are later used to tractably model the structure of the labor force, making the binary allocation constraints less of an additional burden on solving the problem. However, later in the study, the firm's software could handle more granular commands, such as filters by zip code or product families. Demand could then be distributed among factories according to a fractional x by allocating a proportional number of zip codes (weighted by their relative demands) within that destination without significant cost impact. Allowing x to be fractional also permits us to eliminate x and re-write many constraints in a much more convenient format that enables the Affinely Adjustable Robust Counterpart technique to capture the firm's supply chain flexibility in §6.2.3. We present the formulation in terms of x here because it is useful for interpreting the problem, solutions, and results and it also exposes difficulties in solving the problem and some necessary transformations to make Robust Optimization formulations feasible.

Production plans must be developed within a few days and may require running the model under multiple instances to understand implications of different user-input parameters. Note that the above formulation is not linear; in particular, the objective function (6.1) and the labor-capacity constraint (6.6) contain the product of $o_{t,l}$ and $(p_l + r_{t,l})$. Because the firm's management and information technology required that the components of x be binary for implementation of the model, the problem becomes a non-linear integer program which can be difficult to solve quickly without additional structure. Because CPLEX 10.1.1 did not return satisfactory solutions in an acceptable amount of time under a variety of options settings, we exploit the structure of the firm's labor force to develop a more tractable mixed integer linear program.

Production gradually rises over the course of a quarter due to end-of-quarter sales and the firm's requirement that all orders are satisfied by the end of the quarter, which is referred to as "The Hockey Stick Effect." The firm's production capacity follows a similar pattern, as illustrated in Figure 6-1 which displays the firm's historical production capacity $(p_l + r_{t,l})(S_l + o_{t,l})$ at TX, TN, and NC during the Fall quarter (Q3) of 2007. Theoretically, time can be divided into three time intervals, which can shift from quarter to quarter: 1) when only permanent straight-time labor is used (Baseline), 2) when temporary labor is "ramped up" or when permanent and temporary straight-time labor is used (Ramp-up), and 3) when permanent and temporary workers do overtime in addition to straight-time (Overtime). In the figure, it can be seen that each factory begins with below-average production capacity, ramps-

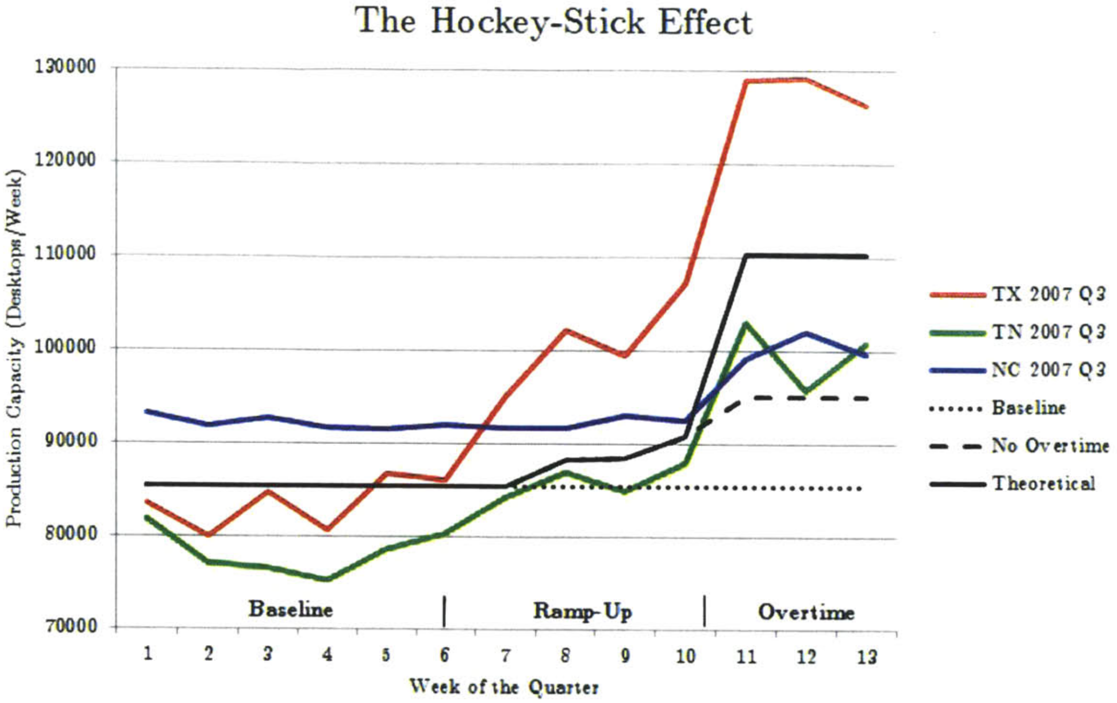


Figure 6-1: The Hockey Stick Effect - The historical production capacity at TX, TN, and NC during Q3 2007 and the model's theoretical production capacity are depicted rising throughout the quarter in three intervals.

up to higher production levels, and then stabilizes once near the maximum possible production capacity. Overtime tends to only occur after the temporary workforce is at its maximum $M_l p_l$. We re-structure the labor constraints into these three intervals.

At the beginning of the horizon, a permanent baseline workforce is chosen and temporary labor levels vary but overtime is not used. Near the end of the horizon, temporary labor is fixed to its maximum but overtime can vary. Because there are only three factories, a discrete search over a small set of baseline permanent labor levels $P_{l,s}$, where s indexes the options available at each factory l, is fairly tractable. Using this approach, either overtime is zero ($o_{t,l} = 0$) or the labor force is one of a few discrete values, i.e. there exists an s such that $(p_l + r_{t,l}) = P_{l,s}(1 + M_l)$. In order to enforce this logic, we introduce two new binary decision variables, $q_{l,s}$ and $w_{t,l}$, which respectively represent the choice of permanent labor level $P_{l,s}$ and the decision to use (1) or not use (0) overtime. We then transform the above formulation into an MILP by replacing the original objective (6.1) with

$$\text{minimize:} \qquad \sum_{t,d,l} C_{d,l} y_{t,d,l} + \sum_{t,l} \left[\frac{S_l}{U_l} (C_l^{pS} p_{t,l} + C_l^{rS} r_{t,l}) + v_{t,l} \right] \tag{6.8}$$

and inequality (6.6) with the following inequalities:

$$\sum_s q_{l,s} = 1 \qquad \forall l \tag{6.9}$$

$$p_l = \sum_s P_{l,s} q_{l,s} \qquad \forall l \tag{6.10}$$

$$r_{t,l} \geq M_l P_{l,s} (w_{t,l} + q_{l,s} - 1) \qquad \forall t, l, s \tag{6.11}$$

$$o_{t,l} \leq w_{t,l} O_l \qquad \forall t, l \tag{6.12}$$

$$\sum_d y_{t,d,l} \leq (1 + M_l)(P_{l,s}(S_l + o_{t,l}) + \overline{P}_l (S_l + O_l)(1 - q_{l,s})) \qquad \forall t, l, s \tag{6.13}$$

$$\sum_d y_{t,d,l} \leq (p_l + r_{t,l}) S_l + \overline{P}_l (1 + M_l) o_{t,l} \qquad \forall t, l \tag{6.14}$$

$$v_{t,l} \geq \frac{1}{U_l} (C_l^{p,OT} + C_l^{r,OT} M_l)(o_{t,l} P_{l,s} - (1 - q_{l,s}) O_l \bar{P}_l) \qquad \forall t, l, s \tag{6.15}$$

$$w_{t,l},\ q_{l,s} \in \{0, 1\}$$

$$v_{t,l} \geq 0.$$

Here, O_l is an upper bound on the amount of overtime that can be used and $\bar{P}_l = max_s\{P_{l,s}\}$ is the largest allowable permanent labor force; both are used in disjunctive "big-M" constraints to provide upper bounds when the constraint is not intended to be active. Equations (6.9) and (6.10) discretize the permanent labor decision. Inequalities (6.11) and (6.12) force temporary labor to be at its maximum or overtime to be zero. Inequalities (6.13) and (6.14) provide production capacity constraints when there is overtime and when there is not overtime, respectively. At optimality, inequality (6.15) forces auxiliary decision $v_{t,l}$ to be the cost of overtime at factory l during week t. Figure 6-1 also contains an illustration of how we have modeled production capacity theoretically, through the lines for 1) the (Baseline) permanent labor capacity $p_l S_l$, 2) the (No Overtime) capacity during the labor ramp-up (p_l +

$r_{t,l})S_l$, and 3) the (Theoretical) total production capacity $(p_l + r_{t,l})(S_l + o_{t,l})$.

In order to be useful to the firm, the model must address some managerial concerns described in 4.3.3, regarding fairness in balancing factory workloads. The firm preferred solutions that do not give overtime to some factories while under-utilizing others; enforcing this was not necessary because in every scenario that we analyzed it was optimal to do so. The firm also desired a limit on the average cycle-time from customer order to delivery (the total time spent in backlog, manufacturing, and shipping); however, these constraints were never active. Lastly, arrangements with the local government led to minimum staffing requirements at NC; we include these lower bounds on p_l in some of our analysis, providing interesting insights into the cost of having made those agreements.

"Non-geo-eligible" demand, which is described in 4.3.1, cannot be re-allocated to other factories. Sometimes, major corporate customers provide their own outbound transportation but will pick up their order from a particular factory. Alternatively, some rarer parts are not stocked at all locations and can only be fulfilled by one factory. Non-geo-eligible demand has its own forecast and production decisions but we fix its allocation decision. In a typical solution, non-geo-eligible demand consumes a fixed amount of capacity in the week it becomes known.

6.2 Robust Formulation

Although the formulation in §6.1 is an accurate representation of the firm's problem and its solutions suggested significant cost savings, as discussed in §6.6, the firm's management was concerned that it was overly reliant on the firm's forecasts $F_{t,d}$, which often varied significantly from the true demand. Make-to-order manufacturers often maintain excess or flexible production capacity to quickly meet variations in demand. Because the formulation is deterministic, its solutions suggest that the firm can meet demand with significantly less capacity than the firm historically used. In order to address this and protect the above MILP formulation against uncertainty in demand, we investigate approaches to incorporating demand uncertainty which can exploit the manufacturing network's flexibility by planning for factories to compensate for each other. Because demand data was available for less than three quarters of a year, we did not have sufficient data to employ techniques similar to the detailed demand model and simulation that we did for the execution problem in Chapter 4, which had a smaller temporal scope. The lack of demand data also makes stochastic optimization limited in its usefulness. Instead, we use methods that can protect against uncertainty with very limited demand information. In particular, we adopt Robust Optimization. Formulating the robust counterpart of a MILP that is already difficult to solve requires significantly more elaboration. We review the Robust Optimization literature in 6.2.1. We develop an uncertainty set that contains the demand values that our solution must be able to satisfy in §6.2.2. We then introduce an affinely adjustable policy that allows the nominal model to respond to fluctuations in demand and show how to incorporate this without making the problem size too large, in §6.2.3. Finally, we develop the robust counterpart to all other constraints in §6.2.4.

6.2.1 Robust Optimization Literature

Because of the large-scale nature of instances for our deterministic formulation, and because multi-period problems tend to allow for recourse after uncertainty is realized, we restrict ourselves to simple, tractable approaches to dealing with uncertainty. Dynamic Programming would suffer from the curse of dimensionality due to the large decision and state spaces. Though we could attempt a multi-stage Stochastic Program, sampling from the demand distribution and solving a large number of deterministic instances with the additional constraint that early stage-decisions required to be the same across instances, the sample space grows far too quickly, even for simple, crude approximations of the uncertainty. Furthermore, the lack of demand data at this scope makes determining a demand distribution difficult. Many other approaches have similar challenges. Robust Optimization seemed to be most tractable method available and provided a reasonable interpretation of uncertainty.

Robust Optimization (RO) addresses the problem

$$max_x\{c'x \,:\, Ax \leq 0 \; \forall A \in D\}$$

where $x \in \mathbb{R}^n$ and A is a matrix of uncertain coefficients that lie in some uncertainty set D. The major contrast between RO and Stochastic Programming (SP) is that RO addresses *hard* constraints that must be satisfied for any realization of the data (and hence limits itself to bounded uncertainty sets) whereas SP tends to employ *soft* constraints that can either be violated with an objective penalty cost (recourse) or with at most some desired probability (chance constraints).

Soyster [Soy73] presented the first RO approach which, along with other early works, was very conservative and hard to generalize. In the late 1990s, Ben-Tal and Nemirovsky [BTN99] [BTN98] as well as El Ghaoui, Oustry, and Lebret [EGL97] [EGOL+98], independently introduced many important RO formulations, results, and applications. The most notable result is that the robust counterpart (RC) of convex optimization problems that have ellipsoidal uncertainty sets can be formulated as optimization problems that are (approximately) tractable. Bertsimas and Sim [BS04] develop a robust approach for LPs and Integer Programs (IP) using linear constraint-wise uncertainty sets whose robust counterpart remains an LP or IP respectively. They introduce the notion of a "Budget of Uncertainty" that limits the number of parameters that can deviate from their nominal value.

Bertsimas, Pachamanova, and Sim [BPS04] elegantly characterize robust counterparts of LPs with uncertainty sets described by arbitrary norms, showing that they are convex optimization problems whose constraints are defined in terms of their dual norms. "The dual of the L_p norm

$$||\mathbf{x}||_p = (\sum_{j=1}^{n} |x_j|^p)^{1/p},$$

is the L_q norm $||\cdot||_q$ with $q = 1+\frac{1}{p-1}$." Most notably, $||\cdot||_\infty$ and $||\cdot||_1$ are dual to each other and the dual of $||\cdot||_2$ is itself. They compare the most popular special cases, in particular, the ellipsoidal ($||\cdot||_2$) uncertainty sets of Ben-Tal and Nemirovsky and those of El Ghaoui, Oustry, and Lebret against the D-Norm (a combination of $||\cdot||_1$and $||\cdot||_\infty$) of Bertsimas and Sim. [BPS04] also gives loose but very general probability guarantees against constraint violations for the above special cases. Importantly, having parameters depend on other parameters in multiple rows and columns can be modeled by the uncertainty set

$$D = \{A : ||M(vec(A) - vec(\bar{A}))|| \leq \Delta\}$$

where M is an invertible matrix and $vec(\bar{A})$ is a constant vector in $\mathbb{R}^{mn\times 1}$. However, the dual norm may be extremely difficult to optimize over and most practical implementations restrict their attention to row-wise uncertainty.

The RO papers above focus on what are typically referred to as single-stage programs, i.e. where none of the uncertainty is known before the decisions are made. Implementing them with a rolling horizon is akin to Open-Loop Feedback Control which can often be overly conservative. In many contexts, including ours, some decisions are made before any uncertain values are known, while other decisions are made after some uncertain values are known. Additionally, decisions such as slack and surplus variables, as well as, auxiliary variables used to transform piecewise-linear functions such as $max\{0, x_i\}$ or $|x_i|$ into linear functions, do not correspond to actual decisions and should be allowed to tune themselves to varying data. Multi-stage SP often addresses these issues but is not tractable for large-scale problems. The RO equivalent of multistage SP is called the Adjustable Robust Counterpart (ARC), wherein some variables can adjust to (depend on) the uncertainty. These are both akin to Closed-Loop Feedback Control. Ben-Tal, Goryashko, Guslitzer, and Nemirovsky [BTGGN04] propose the Affinely Adjustable Robust Counterpart (AARC) as a tractable approximation to the (ARC). The AARC relies on a notion from Control Theory, that of restricting the search to policies that depend affinely on the uncertain parameters, in order to gain tractability . It effectively allows for recourse decisions that are affine functions of the uncertain parameters. Chen, Sim and Sun [CSS07] compare RO against chance-constrained SP, including multi-stage problems and the AARC, and discuss why RO is often much more tractable and tends to satisfy soft probabilistic constraints well.

Ben-Tal, Golany, Nemirovski and Vial [BTGNV05] implement the AARC for a two-echelon multi-period supply chain problem known as the retailer-supplier flexible commitment using a min-max cost function and test the AARC's probability guarantees through simulation. Bertsimas and Thiele [BT04] apply the approach in [BS04] to a discrete-time inventory-network management problem with some of the typical extensions. They show that the robust solution is identical to the optimal solution (under some conditions) when the character of the optimal solution is known (via analysis of the Dynamic Program) and show that the robust approach is efficient to implement and performs well in scenarios when the optimal solution is not known. Bertsimas, Brown, and Caramanis [BBC10] surveys a large portion of the available Robust Optimization literature and results. It also has a large number of applications, including a large section on multi-stage problems that use the AARC developed in [BTGGN04]. Bertsimas, Iancu, and Parrilo [BIP10] prove the near-optimality of affinely adjustable policies in multi-stage optimization problems and provide an example of its use in inventory management. Although we are not aware of any other implementations of multi-period Robust Optimization in supply chain planning problems, the above papers suggest strong potential for tractably modeling the firm's unique planning problem.

6.2.2 Uncertainty Model

Constructing an uncertainty set for a Robust Optimization (RO) model can be difficult. Much of the literature either assumes that the set is known a priori or gives simple examples. For the planning problem, we restrict our attention to uncertainty in the demand vector F and investigate the interaction of its elements with each other. Furthermore, though it is not necessary beyond this section, we adopt a natural probabilistic interpretation of uncertainty to develop a bounded uncertainty set.

The firm's long-term sales forecasting group generates $\hat{F}_{t,d}$, a point-estimate of the number of systems that will be sold in each week t to customers at destination d, which we assume to be unbiased. The firm's historical production quantities per week for a few quarters were also available, which, when aggregated over factories, provide an estimate, $\sigma_{t,d}^2$, of actual demand variance for each destination.

Using uncertainty sets that only consider errors for individual forecasts $\hat{F}_{t,d}$ would not capture the impact of uncertainty on production capacity which deals with large subsets of demand; the robust counterpart would be equivalent to adjusting each component of the forecast upward. Instead, we consider "groups" $G \subset D$ of destinations, which are subsets of the set of all demand destinations D, to protect against; that is, our uncertainty sets will limit the total demand within a group G. Using groups of destinations allows us to make stronger guarantees regarding demand satisfaction without generating more conservative solutions because aggregate demand will have less variability than demand for individual destinations.

Uncertainty over time is treated more simply. Note that only constraints (6.3) and (6.4) involve more than one time period; because most orders are satisfied in the same week that they are made, i.e. the backlog of orders turns in less than a week, the problem is essentially decoupled across time periods, excepting the permanent labor decisions. Because the key constraints we wish to protect, those regarding production capacity such as (6.13) and (6.14), consider only one week at a time, we can ignore aggregation of uncertainty over time, as otherwise the robust counterparts would ignore uncertainty in other time periods anyways.

Consider a particular grouping G. Let $\tilde{F}_{t,G}$ be the random variable representing the total demand from group G during week t. For developing uncertainty sets, we assume that the demand $\tilde{F}$ is Normally distributed according to

$$\tilde{F}_{t,G} \sim N(\hat{F}_{t,G}, \sigma_{t,G}^2)$$

where a group's demand variance is defined by $\sigma_{t,G}^2 \triangleq \sum_{d \in G} \sigma_{t,d}^2$. This differs from our Log-Normal fit for demand in §4.5 because the scope is larger for a few reasons: 1) the Central Limit Theorem more readily applies to the sum of a large number of somewhat independent consumer decisions or the sum of each many days worth of demand; 2) the distribution parameters are sufficiently far from zero, making the Normal and Log-Normal decisions appear similar; 3) limited data availability makes estimates of the tail of the Log-Normal distribution inaccurate. Results for other reasonable distributions were similar.

Consider the probability

$$\mathbb{P}(|\tilde{F}_{t,G} - \hat{F}_{t,G}| \leq \overline{F}_{t,G}) \geq \alpha_G \tag{6.16}$$

where α_G is some desired service level for group G and $\hat{F}_{t,G} := \sum_{d \in G} \hat{F}_{t,d}$ is the forecast or mean demand for G. The choice of αG determines the constant $\overline{F}_{t,G}$ which defines the support of (or uncertainty set for) F in our robust counterpart to the nominal problem; shortly, we show how to determine $\overline{F}_{t,G}$. In (6.16), we have used a linear (effectively $||\cdot||_1$) constraint on $\tilde{F}$ to ensure that our robust counterpart is also a mixed-integer linear-program. We now solve for $\bar{F}_{t,G,}$ to obtain

$$\overline{F}_{t,G} \triangleq \hat{F}_{t,G} + z(\frac{1+\alpha_G}{2})\sigma_{t,G},$$

where $z(\cdot)$ is the inverse of the Standard Normal cumulative distribution function. Requiring that we can produce for every demand in the range $[\hat{F}_{t,G} - \overline{F}_{t,G},\ \hat{F}_{t,G} + \overline{F}_{t,G}]$, guarantees that demand in $(t,\ G)$ is met with probability α_G, where α_G can be chosen by the user to ensure any desired service level. However, if the groupings G are not disjoint and constraints are applied to $\tilde{F}$ for all groups simultaneously, the probability of violation is much more difficult to compute. A lower bound on the guarantee is the product of their service values, $\alpha = \prod_G \alpha_G$, which for identical α_G values decays geometrically with the number of groups. We later avoid this issue by testing the model with a single group G and finding that it is not significantly different from solutions when some overlapping groups are used.

6.2.3 Affinely-Adjustable Production

Before we can begin incorporating the uncertainty set into the nominal formulation, we first simplify the model structure. Note that only (6.3) and (6.4) contain the uncertain parameter $F_{t,d}$. Because we would like factories to help each other compensate for fluctuations in demand, we aggregate these constraints across factories l; we change

$$\sum_{\tau=1}^{t}(y_{\tau,d,l} - F_{\tau,d}x_{\tau,d,l}) \leq 0 \qquad \forall t, d, l \qquad (6.17)$$

with equality for $t = T$ to

$$\sum_{\tau=1}^{t}(\sum_{l} y_{\tau,d,l} - F_{\tau,d}) \leq 0 \qquad \forall t, d \qquad (6.18)$$

with equality for $t = T$, by relaxing the binary constraint on x and using the fact that $\sum_l x_{t,d,l} = 1$. Note that, in (6.18), the only decision variable is the production quantity y and the uncertain parameter F is not the coefficient of a variable. Now that we have developed uncertainty sets and have isolated the uncertainty F, we are ready to develop the robust counterpart to this slight re-formulation of the nominal problem presented in §6.1.

Herein, let $f_{t,d} := \tilde{F}_{t,d} - \hat{F}_{t,d}$ and $f_{t,G} := \sum_{d\in G} f_{t,d}$ be the deviation of demand from its mean. As discussed in §6.2.2, we must produce a solution that satisfies the nominal constraints for any demand vector F such that

$$F \in \{\hat{F} + f \,:\, |f_{t,G}| \leq \overline{F}_{t,G} \;\; \forall t, G\} \subset \mathbb{Q}^{T\times|D|} \qquad (6.19)$$

We also define

$$\mathbf{F} \triangleq \{f : \; |f_{t,G}| \leq \overline{F}_{t,G} \;\; \forall t, G\}.$$

Because the termination criteria ($t = T$) in (6.18) must hold with equality $\forall f \in \mathbf{F}$,

f must be constant almost surely or y must depend on f. The prior case corresponds to the nominal model of §6.1. We now address the latter, which the literature refers to as the Adjustable Robust Counterpart. In order to maintain tractability, we will use the Affinely Adjustable Robust Counterpart (AARC) approach developed in [BTGGN04], wherein we restrict our attention to decisions y that are affine functions of f. The most general dependency would allow each production decision $y_{t,d,l}$ to depend on $\{f_{t',d'} : t' \leq t, d \in D\}$. However, it seems unreasonable to vary production for a particular destination based on demand for some other destination; furthermore, allowing the most general dependency would cause the problem's temporal dimension to increase by a factor on the order of T^2 after taking the robust counterparts. It may be the case that demand for a particular period t' is satisfied during a different period t; however because orders tend to be fulfilled in the same week that they are made, it would be reasonable to assume that production decision $y_{t,d,l}$ depends only on $f_{t,d}$. This will be key to keeping the robust counterparts simple. Thus, we replace every instance of $y_{t,d,l}$ in the nominal formulation with the simple affine function

$$\hat{y}_{t,d,l} + f_{t,d} y_{t,d,l}. \tag{6.20}$$

We now return to the constraints in (6.18) and write their robust counterparts, using $F_{t,d} = \hat{F}_{t,d} + f_{t,d}$ with $f \in \mathbb{F}$ and y replaced by (6.20), as

$$max_{f \in \mathbf{F}} \sum_{\tau=1}^{t} f_{\tau,d}(\sum_{l} y_{\tau,d,l} - 1) \leq \sum_{\tau=1}^{t} (\hat{F}_{\tau,d} - \sum_{l} \hat{y}_{\tau,d,l}) \qquad \forall t, d \tag{6.21}$$

with equality required when $t = T$. Note that we do not care solely about f that maximize the LHS of (6.21), but in the case of an inequality, if it is satisfied for a maximal f, it is satisfied $\forall f \in \mathbb{F}$. We now prove and then explain a theorem that lets us avoid solving this sub-problem.

Theorem 1. *For a full-dimensional, convex uncertainty set* $\mathbf{F}$, *the constraints in (6.21) are satisfied* $\forall f \in \mathbf{F}$ *if and only if* y *and* $\hat{y}$ *satisfy*

$$\sum_l y_{t,d,l} = 1 \qquad \forall t, d \qquad (6.22)$$

$$\sum_{\tau=1}^{t} \sum_l \hat{y}_{\tau,d,l} \leq \sum_{\tau=1}^{t} \hat{F}_{\tau,d} \qquad \forall t, d \qquad (6.23)$$

where the last inequality holds with equality for $t = T$.

Proof. If (6.22) holds, then substituting it into (6.21) yields $0 \leq \sum_{\tau=1}^{t}(\hat{F}_{\tau,d} - \sum_l \hat{y}_{\tau,d,l})$ which is exactly what (6.23) guarantees.

Conversely, suppose (6.23) does not hold and substitute $f = \mathbf{0} \in \mathbb{F}$ into (6.21) to yield a contradiction. Instead, suppose y and $\hat{y}$ satisfy (6.23) but violate (6.22) for some (t', d'), i.e $\sum_l y_{t',d',l} - 1 \neq 0$. Consider (6.21) for $t = T$ and $d = d'$ and note that the right-hand-side is zero because (6.23) holds with equality for $t = T$. Then for (6.21) to hold, we need

$$\sum_{\tau=1}^{T} f_{\tau,d'}(\sum_l y_{\tau,d',l} - 1) \quad = 0 \quad \forall f \in \mathbb{F} \qquad (6.24)$$

Because $\mathbb{F}$ is full dimensional and convex,

$$\exists f' \in \mathbb{F},\ f'' \in \mathbb{F} \text{ s.t. } f'_{t',d'} \neq f''_{t',d'} \text{ but } f'_{t,d} = f''_{t,d} \forall (t, d) \neq (t', d'). \qquad (6.25)$$

Suppose that (6.24) holds for both f' and f'' and subtract (6.24) for f'' from (6.24) for f', noting that all terms except (t', d') cancel, yielding $(f'_{t',d'} - f''_{t'd'})(\sum_l y_{t',d',l} - 1) = 0$. Then $\sum_l y_{t',d',l} - 1 \neq 0$ implies $f'_{t',d'} = f''_{t'd'}$ which contradicts (6.25). □

We now interpret Theorem 1. Note that (6.23) is the same as (6.18) but with the nominal y replaced by $\hat{y}$, the constant of the affine function, and the nominal F replaced by its mean, $\hat{F}$, indicating that $\hat{y}$ represents the average production level. Additionally, (6.22) indicates that $y_{t,d,l}$ represents the fraction of $f_{t,d}$, the deviation in demand from destination d during period t, that factory l will produce. By not requiring non-negativity of y (but still of $y + \hat{y}f$), some production can be negatively correlated with demand deviations when $\hat{F} - \overline{F} > 0$. We use Theorem 1 to avoid actually having to implement the Robust Counterpart in (6.21) which would otherwise involve dealing with demand uncertainty over time; this is crucial to keeping the size of the Robust Counterpart tractable.

Note that our definition of affinely adjustable in (6.20) was not unique; instead, had we used $\hat{y} + yF$ for $F \in [\hat{F} - \overline{F}, \hat{F} + \overline{F}]$ we would get $\hat{y} = 0$ along with (6.22) and $y_{t,d,l}$ would be interpreted as the fraction of total demand for (t, d) served by l. This would correspond to an affine function through 0 instead of $\sum_T \hat{F}_T$. This alternate approach corresponds to the intuitive idea of allocating demand to factories proportionally ($y_l \propto x_l$) and hence it is easier to implement in practice, e.g. there is no need to wait until the end of the week for computations relative to the mean. However, it forces factories that help with worst-case demand to also help with typical demand. We choose (6.20) instead because it is centered and symmetric about the mean, median, and mode of the demand distribution and, at the typical values, it allows some factories to commit to a portion of typical demand levels without concern for uncertainty while allowing other factories to take on most of the uncertainty. In essence, our choice of affine dependency models contingency plans; if demand deviates significantly from the mean, our solution indicates how to respond.

Because we allow $f < 0$, the definition in (6.20) cannot assign $y_l > 0$ unless $\hat{y}_l > 0$ when we take the Robust Counterpart of the original non-negativity constraint $\hat{y} + yf \geq 0$; if instead we only consider $f \geq 0$, which will simply change the dual constraints we re-inject from inequalities to equalities but change the meaning of our uncertainty set, we can have a factory aid another in production only for demand above the nominal value $\hat{F}$. By symmetry of the Normal distribution,

$$\mathbb{P}(|\tilde{f}_{t,G}| \leq \overline{F}_{t,G}) = \mathbb{P}(\tilde{f}_{t,G} \leq \overline{F}_{t,G} \mid f \geq 0),$$

so our service level α_G gives the same $\bar{F}_{t,G}$. In fact, using $f \geq 0$ with the same $\overline{F}$ will de-emphasize negative correlations in demand and lead to more emphasis in protecting against demand deviations from destinations with lower shipping costs because their demand cannot be "given" (by the inner maximization problem) to states with higher costs. In fact, it will even avoid the complication of negative demand that arises when using a Guassian distribution for modeling demand. We investigate both of these approaches ($f \geq 0$ and f unrestricted in sign) simultaneously due to the minimal effort required to alter the uncertainty set of (6.19).

6.2.4 Robust Counterparts

Although Theorem 1 lets us avoid solving the maximization subproblem (6.21) for each of those constraints, the affine production function (6.20) introduces the uncertain term f into several other constraints. We must now solve maximization subproblems for these constraints; however, these subproblems do not involve the aggregation of uncertainty over time, making them much smaller in size and less conservative in protecting solutions.

We now develop the Affinely Adjustable Robust Counterparts for constraints involving the production variables, the only remaining constraints containing uncertain parameters. Each constraint involving uncertain parameters will be enforced by taking the dual of the maximization (with respect to uncertainty) problem and re-injecting the dual objective with additional constraints on the dual variables. We denote all dual variables to Robust Counterpart subproblems by either $\overline{\mu}$ or $\underline{\mu}$ which will be associated with the upper and lower bounds on the deviations of f, respectively. For the most part, we need separate dual variables for every constraint containing uncertainty (and hence a maximization problem in which dual variables will appear) in the nominal problem, but our notation will ignore this for simplicity of presentation. However, it is important in implementation to index which dual variables pertain to which constraint.

The dual variables $\overline{\mu}$ or $\underline{\mu}$ are shadow prices for $-\overline{F} \leq f \leq \overline{F}$ and therefore only one of each pair can be non-zero in any solution. Each μ is only involved in constraints for its own subproblem. We could re-write them as the absolute value of a single dual variable and if we desire use the $max\{\mu, -\mu\}$ approach to linearizing absolute value instead of separating it into two non-negative variables. However, to maintain the ability to adopt $f \geq 0$, we refrain from doing so.

We begin with protecting the capacity constraints (6.7), (6.13), and (6.14), which all involve $\sum_{k,d} y_{t,d,l}$ as the only source of uncertainty and hence have the same relevant uncertainty sets

$$\mathbb{F}_t \triangleq \{f_t : |\sum_{d \in G} f_{t,d}| \leq \overline{F}_{t,G} \forall G\}$$

for each $t \in \{1..T\}$. In fact, for fixed (t,l), the coefficients of the uncertain parameters in (6.7), (6.13), and (6.14) are the same and therefore they will have the exact same dual solution (including for each s), which will slightly reduce our computational burden. Constraints (6.7), (6.13), and (6.14), can be re-written as

$$\sum_d y_{t,d,l} f_{t,d} \leq (S_l + o_{t,l})A_l - \sum_d \hat{y}_{t,d,l} \qquad \forall t, l \qquad (6.26)$$

$$\sum_d y_{t,d,l} f_{t,d} \leq (1 + M_l)(P_{l,s}(S_l + o_{t,l})$$

$$+ \overline{P}_l(S_l + O_l)(1 - q_{l,s}) - \sum_d \hat{y}_{t,d,l} \qquad \forall t, l, s \qquad (6.27)$$

$$\sum_d y_{t,d,l} f_{t,d} \leq (p_l + r_{t,l})S_l + \overline{P}_l(1 + M_l)o_{t,l} - \sum_d \hat{y}_{t,d,l} \qquad \forall t, l \qquad (6.28)$$

Choose one of the three sets of constraints and fix t, l, and where appropriate, s. Let RHS denote the right-hand-side of the constraint and let β_d be the coefficient of $f_{t,d}$. Then the robust counterpart for that constraint is simply

$$max_{f \in \mathbb{F}_t} \sum_d \beta_d f_{t,d} \leq RHS$$

which, by re-injecting the dual and applying strong duality ($\mathbb{F}$ is bounded and $f = 0 \in \mathbb{F}$ is always feasible), will hold if

$$\exists \overline{\mu}, \underline{\mu} \geq 0 : \sum_{G \ni d} (\overline{\mu}_G - \underline{\mu}_G) = \beta_d \ \forall d \text{ and } \sum_G \overline{F}_{t,G}(\overline{\mu}_G + \underline{\mu}_G) \leq RHS \qquad (6.29)$$

where $G \ni d$ is the set of groups G that contain destination d.

In order to protect the objective function (6.8) against uncertainty , we rewrite it as

$$\text{minimize } \zeta \text{ subject to } \sum_{t,d} f_{t,d}(\sum_{l} C_{d,l} y_{t,d,l}) \leq \zeta - \eta \qquad (6.30)$$

where

$$\eta \triangleq \sum_{t,d,l} C_{d,l} \hat{y}_{t,d,l} + \sum_{t,l} \left[\frac{S_l}{U_l} (C_l^{pS} p_{t,l} + C_l^{rS} r_{t,l}) + v_{t,l} \right].$$

If we maximize the left-hand-side of the constraint in (6.30) over $f \in \mathbb{F}$, take the dual, and re-inject it, we get

$$\text{minimize } \zeta \text{ subject to } \sum_{t,G} \overline{F}_{t,G}(\overline{\mu}_{t,G} + \underline{\mu}_{t,G}) \leq \zeta - \eta \qquad (6.31)$$

with non-negative dual variables $\overline{\mu}, \underline{\mu} \geq 0$ and the following constraints

$$\sum_{G \ni d} (\overline{\mu}_{t,G} - \underline{\mu}_{t,G}) = \sum_{l} C_{d,l} y_{t,d,l} \qquad \forall t, d. \qquad (6.32)$$

The robust counterparts of the non-negativity constraints are

$$\begin{aligned} & min_{|f| \leq \overline{F}} \ \hat{y} + yf \geq 0 \\ \iff & |y| \overline{F} \leq \hat{y} \\ \iff & \exists \mu : \mu \overline{F} \leq \hat{y}, \ \mu \geq y, \ \mu \geq -y. \end{aligned}$$

More explicitly, and using the same notation as before, we need $\overline{\mu}, \underline{\mu} \geq 0$ to satisfy

$$\begin{aligned} \bar{F}(\overline{\mu} + \underline{\mu}) &\leq \hat{y} \\ \overline{\mu} - \underline{\mu} &= -y. \end{aligned}$$

If we were to have $f \geq 0$, the robust counterpart would be

$$\begin{aligned} & min_{0 \leq f \leq \overline{F}}\ \hat{y} + yf \geq 0 \\ \iff & \hat{y} \geq 0,\ -y\overline{F} \leq \hat{y} \\ \iff & \exists \mu \geq 0 : \mu\overline{F} \leq \hat{y},\ \mu \geq -y \end{aligned} \tag{6.33}$$

which is clearly less restrictive on y and $\hat{y}$. For the other constraints above, if $f \geq 0$, we would simply delete $\underline{\mu}$ and the additional equality constraints would become "greater than or equal to" ($\geq$) constraints.

In summary, we do the following: we replace each constraint in (6.7), (6.13), and (6.14) with the new set of decisions and constraints in (6.29); similarly, the objective is replaced by (6.31) and (6.32); the non-negativity constraints are replaced by (6.33). By doing so, we have protected against uncertainty in the nominal formulation and developed a robust mixed-integer linear formulation.

6.3 Extension to Multi-Product Setting

Although the firm builds customized desktops for each individual customer order, the above formulation does not model this for clarity in the exposition of Robust Optimization. However, the model made for use by the firm and our analysis in the next section distinguish between the firm's "Lines of Business." For the time period of the available data, the firm had two desktop product lines, one for large corporate customers and one for individual consumers. The major differences are that their demand is forecast separately, the chassis are of different weight and size, they have different shipping requirements, and factory bottlenecks depend on the proportion of different products being produced. The latter, which is illustrated in §2.3, is shown to be of utmost importance to efficient production in §6.6. The Lines of Business do have some similarities; it takes each laborer approximately the same amount of time to assemble desktops from different product lines, avoiding further complications in the labor capacity constraints. In this subsection, we describe how we extend the above formulation to a multi-product setting.

The implication for the model is that we introduce an additional index k for the production decisions y and $\hat{y}$ and the dual variables $\overline{\mu}$ and $\underline{\mu}$, along with the demand parameters F and $\hat{F}$ and the physical production capacity parameter A. Furthermore, we must divide the production variables $y_{t,d,l,k}$ by the corresponding $A_{l,k}$ in the physical production capacity constraint to normalize them properly. This will then necessitate separate dual variables for the physical capacity constraints. We enforce the build-to-order and dual-defining equality constraints for each product line k and we aggregate production over k in the capacity constraints. The labor production rate U is independent of k. Because demand for various lines of business is fairly independent, we do not aggregate demand uncertainty over them.

Although this extension seems rather simple, it allows for a very interesting additional dimension in the computational analysis of the results, which we discuss next, by showing how the firm can exploit differences among their product lines, factory capabilities, and customer segments.

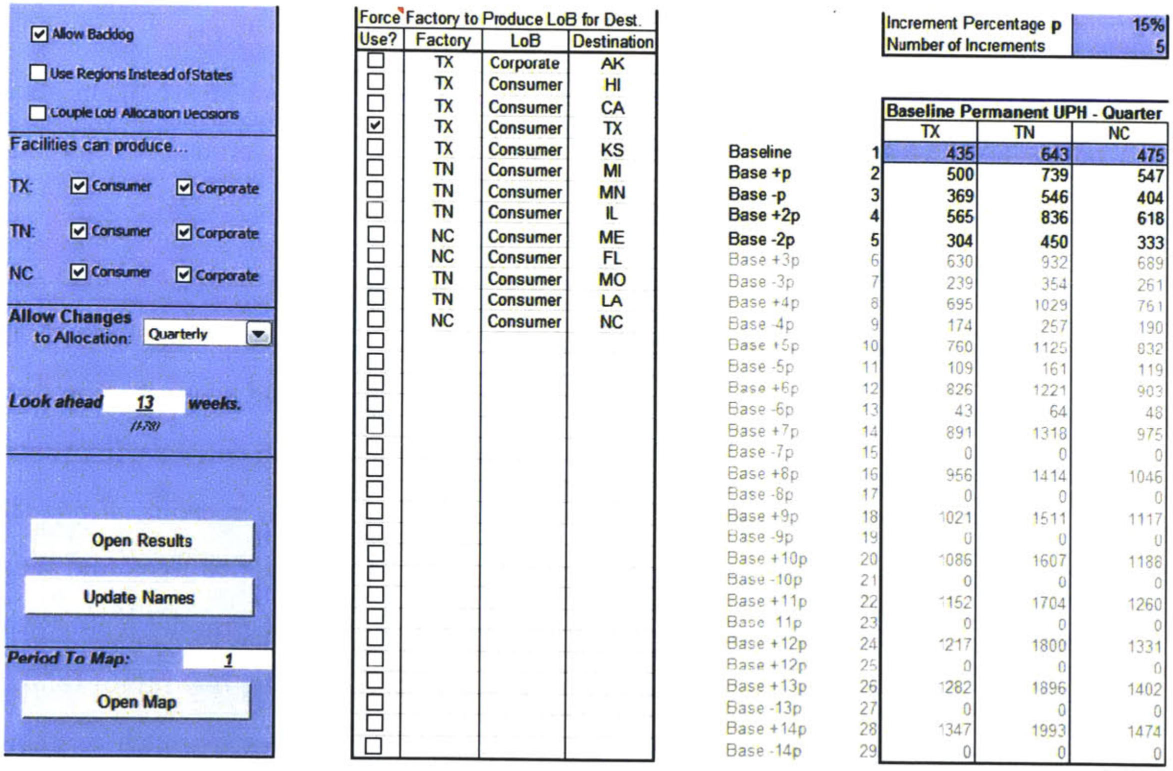
6.4 Implementation Details

We used a planning horizon of one fiscal quarter, which begins and ends with no backlog of orders, with time discretized into thirteen weeks. At the time, the firm's three US desktop manufacturing factories were TX in Austin, Texas; TN in Nashville, Tennessee; and NC in Winston Salem, North Carolina. We used the firm's two major desktop lines of business, consumer and corporate. The demand destinations were the 50 US states along with Washington D.C. We protected against demand uncertainty for the following groups G: 1) the 50 US States and Washington D.C., 2) the thirteen regions identified by the firm on the geo-manufacturing map, depicted in Figure 2-4, 3) Western, Central, and Eastern US, defined by the bold lines in the same map, and 4) nationwide, i.e. the sum of all demand.

Almost all data used for this model is described in §4.4 and §4.5, although some conversions were necessary to match the scope. Many input parameters were collected internally by the firm's planners, including the sales forecast F, inbound shipping, outbound shipping, and labor costs C, units-per-labor-hour U, and the maximum permanent-to-temporary labor ratio M. The maximum number of straight-time and overtime hours that a factory can effectively operate, S and O, respectively, along with the factory bottlenecks A were acquired in interviews with factory managers, described in §2.3.

This led to implementations of the robust formulation with about 25,000 variables and constraints. On a Intel Pentium 1.6GHz processor with 768Mb of RAM, CPLEX 10.1.1 typically required one to twenty minutes of computation time. However, for some choices of the discrete set of permanent labor levels P, such as providing a wide range of choices that were far from optimal or empirical values and would make the "Big-M" constraints (6.13) and (6.14) have poor upper bounds, the optimization software could require weeks to close the optimality gap to within 0.1%. Even after much tweaking of algorithm parameters, it had trouble developing strong lower bounds on the optimal solution and would exhaustively branch on the binary overtime decisions w and permanent labor levels q. To address this, we terminated the solver early, removed poor choices of P after testing that they did lead to sub-optimal solutions, and then the model ran smoothly.

As in §5.2.9, challenges arose in making the model user-friendly. Given that the planning model uses significantly less data and is used less frequently, data management issues were not such a large concern. However, significant effort was still required to make the interface accessible and informative. For example, Figure 6-2 displays a screen shot of the interface that the firm's employees used to control the model. The top left most part of the interface allows the user to specify whether the backlog of orders is allowed to accumulate between weeks, whether the model should use the larger regions of Figure 2-4 instead of states, whether the model must give the same allocation decisions x for both Lines of Business, and which factories can produce which Lines of Business. Below this, the user can change the time horizon and can let x change at different frequencies, from weekly to annually. Below that are buttons that help the user quickly access output data, update data parameters, and access automatically-generated visual maps of the most recent solution, which looks like the map pairs in Figure 6-3 which we will discuss later. The center part of the control interface in Figure 6-2 allows the user to force the model to assign any LoB-destination pairs to a particular factory; in the screen shot, TX has been forced to produce all Texas orders for consumer desktops; this was useful to enforce non-geo-eligibility or disallow parts of the solution that the firm did not agree with, such as shipping to Alaska and Hawaii from NC. The right-most part of the interface shows how the permanent labor levels were discretized into the values of $P_{l,s}$; occasionally, multiple model runs were necessary to find a satisfactory set of choices. Similar to the Execution Model, model output was displayed in a visually-informative format with summary statistics and comparisons to a baseline solution.



☑ Allow Backlog

□ Use Regions Instead of States

□ Couple LoB Allocation Decisions

Facilities can produce...

TX: ☑ Consumer ☑ Corporate

TN: ☑ Consumer ☑ Corporate

NC ☑ Consumer ☑ Corporate

Allow Changes to Allocation: Quarterly

Look ahead 13 weeks. (1-78)

Period To Map: 1

Force Factory to Produce LoB for Dest.

Use?	Factory	LoB	Destination
□	TX	Corporate	AK
□	TX	Consumer	HI
□	TX	Consumer	CA
☑	TX	Consumer	TX
□	TX	Consumer	KS
□	TN	Consumer	MI
□	TN	Consumer	MN
□	TN	Consumer	IL
□	NC	Consumer	ME
□	NC	Consumer	FL
□	TN	Consumer	MO
□	TN	Consumer	LA
□	NC	Consumer	NC
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			
□			

Increment Percentage p	15%
Number of Increments	5

		Baseline Permanent UPH - Quarter		
		TX	TN	NC
Baseline	1	435	643	475
Base +p	2	500	739	547
Base -p	3	369	546	404
Base +2p	4	565	836	618
Base -2p	5	304	450	333
Base +3p	6	630	932	689
Base -3p	7	239	354	261
Base +4p	8	695	1029	761
Base -4p	9	174	257	190
Base +5p	10	760	1125	832
Base -5p	11	109	161	119
Base +6p	12	826	1221	903
Base -6p	13	43	64	48
Base +7p	14	891	1318	975
Base -7p	15	0	0	0
Base +8p	16	956	1414	1046
Base -8p	17	0	0	0
Base +9p	18	1021	1511	1117
Base -9p	19	0	0	0
Base +10p	20	1086	1607	1188
Base -10p	21	0	0	0
Base +11p	22	1152	1704	1260
Base 11p	23	0	0	0
Base +12p	24	1217	1800	1331
Base +12p	25	0	0	0
Base +13p	26	1282	1896	1402
Base -13p	27	0	0	0
Base +14p	28	1347	1993	1474
Base -14p	29	0	0	0

Figure 6-2: User interface for the planning model, including controls to make popular changes to input parameters and easily access the model output.

6.5 Validation and Analysis Methodology

The purpose of the following validation effort is to understand any discrepancies between the planning model's prediction of firm's decisions (and their monetary implications) using the collected input data compared against actual decisions made and financial costs incurred by the firm. In addition, we wish to understand how the model's suggestions differ from the actual decisions the firm made, what managerial insights can be extracted, and how much cost savings dynamic order allocation can offer.

In addition to the input parameters described in §6.4, the following financial figures from Q3 (August to October) and Q4 (November to January) 2007 were available from the firm: 1) total outbound and labor costs per factory per week, with benefits and wages separate and distinctions between wages made in straight-time versus overtime and by permanent employees versus temporary employees; 2) total production quantities for each line of business (LoB) at each factory for each destination state during each week; and 3) total labor hours split by straight-time versus overtime and permanent versus temporary at each factory during each week.

Though the total cost figures and production quantities given by the firm's financial figures and the model can be compared directly, the model's labor decisions and assumptions do not fit the firm's data quantities perfectly and require some data conversions. As can be seen in 6-1, the firm's historical production capacity fluctuates more than our theoretical model of it; it does so for the following reasons. First, the firm's straight-time permanent labor level varies slightly from week to week and must be smoothed to an average value to match the model. Second, the model ignores the fact that the firm occasionally uses overtime before saturating the maximum temporary labor bound and that temporary laborers might not stay for as much overtime as permanent laborers. Last, the units-per-labor-hour rate U varies with labor type and even between laborers of the same type, but the model treats it as fixed, possibly causing bias in labor cost calculations. In order to compare similar objects, we made reasonable conversions of the firm's historical decisions to ones that fit the model's structure so that we may inject some or all of their solution into the model, optimize over the remaining decisions, and make comparisons. The firm considers these conversions to either be consistent with their current planning considerations or not significantly far from reality. In addition to comparing the model's cost prediction for historical decisions, we are also interested in comparing the model's ability to predict the firm's decisions and find better ones given that only some decisions are already set.

We report and compare costs and decisions from the following scenarios:

1. The firm's reported financial cost totals.

2. The firm's *production* quantities multiplied by the appropriate *outbound costs* and the firm's total *labor hour* figures multiplied by *average wage rate per hour.*

3. Using, the conversions just described, fix some decisions to their empirical value and optimize any other decisions as follows:

    (a) Fix all decisions (*production*, *permanent* and *temporary labor*, and *overtime factory hours).*

    (b) Fix *permanent labor* and optimize *production*, *temporary UPH*, and *overtime factory hours.*

    (c) Repeat (b) but use the robust formulation and vary α_{Nat}.

    (d) Fix total demand (sum of historical *production* levels across factories) and optimize *all* decisions.

    (e) Repeat (d) but require a minimum permanent labor force at the North Carolina factory.

    (f) Repeat (d) but use the robust formulation with and without $f \geq 0$ and (i) $\alpha_{Nat} = 95\%$, (ii) $\alpha_{Nat} = 99\%$, and (iii) $\alpha_G = 99\% \forall G$.

We evaluate shipping costs at the mean demand $\hat{F}$. For robust solutions, labor costs are based on the labor levels necessary to meet the worst-case demand, as labor was not adjustable in the robust formulation. This provides an even basis for comparison. It allows us to understand how much additional staffing may be necessary and how much more we are willing to pay in outbound shipping for a particular level of customer service. α_{Nat} refers to the service guarantee for the sum of all demand, i.e. the national *(Nat)* group.

Scenario (a) serves as the baseline for almost all comparisons and represents the model's projection of the firm's historical decisions. Scenarios (1) and (2), when compared with (a), depict the model's accuracy. The remaining scenarios demonstrate the cost savings potential of various dynamic order allocation strategies.

6.6 Results and Insights

A summary of the total supply chain costs for the ten scenarios above, computed for Q4 2007, is included in Table 6.1. Similar results held for Q3 2007 but contained no additional interesting insights. On average, of the total relevant supply chain cost, inbound shipping constituted 15%, outbound shipping 45%, and labor 40%.

Scenario	Inbound Cost	Outbound Cost	Labor Cost	Total Cost
(1)	0.0%	0.0%	3.4%	1.40%
(2)	0.0%	0.0%	-0.1%	-0.04%
3(a)	0.0%	0.0%	0.0%	0.0%
3(b)	-0.2%	3.5%	13.2%	6.98%
3(c)	0.4%	3.6%	11.9%	6.57%
3(d)	3.9%	2.7%	21.5%	10.66%
3(e)	-1.5%	2.9%	19.9%	9.27%
3f(i)	2.1%	3.3%	15.8%	8.34%
3f(ii)	3.3%	3.4%	12.8%	7.08%
3f(iii)	2.5%	3.1%	12.6%	7.05%

Table 6.1: Relative inbound, outbound, and labor cost savings for each scenario as a percent of the cost of baseline scenario (a) in Q4 2007.

We now consider how well the model predicts the firm's costs to validate the model's accuracy. First, the product of the labor cost parameters with the firm's total reported labor hours mismatched the firm's reported labor costs by 3.4%. Second, the assumption that permanent and temporary workers stay for the same amount of overtime added an additional 0.1% error of the opposite sign. For Q3 2007, because we smoothed the permanent labor level to its mean, there was insufficient labor available for the model to find a feasible allocation if permanent labor was fixed, indicating that the conversion assumption for permanent labor is rather conservative. Taken together, these errors partially mitigate each other, because the model estimates that labor costs less than it actually does but requires more labor than is actually needed. The outbound shipping costs matched financial figures almost exactly. Although we had no basis for comparing the model's inbound shipping costs to financial figures, they are likely to be accurate on average because the model's costs were constructed by averaging historical inbound shipping costs. Because typical labor cost savings in Table 6.1 range from 10-20% with a bias of at most 3.4%, along with the fairly accurate shipping costs, we are fairly confident in the model's total supply chain cost predictions.

We now examine the model's suggested solutions to the firm's problems in scenarios (a) through (iii). Table 6.2 presents the suggested permanent labor levels as percentages of the total baseline permanent labor force in scenario (a). Scenarios (a), (b), and (c) have their labor force fixed to the model's estimate of the firm's Q4 2007 permanent labor force. Figure 6-3 contains eight U.S. maps, one for the consumer product line and one for the corporate product line, for the four nominal scenarios (a), (b), (d), (e), ordered from top to bottom. The map depicts which factory produced the most desktops for that product line in Q4 2007. Note that in (a), other than in Alaska, the production followed the static geo-manufacturing map in Figure 2-4, using the same allocation decisions for both product lines, making this scenario highly representative of the firm's default download rules. These results are consistent with our analysis of the Greedy policy in §4.6.1 but are contrary to the findings in §5.3.1 where the execution model's "actual" solution did not match the map in Figure 2-4 because the firm had actually moved orders. Scenario (a) matches this map and the firm's historical costs; in Q4 2007, the firm was producing desktops largely according to that map which was within a few percent of the optimal shipping costs. The Greedy policy in Chapter 4 also did so. However, in Q3 2008, the Operations Center moved orders in a sub-optimal manner, which is reflected in the results of Chapter 5.

Scenario	TX	TN	NC
3(a)(b)&(c)	28.8%	24.1%	47.1%
3(d)	22.8%	30.5%	16.3%
3(e)	21.6%	14.4%	41.3%
3f(i)	27.9%	36.8%	24.8%
3f(ii)	28.6%	37.0%	27.7%
3f(iii)	31.4%	35.8%	25.3%

Table 6.2: Permanent labor capacity at each factory as a percent of the baseline (a) total permanent labor force.

Scenario (b), which uses the same permanent labor force, suggests significant labor savings and some outbound cost savings potential from timely and optimal allocation of demand. In particular, as seen by comparing (a) and (b) in Figure 6-3, it suggests having TN cover more of the consumer desktops while having TX and NC cover more of the corporate desktops, resulting in 6.98% cost savings. The major underlying cause for this shift is that production rates depends on product mix, which is discussed in §2.3 and §6.3, and the associated production capacity parameters $A_{l,k}$. TN was originally built to handle the more variable and customized consumer desktops, while TX has better technology to sort, bundle, and shrink-wrap large corporate orders; NC was flexible. With less factory hours necessary when running a higher consumer product mix at TN and the opposite at TX, making allocation decisions for each product line accordingly offers the firm significant savings potential.

Another cost savings opportunity is exploiting anomalies in the shipping cost structure. Domestic shipping costs are usually proportional to the shipping distance. However, some third-party logistics network configurations defy this rule of thumb. For example, as we verified in the cost data and illustrate for consumer desktops in Figure 6-3 (b), shipping to Utah, Vermont, and New Hampshire from TN was more cost effective than first shipping to intermediate states, such as Colorado or Pennsylvania. Optimization software facilitates exploiting such cost discrepancies while maintaining balanced factory loads.

In scenario (c), we were able to protect against up to twelve standard deviations in national demand above its mean ($f \geq 0$, $\alpha_{Nat} \approx 1$) before the firm's permanent staffing level was insufficient. Results are shown for $\alpha_{Nat} = 99\%$ and suggest that the additional temporary and overtime staff necessary to serve any of all but 1% of all possible demand realizations, at that relatively high permanent staffing level, would cost 1.3% of the labor cost or 0.41% of the total supply chain costs. For many managers, this is a small price to pay for such strong service guarantees.

Scenario (d) represents the optimal nominal solution, wherein the model is allowed to re-allocate production and make all labor decisions. In Table 6.2 we see a dramatic reduction in NC's permanent workforce along with a significant labor decrease in TX and increase at TN. The solution for scenario (d), represented by the third pair of maps in Figure 6-3, suggests TN should cover 75% of the consumer products and 30% of corporate products and that NC should not produce any consumer products and only 30% of corporate products, even though it is the most flexible factory in terms of physical manufacturing lines and production capacity. Although scenario (d) suggests 10.66% savings in total supply chain costs, the firm would find this imbalanced arrangement unsatisfactory. We rectify this in scenario (e) by showing that even if the NC permanent labor force remains at its minimum (89% of the NC baseline or 41.7% of the total baseline), which is considerably higher than the staffing levels at TX and TN, total savings of up to 9.27% are still possible by reducing the TN workforce and letting NC and TX handle 90% of the consumer desktops. This scenario illustrates both the cost of such arrangements with local governments and the firm's supply chain's flexibility to adapt to such restrictions. The immense savings from two very different solutions, along with the desire to both balance and maintain somewhat higher staffing levels, indicate that the Robust Optimization solutions, which we discuss next, should simultaneously provide strong service guarantees and cost savings.

Scenarios (i), (ii), and (iii) introduce protection against uncertainty in the demand. For all three, without $f \geq 0$, the non-negativity constraints $\hat{y} + yf \geq 0$ resulted in infeasible instances. One workaround would be to truncate $\tilde{F}$ from below at 0, but this heavily complicates the probability of constraint violation. Instead, as discussed in §6.2, the interpretation for $f \geq 0$ is simple and only requires $\alpha_{Nat} < 1$ which implies $\sum_d f_{t,d} \leq \bar{F}_{t,Nat} < \infty \ \forall t$ to generate a bounded uncertainty set. Hence, results are shown only for $f \geq 0$.

The solutions for scenarios (i), (ii), and (iii) suggest significant increases in the labor force at TX and decreases at NC. Interestingly, as is apparent in Table 6.1, these three scenarios focus more on inbound costs than other solutions do, mostly because shipping costs for large demand deviations are emphasized by the maximization in the lower bound on the objective, but scenario (iii) does so to a lesser extent by applying less protection against extreme demand. Demanding 95% protection against nationwide uncertainty costs 2.32% of the optimum but still saved 8.34% over the firm's decisions. Raising the service level to 99% cost an additional 1.2% of the optimum but retained 7.08% savings.

The solution for scenario (ii) assigns all uncertain demand to TN via the y in (6.20), excluding the corporate uncertainty which it allocates to TX in week 2 and 11, while (iii) assigns all uncertain consumer demand to TX but splits consumer desktop uncertainty (which was much less variable) evenly across the nation. The negligible cost difference between (ii) and (iii) suggests little need to apply upper bounds on individual demand uncertainties $F_{t,d}$. Additionally, the predicted cost savings in Table 6.1 is uniformly lower for (iii) than (ii) even though the uncertainty set in (iii) is a strict subset of the set for (ii). Although this is at odds with intuition, in Table 6.1, the shipping cost is evaluated at the mean value of the demand distribution; this phenomena arises from non-linearities in the objective value as a function of uncertainty.

In collaboration with this large build-to-order desktop manufacturing firm, we developed a realistic large-scale planning model that is robust to fluctuations in demand and offers significant cost savings in an industry with rapidly shrinking profit margins. The model optimizes the process of dynamically balancing factory loads and geographic transportation costs, helping to reduce both transportation and labor costs, while maintaining quantifiable service level guarantees.

The model indicates that making separate allocation decisions for each product line while considering their joint impact on factory bottlenecks and staffing requirements, alongside shipping costs, appears to be a major venue for cost savings. Additionally, it shows that setting minimum staffing requirements or service guarantees can cost 1-4% percent of the "optimal" total supply chain cost but also offer immense benefits, such as government incentives or satisfied customers, respectively. Even then, with the model's suggestions, one could still save 7-9% in total supply chain costs.

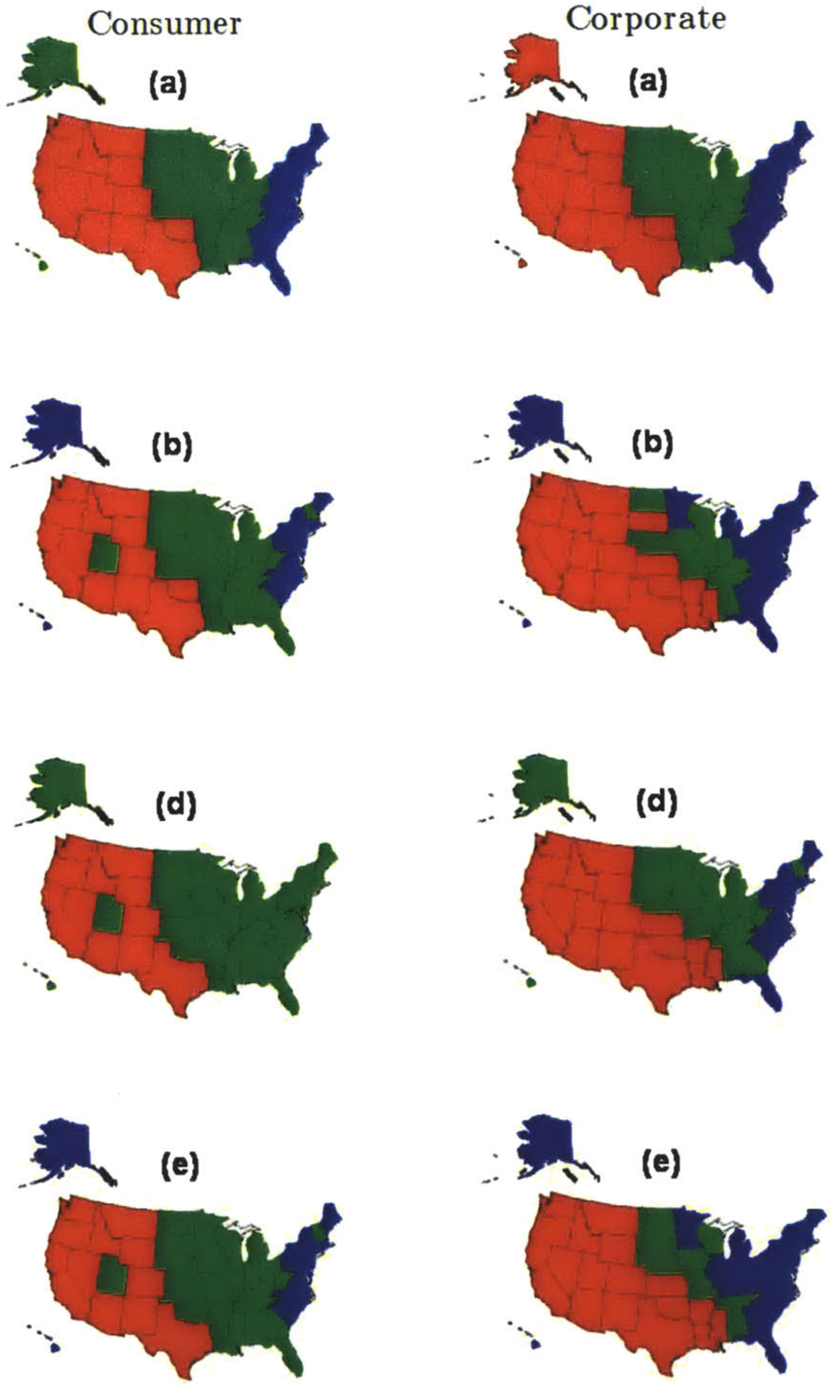


Figure 6-3: Maps of the United States colored by which factory produces the most desktops for that destination state in scenarios (a), (b), (d), and (e), for both Lines of Business. TX is red; TN is green; NC is blue.

Chapter 7

Conclusion

In collaboration with a large make-to-order desktop manufacturing firm that has multiple production facilities, we explored the intricacies, practical difficulties, and cost saving opportunities of using optimization techniques to solve complex production planning and control problems. The decision of which orders are allocated to each factory plays a significant role in shipping costs, expensive labor decisions, and customer service. Throughout the development of three distinct mathematical formulations of the problem, which were grounded by the firm's data, testing, and feedback, we discussed practical challenges to balancing tractability and realism. All three models consistently indicate that further optimization beyond the firm's current practice offers several hundred thousand dollars in daily cost saving opportunities.

We first modeled the execution problem as a stochastic Dynamic Program with a detailed demand model and evaluated solution policies via simulation. The rolling-horizon, certainty-equivalent Linear Programming policy solved the problem to near-optimality despite being deterministic and offered significant savings over the firm's Historical policy, which itself outperformed the static geoman map of the Greedy policy by a large margin. Solutions were shown to be insensitive to forecast errors and the cost of order lateness. Although the cost of each policy varied with demand, their relative efficacy did not.

We then developed a much more detailed Mixed Integer Linear Programming formulation of the execution problem, sacrificing the stochastic demand model in order to incorporate knowledge of individual orders, product lines, parts availability, limitations of the labor force, and geo-eligibility. Data acquisition and estimation, model usability, managerial issues, and the flexibility of labor were discussed at length. Results from two time periods that had different manufacturing network configurations both show significant savings potential, similar to the previous model.

For the planning problem, a Mixed Integer Linear Program formulation determines how to allocate geographic regions to factories. The joint capacity decision of how to hire labor over the horizon and how long the factory should operate is handled by breaking the hockey-stick-effect of production capacity into simpler intervals. Robust Optimization is introduced and uncertainty sets are developed to give probabilistic service guarantees despite having limited demand data. Production becomes an affine function of demand and constraints are reformulated to maintain realism and make robust counterparts tractable. The model and data is validated in comparisons to the firm's financial figures. Again, results suggest significant cost savings potential even when strong service guarantees are made.

Because the firm began focusing on retail rather than make-to-order sales channels and it began closing its North American factories while we studied this problem, there was insufficient time to fully integrate these solutions into the firm's operating environment. Although we linked software prototypes to live data, the firm's information technology department would require several months to two years to develop secure, fail-proof software implementations of the solutions that we recommended. Future work in similar settings should strive to make the transition from modeling and solving the problem off-line to regular on-line use easier and faster, including robust data acquisition and estimation, making input and output understandable and adjustable, handling exceptions to almost every rule, and early incorporation of managerial concerns such as fairness.

Nonetheless, the firm did find many of the models' managerial insights useful. The insight that the firm was most excited about was that differentiating the geoman map by line of business could take advantage of each factory's distinctive production capabilities while still focusing on outbound costs; most of our recommended solutions moved orders for corporate desktops to the Texas factory while emphasizing consumer desktop production at the factory in Tennessee, mitigating the impact of factory bottlenecks. When balancing loads across distant factories, moving orders through an intermediate factory can help maintain low shipping costs. Because direct shipping costs are not always proportional to the geographic distance between factories and customer addresses, the geoman map need not assign contiguous regions to factories; for example, in some cases, New Hampshire, Vermont, and Utah were better served by the factory in Tennessee even though all surrounding states were assigned to other factories. Another important insight was that producing as much as possible at each factory every day can be sub-optimal; delaying production today, by not extending shift lengths beyond their nominal value, when forecasts predict low demand relative to capacity in the near future, can reduce labor costs. The insight worth the most in cost savings was recognizing that the ATB to capacity ratio does not convey imbalances in order lateness; at times, all factories may have ATB in excess of their maximum capacity, but if one factory has so many past-due orders that it cannot produce them within the day, other factories should help. These lessons learned from the analysis of our models are not only quick fixes to drastically improve the performance of the firm's current practices but are also applicable to many other production planning and control settings.

Accounting for all of these opportunities throughout the network and weighing them simultaneously is difficult. As information becomes more available and competitive advantages such as proprietary technology, geographic specialization, and legal protections deteriorate in many industries, competing on cost, customer service, and network-wide solutions becomes more important. This work demonstrates that make-to-order manufacturers with multiple production facilities have much to gain by adopting optimization techniques to quickly and effectively respond to customer demands by leveraging their network's flexibility.

Appendix A

Glossary of Terms and Abbreviations

AARC Affinely Adjustable Robust Counterpart

ARC Adjustable Robust Counterpart

ASRS Automated Storage and Retrieval System

ATB Available-to-Build

BTO Build-to-Order

Cap Capacity (for producing desktops)

DP Dynamic Programming

G Greedy Policy

Geoman Geographic Manufacturing

H Historical Policy

IP Integer Programming

JM Outsourced factories that builds desktops for the firm, located in Juarez and San Jeronimo, Mexico

LN Log-Normal(ly distributed random variable)

LP Linear Programming or Linear Programming Policy

LoB Line of Business

MRP Material Requirements Planning

MILP Mixed Integer Linear Programming

N Normal(ly distributed random variable)

NC The firm's factory in Winston-Salem, North Carolina (NC), USA

NCLL Lean Lines (efficient high-volume production lines) at NC

PH Perfect Hindsight

Prod Production (of desktops)

RO Robust Optimization

TN The firm's factory in Nashville, Tennessee (TN), USA

TX The firm's factory in Austin, Texas (TX), USA

ULH Units-per-Labor Hour, production capacity from one person working one hour

UPH Units-per-Hour, production capacity from a the factory operating for one hour

Bibliography

[BBC10] D. Bertsimas, D.B. Brown, and C. Caramanis. Theory and applications of robust optimization. *Arxiv preprint arXiv:1010.5445*, 2010.

[BEdV04] S. Benjaafar, M. ElHafsi, and F. de Véricourt. Demand allocation in multiple-product, multiple-facility, make-to-stock systems. *Management Science*, pages 1431–1448, 2004.

[Ber05a] D. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1, section 6.1, pages 283–299. Athena Scientific, Nashua, New Hampshire, third edition, 2005.

[Ber05b] D. Bertsekas. Dynamic programming and suboptimal control: A survey from ADP to MPC. *European Journal of Control*, 11(4-5):310–334, 2005.

[BIP10] D. Bertsimas, D.A. Iancu, and P.A. Parrilo. Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research*, 35(2):363–394, 2010.

[BLXE08] S. Benjaafar, Y. Li, D. Xu, and S. Elhedhli. Demand allocation in systems with multiple inventory locations and multiple demand sources. *Manufacturing and Service Operations Management*, 10(1):43–60, 2008.

[BN04] S. Biswas and Y. Narahari. Object oriented modeling and decision support for supply chains. *European Journal of Operational Research*, 153(3):704–726, 2004.

[BPS04] D. Bertsimas, D. Pachamanova, and M. Sim. Robust linear optimization under general norms. *Operations Research Letters*, 32(6):510–516, 2004.

[BS04] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, pages 35–53, 2004.

[BT04] D. Bertsimas and A. Thiele. A robust optimization approach to supply chain management. *Integer Programming and Combinatorial Optimization*, pages 145–156, 2004.

[BTGGN04] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, 2004.

[BTGNV05] A. Ben-Tal, B. Golany, A. Nemirovski, and J.P. Vial. Retailer-supplier flexible commitments contracts: A robust optimization approach. *Manufacturing & Service Operations Management*, 7(3):248–271, 2005.

[BTN98] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, pages 769–805, 1998.

[BTN99] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–14, 1999.

[CHS02] S. Chand, V.N. Hsu, and S. Sethi. Forecast, solution, and rolling horizons in operations management problems: A classified bibliography. *Manufacturing & Service Operations Management*, 4(1):25–43, 2002.

[CL88] M.A. Cohen and H.L. Lee. Strategic analysis of integrated production-distribution systems: models and methods. *Operations Research*, 36(2):216–228, 1988.

[CSS07] X. Chen, M. Sim, and P. Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058, 2007.

[CV05] Z.L. Chen and G.L. Vairaktarakis. Integrated scheduling of production and distribution operations. *Management Science*, pages 614–628, 2005.

[CZB01] C.Y. Chen, Z.Y. Zhao, and M.O. Ball. Quantity and due date quoting available to promise. *Information Systems Frontiers*, 3(4):477–488, 2001.

[DF00] C. Dhaenens-Flipo. Spatial decomposition for a multi-facility production and distribution problem. *International Journal of Production Economics*, 64(1-3):177–186, 2000.

[DFF01] C. Dhaenens-Flipo and G. Finke. An integrated model for an industrial production-distribution problem. *IIE Transactions*, 33(9):705–715, 2001.

[DH06] E. Dolak and M. Hoag. Bursting effects on breaking the plant. Interview and Email Correspondence, November 2006.

[Dha08] N. Dhalla. Evaluating shortage costs in a dynamic environment. Master's thesis, Massachusetts Institute of Technology, 2008.

[Dol06] E. Dolak. Clarification for capacity study. Interview and Email Correspondence, December 2006.

[EGL97] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.

[EGOL$^+$98] L. El Ghaoui, F. Oustry, H. Lebret, et al. Robust solutions to uncertain semidefinite programs. *SIAM Journal of Optimization*, 9:33–52, 1998.

[Ein98] M. Einhorn. Managing forecast variability in a build-to-order environment. Master's thesis, Massachusetts Institute of Technology, 1998.

[ET03] F. Erhun and S. Tayur. Enterprise-wide optimization of total landed cost at a grocery retailer. *Operations Research*, pages 343–353, 2003.

[Fel06] J. Felch. Achievable production output. Interview and Email Correspondence, November 2006.

[Fen60] L. Fenton. The sum of log-normal probability distributions in scatter transmission systems. *Communications Systems, IRE Transactions on*, 8(1):57–67, 1960.

[FGA$^+$10] J. Foreman, J. Gallien, J. Alspaugh, F. Lopez, R. Bhatnagar, C.C. Teo, and C. Dubois. Implementing Supply-Routing Optimization in a Make-to-Order Manufacturing Network. *Manufacturing & Service Operations Management*, 12(4):547–568, 2010.

[FH06] R. Fearing and S. Holly. Interview on desktop staffing. Interview and Email Correspondence, December 2006.

[For08] J. Foreman. Optimized supply routing at dell under non-stationary demand. Master's thesis, Massachusetts Institute of Technology, 2008.

[GN05] A. Gunasekaran and E. W. T. Ngai. Build-to-order supply chain management: a literature review and framework for development. *Journal of Operations Management*, 23(5):423–451, 2005.

[Gra02] S.C. Graves. *Manufacturing planning and control*, pages 728–746. Handbook of applied optimization. Oxford University Press, New York, 2002.

[Gup08] K. Gupte. Impact of retail sales and outsourced manufacturing on a build-to-order supply chain. Master's thesis, Massachusetts Institute of Technology, 2008.

[HMMS60] C. Holt, F. Modigliani, J. Muth, and H. Simon. *Planning Production, Inventories, and Work Force.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1960.

[Hof09] L. Hoffman. The role of flexibility in configure-to-order manufacturing: a framework for managing variation. Master's thesis, Massachusetts Institute of Technology, 2009.

[HZ95] R. Hariharan and P. Zipkin. Customer-order information, leadtimes, and inventories. *Management Science*, pages 1599–1607, 1995.

[IDW03] A.V. Iyer, V. Deshpande, and Z. Wu. A postponement model for demand management. *Management Science*, pages 983–1002, 2003.

[JG95] W.C. Jordan and S.C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–594, 1995.

[Kah87] J.A. Kahn. Inventories and the Volatility of Production. *The American Economic Review*, 77(4):667–679, 1987.

[Kap69] A. Kaplan. Stock rationing. *Management Science*, 15(5):260–267, 1969.

[KMP88] D. Klingman, J. Mote, and N.V. Phillips. A logistics planning system at WR Grace. *Operations research*, 36(6):811–822, 1988.

[Lad09] K. Ladendorf. Dell closing its last large u.s. plant. `http://www.statesman.com/business/content/business/stories/technology/2009/10/08/1008Dell.html`, October 2009.

[Law] A. Law. *Simulation Modeling & Analysis*. McGraw-Hill, New York, N.Y., 4 edition.

[LS98] F.R. Lin and M.J. Shaw. Reengineering the order fulfillment process in supply chain networks. *International Journal of Flexible Manufacturing Systems*, 10(3):197–229, 1998.

[McN05] J.T. McNeil. Order Assignment and Resource Reservation: An Optimization Model and Policy Analysis. Master's thesis, University of Maryland, College Park, 2005.

[MGGP04] S. Moses, H. Grant, L. Gruenwald, and S. Pulat. Real-time due-date promising by build-to-order environments. *International journal of production research*, 42(20):4353–4375, 2004.

[MRRS00] D.Q. Mayne, J.B. Rawlings, C.V. Rao, and PO Scokaert. Constrained model predictive control: Stability and optimality. *Automatica, Oxford*, 36:789–814, 2000.

[MU11] H. Missbauer and R. Uzsoy. Optimization models of production planning problems. *Planning Production and Inventories in the Extended Enterprise*, pages 437–507, 2011.

[MWMZ07] N.B. Mehta, J. Wu, A.F. Molisch, and J. Zhang. Approximating a sum of random variables with a lognormal. *Wireless Communications, IEEE Transactions on*, 6(7):2690–2699, 2007.

[MZ96] L.J. Maccini and E. Zabel. Serial correlation in demand, backlogging and production volatility. *International Economic Review*, 37(2):423–452, 1996.

[PMM08] M. Paquet, A. Martel, and B. Montreuil. A manufacturing network design model based on processor and worker capabilities. *International Journal of Production Research*, 46(7):2009–2030, 2008.

[QB03] S.J. Qin and T.A. Badgwell. A survey of industrial model predictive control technology. *Control engineering practice*, 11(7):733–764, 2003.

[Rey06] A. Reyner. Multi-site inventory balancing in an extended global supply chain. Master's thesis, Massachusetts Institute of Technology, 2006.

[Sch08] J. Scheck. Dell plans to sell factories in effort to cut costs, September 2008.

[SHZZ02] S. Sethi, Y. Houmin, H. Zhang, and Q. Zhang. Optimal and hierarchical controls in dynamic stochastic manufacturing systems: A review. *Manufacturing and Service Operations Management*, 4(2):133–170, 2002.

[SN99] A.M. Sarmiento and R. Nagi. A review of integrated analysis of production–distribution systems. *IIE transactions*, 31(11):1061–1074, 1999.

[Soy73] A.L. Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973.

[SPP+98] E.A. Silver, D.F. Pyke, R. Peterson, et al. *Inventory management and production planning and scheduling*, volume 2. Wiley New York, NY, 1998.

[SS91] S. Sethi and G. Sorger. A theory of rolling horizon decision making. *Annals of Operations Research*, 29(1):387–415, 1991.

[SZ03] J.S. Song and P. Zipkin. Supply chain operations: Assemble-to-order systems. *Handbooks in Operations Research and Management Science*, 11:561–596, 2003.

[SZ07] K.E. Stecke and X. Zhao. Production and transportation integration for a make-to-order manufacturing company with a commit-to-delivery business mode. *Manufacturing & Service Operations Management*, 9(2):206–224, 2007.

[Top68] D.M. Topkis. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science*, 15(3):160–176, 1968.

[UHN10] F. Ul-Haq and M. Nadeem. Build-to-Order Supply Chain in Automotive Industry. Master's thesis, Jönköping International Business School, 2010.

[Vai04] T. Vainio. Intelligent order scheduling and release in a build to order environment. Master's thesis, Massachusetts Institute of Technology, 2004.

[VG97] C.J. Vidal and M. Goetschalckx. Strategic production-distribution models: A critical review with emphasis on global supply chain models. *European Journal of Operational Research*, 98(1):1–18, 1997.

[Wal04] B. Waller. Market responsive manufacturing for the automotive supply chain. *Journal of Manufacturing Technology Management*, 15(1):10–19, 2004.

[WG04] S.D. Wu and H. Golbasi. Multi-Item, multi-facility supply chain planning: Models, complexities, and algorithms. *Computational Optimization and Applications*, 28(3):325–356, 2004.

[WGP03] T. Wagner, V. Guralnik, and J. Phelps. TAEMS agents: Enabling dynamic distributed supply chain management. *Electronic Commerce Research and Applications*, 2(2):114–132, 2003.

[XAG06] P.J. Xu, R. Allgor, and S. Graves. The benefits of re-evaluating the real-time fulfillment decisions. *Manufacturing and Service Operations Management*, 8(1):104–107, 2006.

[Xu05] P.J. Xu. *Order fulfillment in online retailing: What goes where.* PhD thesis, Massachusetts Institute of Technology, 2005.

[ZKM91] M. Zuo, W. Kuo, and K.L. McRoberts. Application of mathematical programming to a large-scale agricultural production and distribution system. *Journal of the Operational Research Society*, pages 639–648, 1991.