2014

Multi-Echelon Inventory Optimization and Demand-Side Management: Models and Algorithms

Lin He
Lehigh University

Multi-Echelon Inventory Optimization and Demand-Side Management: Models and Algorithms

by

Lin He

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Industrial Engineering

Lehigh University

May 2014

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

_____
Dissertation Advisor

Committee Members:

_____
Larry Snyder, Committee Chair

_____
Gregory DeCroix

_____
Shalinee Kishore

_____
Katya Scheinberg

Acknowledgments

Over the past five and a half year, I received lots of support and encouragement from a great number of kind people. Hereby I thank all the great people who helped me to finish my disseration.

First and foremost, I would like to express the greatest appreciation to my advisor, and dissertation committee chair, Professor Larry Snyder, in providing me guidance and support from the beginning of my Ph.D. study to the final step. He always encourages me to overcome difficulties and challenges, supports me whenever there is setback, and guides me to explore unknown territories of knowledges. He is an amazing advisor and admirable role model for me, and he always inspires me to reach for a higher level.

I would like to thank my committee members, Professor Gregory DeCroix, Professor Shalinee Kishore and Professor Katya Scheinberg for providing insightful comments and constructive advices. Professor DeCroix provided me theoretical foundation and inspiration for my work on assembly system. Professor Kishore introduced me to smart grid, a brand new area through our collaboration. Professor Scheinberg gave me lots of suggestions in solving bilevel programming problem. Without their guidance and persistent help, this dissertation would not have been possible.

Lehigh is my favorite school. I owe it to all my Professors at Lehigh, espeically Tamás Terlaky, Ted Ralphs, Frank Curtis, Aurélie Thiele, and Imre Pólik. They gave me wonderful lectures, and I learnt a great deal from them. I'm also thanksful to Rita Frey and Kathy Rambo, who welcomed me with a big smile since my first day at Lehigh, and always helps me to go through all administration challenges.

Lehigh is my second home. I'm grateful to many dear friends at Lehigh for this. I'm

thankful to Dan Li, Jiadong Wang, Ying Bai, Hao Wang, Chen Chen, Choat Inthawongse, Xiaocheng Tang, Xi Bai, Gengyang Sun, Fang Chen, Ruobin Chen, Tengjiao Xiao, Yunfei Song, Yang Dong, and all my other friends who I am not able to list here. Those wonderful time and great experiences we had together will always be a cherishable memory to me. I'm also thankful to Charles Fisher, Aydin Gerek, Tom Necker and Joanne Gerontidis. Living with you guys brought lots of happiness to me too.

Last but not least, I thank my parents for the love and consistent support all these years. Thousands of miles away, I can still feel your accompany every day. I'm lucky to have you in my life all the time.

Contents

List of Tables

List of Figures

Abstract

Inventory management is a fudamental problem in supply chain management. It is widely used in practice, but it is also intrinsically hard to optimize, even for relatively simple inventory system structures. This challenge has also been heightened under the threat of supply disruptions. Whenever a supply source is disrupted, the inventory system is paralyzed, and tremenduous costs can occur as a consequence. Designing a reliable and robust inventory system that can withstand supply disruptions is vital for an inventory system's performance.

First we consider a basic type of inventory network, an assembly system, which produces a single end product from one or several components. A property called long-run balance allows an assembly system to be reduced to a serial system when disruptions are not present. We show that a modified version is still true under disruption risk. Based on this property, we propose a method for reducing the system into a serial system with extra inventory at certain stages that face supply disruptions. We also propose a heuristic for solving the reduced system. A numerical study shows that this heuristic performs very well, yielding significant cost savings when compared with the best-known algorithm.

Next we study another basic inventory network structure, a distribution system. We study continuous-review, multi-echelon distribution systems subject to supply disruptions, with Poisson customer demands under a first-come, first-served allocation policy. We develop a recursive optimization heuristic, which applies a bottom-up approach that sequentially approximates the base-stock levels of all the locations. Our numerical study shows that it performs very well.

Finally we consider a problem related to smart grids, an area where supply and demand

are still decisive factors. Instead of matching supply with demand, as in the first two parts of the dissertation, now we concentrate on the interaction between supply and demand. We consider an electricity service provider that wishes to set prices for a large customer (user or aggregator) with flexible loads so that the resulting load profile matches a predetermined profile as closely as possible. We model the deterministic demand case as a bilevel problem in which the service provider sets price coefficients and the customer responds by shifting loads forward in time. We derive optimality conditions for the lower-level problem to obtain a single-level problem that can be solved efficiently. For the stochastic-demand case, we approximate the consumer's best response function and use this approximation to calculate the service provider's optimal strategy. Our numerical study shows the tractability of the new models for both the deterministic and stochastic cases, and that our pricing scheme is very effective for the service provider to shape consumer demand.

Chapter 1

Introduction

Supply chains have expanded around the world along with the globalization of the economy. Supply chain have reached a greater level of complexity and vastness than ever before. This change has brought many advantages, as well as new challenges. Supply chains have become more vulnerable to disruptins, such as those caused by extreme weather, natural disasters, or labor strikes. Supply disruptions have brought a huge negative impact on the supply chain itself. Multi-echelon inventory systems are hard systems to analyze when compared to a single echelon system. The optimal inventory policy, and allocation policy if required, are still unknown for many types of multi-echelon inventory systems. Part of our research aims to contribute to the understanding of multi-echelon systems under the threat of disruptions by proposing simple and efficient algorithms which can approximate the optimal inventory policies for these systems.

In our first topic, we study assembly systems. These systems have been studied relatively thoroughly, and there are existing algorithms for finding the optimal base-stock levels for them. However, assembly systems subject to supply disruptions are a relatively less studied area. The optimal inventory policy is unknown, and no exact algorithm is available currently. We propose an inventory policy for assembly systems subject to supply disruptions, and develop a simple heurisitic algorithm for optimizing base-stock levels of this policy. The main idea behind our proposed inventory policy and algorithm is the non-equivalence of assembly systems and serial systems under disruptions. When disruptions are not present, it is well known that every assembly system is equivalent to a corresponding

serial system when operated optimally. But this equivalence is broken when the supply is under disruption risk. We study another property of assembly systems under disruption risk when operated optimally, which is similar to the original equivalence between assembly and serial systems. With this new property, we reduce the assembly system under disruption risk into an "almost serial" system, which has a simpler structure to facilitate our analysis. We develop a simple heuristic procedure for optimizing the base-stock levels in this "almost serial" system, to approximate the base-stock levels for the original assembly system under supply risk.

In our second topic, we study distribution systems subject to supply disruptions. Distribution systems are one of the most difficult inventory network topologies to analyze and most previous approaches in the existing literature are heuristics. For distribution systems without supply disruption risk, the optimal inventory and allocation policies are unknown; exact algorithms are computationally expensive; and most exisiting work focuses on a relatively simple structure, one warehouse multiple retailer (OWMR) system. On top of these difficulties, supply disruptions add another layer of difficulty. We consider distribution systems with more than just two echelons, and we also take supply disruptions into consideration. We analyze the effects of supply disruptions on inventory levels, and develop a heuristic algorithm utilizing these effects to optimize base-stock levels in a multi-echelon distribution system under supply risks, assuming a FIFO allocation policy and a base-stock inventory policy. The main idea behind our recursive algorithm is to analyze the effect of one stage's inventory level on its successor's inventory level, so that we can incorporate its successor's cost when calculating its base-stock level. This algorithm is easy to implement, and it also yields good performance.

In the last chapter, we turn our attention to energy systems. Instead of merely providing enough supply to meet the demand with as little cost as possible as in the first two parts of the dissertation, now we concentrate on the interaction between supply and demand. The previous two chapters study a problem in which items can be stored to satisify future demand, whereas in an electric grid, supply and demand have to be exactly matched at all times. Since electricity service providers purchase electricity from the day-ahead market, any deviation in load from this purchased amount might result in some extra costs. Thus

we focus on the problem of how an electricity service provider should set prices so that the consumer reacts to this price and the resulting load profile matches a predetermined profile as closely as possible. We study both deterministic and stochastic demand cases. We analyze the consumer's best response function, and use it to calculate the service provider's optimal strategy. Our numerical study shows the tractability of our new models for both cases.

Chapter 2

Management of an Assembly System Subject to Supply Disruptions

2.1 Introduction

2.1.1 Assembly Inventory Systems

Inventory systems widely exist in our daily life, supermarkets, online store warehouses, and factories. It has become an indispensable part of modern life. Inventory systems can take various structures. The most comon ones are serial systems, assembly systems, and distribution systems. In a *serial system*, each location has one customer and one supplier. The system is connected like a string of locations with an outside supplier in the beginning, and an outside customer at the end. In an *assembly system*, there can be several different suppliers for each location, each supplying a different subcomponent or raw material. And these subcomponents are assembled into components to satisfy customer demand. There is a single final product at the end location of the assembly system. In contrast to an assembly system, in a *distribution system*, there can be several different customers for each location, and only one supplier. It is the exact mirror image of an assembly system. All three network structures can represent the configuration of either a production system or a

transportation system. They are building blocks for more complicated network structures. Refer to Figures 2.1, 2.2 and 2.3 for examples of a serial system, assembly system and distribution system. The serial system has the simplest network structure among all three.

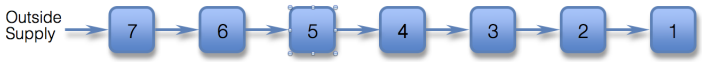

Figure 2.1: Example of a serial system

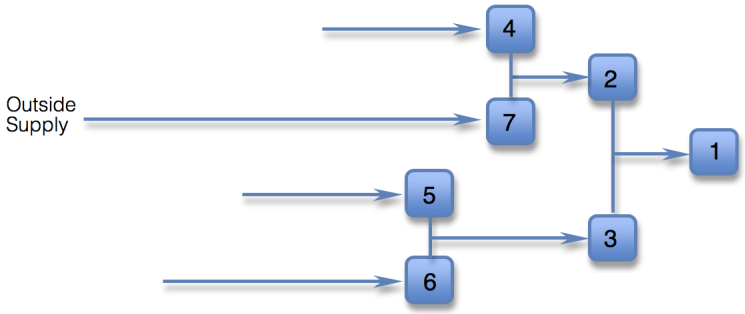

Figure 2.2: Example of an assembly system

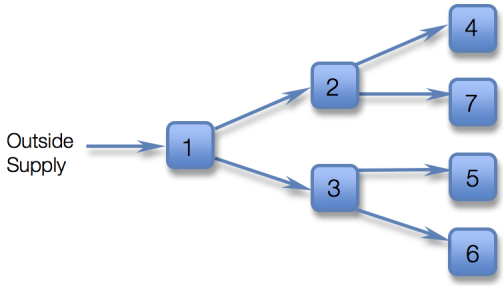

Figure 2.3: Example of a distribution system

It has been extensively studied and relatively well understood. Clark and Scarf [16] studied the problem of determining optimal policies in a multi-echelon model. For a serial system with N echelons, their algorithm recursively optimizes N nested convex functions. More research followed this algorithm. We provide a brief summary of the literature on serial inventory systems in section 2.2.2. van Houtum [84] provides an excellent review of multi-echelon inventory theory. We refer the reader to it for more details.

In this chapter, we consider an assembly system, where exactly one location receives demand from an outside source, and orders from multiple other locations. This location assembles components from other locations into final products to satisfy the outside demand. In order to produce one final product, this location needs one unit of each component from all of its suppliers. These other locations also order from their multiple suppliers, assemble

subcomponents from their respective suppliers into components to meet the demand from the final single location, and similarly for all other upstream locations. The most upstream locations order from outside sources which are assumed to have sufficient supply capacity. Hence, the goods are assembled during their flow from upstream to downstream. In contrast, information flows in the opposite direction, from downstream to upstream. We call a location that satisfies the demand of location i a predecessor of i, and the location that places an order to i as the successor of i. In an assembly system, each location has only one successor and has at least one predecessor. The outside demand in each period occuring at the end location is assumed to be a random variable.

One issue for inventory systems with multiple locations is the control mechanism. It is mostly about information: which location knows what information, when do they know it, and how do they use it. Typically, there are two types of control mechanisms, *centralized* control and *decentralized* control. In centralized control, there is only one decision maker who knows all the information, and makes all the decisions for every location. It could be either an outside entity or one of the locations in the inventory system. The other mechanism is decentralized control, in which every participants is its own decision maker. Each location in the supply chain determines when to order, and how much to order for itself. In our work, we assume centralized control for this assembly system, as in a factory, it is more common that the manufacturing of every component is under centralized control.

2.1.2 Supply Disruptions

As companies have increasingly started sourcing globally, supply chains have become more widely spread around the globe. This benefits companies in various ways, but it also increases the risk of possible supply disruptions. Despite how much effort is spent on keeping the supply chain functioning normally, supply disruptions inevitably happen from time to time. Whenever a disruption happens, it has a major impact on the whole supply chain. For example, when a massive earthquake and tsunami happened in Japan in 2011 [43][77], the Sendai Nikon plant was severely damaged in this disaster. The plant closure had an impact directly on camera and lens production, as well as further afield. To deal with this, Nikon had to react proactively. It shifted this production to Notion VTEC to resume its

normal operations with minimum delay. There are also countless other examples showing the profound impact of supply disruptions. This shows the importance for companies to have better strategies in preparing for supply disruptions.

There are three typical categories of supply uncertainty. The first one is *supply disruptions*, such as the Nikon's production in the Japan 2011 earthquake and tsuanami. When a company's supply is disrupted, its supply process comes to a halt, or it won't receive any new items that were supposed to be delivered until the supply disruption process is completely over. The second one is *yield uncertainty*. It means the actual amount of items delivered by the supplier could be a random number dependent on the ordered quantity. For example, among each batch of products delivered, some products might have defects, which makes them unwanted. The number of defective products could be a random variable. The third form of supply uncertainty is *leadtime uncertainty*. When the leadtime is stochastic, the delivery of products takes a random amount of time, but the exact amount of products ordered would arrive. For our study, we only consider the first type, supply disruptions.

Disruptions can happen routinely due to various reasons such as natural disasters, machine breakdowns, labor strikes and so on. Disruptions could either halt the transportation process, or cause malfunction in the production process in some facilities. Most inventory models can handle demand uncertainty relatively well. However, they do not provide same level of protection against supply uncertainty, especially supply disruptions. Even though supply disruptions can not be prevented, it is still cruicial for companies to plan ahead, so that possible damage is minimized. There are different tactics to mitigate supply disruptions as summarized in Tomlin [82].

- *Passive Acceptance:* This is the default strategy against supply disruption risks. The company passively accepts supply risks, and it sources from the unreliable supplier exclusively without any extra inventory to protect against supply disruptions.

- *Inventory Mitigation:* The company sources from the unreliable supplier exclusively, but it carries some extra inventory to mitigate supply disruptions, in addition to the inventory for regular demand.

- *Sourcing Mitigation:* The company sources exclusively from a reliable supplier. Then

there is no risk of supply disruptions.

- *Contingent Rerouting:* The company sources exclusively from an unrelialbe supplier when it is available, and it carries no extra inventory to mitigate supply disruptions in addition to the inventory for regular demand. When the supplier becomes unavailable due to supply disruptions, it reroutes to a reliable supplier to replenish its inventory.

- *Mixed Strategy:* The company could choose a combination of tactics to mitigate the effect of supply disruptions. For example, the company could source exclusively from an unrelialbe supplier when it is available, and carry some extra inventory to mitigate supply disruptions in addition to the inventory for regular demand. It could also reroute to a reliable supplier to replenish its inventory during a supply disruption.

The best tactic for each company is determined by the nature of the disruptions, as well as the company's goals. In this work we only focus on the use of inventory mitigation to tackle supply disruptions. With sufficient inventory on hand, excess stockout cost, expediting cost, and loss of goodwill cost can be reduced, but too much inventory can also cost a fortune. It is necessary to reach a balance between stocking enough inventory to protect against supply disruptions and keeping a low inventory level to save cost.

In this work, we address this issue, for an assembly system with a single end product, for which one or several components, or the end product itself, is under supply risk. Disruption of assembly components or subassemblies could be a worse problem than that of an end product to be sold at a retail store. In an assembly system, every component is necessary, and a shortage of one component could shut down the whole manufacturing system. It can not only disrupt final product delivery, but also incur extra cost for other components or subassemblies as they would have no option but to wait. It would be beneficial to find a good inventory policy to deal with disruption risks in assembly systems.

We build a model of an assembly system that produces one final end product to satisfy stochastic outside demand, with some locations in the system under supply risk. If there is no disruption and every supplier is perfectly reliable, it is well known that the system can be reduced to an equivalent serial system (Rosling [65]). However, this simple and exact equivalence does not hold if there is a potential supply disruption (DeCroix [18]). The

presence of disruptions interferes with a key concept called long-run balance. Inspired by the violation of long-run balance, we explore what other properties hold for the optimal policy. The new property we identify suggests a way to reduce the assembly system into a serial system with some extra stages. Even though the original system and the reduced system are not strictly equivalent, this reduced system still provides us with an easier approximate way to deal with the assembly system. We propose an inventory policy for the original system based on the reduced system. We also propose an algorithm for this inventory policy based on the classical recursive algorithm for serial systems, so that an approximation of the inventory policy parameters for the original assembly system can be obtained easily. We test it on a set of instances, and find it yields good solutions, with significant cost savings achieved on the examples tested.

2.2 Literature Review

In this section, we review the existing work on serial systems, assembly systems, and inventory models with supply risks.

2.2.1 Serial Systems

A serial system is the simplest network structure among all multi-echelon systems. We want to determine optimal inventory policies for all locations so that the average total inventory cost is minimized. This research was initiated by Clark and Scarf [16]. The authors consider a serial system with stochastic demand but no supply disruptions under periodic review with constant leadtime. They determine the optimality of order-up-to (base-stock) policies based on echelon stock calculations to minimize the total inventory cost, and propose a recursive algorithm for computing the optimal base-stock levels. They argue that the optimal ordering decisions are such that they keep the echelon inventory position at a constant level. The echelon inventory position is the amount obtained by adding the local inventory positions at the location and all its downstream locations. Axsater and Rosling [7] compare local and echelon inventory policies for multi-echelon serial systems. They argue that a local stock policy can always be replaced by an equivalent echelon stock policy.

Following the study by Clark and Scarf [16], many other researchers have contributed to serial inventory theory. Federgruen and Zipkin [23] extend the result of Clarf and Scarf [16] to the infinite horizon case, for both discounted costs, as well as average costs. They show that the infinite horizon problem is much easier to solve compared to the finite horizon problem. The authors also study normally distributed demand, and acheive further simplifications. Chen and Zheng [14] establish a lower bound and evaluate its performance for a one-warehouse multi-retailer system by comparing it with simple, heuristic policies. It proves to be good for small sized problems. This paper also offers a simplified proof for the optimality results of serial and assembly systems.

The algorithm of Clark and Scarf [16] calculates the exact base-stock levels for serial systems. It requires minimizing N nested convex functions recursively. It can only be implemented numerically, and closed form solutions are not available. Shang and Song [73] develop a simple but accurate heuristic to identify the key determinants of the optimal policy. Their heuristic minimizes the lower and upper bounds functions for the echelon cost functions, and the minimizer forms the upper and lower bounds for the optimal solutions. Their heuristic takes the simple average of the solution bounds, but the average error of the heuristics is surprisingly good, with only 0.24% gap.

2.2.2 Assembly Systems

Schmidt and Nahmias [71] consider an assembly system where two components are assembled into one final product. They find the optimal ordering policy for both the components and the end product, using dynamic programming. The optimal policy has a complex structure, where the optimal order for one componenet depends on the inventory status of the other.

Rosling [65] studies a general assembly system over an infinite horizon. He shows that the system should satisfy a condition called long-run balance when operated optimally. This condition allows the system to be reduced to an equivalent serial system. The resulting serial system can be solved optimally with the algorithms developed by Clark and Scarf [16], Federgruen and Zipkin [23], Chen and Zheng [14].

Bollapragada et al. [9] study an assembly system with local base-stock policies where

component suppliers has random production capacity. Their model minimizes the cost under service onstraints. They propose a decomposition approach to find near-optimal base-stock levels, and reaches only 0.66% average error across the instances tested.

2.2.3 Supply Disruption

There are many papers on supply disruption. We briefly review the exisiting literature on inventory models under supply disruption risks. They can be categorized into two groups: single supplier models and multiple supplier models.

Single Supplier Models

Single supplier models assume there is only one unreliable supplier to order from, and no sourcing mitigation is available. Meyer et al. [53] are the first to consider supply disruptions. They consider a single stage production/storage system facing constant demand, and subject to stochastic failure and repair. They formulate the average inventory level in the storage tank for the general case, as well as Poisson failures with exponential repair times, and Poisson failures with constant repair times. Posner et al. [62] consider the same problem with a compound Poisson demand process, and derive an explicit closed form solution for the steady-state distribution of the inventory level.

Parlar et al. [61] model the disruption as a two-state continuous time Markov chain. They determine the optimal reorder point, as well as the optimal order quantity. In their work, it was assumed that the state of the system was identified at a cost, so how long to wait before the next order during the off state is another decision variable. The objective cost function is constructed by the renewal reward theorem. This paper also considers a random yield problem.

Ross et al. [66] considers a problem with time dependent disruption probability, as well as time dependent demand. It is modeled as a two-dimensional non-homogeneous continuous-time Markov chain, and solved numerically to evaluate the total cost of different ordering policies, some of which are time dependent while others are not. They compare the proposed policies under different cost, demand and disruption parameter settings. They find non-stationary policies have a better balance between cost and robustness.

Moinzadeh et al. [56] study a system with a constant production rate and demand rate, but subject to random disruption. The time interval between disruptions is exponentially distributed, and there is positive production setup cost and/or setup time. The authors propose a procedure to find optimal values for an (s, S) policy. This work also indicates that setup cost reduction is more effective when the system is more reliable, and setup cost reduction in an unreliable system would lead to higher safety stock level.

Liu and Cao [51] examine a production-inventory system with a compound Poisson process and general demand size distribution. They suggest one condition to ensure that the steady state distribution of the inventory process exists, and derive an expression for it, then compute the cost for exponentially distributed demand sizes.

Gupta [32] studies a (Q, r) model with Poisson demand and with exponentially distributed lengths of on and off periods. Unmet demands are lost. He analyzes two cases: one with negligible leadtime but arbitrary number of size-Q outstanding orders; and another with constant leadtime but at most one outstanding order. The author formulates an exact expression to minimize total cost. His computation indictates that ignoring supply uncertainty or approximate modeling could be costly.

Mohebbi [54] considers a continuous-review inventory system with compound Poisson demand and Erlang distributed leadtime under loast sales. He calculates exact analytical expressions for the case when demand sizes are exponentially distributed.

Parlar [60] considers a continuous-review inventory model with random demand and random leadtime. The supplier availability is determined by an alernating renewal process. The author develops the average cost objective function using the renewal reward theorem by identifying regnerative cycles of the inventory position process. An algorithm is provided to find the optimal q and r for the (q, r) policy.

Arreola-Risa et al. [2] explore a problem where unmet demand is partially lost and partially backordered. The authors apply an (s, S) policy, and propose an algorithm to compute optimal values of the policy parameters. They also show how the optimal policy parameters would change as the severity of the supply disruption changes, or the behavior of unmet demand changes.

Özekici and Parlar [57] consider an infinite-horizon periodic-review inventory model with

unreliable suppliers where the demand, supply and cost parameters change randomly. Their study shows that an environment dependent base-stock policy is optimal when there is no fixed ordering cost, and a two-parameter environment-dependent (s, S) policy is optimal under some conditions.

Güllu et al. [31] study a periodic review inventory model with deterministic dynamic demand and nonstationary supply unavailability. The authors show the optimality of an order-up-to policy, and provide a newvendor-like formula to calculate the order-up-to level.

Li et al. [50] investigate a periodic review model with random demand and unreliable supply. Both the lost sales case and the backorder case are studied, for the discounted cost criteria, as well as the long-run average cost. They derive structural properties and bounds on the optimal policy for the linear cost model.

Multiple Supplier Models

Parlar and Perry [59] analyze (Q, r) models with supply uncertainty with single and multiple suppliers. The authors calculate the average cost objective function for the case of single and multiple suppliers with concepts from renewal reward processes. They also show that as the number of suppliers increases, the model reduces to the classical EOQ model.

Tomlin and Wang [83] consider a multiple product setting in which a company can invest in product-dedicated resources and totally flexible resources. The authors consider four different strategies: a single-source dedicated strategy, a single-source flexible strategy, a dual-source dedicated, and a dual-source flexible strategy. They investigate how product portfolio, resources, and the firm influence the design strategy through a numerical study. They also show that dual sourcing is preferred when supply chain reliability decreases.

Tomlin [82] studies a single-product setting in which a company can source from one unreliable supplier and one reliable but more expensive supplier. He finds that a supplier's disruption profile-the percentage uptime-disruption length-the uptime length, are key factors in deciding the optimal strategy. He also shows that a mixed mitigation strategy, i.e. partial sourcing from the reliable supplier and carrying inventory, is optimal if the firm is risk averse, or the unreliable supplier only has finite capacity. He also discusses the conditions under which contingent rerouting is optimal.

Dada et al. [17] consider a single period newsvendor problem with multiple suppliers under supply risks. With some probability, unreliable suppliers can deliver an amount strictly less than the amount ordered. The authors show that cost takes precedence over reliability, which means a given supplier is chosen only if all less expensive suppliers are chosen, even if this supplier is more reliable. But the relative size of orders to a given supplier depends on its reliability.

Chopra et al. [15] study a model where a company has a perfectly reliable suplier, and another supplier subject to both random yield and disruption risk. The authors argue the importance of planning for the right forms of supply uncertainty that a company faces. They show that if the supply uncertainty mostly comes from yield uncertainty, the company should order more from the unreliable supplier to acheive lower cost; and if the supply uncertainty mostly comes from disruptions, it is better to order more from the reliable source.

Schmitt and Snyder [72] consider a firm facing both supply disruptions and yield uncertainty. The authors argue the importance of analyzing inventory models under supply risk for a sufficiently long time horizon. They study a problem in which a firm has an unreliable supplier which could be completely disrupted, and is also subject to yield uncertainty. They consider one case where only one unreliable supplier exists, and a second case where a second reliable but more expensive supplier is available. They develop models for both cases and compare the results to those found when a single-period approximation is used. The results demonstrate that a single-period approximation is not accurate, as it causes increases in cost and under-utilizes the unreliable supplier.

Snyder et al. [76] provides an excellent review on supply chain disruptions. They summarize the existing work into six different categories: evaluating supply disruptions, strategic decisions, sourcing decisions, contracts and incentives, inventory, and facility location.

This work is built on the literature on optimal inventory policies for assembly systems with random demand but no supply disruptions. There are several relevant works on this topic. Rosling [65] showed that an assembly system can be reduced to an equivalent serial system for stochastic demand with no supply disruptions. He introduced the notion of long-run balance, which ensure that components arrive at the assembly point in a matched

way, so nothing is left over. It is optimal to preserve this long-run balance under an optimal inventory policy. DeCroix [18] explained why the assembly system cannot be reduced to a serial system if disruptions are present, and he proposed a method to replace some subsystems of an assembly system by a series structure, as well as a heuristic to solve this problem. Clark & Scarf [16] initiated a recursive algorithm for computing the optimal base-stock levels for a serial system with stochastic demand but no supply disruptions. Federgruen & Zipkin [23] and Chen & Zheng [14] kept working on this topics. Zipkin [88] provides a more detailed description of the recursive algorithm and these results.

Our work is closest to that of DeCroix [18]. We consider inventory policies for a periodic review assembly system with stochastic demand under supply risk. We explore why this system can't be reduced to a serial system, and what implication it has for the optimal inventory policy. We build our work on the reason for this non-equivalence, and propose a new policy to deal with it. Based on this policy, we propose a method to partially reduce the assembly system into a serial system. Also, a heuristic algorithm is suggested for determining the parameters of the reduced system.

2.3 Model Basics

We consider an assembly network with N stages which are indexed by $i = 1, 2, \cdots, N$.Stage 1 is the final product. Each stage i orders from its predecessors, or assembles from subcomponents obtained from its predecessors, and meets the demand from its successor. Each stage i might have one immediate successor which is denoted as $\mathcal{S}(i)$, and all of its downstream stages are denoted by the set $\mathcal{A}(i)$. It might also have multiple immediate predecessors denoted as $\mathcal{P}(i)$. Stages without predecessors order from outside suppliers. Without loss of generality, we assume each stage i requires exactly 1 component from each of its immediate predecessors to produce item i. Stage i requires a leadtime l_i for the delivery to stage i, or assembly of item i.

This work considers a periodic review case, and in each period t, the system sees a stochastic demand D_t for the end product. Assume the random demand is stationary over time. In each period, the unmet demand for the end product is backordered with a cost of

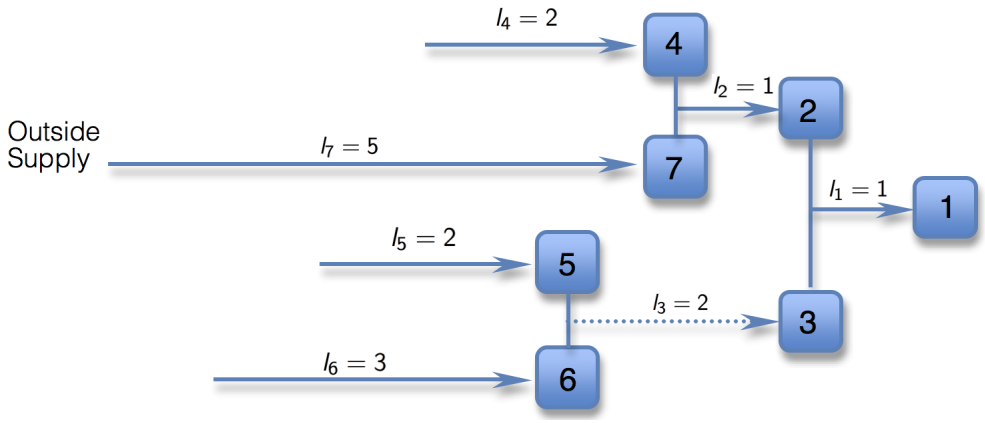

Figure 2.4: Example of general assembly system

p per item per period. In addition, every unit of item i in the system, either in transit from i, or in inventory at stage i, incurs a cost of H_i per period. Define the echelon holding cost as the extra cost incurred when predecessor items are assembled, i.e. $h_i = H_i - \sum_{k \in \mathcal{P}(i)} H_k$. There is no fixed ordering cost in this model.

In this system, a subset J of stages are subject to stochastic supply disruptions among their predecessors. In the beginning of each period, the state of supply availability for each $i \in J$ is examined. If the suppler of any stage i is unavailable, stage i can not place any new order, or the assembly process of new units of item i cannot be initiated. Previously shipped items or items already in the assembly process are not affected, and stage i's predecessors can still observe the whole system's state to make their own decisions. The disruption status for i's suppliers is governed by a two-state discrete-time Markov chain. For stage i, its supply disruption status turns from available in the current period to unavailable in the next period with probability β_i. If it is unavailable, it becomes available in the next period with probability γ_i. The probability of being disrupted for k consecutive periods is denoted as π_k^i. For the sake of simplicity, this work only focuses on the case with only one unreliable supplier. The case with multiple unreliable suppliers would be the same, only differs in some notation and conditions.

The events in each period happen in the following order: (1) pipeline orders arrive; (2) the state of the system is observed; (3) ordering decisions are made; (4) customer demand is observed; (5) costs are charged. See Figure 2.5.

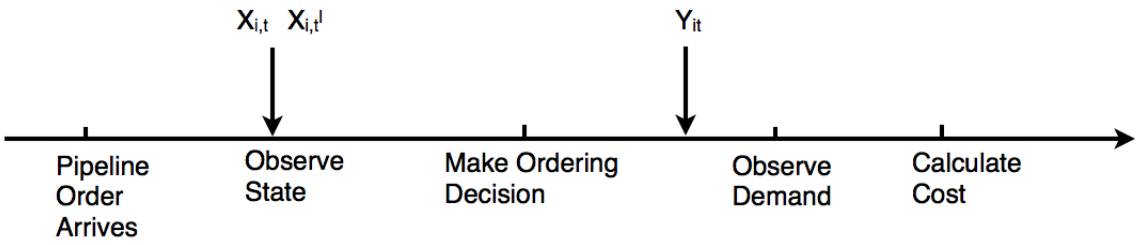

Figure 2.5: Sequence of events

In this work, we use the concept of echelon inventory, which is basically the same as it is in serial systems. I_i includes the inventory that is stored at stage i and all of its successors, and it also includes the unit i that has already been assembled into other product.The notation used in this model, which is based on the notation of Rosling [65], is as follows:

M_i : total leadtime for item i and all its successors: $M_i = l_i + \sum_{k \in \mathcal{A}(i)} l_k, \forall i = 1, 2, \cdots, N$. We assume the stages are indexed in a way such that $M_i \geq M_{i-1}, \forall i$;

L_i : equivalent leadtime for item i: $L_i = M_i - M_{i-1}$;

s_i : echelon base-stock level of item i for the state when every supplier is available;

$X_{i,t}$: echelon inventory position of item i in period t after pipeline order arrives and before the ordering decision is made;

$Y_{i,t}$: echelon inventory position of item i in period t after the ordering decision is made;

$X^l_{i,t}$: echelon on hand inventory of item i in period t after pipeline order arrives and before the ordering decision is made:

$$X^l_{i,t} = Y_{i,t-l_i} - \sum_{k=t-l_i}^{t-1} D_k$$

$X^L_{i,t}$: echelon inventory position of item i in period t ordered L_i periods ago or earlier:

$$X^L_{i,t} = Y_{i,t-L_i} - \sum_{k=t-L_i}^{t-1} D_k$$

$X'_{i,t}$: local on hand inventory of extra item i, which is not assembled due to shortage of other items, and is stored at stage i in period t after pipeline order arrives and before the customer demand is observed, $\forall i \in J$. $X'_{i,t}$ is included in $X_{i,t}$;

α : discount rate for cost in each period.

Here we know that $X_{i,t} = Y_{i,t-1} - D_t$ as the echelon inventory faces outside demand directly, and $Y_{i,t} \geq X_{i,t}$, as negative orders are not allowed. Also, the assembly ordering decision for item i is constrained by its immediate predecessors, $Y_{i,t} \leq X_{j,t}$, $\forall j \in \mathcal{P}(i)$. When the supplier is available, the system functions the same way as an assembly system without supply risk.

Our goal is to find a policy that minimizes the expected inventory cost per period over an infinite horizon. In each period, the inventory level of the end product 1 is $X_{1,t} - D_t$, so there is inventory holding cost $H_i(X_{1,t} - D_t)^+$, as well as backorder cost $p(X^l_{1,t} - D_t)^-$, for item 1. For other stages, the amount of item i at stage i is $X_{i,t} - D_t$, so the respective holding cost is $h_i(X_{i,t} - D_t)$. Summing up all costs in terms of echelon holding cost, and following similar computations about echelon inventory cost and local inventory cost as in Zipkin [88], we have the following total cost in period t:

$$\sum_{i=1}^{N} [h_i(X_{i,t} - D_t)] + (p + H_1)[X_{1,t} - D_t]^-$$

Let $\bar{Y}_{i,t}$ denote an upper bound for $Y_{i,t}$, the echelon inventory position for item i in period t after ordering decisions are made. Let K be the set of stages that order directly from outside suppliers. There are different values for $\bar{Y}_{i,t}$ for different disruption scenarios:

$$\bar{Y}_{i,t} = \begin{cases} \infty & \text{if supply is available in period } t, \text{ and } i \in K \\ \min_{k \in \mathcal{P}(i)}\{X_{k,t}\} & \text{if supply is available in period } t, \text{ and } i \notin K \\ X_{i,t} & \text{if supply is unavailable in period } t \end{cases}$$

Then taking expected values over D_t, and summing over all periods, the **assembly problem** formulation is as follows:

$$\begin{aligned} \min_{Y_{i,t}} \quad & E\left[\sum_{t=1}^{\infty} \alpha^{t-1} \left(\sum_{i=1}^{N} [h_i(X_{i,t} - D_t)] + (p + H_1)[X_{1,t} - D_t]^- \right)\right] \\ \text{s.t.} \quad & X_{i,t} \leq Y_{i,t} \leq \bar{Y}_{i,t} \end{aligned}$$

Here we assume that $h_i > 0$, ensuring the holding costs financially and physically increase from upstream to downstream as product goes through the assembly system. We also assume $\sum_{i=1}^{N} h_i < p + H_1$, which ensures it is always cheaper to hold inventory rather than to stock out.

2.4 Policy Properties

The assembly system as formulated in the previous section is computationally intractable. The expectation is very hard to calculate, and the constraint involves random variables. One possible approach is dynamic programming. However, the state space is trememduous, since it would be necessary to keep track of all order shipments, as well as disruption status. Curse of dimentionality would make this problem prohibitively difficult to solve. As discussed above, Rosling [65] solves the problem of assembly systems without disruptions. He shows that the assembly system is equivalent to a serial system when operated optimally, and each stage follows a base-stock policy. The algorithm to find the base-stock levels of serial system is already well-known. This greatly simplifies the problem of assembly systems without disruptions. If a similar equivalence or modification could be applied to the assembly system with supply disruptions, it could facilitate the analysis a great deal.

In Rosling's work, there is a key concept called *long-run balance* which leads to the equivalence between assembly systems and serial systems. Define $X_{i,t}^{M-\mu}$ to be the echelon inventory position of item i at time t that is ordered $M_i - \mu$ periods ago or earlier:

$$X_{i,t}^{M-\mu} = Y_{i,t-(M_i-\mu)} - \sum_{k=t-(M_i-\mu)}^{t-1} D_k$$

The original *long-run balance* in Rosling [65] is defined as follows: item i is in long-run balance in period t if

$$X_{i,t}^{M-\mu} \leq X_{i+1,t}^{M-\mu}, \quad \forall \mu = 0, 1, \cdots, M_i - 1.$$

If this condition holds for all $i = 1, 2, \cdots, N-1$, we say that the system is in long-run balance. This concept says that in long-run balance the inventory positions equally close

to the end item increase with i, i.e., with total leadtime. This means that for item i and j in the assembly system, if i is upstream from j, then there should be at least as many units of item i as item j, since item i is cheaper and also takes a longer time to obtain. Moreover, even though long-run balance does not state this explicitly, item i and $i+1$ should be ordered in a perfectly matched way when the system is operated optimally. This means the same amount of both items should reach their common successor in the same period, since any extra units would incur some unnecessary cost. According to Rosling's argument, the optimal inventory policy for an assembly system without disruptions would follow this long-run balance, and as a consequence, the assembly system is equivalent to a serial system.

However, this might not be true for some items if the assembly system is under disruption risk. To demonstrate this, consider the three-stage system in Figure 2.6.

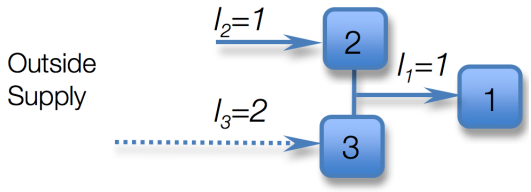

Figure 2.6: Assembly system example

Stage 1 needs both item 2 and item 3 to assemble item 1, while stages 2 and 3 order from outside suppliers. If we want to keep the long-run balance for stage 2 and 3 for $\mu = M_2$, we would require:

$$Y_{2,t} = X_{2,t}^{M-M_2} \leq X_{3,t}^{M-M_2} = Y_{3,t-(M_3-M_2)} - \sum_{k=t-(M_3-M_2)}^{t-1} D_k = Y_{3,t-1} - D_{t-1}$$

This implies that the inventory position of item 2 should never exceed the inventory position of item 3 ordered 1 period ago. Otherwise, there would be some extra units of item 3, as no more units of item 2 would arrive during the same period. Without disruptions, it would be optimal to keep this long run balance as it would prevent storing extra units. But under disruption risk, this might not hold. For the case when stage 2 is disrupted and cannot meet the demand from stage 1, it might be useful to have some extra units of item 2 at

stage 2. When a disruption happens, stage 2 might not place any new order to its supplier, but it could satisfy the demand from stage 1 using this extra units, so that stage 1 could keep assembling item 1 as it can still receive components from stage 3. In this case, $Y_{2,t}$ could be larger than $X_{3,t}^{M-M_2}$, which violates long-run balance.

Due to the lack of long-run balance, an assembly system subject to supply disruptions cannot be reduced into an equivalent serial system. However, this concept still provides us some insight on how to find good solutions for the base-stock levels. Although the optimal policy does not preserve long-run balance for the entire system, the optimal policy still preserves *item-specific* long-run balance for some stages, as stated in the following Proposition.

Proposition 2.1 (DeCroix [18]): For each item i such that $\{\{i\} \cup \mathcal{A}(i)\} \cap J = \emptyset$, the optimal ordering policy in each period satisfies $Y_{i,t} \leq \max\left\{X_{k,t}^{M-M_i}, X_{i,t}\right\}$ for all $k > i$. For each item i such that $\{\{i\} \cup A\} \cap J \neq \emptyset$, the optimal ordering policy in each period satisfies $Y_{i,t} \leq \max\left\{X_{j,t}^{M-M_i}, X_{i,t}\right\}, \forall j > i$, such that $\max\{k :\in \{\{i\} \cup \mathcal{A}(i)\} \cap J\} = \max\{k :\in \{\{j\} \cup A(j)\} \cap J\}$.

Proof. See DeCroix [18]. □

This Proposition states that item-specific long-run balance is preserved by the optimal ordering policy for stages unaffected by disruptions. As for stages affected by disruptions, the optimal policy preserves iterm-specific long-run balance relative to other stages which are affected by disruptions at the same downstream stages. DeCroix [18] introduces partial reduction of the assembly system based on this partial preservation of long-run balance.

Our policy differs from DeCroix [18]'s by recognizing that, as explained previously, it might be cheaper to store some extra units i at stage i, if stage i is under supply disruption risk. Consider the system in Figure 2.6. When a disruption happens to stage 2's supplier, stage 2 stops ordering new items, while stage 3 is functioning normally. Stage 1 could still utilize the extra units of item 2 to continue the assembly process to meet the demand, and this extra inventory of item 2 would not be replenished until the disruption is over. The following Proposition characterizes the condition under which it is optimal to hold this extra inventory at stages with supply disruption risk.

Proposition 2.2: For stage i facing supply disruptions, there exists a cost p_i such that it is optimal to keep some extra units of i at stage i if and only if

$$\sum_{k=0}^{\infty} \pi_k^i F^{k+1}(s_i) H_j < \sum_{k=1}^{\infty} \pi_k^i [1 - F^{k+1}(s_i)] p_i;$$

where F^k is the cdf of k consecutive periods' worth of demand, and p_i is a cost related to the shortage of one unit i for one unit time, which is dependant on s_i. Moreover, it is optimal for this extra inventory to follow a base-stock policy.

Proof. Assume the cost incurred by a shortage of item i is p_i per item per unit time. As stage i faces supply disruption risks, it has two choices, either to store some extra units i in addition to s_i or not. If stage i stores one extra unit of item i, stage i will not be disrupted with probability π_0^i, and the extra unit of i will be not used with a probability of $F^1(s_i)$; with probability π_1^i, stage i will be disrupted for one period, and the extra unit of i will be not used with a probability of $F^2(s_i)$; with probability π_2^i, stage i will be disrupted for two periods, and the extra unit of i will be not used with probability $F^3(s_i)$; and so on. So the expected probability of holding this unit is $\sum_{k=0}^{\infty} \pi_k^i F^{k+1}(s_i)$. The expected cost of holding one extra unit is $\sum_{k=0}^{\infty} \pi_k^i F^{k+1}(s_i) H_i$. Similarly, with probability $\sum_{k=1}^{\infty} \pi_k^i [1 - F^{k+1}(s_i)]$, there will be a demand for the extra unit of i. The expected cost for not holding one extra unit is $\sum_{k=1}^{\infty} \pi_k^i [1 - F^k(s_i)] p_i$. So if

$$\sum_{k=0}^{\infty} \pi_k^i F^{k+1}(s_i) H_i < \sum_{k=1}^{\infty} \pi_k^i [1 - F^{k+1}(s_i)] p_i,$$

it is better to accept the stockout, and otherwise it is optimal to carry some extra units of item i.

If the condition holds, it would be better for stage i to hold some extra units of item i. We denote D^k as k consecutive periods' worth of demand. For the extra amount of inventory i, it would only be used when demand is greater than s_i. With probability $\pi_0^i F^1(s_i)$, there will be a demand of $D^1 - s_i$ for this extra inventory i; with probability $\pi_1^i F^2(s_i)$, there will be a demand of $D^2 - s_i$ for it; with probability $\pi_2^i F^3(s_i)$, there will be a demand of $D^3 - s_i$ for it; similarly with probability $\pi_n^i F^{n+1}(s_i)$, there will be a demand of $D^{n+1} - s_i$ for it, and

so on. So this is exactly the same as a single stage system with supply disruptions, except the probabilities are different. The optimality of base-stock policy for this type of system is well known. □

We denote the base-stock level for the extra inventory of item i as s'_i. This extra safety inventory is replenished whenever the disruption is over and the unreliable supplier has enough inventory to replenish. Once the supplier is disrupted and no more new units are coming in, it is utilized to satisfy demand from downstream.

With the possible extra units to protect against supply disruptions, long-run balance does not necessarily hold when the system is operated optimally. But if the amount of extra units is excluded from consideration, everything else functions in a way such that long-run balance is preserved, just as though there is no supply uncertainty. We call this *generalized item-specific long-run balance*, which is stated in the following Proposition:

Proposition 2.3: For any item i, it is optimal to stay in generalized item-specific long run balance:

$$X_{i,t}^{M-\mu} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t} \le X_{k,t}^{M-\mu} - \sum_{j\in A(k)\cup\{k\}} X'_{j,t}, \quad \forall k > i,\, \forall \mu = 0, 1, \cdots, M_i - 1.$$

Proof. The amount of extra units of item i stored at stage i is $X'_{i,t}$, and the amount of all possible extra inventories of stage i's downstream stages is $\sum_{j\in A(i)} X'_{j,t}$. So $X_{i,t}^{M-\mu} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t}$ could be considered as the amount of echelon inventory that is flowing in the system since $M_i - \mu$ periods ago, while all the extra inventory still stays in the system.

During the period when stage i and k's supply is available, they both function the same as the case with no disruptions. This is because $X_{i,t}^{M-\mu} = Y_{i,t-(M_i-\mu)} - \sum_{j=t-(M_i-\mu)}^{t-1} D_j$, and $Y_{i,t-(M_i-\mu)}$ is not affected as there is no supply disruption, and D_j is independant of the system. So the system satisfies long-run balance, if the extra amounts $\sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t}$ and $\sum_{j\in\mathcal{A}(k)\cup\{k\}} X'_{j,t}$ are excluded from consideration.

Next consider the case where stage i's supply is disrupted, and stage k's supply is available. As $M_k > M_i$, stage k is to the upstream of stage i. Its ordering decision is not affected by stage i's supply status, and it can keep ordering normally. Stage i can no longer

make any new order, as its supplier is currently unavailable. So $X_{i,t}^{M-\mu} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t}$ decreases by D_t for period $t+1$, while $X_{k,t}^{M-\mu} - \sum_{j\in\mathcal{A}(k)\cup\{k\}} X'_{j,t}$ varies by $Y_{k,t-(M_i-\mu)} - X_{k,t-(M_i-\mu)} - D_t$. Because $Y_{k,t} \geq X_{k,t}$, we have:

$$
\begin{aligned}
& X_{i,t+1}^{M-\mu} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t+1} \\
=& X_{i,t}^{M-\mu} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t} - D_t \\
\leq& X_{k,t}^{M-\mu} - \sum_{j\in\mathcal{A}(k)\cup\{k\}} X'_{j,t} - D_t \\
\leq& X_{k,t}^{M-\mu} - \sum_{j\in\mathcal{A}(k)\cup\{k\}} X'_{j,t-1} + \left(Y_{k,t-(M_i-\mu)} - X_{k,t-(M_i-\mu)}\right) - D_t \\
=& X_{k,t+1}^{M-\mu} - \sum_{j\in\mathcal{A}(k)\cup\{k\}} X'_{j,t+1}
\end{aligned}
$$

The first equality is true because stage i is facing supply disruption, $X_{i,t}^{M-\mu} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t}$ decreases by D_t for period $t+1$; the first inequality is true based on the assumption that the generalized long-run balance holds for period t; the second inequality holds because $Y_{k,t-(M_i-\mu)} \geq X_{k,t-(M_i-\mu)}$; the second equality holds because $X_{k,t}^{M-\mu} - \sum_{j\in\mathcal{A}(k)\cup\{k\}} X'_{j,t}$ varies by $Y_{k,t-(M_i-\mu)} - X_{k,t-(M_i-\mu)} - D_t$ for period $t+1$.

Next consider the case where stage i's supply is available, and stage k's supply is unavailable. As stage i is to the downstream of stage k (not necessarily j's successor), stage i should make its ordering decision based on stage k's inventory level and pipeline inventory. This is the same as the case where there is no disruption.

Finally, consider the case where both stage i's and stage k's supply are disrupted. This case is trivial as they can not make any new order during disruption, and the generalized item-specific long run balance is preserved from any case it was in. □

Combining both Proposition 2.1 and Proposition 2.3, we can see that every item satisfies long-run balance, excluding the extra inventory for disruption risk. However, these stages also satisfy long-run balance if we exclude all the extra items stored but not assembled yet.

This generalized long-run balance allows us to reduce the assembly system partially into a serial system. The following algorithm explains which parts of the system can be reduced

into a serial system, which parts cannot, and how the reduction works. It is almost the same as Rosling's approach of reducing an assembly system to a serial system, with some modifications.

Partial Series Reduction Algorithm

1. Form a new serial system with stages labeled $i = 1, 2, \cdots, N$.

2. The leadtime for item i is given by $L_i = M_i - M_{i-1}$, and the holding cost coefficient becomes h_i.

3. For stage i with unreliable supplier j, there is one extra stage i' at the same location as i, with the same holding cost. Stage i can order from stage $i+1$ to replenish its inventory when its supplier is not disrupted, and do nothing when its supply is disrupted. It satisfies demands from stage $i-1$. The extra stage i' can only replenish its inventory from j when a disruption is over, and stage $\mathcal{S}(i)$ can order from i' when stage i is out of inventory due to a supply disruption. The leadtime for i' is l_i, and the leadtime for $\mathcal{S}(i)$ is $l_{\mathcal{S}(i)}$.

For the system shown in Figure 2.4, if only stage 3 is under disruption risk, it can be reduced to the system in Figure 2.7.

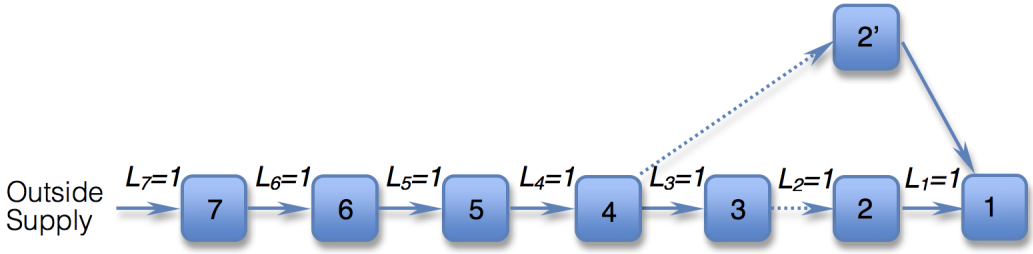

Figure 2.7: Reduced system of assembly system in Figure 2.4.

In the original system, stage 3 stores some extra units of item 3. When there is no supply disruption present in the system, the whole system functions as a regular assembly system does. When there is supply disruption at stage 3's supplier, stage 3 can utilize the extra inventory stored at itself to satisfy the demand from stage 1. In comparison, in the reduced system, stage $2'$ stores some inventory of item 2, and it only orders from stage 4 when the supplier is available. When there is no supply disruption present at the system, the whole system functions as a regular serial system does. When there is a supply disruption present in the system, stage 2 would experience a supply disruption. Once its inventory is depleted,

stage 1 can order from stage $2'$ instead.

With this new way of reducing the assembly system, we can get a simplified inventory system, which acts as a serial system when there is no disruption, and acts as a different serial system when supply disruptions happen. A natural question is whether these two systems are equivalent to each other when operated optimally. Even though equivalence would be a valuealbe insight into the original assemblu system and would facilitate our analysis, unfortunately it does not hold under supply risks. The following Proposition shows the non-equivalence of the optimal policies for the original system and the reduced system.

Proposition 2.4: The reduced system is not equivalent to the original assembly system in general when operated optimally. However, the two systems are equivalent when operated optimally for the special case where stage i is under supply disruption risk, and every other stage k satisfies one of the following three conditions: 1) $M_i - l_i \geq M_k$, 2) $M_i \leq M_k - l_k$, 3) $M_i = M_k$ and $l_i = l_k$

Proof. The equivalence is not true, because the reduced system provides an intermediate stage for storing inventories during disruptions. Because of the assumption that the holding cost at downstream stages is always higher than that at upstream stages, one intermediate stage would save some cost versus storing inventory at downstream stages. For our example in Figure 2.4., when a disruption for stage 2's supplier happens, stage 3 could either keep ordering and receiving new inventories, or it could stop ordering until the disruption is over.

For the reduced system in Figure 2.7., stage 4 in the reduced system is stage 5 and 6 in the original system, and stage 2 in the reduced system is stage 3 in the orginal system. In the reduced system, item 4 could be shipped to stage 3 and assembled into item 3. The holding cost for item 3 at stage 3 is cheaper than the holding cost for item 2 at stage 2. Correspondingly, this means in the original system, the inventory shipped from stage 5 and 6 could be stored somewhere before it reaches stage 3, at a lower cost than H_3. So the reduced system adds this intermediate stage to acheive a lower holding cost, which does not exist in the original system. Thus these two systems generally are not equivalent to each other.

The equivalence holds for the special case where stage i faces supply disruption risks, and if stage k has a leadtime which overlaps with stage i's, then they should be the exact same distance away from the end product, and their suppliers have to be at the same distance away from the end product too. For this case, when a disruption happens, stage k would stop ordering, because it is cheaper to store inventory at its predecessor by assumption. And once the disruption is over, stage i's and k's suppliers resume shipping in a perfectly matched manner, or resume the assembly process. For this case, this two systems are equivalent. □

This Proposition gives us a sense of how far away our inventory policy is from the optimal inventory policy. We know that for unreliable suppliers, it would be optimal to hold extra inventory, and the extra inventory should follow a base-stock policy. One unanswered question is: for any stage k that overlaps with stage i's leadtime, should stage k keep ordering or not when stage i is experiencing supply disruption? The answer depends on the specific model settings. It might be better for stage k to keep ordering if the leadtime for k is much longer than that of item i and the disruption is short. Also it might be better for stage k to wait to order until i's supply disruption is over if the leadtime for both items k and i are close, and the disruption is long. This is relative to both the disruption profile and the system configuration, and we do not yet have a definite answer to this question.

2.5 Order-up-to Levels

2.5.1 Order-up-to Quantity

For an assembly system with supply risks, the optimal ordering policy involves determining whether locations thatare not affected by a current disruption should continue to order during the disruption. From our numerical experiments (see section 2.7.2), delaying ordering seems to happen relatively rarely. So we assume all locations continue ordering during supply disruptions. This section considers what the optimal order-up-to levels are under this assumption.

As stated previously, $X'_{i,t}$ follows a base-stock policy with a base-stock level of s'_i, and

$Y_{i,t} - \sum_{j\in\mathcal{A}(i)\cup\{i\}} X'_{j,t}$ follows the generalized item-specific long-run balance. The modified base-stock policy is as follows:

$$Y_{i,t} = \begin{cases} \min\{s_i + \sum_{j\in\mathcal{A}(i)} s'_j, X^L_{i+1,t}\} & \text{if } X_{i,t} \le s_i + \sum_{j\in\mathcal{A}(i)} s'_j \\ X_{i,t} & \text{if } X_{i,t} > s_i + \sum_{j\in\mathcal{A}(i)} s'_j \end{cases}$$

This policy also uses the upper bound $X^L_{i+1,t}$ to preserve generalized item-specific long-run balance, which is required for the system reduction. We shall discuss how to calculate s_i and s'_i using the reduced system in section 2.6.

2.5.2 Delay Ordering Decision

Normally a stage orders from its supplier to increase its echelon inventory to the optimal order-up-to level. But in some instances it might order 0 from its supplier, even though its supplier may have enough inventory on-hand. As explained previously, this happens if its leadtime overlaps with a stage under supply risk. Such a stage needs to decide whether it should order any new inventory when a supply disruption is present in the system. There is no definite answer to this question. It depends on the circumstances respective. This decision comes second to calculating the order-up-to levels.

2.6 Heuristic Method

As stated above, inventory optimization for an assembly system under supply risk is a large scale problem. Solving it exactly would require dynamic programming over the whole time horizon. Calculating the optimal policy parameters would require keeping a record of outstanding inventories, on hand inventories, and backorders in each period in the dynamic programming algorithm. A one-unit increase in leadtime would lead to the dimension of the state space increasing by one. This would result in a huge stage space that is prohibitively large to solve. Even worse, due to the disruption process, there might be some items shipped to the next stage but not being assembled. This would further increase the dimension of the state space. It is impossible to solve this exactly by dynamic programming within a reasonable time. As a result, a heuristic policy that is easy to compute and implement is

more desireable.

As mentioned previously, there is some connection between assembly systems with disruptions and serial systems. Although these two systems are not equivalent to each other, we still can use the base-stock levels for the reduced "almost-serial" system to approximate the base-stock levels for the original assembly system with supply disruption risks. The algorithm for serial systems without disruptions is not directly applicable to this problem. However, a variation would make a good heuristic for assembly systems with disruptions, due to the resemblance. Recall that for an assembly system without disruptions, the optimal policy is a modified base-stock policy, and the corresponding base-stock levels are obtained by a recursive algorithm. We adapt this algorithm to our "almost-serial" system with some extra stages. For convenience of notation, we use k to denote the smallest-indexed stage under supply disruption risk.

2.6.1 Recursion for Regular Stages

For the reduced system, we propose a recursive heuristic, based on the recursive algorithm for serial systems, which provides a method to obtain $s_i, \forall i = 1, \cdots, N$. Let D^{L_i} be the random demand in L_i consecutive periods. A *regular stage* refers to any stage that does not have any unreliable suppliers. Then for stages $i = 1, 2, \cdots, k-1$, we follow the same recursive algorithm as for a serial system, as in Zipkin [88]:

$$\begin{aligned}
\underline{C}_0^*(x) &= (p + H_1)[x]^- \\
\hat{C}_i^*(x) &= h_i x + \underline{C}_{i-1}^*(x) \\
C_i^*(y) &= E[\hat{C}_i^*(y - D^{L_i})] \\
s_i^* &= \arg\min\{C_i^*(y)\}, \forall i \in N \\
\underline{C}_i^*(x) &= C_i^*(\min\{s_i^*, x\}), \forall i \in N
\end{aligned}$$

2.6.2 Recursion for Stages with Unreliable Supplier

Stage k has an unreliable supplier from our assumption. We will need s_k^* as well as $s_k'^*$, the extra-units base-stock level. When a disruption happens and s_k gets depleted, or after

the disruption ends and while s_k is being replenished, s'_k is in use. This happens with a probability of $\frac{\beta_k}{\beta_k+\gamma_k}$, which is determined by stage k's supply disruption profile. When s'_k is in use, stage k works like a single stage inventory system with disruptions. As a modification, we have:

$$\hat{C}_k^*(x) = h'_k x + \frac{\beta_k}{\beta_k+\gamma_k}\underline{C}_{k-1}^*(x)$$
$$C_k^*(y) = \sum_{n=0}^{\infty} \pi_n^k E_{D^{n+L_k}}[\hat{C}_k^*(y - D^{n+L_k})]$$
$$s'^*_k = \arg\min_y\{C_k^*(y)\}$$

where $\pi_0^k = \frac{\gamma_k}{\beta_k+\gamma_k}$ and $\pi_m^k = \frac{\beta_k\gamma_k}{\beta_k+\gamma_k}(1-\gamma_k)^{m-1}$.

Similarly with a probability of $\frac{\beta_k}{\beta_k+\gamma_k}$ there is no disruption for stage k. So the algorithm for s_k^* is mostly the same as for non-disrupted stages.

$$\hat{C}_k^*(x) = h_k x + \frac{\gamma_k}{\beta_k+\gamma_k}\underline{C}_{k-1}^*(x)$$
$$C_k^*(y) = E[\hat{C}_k^*(y - D^{L_k})]$$
$$s_k^* = \arg\min_y\{C_k^*(y)\}$$
$$\underline{C}_k^*(x) = C_k^*(\min\{s_k^*, x\})$$

2.6.3 Recursion for Unreliable Suppliers

After obtaining s_k^* and s'^*_k from section 2.6.1 and 2.6.2, we proceed to stage k's predecessor, stage i. This stage follows the same algorithm as other reliable stages. But as i is an unreliable supplier, it places orders in a different manner than its downstream stages does because the demand it receives follows a different distribution due to the disruption. The demand mean would remain the same, while the variance would change for its upstream suppliers. To illustrate how the variance changes due to the supply uncertainty, here we show how it changes assuming the demand for the final product has a normal distribution $N(\mu, \sigma^2)$. If stage i is not a supplier for k, the calculation for s_i^* would use the same demand variance σ.

Stage k orders from i in a different way due to the disruption. This disruption is a Markov process, so as calculated before, the pmf of the disruption process stage k is: $\pi_0^k = \frac{\gamma_k}{\beta_k+\gamma_k}$, and $\pi_n^k = \frac{\beta_k\gamma_k}{\beta_k+\gamma_k}(1-\gamma_k)^{n-1}$. So the probability P_n of seeing n consecutive periods' demand is:

$$P_n = \pi_n^k$$

As the demand in each period has a normal distribution $N(\mu, \sigma^2)$, the demand which stage i sees in one period is $D \sim N(\mu, \sum_{n=1}^{\infty} n^2 P_n^2 \sigma^2)$.

This subsection discusses the demand seen by an unreliable supplier. With the assumption of normally distributed demand, this computation yields a better approximation of the demand it sees. For general demand, we can make the approximation that the unreliable supplier would see the same demand as its successor.

2.7 Numerical Experiments

In this section we conduct a numerical study. It first explores how well the inventory policy and the heuristic recursive algorithm perform when compared to the solutions generated by DeCroix's inventory policy and algorithm for modest-sized problems.

2.7.1 Comparison to DeCroix's Policy

The ordering policy proposed in this work differs from DeCroix's policy mostly by its inventory policy structure. Here we test the performance of these two ordering policies, with parameters obtained from the corresponding algorithms.

In order to show the effect of different assembly system structures, we consider several different systems. Here we assume demand in each period is normally distributed, $D \sim N(\mu = 20, \sigma^2 = 4)$. System 1 is shown in Figure 2.8.

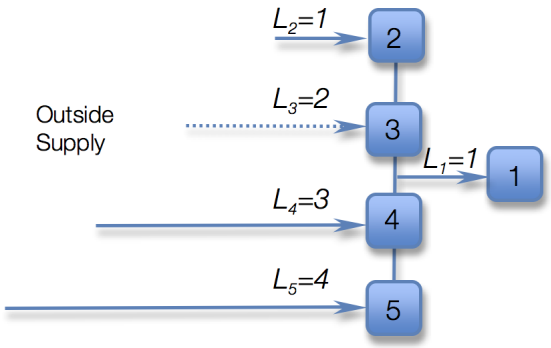

Figure 2.8: Assembly system 1 for comparison to DeCroix's Policy

For this two echelon system, we assume $l = [1, 1, 2, 3, 4]$, and assume that only stage 3 faces supply risk, i.e., $J = \{3\}$. In this system, we use different echelon holding costs, as well as different disruption parameters. The unit backorder cost is fixed at $p = 100$ for all cases. We test all 60 combinations of holding cost and disruption parameters for system 1 are shown in Table 2.1.

$(h_1, h_2, h_3, h_4, h_5)$	Average disruption length $\frac{1}{\gamma_i}$	Average time between disruption $\frac{1}{\beta_i}$
(1, 1, 1, 1, 1)	1.33	2
(1, 1, 1, 5, 1)	2	5
(1, 1, 5, 1, 1)	4	10
(1, 5, 1, 1, 1)	8	
(5, 1, 1, 1, 1)		

Table 2.1: System 1 parameters for performance test

We also consider system 2, with three echelons shown in Figure 2.9. We assume demand in each period is normally distributed $D \sim N(\mu = 20, \sigma^2 = 4)$, $l = [1, 1, 1, 1, 1]$, and again assume $J = \{3\}$.

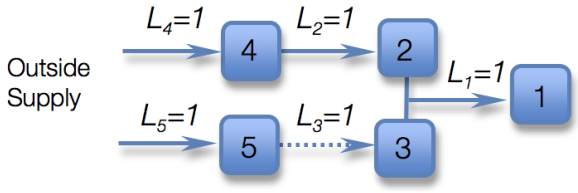

Figure 2.9: Assembly system 2 for comparison to DeCroix's Policy

We test all 48 combinations of the holding cost and disruption parameters for system 2

shown in Table 2.2.

(h_1, h_2, h_3)	Average disruption length $\frac{1}{\gamma_i}$	Average time between disruption $\frac{1}{\beta_i}$
(1, 1, 1)	1.33	2
(1, 1, 5)	2	5
(1, 5, 1)	4	10
(5, 1, 1)	8	

Table 2.2: System 2 parameters for performance test

We also test this policy on a more general assembly system, system 3 in Figure 2.4. For this general system, we assume $l = [1, 1, 2, 2, 2, 3, 5]$, and again assume $J = \{3\}$. Demand in each period is normally distributed, $D \sim N(\mu = 20, \sigma^2 = 4)$ as well, and the backorder cost remains the same at $p = 100$ among all cases. We also test different echelon holding cost and disruption parameters for this system. We test all 84 possible combinations of parameters shown in Table 2.3. For each of the 60 instances for system 1, and 48 instances

$(h_1, h_2, h_3, h_4, h_5, h_6, h_7)$	Average disruption length	Average time between disruption
(1, 1, 1, 1, 1, 1, 1)	1.33	2
(1, 1, 1, 1, 1, 5, 1)	2	5
(1, 1, 1, 1, 5, 1, 1)	4	10
(1, 1, 1, 5, 1, 1, 1)	8	
(1, 1, 5, 1, 1, 1, 1)		
(1, 5, 1, 1, 1, 1, 1)		
(5, 1, 1, 1, 1, 1, 1)		

Table 2.3: System 3 parameters for performance test

for system 2, and 84 instances for the system 3, we compute the base-stock levels using the recusrive algorithm proposed in section 2.6 as well as DeCroix's algorithm, and calculate the respective average inventory cost per period by simulation. For each instance tested, the system is simulated for 5 trials, 2000 periods per trial, for both DeCroix's policy and our heuristic policy. For both policies, we subtract the averge pipeline inventory holding cost, as it is a constant for both policies. We focus on the inventory holding cost at all stages, as well as the backorder cost at the final stage.

A comparison of the results of the 2 policies for these 3 systems with a total 192 different parameter settings, is given in Fig 2.10.

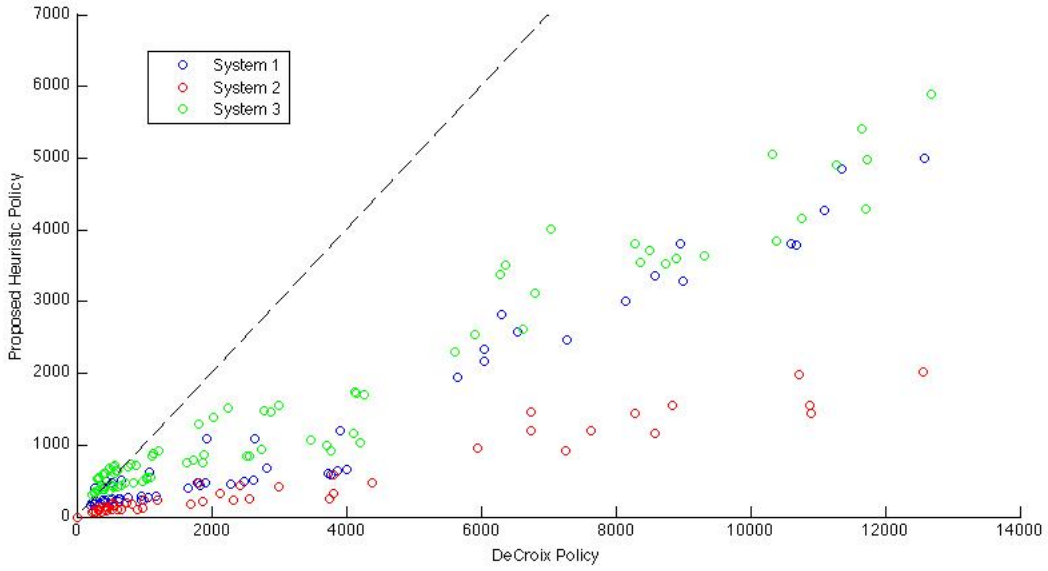

Figure 2.10: Assembly system for comparision to DeCroix's algorithm

The horizontal axis shows the average cost per period for DeCroix's policy, while the vertical axis shows the average cost per period for our heuristic policy. The dotted line in the graph is $y = x$. Any markers below this dotted line represent instances for which our heuristic policy outperforms DeCroix's policy, and vice versa for markers above the line. The graph clearly shows that for most of the instances tested, it is beneficial to have some extra inventory to prepare for supply disruptions. Some instances are further away from the origin, and they tend to stay far away from the dotted line. These are the instances where supply disruptions have a large impact on total cost, and our inventory policy and parameters have much better performance for these instances. Some instances are located closer to the origins; they are less affected by supply disruptions.

To explore this comparison further, we calculate the performance as measured by the ratio between the two policies:

$$\text{Ratio} = \frac{\text{Average cost per period for our heuristic policy}}{\text{Average cost per period for DeCroix's policy}}$$

If the ratio is less than 1, it means that our policy achieves lower cost. It also suggests that it is better to have extra inventory of the unreliable item. The statistics for the performance

ratio are reported in Table 1.4., which gives the average, minimum, and maximum ratio between the average cost of solutions returned by the two policies, as well as the percentage of instances in which our policy performs better. One can see that our policy performs

	System 1	System 2	System 3
Average	0.470	0.197	0.723
Min	0.158	0.069	0.242
Max	1.515	0.410	1.720
Percentage of ratios less than 1	95.0%	100.0%	76.2%

Table 2.4: Parameters for performance test 1

generally better. For system 1, 3 out of 60 tested instances indicate that the system benefits from extra inventory. For system 2, surprisingly, every instances shows it is better to have extra inventory. And for system 3, 64 out of 84 instances holds extra inventory.

The value of extra inventory

We tested three different assembly system. Each shows a different likelihood that extra inventories will be used to buffer against disruptions. The amount of savings achieved from extra inventory also varies. For system 1, the average savings is 53.0%, while the savings for system 2 reaches a significant 80.3%. For system 3, the average savings for all instances reaches a savinsg of 27.7%. If we only consider the cases in which no extra inventory is prefered for system 3, the saving reaches 46.0%. In general, if extra inventory is suitable for a system, it would result in a significant amount of cost savings.

Factors in choosing extra inventory

Even though more than 88.0% of instances tested show a benefit from extra inventories, there are still some instances which do not. It is necessary to understand under which circumstances extra inventories is undesirable.

The biggest factor in determining the value of extra inventory is inventory holding cost. There are three instances in which system 1 holds no extra inventory. All of them occur when the stage under supply risk has a high holding cost.

The second factor is the assembly system structure. As shown in the Figure 2.10., system 2 requires extra inventory for all of its instances. Both system 1 and 3 show that

when the stage 3 which is under supply risk, has a large holding cost, it is more likely to have no extra inventory.

The third factor is the average time between disruptions. Our numerical experiments show that no extra inventory is prefered for the instances with longer average time between disruptions. This is equivalent to saying if supply disruptions happen less frequently, it would be less likely to require extra inventory.

The fourth factor is the average disruption length. The numerical experiments also show that instances with shorter average disruption length are less likely to require extra inventory.

All of these factors coincide with our intuition. If supply disruptions happen more frequently, or last longer, it is more likely to require extra inventory to mitigate the disruptions. The conditions under which it is optimal to hold extra inventory are not easy to characterize explicitly. Numerical experiments are recommended to determine whether it is better to hold extra inventory or not.

2.7.2 The Value of Delaying Ordering

According to Proposition 2.4, the reduced "almost-serial" system is not equivalent to the original system. For the case where stage i faces supply disruptions, stage j needs to make a decision on whether it should keep ordering when stage i's supply is disrupted, if i and j's leadtimes overlap with each other. Consider the simple case where stage i and j are at equal distance from the end product, in terms of total leadtime M_i and M_j. If $l_j \leq l_i$, then stage j could just make its ordering decision based on pipeline inventory from stage i. If $l_j > l_i$, and stage i's supply is disrupted, stage j needs to consider whether it should keep ordering or not. This is what we explore in this numerical experiment.

In this numerical experiment, our goald is to show that instances exist for which it is better to delay ordering item j when stage i's supply is disrupted. This happens rarely, but there are some examples. To find one instance, we still work with system 1 in Figure 2.8., with $h = (1, 1, 1, 5, 1)$, $p = 15$. We consider the case where $J = \{3\}$, $\beta_3 = 1/51$ and $\gamma_3 = 1/50$. As discussed, we only need to decide if stage 4 and stage 5's predecessor should postpone shipping or not, if stage 3's supply is disrupted.

We compute the average inventory cost per period by simulation for both cases: postpone ordering or continue ordering. The system is simulated for 25 trials, 2000 periods per trial. The average cost per period for the case with continued ordering is 21569.9, while the average cost per period for the case with delayed ordering is 19399.6. Continuing ordering incurs 11% more cost than delaying ordering. This verifies our argument in Proposition 2.4 that the reduced system is not equivalent to the original system.

This example provides a few insights about when this could happen. First, the important factor affecting the decision is the relative ratio of the holding cost to the stockout cost. If the stockout cost is very large compared to the holding cost, very likely it is better to continue ordering when a disruption happens, so that any stockouts would be mitigated as soon as the disruption is over. In our example, the local holding cost for item 1 is 9, while the stockout cost is 15, which is relatively small. So delayed ordering is prefered in this example.

Second, another factor is the disruption length. If a disruption lasts very long, it is possible to delay ordering, as any new inventory would sit in the system. The cost for holding this new inventory during a disruption might overwhelm the benefits from meeting backorders when the disruption is over. Our example chooses an exaggerated expected disruption length of 50 to emphasize this point.

Third, instances for which delaying ordering is optimal are rare. Normally they can could be ignored. To confirm this, we calculated the value of delaying ordering for the system with the settings in Table 2.1, and $p = 100$. All instances are shown to be better if all stages continue shipping when a disruption is present in the system. More important, because our policy encourages any stage with unreliable supply to store some extra inventory, natually it tends to be better to keep ordering during disruptions as we want to utilize this extra inventory to mitigate disruptions.

In conclusion, it is usually better to keep ordering with supply disruption's presence at the system. We only need to examine the instances where the stockout cost is relatively small compared to the holding cost, or the average disruption length is very long.

2.8 Conclusion and Future Work

In this chapter, we study inventory management for an assembly system where some stages are under supply risk. An assembly system without supply disruptions can be reduced to a serial system. However this is not true for the case with supply disruptions. We explain the reasons for this and build our policy accordingly. It might be optimal for stages to store extra units of inventory if they are under supply risk, and we explore how to manage this extra inventory. This extra inventory should also follow a base-stock policy, be utilized when a disruption happens in the system, and get replenished when the disruption is over. Thus we have our heuristic policy for assembly systems with supply risk. The long-run balance property through the entire system is destroyed by the disruption, but if this amount of extra inventory is excluded from consideration, we have a generalized item-specific long-run balance. These results allow us to present a method for reducing an assembly system with supply risk to a serial system plus some extra stages. Although these two systems are generally not equivalent to each other, the reduced system still provides an approach to calculate the base-stock levels for our heuristic policy. Due to the inherent difficulty in calculating the optimal parameters for assembly systems with disruption, we build our heuristic algorithm for it based on the recursion for serial systems with no disruptions.

Based on our numerical results, we study several questions related to our heuristic policy. First we show that our heuristic policy can generate cost that are significantly smaller than those from the current state-of-the-art policy. This makes extra inventory a very important factor in assembly systems with disruptions. We also explore what might affect the decision of whether to hold extra inventory or not. If the holding cost for stages with unreliable suppliers is relatively low compared to the stockout cost, it is better to hold extra inventory. And if the disruptions last longer, or the disruptions happen more frequently, extra inventory is prefered. Our numerical experiments show that extra inventory is optimal for a majority of instances tested. It is therefore beneficial to consider extra inventory when managing an assembly system under supply risk.

Chapter 3

Management of a Distribution System under Supply Risk

3.1 Introduction

As mentioned in the previous chapter, the three most commonly studied types of inventory systems are: serial systems, assembly systems, and distribution systems. The previous chapter focuses on inventory policies for an assembly system under supply risks. In this chapter, we move onto another important structure: distribution systems under supply risks.

A distribution system is the mirror image of an assembly system. One location, which is called the root location, receives inventory from an outside source, and send inventories to other locations. Other locations that have received inventories send inventories to their own customer locations. This goes on until inventories reach the last locations, which are called leaves, where outside customer demands are satisfied. Information flows upwards from leaf nodes to the root node. The outside demand at each leaf node is assumed to be a random variable that follows a stationary Poisson distribution. The root node replenishes its inventory from an outside source which is assumed to have sufficient inventories. All the demand from downstream stages consititues the demand process of their suppliers. As before, we refer to a location that satisifies the demand of location i as the predecessor of

i, and the location that places an order to i as the successor of i. In a distribution system, each location has only one predecessor and at least one successors, except for the leaves.

One aspect of distribution systems that is not present in assembly and serial systems is the allocation policy. When a stage has sufficient inventory on hand, it can meet the demands from its successors. However, if it does not have enough inventory on hand, some of the demand it faces won't be satisfied. Among all demands, which one should be satisfied first needs to be decided. The allocation policy becomes an important question for distribution systems. However, the optimal allocation policy is unknown (Gurbuz et al. [34]). Our model is a continuous review model, and we assume demand at any node can occur at any time. When a stage receives an order from one of its successors, we assume it fills the orders sequentially, i.e. a First-Come-First-Served policy. This means that when a stage has one unit of product on hand, it will send this unit to the successor who requested it first, even though there might be some other successors which might need this unit more. If the system is centralized, the decision maker needs to know which successor has priority or more severe consequences, and allocate units accordingly. Since the optimal allocation policy is hard to determine, we consider a decentralized mechanism, where FCFS allocation policy is a direct consequence.

One of the simplest distribution systems is the so-called one-warehouse, multiple-retailers (OWMR) system. It has only two echelons, where the root node is called warehouse, with the remaining nodes known as retailers. Even for the OWMR system, the simplest distribution network structure, there is no simple, elegant method to solve this problem. The most popular method for the OWMR is the projection method. The algorithm searches over all possible values s_0' of the base-stock level for the warehouse. Assuming there are N retailers, for each given s_0', the algorithm chooses the optimal $s_j', j = 1, \cdots, N$, for all retailers. This algorithm divides the total system cost function into N convex single variable functions, which enables fast computation of the base-stock levels at the retailers. However, the system cost function is not a convex function of s_0. Consequently, the algorithm has to conduct exhaustive search to find the optimal warehouse base-stock levels. There are some other heuristic procedures in the literature related to the projection method, and we provide a short review of them.

The projection method can be generalized to general distribution systems. But the computational cost becomes prohibitively expensive. This chapter considers distribution systems under supply risks. This supply risk makes the system even harder to solve. In this chapter, we assume leadtimes for all stages are deterministic, but some locations are under supply disruption risks. For a stage with an unreliable supplier, supply disruptions occur according to a stationary Poisson process, and during a disruption no more new orders can be placed to its supplier. However, previous shipments are not affected. The end of this supply disruption also occurs according to a stationary Poisson process. Once the supplier becomes available, it uses its on-hand inventory to meet the backorders. If there is more demand than the on hand inventory, it still applies an FCFS allocation policy to meet the backordered demands. Hence this system is a continuous review decentralized distribution system under supply risks, and each stage implements a base-stock policy for its inventory. There is a linear ordering cost with no fixed ordering cost. There is an be inventory holding cost at all locations, and a backorder cost only at all the leaf nodes. Unsatisfied demands at non-leaf nodes are backordered without any cost. Assuming an infinite horizon, our objective is to minimize the long run expected inventory cost per unit time. For this chapter, we also assume a decentralized control mechanism. Thus each location in the distribution system observes its own inventory position, and makes ordering decisions following a local base-stock policy.

We consider distribution systems with more than two echelons, and we also consider supply disruptions in our model. We analyze the effects of supply disruptions on inventory levels, and develop a heuristic algorithm utilizing these effects to optimize base-stock levels in a multi-echelon distribution system under supply risks, assuming FIFO allocation policy and base-stock inventory policy. The main idea behind our recursive algorithm is to analyze the inventory shortage at each stage due to supply disruptions as well as stockouts at its predecessor, so that we can incorporate its successor's costs when calculating its base-stock level. This algorithm is easy to implement, and it also yields solutions within a 3% optimality gap on the instances tested.

3.2 Literature Review

In this section, we briefly review the literature related to our work. Since we discussed the literature on serial systems, assembly systems and inventory models with supply uncertainty in the previous chapter, we skip them in this part of the review even though they are closely related to distribution systems. We only focus on distribution systems in this review.

3.2.1 One-Warehouse, Multiple-Retailer Systems

The one-warehouse, multiple-retailer (OWMR) system is a special case of general distribution systems. It has only two echelons: one warehouse, and multiple retailers. It is the simplest distribution system. Thus many paper have focused on it. There are two main difficulties related to this problem: determining the optimal replenishment policy, as well as finding the most cost-effective allocation policy in case of insufficient supply at the warehouse.

The first work on the decentralized OWMR problem, known as METRIC, is by Sherbrook [74] and was applied by the US Air Force to manage aircraft engine inventories. The metric model approximates outstanding orders by Poisson random variables, which means the backorder level at the warehouse can be approximated as Possion too, and permits a simple evaluation of different policies. Simon [75] as well as Kruse and Kaplan [42] characterize exact expressions for the stationary distributions of stock on hand, stock in repair, and backorders at both warehouse and retailers.

Graves [28] presents an exact model for finding the steady-state distribution of the net inventory level at each site, assuming compound Poisson demand and deterministic shipment time. He approximates the outstanding orders by a negative binomial distribution using a two-moment approximation, which is more accurate than the METRIC approximation in general. Axsäter [3] provides simple recursive procedures to calculate the holding cost and backorder cost for different policies. He uses an inventory cost function to reflect the inventory holding cost and bakcorder cost incurred on an average unit.

Gallego et al. [25] develop an approximation algorithm called the restriction-decomposition heuristic. It decomposes the system into more manageable newsvendor-type subsystems. It

works in a top-down manner, starting by computing the base-stock levels at the root location, or the warehouse, first, and then working on retailer locations. In contrast, Rong et al. [64] propose a recursive optimization heuristic and a decomposition-aggregation heuristic, which act in a bottom-up way.

Svoronos and Zipkin [78] study the same problem but under slightly different assumptions: they assume stochastic leadtimes, and that orders do not cross in time, which means the queues in each in-transit system observe FIFO. They describe methods for approximating the steady-state behavior of the system. The results indicate that the transit time variance has a large impact on system performance. A system with large transit-time variances will need larger inventories to prepare against stockouts when compare with a system that has fixed leadtime.

In addition to the work on base-stock policies mentioned above, there is some work focused on other inventory policies. Svoronos and Zipkin [78] study the problem where all locations follow a continuous review (R, Q) policy. They apply a decomposition technique for this system: approximating each location as a single location inventory system. They estimate the long-run average inventory at each location, and the backorders at the retailers, then use these estimates to minimize total cost. The results suggest the approximation works very well. Andersson et al. [1] study the same problem. The authors investigate a coordinated but still decentralized control procedure. The procedure is based on an approximation: replacing the stochastic leadtime observed by retailers by their averages. This enables the decomposition of the multi-echelon inventory problem into a number of coordinated single-echelon inventory problems. The decision at the warehouse affects the retailers through the marginal cost increase with respect to the increase in leadtime. This information can be used as a shortage cost at the warehouse to determine its near-optimal reorder point. Axsäter [4] considers a problem in which each location implements a continuous review (R, Q) policy, and each retailer faces a compound Poisson demand process and constant leadtime. The author calculates the complete probability distributions of the retailer inventory levels to evaluate control policies exactly.

Other authors consider periodic-review distribution networks. These models can be formulated using dynamic programming. However, such models becomes hard to solve due

to the huge state space. Consequently, approximations for the optimal policy parameters have been developed for this type of problem. Eppen and Schrage [20] consider the OWMR problem under periodic review assuming the warehouse does not hold inventory. Under the assumption that every incoming order at the warehouse is large enough that it can satisify the same percentage of demands at all retailers, they derives approximately optimal policies and inventory costs. Erkip et al. [21] study the same problem with correlated demand, not only across retailers, but also in time. The authors formulate the optimal safety stock explicitly as a function of the level of correlation through time. Jackson [40] studies the same problem, relaxing the assumption that the warehouse does not hold inventory. The author considers a single order cycle, and develops the exact cost model, as well as an approximate of the cost model which is similar to the projection algorithm in the case of identical retailers. Federgruen and Zipkin [23] consider the periodic review OWMR problem with the assumption that the warehouse does not hold inventory. This problem natually leads to a dynamic program which has a huge state space. The authors approximate it systematically with a single location inventory problem using a key concept called myopic allocation.

The second important question to be addressed for distribution systems is the allocation policy. Axsäter and Rosling [7] compare local and echelon stock policies for multi-echelon inventory control. Their major results are for serial and assembly systems. The authors show that for (Q, r)-rules, echelon stock policies are in general better than local stock policies. But this is not the case for distribution systems. Axsäter and Juntti [5] use simulation to study distribution systems with stochastic demand. The results show that echelon stock policies do not always outperforms local stock policies. It depends on the structure of the system. One standard allocation method widely used in the literature assumes that the incoming orders are always large enough that stockouts can be acheived with equal probabilities at all successors, and negative allocations to retailers are permitted. Many existing papers adopt this allocation policy, including Eppen and Schrage [20]. Federgruen and Zipkin [23] use an allocation policy called myopic allocation which minimizes the expected cost in the current period, ignoring costs in all subsequent periods. Federgruen and Zipkin [24] adopt same allocation policy. van Houtum et al. [84] use a similar idea called relaxed

myopic allocation. Verrijdt and de Kok [85] also use this allocation method.

However, the standard allocation policy does not always work well. Axsäter et al. [6] argue the standard allocation policy might not be a good approach, especially when there are significant differences among retailers, in terms of their demand characteristics and service requirements. The authors propose a virtual assignment ordering rule for warehouse replenishments, and a two-step allocation rule for allocating inventories to retailers, which yield great improvements. There are also some other allocation policies. Jackson [40] examines a "ship-up-to-S" policy in which the warehouse restores the retailer's inventory position to a preset value S if there is sufficient inventory. McGavin et al. [52] study a two-interval allocation policy to minimize expected lost sales per retailer while ignoring inventory holding costs. Graves [29] introduces an allocation scheme called virtual allocation, which can be viewed as an equitable allocation policy in which the order that has been outstanding longest is satisfied first. Cachon [12] uses random allocation in which all orders received by warehouse are shuffled randomly, then satisfied in that sequence. Gurbuz et al. [33] present a centralized ordering policy which orders for all retailers simultaneously. It works just as though only the warehouse orders, allocates, and distribute inventories to retailers, but does not carry any inventory itself. One retailer's order might be postponed or expedited to save total ordering and shipping cost. Our model use a first-come, first-served policy in a continuous review setting to facilitate our analysis.

3.2.2 Distribution Systems

There are two typical approaches in multi-echelon inventory theory: *stochastic service* and *guaranteed service* models. The stochastic service model assumes the delivery time between stages can vary depending on the availability at the supplier stage, while the guaranteed service model assumes that each stage has a guaranteed service or delivery time by assuming the demand is bounded. Graves and Willems [30] discuss the differences between them, and compare them in terms of their underlying assuptions, computational and modeling implications, as well as the nature of results produced.

First we discuss the stochastic service approach. Lee and Billington [45] develop a multi-echelon inventory model for a Hewlett-Packard supply chain. The authors model a supply

chain as a set of SKU-locations where each location could uses a single stage base-stock level as an input. The single stage base-stock level is a function of the replenishment leadtime, which can be obtained with approximate expressions developed in the paper. Ettl et al. [22] consider a similar problem, making a clear distinction between nominal and actual leadtimes, where the nominal time is the replenishment time a stage experiences when its supplier carries enough inventory on hand, while the actual leadtime is the replenishment time a stage experiences accounting for possible supplier backorders. The authors develop an approximation of the actual leadtime random variable, and use this approximation to determine the base-stock levels to acheive service level targets. Glasserman and Tayur [27] consider the same problem with different assumptions on capacity limits, which require the system to follow a modified base-stock policy. The authors derive the problem formulation, and develop estimates of the objective derivatives to conduct a gradient based search to find the optimal base-stock levels. Zhao [86] studies the problem with compound Poisson demand processes. The author characterizes backorder delays for each unit of demand, and proposes approximations and decompositions to evaluate the system efficiency and optimize base-stock levels.

The guaranteed service model is studied by Inderfurth and Minner [39]. The authors study multi-stage inventory systems under a periodic review base-stock policy. The authors assume every location satisifies a service level constraint, and no internal delays occur. They formulate the problem and derive optimal policy properties for it. They observe that safety stock coverage times can only take values from extreme points of the solution set.

In addition, we note that some papers mentioned previously in the context of OWMR systems also apply to general distribution systems, including Sherbrook [74], Grave [28], Lee and Moinzadeh [47][46], Svoronos and Zipkin [78][79].

In this work, we consider a distribution network, with a first-come, first-served (FCFS) allocation policy employed at all stages, operating under the stochastic service model. This problem is inherently difficult to solve, as most algorithms give a heuristic solution and the best known exact algorithm requires an exhaustive search over the solution space. Since the optimal policy structure is unknown, we will choose a reasonable policy, a base-stock policy, and then find near optimal parameters for this policy. In this work, we minimize the

average inventory cost per unit time for an infinite-horizon continuous-review model with Poisson customer demand. We suggest a heuristic algorithm to find the base-stock levels at each inventory location.

3.3 Preliminaries

We consider a distribution system (V, E) with $N = |V|$ stages, indexed from upstream to downstream with stage 1 being the furthest upstream supplier. V is the collection of stages or locations in this distribution system, while E is the collection of supplier-customer relationships. Each stage $i \in V$ orders from its predecessor $\mathcal{P}(i) = \{j \in V : (j, i) \in E\}$ (or outside supplier) and meets the demand from its successors $\mathcal{S}(i) = \{j \in V : (i, j) \in E\}$. A stage might have multiple immediate successors, but at most one immediate predecessor. We define a set $\mathcal{L} := \{i \in V : \mathcal{S}(i) = \emptyset\}$ to represent the leaf nodes, which meet outside demands. Each stage i requires products from its immediate predecessor to satisfy demands observed from downstream customers, and delivery from $\mathcal{P}(i)$ to i takes a deterministic leadtime of l_i. However, due to stock outs and possible supply disruptions, it might take longer than l_i for an order to actually arrive from the time that it was placed. We also use D_i to denote the leadtime demand seen by stage i, which is the demand seen by stage i during a period with length l_i.

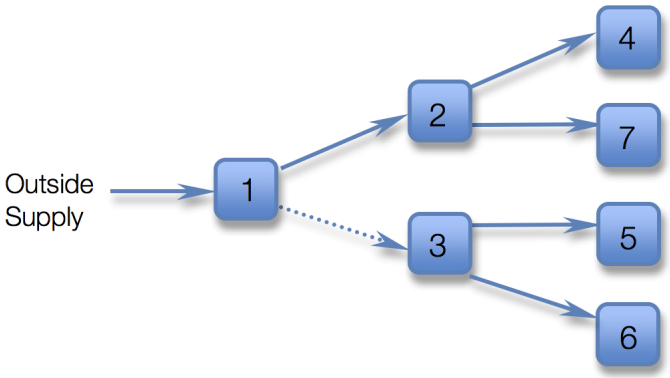

Figure 3.1: Example of Distribution System

This work considers a continuous review model without fixed ordering cost. Demand at leaf node i is a Poisson process with rate λ_i, and for non-leaf nodes $\lambda_i = \sum_{j \in \mathcal{S}(i)} \lambda_j$.

During a stockout or supply disruption, unsatisfied demands at each stage are backordered, but only backorders at leaf nodes are charged a penalty cost of b_i per item per unit time. All locations can carry inventories, and every unit at stage i or in transit to i incurs a cost of H_i per unit per unit time. Define the echelon holding cost incurred at stage i as $h_i = H_i - H_{\mathcal{P}(i)}$ per unit per unit time, where H_i is the local holding cost for stage i. h_i is the holding cost increment due to storing it at stage i from storing it at i's supplier.

In this model, a subset $J \subseteq V$ is under supply disruption risk, which means that only the supply of the nodes in J may be disrupted. Whenever $i \in J$ wants to order a new item, it checks the disruption status of its supplier $\mathcal{P}(i)$. If it is experiencing a supply disruption, the order to $\mathcal{P}(i)$ is backordered, but previous shipments are not affected. The occurrence of a supply disruption at stage i is a Poisson process with rate β_i, and the end (or cease) of a supply disruption is also a Poisson process with rate γ_i. Equivalently, disruptions follow a continuous-time Markov process with two states representing normal and disrupted operation. For node $j \notin J$, its supply is still available even if it shares the same supplier with node $i \in J$ and i's supply is disrupted.

We use I_i to denote the local on-hand inventory level at node i, and denote the backorder level by B_i. We denote the local inventory level at stage by IL_i, which is $IL_i = I_i - B_i$. We also denote the shortage at stage i only due to a current supply disruption at $\mathcal{P}(i)$ by $W_{\mathcal{P}(i),i}$. $B_{\mathcal{P}(i),i}$ reprsents the shortage at stage i only due to backorders at stage $\mathcal{P}(i)$, excluding the portion of shortage due to supply disruptions.

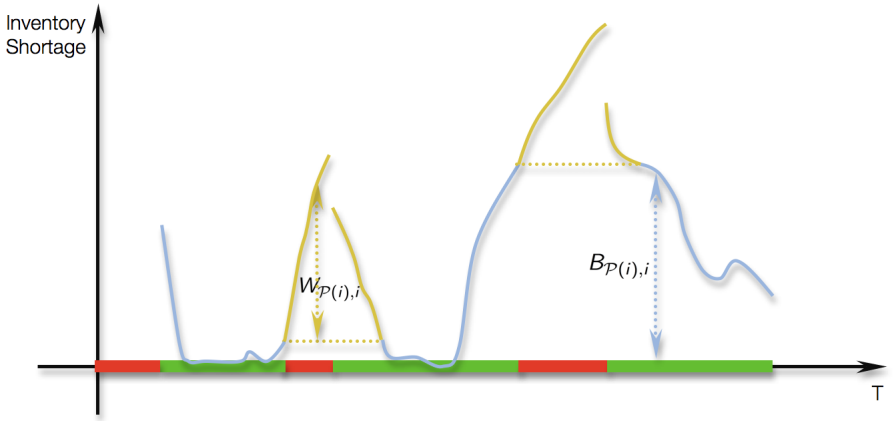

Figure 3.2: Inventory shortages due to supply disruption and supply backorder

In this graph, the red section on the horizontal axis represents supply disrupted interval, and the green section represents an available interval. There are some differences between $B_{\mathcal{P}(i),i}$ and $W_{\mathcal{P}(i),i}$. As shown in the graph, $W_{\mathcal{P}(i),i}$ is the shortage increment when supply is disrupted, and $B_{\mathcal{P}(i),i}$ is the shortage increment when supply is available. For example, if stage i's supply is disrupted from t_0, and it has been disrupted untill the current time t_1, then $W_{\mathcal{P}(i),i}(t_1)$ equals the demand quantity i sees from t_0 to t_1. And when this supply disruption ends at t_1, all the inventory at $\mathcal{P}(i)$ is used to mitigate $W_{\mathcal{P}(i),i}(t_1)$. During this supply disruption interval, $B_{\mathcal{P}(i),i}(t_1)$ could be non-zero. It equals $B_{\mathcal{P}(i),i}(t_0)$ in this supply disruption interval. When the supply disruption is over at t_1, $B_{\mathcal{P}(i),i}$ will not decrease until $W_{\mathcal{P}(i),i}$ decreases to 0 first. The inventory shortage at i for any given time t is $W_{\mathcal{P}(i),i}(t) + B_{\mathcal{P}(i),i}(t)$.

As the optimal ordering policy structure is unknown, we study a reasonable alternative, a base-stock policy, in this work. The decision variables are s_i, the local inventory base-stock level at stage $i, \forall i \in V$. We also denote the echelon base-stock level as $S_i = \sum_{j \in \mathcal{S}(i)} s_j$. According to "conservation of flow", as in Zipkin [88], the inventory level IL_i is composed of several parts. Every time stage i orders, it brings the echelon inventory position to S_i. But not all of these items are on hand at stage i. There is a shortage of $W_{\mathcal{P}(i),i} + B_{\mathcal{P}(i),i}$ due to either disruption or backorders at its predecessor; part of the inventory IL_i is still en route to i, which is accounted for the leadtime demand D_i; and part of the inventory is located in its downstream stages, and this amount is $\sum_{j \in \mathcal{S}(i)} S_j - \sum_{j \in \mathcal{S}(i)} W_{i,j}$, since the total local inventory positions at its downstream stages are $\sum_{j \in \mathcal{S}(i)} S_j$ and $\sum_{j \in \mathcal{S}(i)} W_{i,j}$ is not shipped due to supply disruptions. Thus we have the following equations for IL_i:

$$IL_i = S_i - B_{\mathcal{P}(i),i} - W_{\mathcal{P}(i),i} + \sum_{j \in \mathcal{S}(i)} W_{i,j} - D_i - \sum_{j \in \mathcal{S}(i)} S_j \tag{3.1}$$

$$I_i = \left[S_i - B_{\mathcal{P}(i),i} - W_{\mathcal{P}(i),i} + \sum_{j \in \mathcal{S}(i)} W_{i,j} - D_i - \sum_{j \in \mathcal{S}(i)} S_j \right]^+ \tag{3.2}$$

$$B_i = \left[B_{\mathcal{P}(i),i} + W_{\mathcal{P}(i),i} - \sum_{j \in \mathcal{S}(i)} W_{i,j} + D_i + \sum_{j \in \mathcal{S}(i)} S_j - S_i \right]^+ \tag{3.3}$$

In this work, we are trying to find the base-stock levels at all stages so that the total

expected inventory cost per unit time is minimized. There is a purchasing cost for each item, and a holding cost for pipeline inventory, but we omit these costs from the calculation, as their expectations are constant. We let $\mathbf{S} = \{S_i\}_{i \in V}$. So the long run expected inventory cost per unit time is:

$$C(\mathbf{S}) = \sum_{i \in V} H_i E[I_i] + \sum_{i \in \mathcal{L}} b_i E[B_i]$$

3.4 Heuristic

The heuristic proposed in this work is similar to the recursive algorithm for finding the optimal base-stock levels in serial systems. The recursive algorithm starts from the stage facing customer demands and solves for the optimal base-stock level at each stage in a recursive manner. We propose a heuristic for distribution systems in a similar way, which incorporates the effects of supply disruptions. Before we introduce this heuristic, we analyze the key random variables related to supply disruptions.

3.4.1 Estimation of $W_{\mathcal{P}(i),i}$

$W_{\mathcal{P}(i),i}$ represents the shortage of items at stage i due to a current disruption from supplier $\mathcal{P}(i)$. It captures the effect of supply disruption on inventory level and backorder level. The distribution of $W_{\mathcal{P}(i),i}$ is one key aspect of our heuristic algorithm.

Since $W_{\mathcal{P}(i),i}$ only increases during disrupted intervals, and decreases during available intervals, it is natural to consider the case when stage i's supply is disrupted and the case when it is not separately.

Disrupted Interval

For the case when the supply is disrupted, we have the following results.

Proposition 2.1. Assume stage i's supply stage $\mathcal{P}(i)$ is disrupted at time t. Then

$$P(W_{\mathcal{P}(i),i}(t) = n | \text{stage } \mathcal{P}(i) \text{ is disrupted at time } t) = \frac{\gamma_i \lambda_i^n}{(\gamma_i + \lambda_i)^{n+1}}.$$

Proof. Let t_0 be the unknown start time of the disruption. Because the occurrence of a

supply disruption at stage i is a Poisson process, the probability density function of the current disruption starting from time t_0 is given by $\gamma_i e^{-\gamma_i(t-t_0)}$ due to the Poisson process's memoryless property.

Denote $D(t)$ as the cumulative demand in $[0,t]$. Since the demand seen by stage i has a Poisson distribution with rate λ_i, the demand during $t-t_0$ time units has a Poisson distribution with rate $\lambda_i(t-t_0)$, i.e.,

$$P(D(t)-D(t_0)=n)=e^{-\lambda_i(t-t_0)}\frac{[\lambda_i(t-t_0)]^n}{n!}.$$

Therefore, we have:

$$\begin{aligned}
&P(W_{\mathcal{P}(i),i}(t)=n|\text{stage } \mathcal{P}(i) \text{ is disrupted at time } t)\\
=&\int_{t_0=-\infty}^{t} P(D(t)-D(t_0)=n|\text{disruption started at time } t_0)\gamma_i e^{-\gamma_i(t-t_0)}dt_0\\
=&\int_{t_0=-\infty}^{t} \gamma_i e^{-(\gamma_i+\lambda_i)(t-t_0)}\frac{[\lambda_i(t-t_0)]^n}{n!}dt_0\\
=&\int_{t_0=-\infty}^{t} -\frac{\gamma_i\lambda_i^n}{(\gamma_i+\lambda_i)^{n+1}}e^{-(\gamma_i+\lambda_i)(t-t_0)}\frac{[(\gamma_i+\lambda_i)(t-t_0)]^n}{n!}d(\gamma_i+\lambda_i)(t-t_0)\\
=&-\frac{\gamma_i\lambda_i^n}{(\gamma_i+\lambda_i)^{n+1}}\int_{x=\infty}^{0} e^{-x}\frac{x^n}{n!}dx\\
=&\frac{\gamma_i\lambda_i^n}{(\gamma_i+\lambda_i)^{n+1}}
\end{aligned}$$

The last equality holds because $e^{-x}\frac{x^n}{n!}$ is the pdf of the Erlang distribution with $k=n+1, \lambda=1$. □

Recovery Interval and Regular Interval

The case when the supply is available is more complicated. For the sake of simplicity, here we only discuss the case where $l_i=0$ and assume the last disruption ended at time t_1. The case in which $l_i>0$ is similar except that time is shifted. Here we define $W_{\mathcal{P}(i),i}(t^+)=\lim_{\epsilon\to 0} W_{\mathcal{P}(i),i}(t+\epsilon)$.

Assume stage i's supply has been disrupted from t_0 untill the current time t_1. There are two cases for times with no disruption.

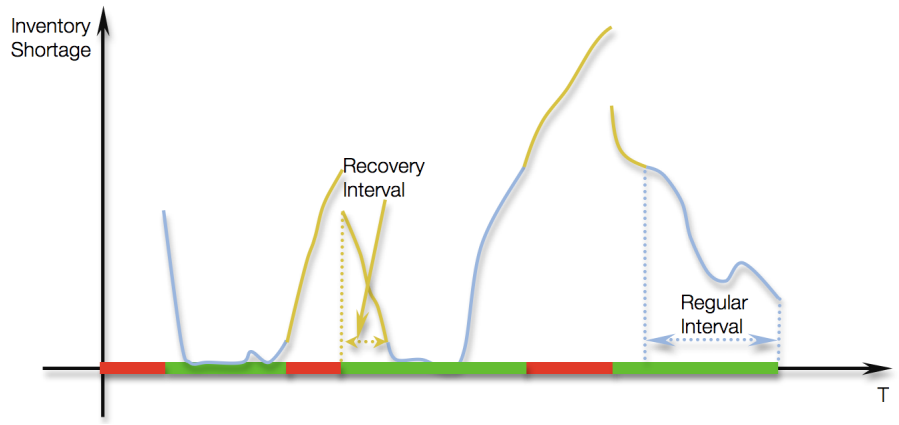

Figure 3.3: Recovery intervals and regular intervals in supply available intervals

Case 1 is the regular interval, which refers to the time interval where $W_{\mathcal{P}(i),i}(t_1)$ has been fully replenished, and everything performs just like the system is reliable and no disruption has ever happened, or $W_{\mathcal{P}(i),i}(t_1^+) = 0$. Case 2 is the recovery interval, which refers to the time interval in which a disruption has just ended, but stage i has not yet fully replenished $W_{\mathcal{P}(i),i}(t_1)$ yet, i.e. $W_{\mathcal{P}(i),i}(t_1^+) > 0$. We let P' denote the long-run probability that stage i is in recovery interval.

When the supply disruption ends at t_1, there are two different possibilities for $W_{\mathcal{P}(i),i}(t_1^+)$. If $IL_{\mathcal{P}(i)}(t_1) \geq W_{\mathcal{P}(i),i}(t_1)$, stage i goes into a regular interval immediately. The shortage due to the supply disruption is immediately mitigated, and $W_{\mathcal{P}(i),i}(t_1^+) = 0$. $W_{\mathcal{P}(i),i}$ has the probability mass function:

$$P(W_{\mathcal{P}(i),i}(t_1^+) = n | \text{stage } i \text{ is in regular interval}) = 0, \quad \forall n > 0$$

With probability P', $IL_{\mathcal{P}(i)}(t_1) < W_{\mathcal{P}(i),i}(t_1)$, and stage i enters a recovery interval. If $IL_{\mathcal{P}(i)}(t_1) > 0$, only an amount $IL_{\mathcal{P}(i)}(t_1)$ of inventory is available to replenish the shortage $W_{\mathcal{P}(i),i}(t_1)$, and

$$W_{\mathcal{P}(i),i}(t_1^+) = W_{\mathcal{P}(i),i}(t_1) - IL_{\mathcal{P}(i)}(t_1)$$

If $IL_{\mathcal{P}(i)}(t_1) \leq 0$, there is no inventory on hand to do anything, and

$$W_{\mathcal{P}(i),i}(t_1^+) = W_{\mathcal{P}(i),i}(t_1)$$

$W_{\mathcal{P}(i),i}(t_1^+)$ does not have a simple, concise pmf during a recovery interval, as it involves $W_{\mathcal{P}(i),i}(t_1)$ and $IL_{\mathcal{P}(i)}(t_1)$. This complicates the calculation for $W_{\mathcal{P}(i),i}$. We would prefer an easier approach for its pmf, which we develop next.

The Long-Run Probability of Being in a Recovery Interval

With probability P' stage i is in a recovery interval, and $W_{\mathcal{P}(i),i}$ is relatively hard to compute in this interval. Let us to consider the probability of being in the recovery interval.

As discussed in last paragraph, this recovery interval depends on $IL_{\mathcal{P}(i)}(t_1)$ and $W_{\mathcal{P}(i),i}(t_1)$. We explore their relationship here. Assume the local inventory transit position of stage $\mathcal{P}(i)$ equals y. According to "conservation of flow" (3.1), we need to consider

$$IL_{\mathcal{P}(i)} - W_{\mathcal{P}(i),i} = y - \sum_{j\in\mathcal{S}(\mathcal{P}(i))} S_j - D_{\mathcal{P}(i)} - B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)} + \sum_{j\in\mathcal{S}(\mathcal{P}(i)),j\neq i} W_{\mathcal{P}(j),j} - W_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)}$$

If this is positive and the disruption ends at t_1, stage i will go into a the regular interval, otherwise it goes into a recovery interval. The first part $y-\sum_{j\in\mathcal{S}(\mathcal{P}(i))} S_j - D_i - B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)}$ is the inventory level for stage i during times with no disruption, and it is assumed to have a much greater probability of being positive than being negative, since the stockout cost is normally higher than the holding cost. The remaining terms, $\sum_{j\in\mathcal{S}(\mathcal{P}(i)),j\neq i} W_{\mathcal{P}(j),j}$ and $W_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)}(t_1)$, are random variables depending on the disruption profiles, and their distributions are relatively hard to analyze, as they are interconnected with each other, especially during this recovery interval.

One simple way to address this difficulty is to assume

$$P' = 0.$$

This assumption is saying $IL_{\mathcal{P}(i)} - W_{\mathcal{P}(i),i} > 0$ at all times, and when a supply disruption for stage i is over, it enters a regular period immediately. In other words, there is no more recovery period. This assumption can also be interpreted in another way: when stage i's supply is disrupted, its supplier $\mathcal{P}(i)$ can not ship any new inventory to i, but $\mathcal{P}(i)$ can keep ordering to prepare for the large demand caused by $W_{\mathcal{P}(i),i}$; once the disruption is over,

$\mathcal{P}(i)$ can mitigate the shortage caused by the disruption immediately. This simplifies our computation, and we have

$$P(W_{\mathcal{P}(i),i}(t) = n|\text{stage } i \text{ is not disrupted at time } t) = 0, \forall n > 0.$$

This concludes the computation for the probability mass function of $W_{\mathcal{P}(i),i}$. Obviously, our approach for calculating P', and thus the distribution of $W_{\mathcal{P}(i),i}$, is approximate.

3.4.2 Estimation of $B_{\mathcal{P}(i),i}$

The computation of $B_{\mathcal{P}(i)}$, and therefore $B_{\mathcal{P}(i),i}$, is another key part to our heuristic algorithm. For stage $\mathcal{P}(i)$, assume its echelon inventory transit position equals y, and the amount of inventory in transit to $\mathcal{P}(i)$ is $D_{\mathcal{P}(i)}$. According to "conservation of flow" (3.1), we have

$$IL_{\mathcal{P}(i)} = y - D_{\mathcal{P}(i)} - B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)} + \sum_{j \in \mathcal{S}(\mathcal{P}(i))} W_{\mathcal{P}(i),j} - W_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)} - \sum_{j \in \mathcal{S}(\mathcal{P}(i))} S_j.$$

This gives us a way to calculate $B_{\mathcal{P}(i)}$:

$$B_{\mathcal{P}(i)} = \left[y - D_{\mathcal{P}(i)} - B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)} + \sum_{j \in \mathcal{S}(\mathcal{P}(i))} W_{\mathcal{P}(i),j} - W_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)} - \sum_{j \in \mathcal{S}(\mathcal{P}(i))} S_j \right]^{-}.$$

In this equation, it is obvious that $B_{\mathcal{P}(i)}$ is related to $B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)}$, while $B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)}$ is related to $S_{\mathcal{P}(\mathcal{P}(i))}$. But $S_{\mathcal{P}(\mathcal{P}(i))}$ is unknown when calculating $B_{\mathcal{P}(i)}$, thus it makes this hard to calculate. A simple way to deal with this difficulty is to approximate $B_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)}$ with 0, which greatly simplifies the problem.

$B_{\mathcal{P}(i),i}$ is defined as the shortage at stage i only due to backorders at stage $\mathcal{P}(i)$, excluding the portion of shortage due to supply disruptions. Therefore the backorder at stage $\mathcal{P}(i)$ is

$$B_{\mathcal{P}(i)} = \left[y - D_{\mathcal{P}(i)} + \sum_{j \in \mathcal{S}(\mathcal{P}(i))} W_{\mathcal{P}(i),j} - W_{\mathcal{P}(\mathcal{P}(i)),\mathcal{P}(i)} - \sum_{j \in \mathcal{S}(\mathcal{P}(i))} S_j \right]^{-}.$$

Moreover, these backorders are shared proportionally among its successors $\mathcal{S}(\mathcal{P}(i))$. Hence

we have

$$B_{\mathcal{P}(i),i} \sim \text{Bin}\left(B_{\mathcal{P}(i)}, \frac{\lambda_i}{\sum_{j\in\mathcal{S}(\mathcal{P}(i))}\lambda_j}\right), \quad \forall i \in 1, 2, \cdots, N.$$

3.4.3 Recursive Optimization Heuristic

The system cost can be evaluated with the introduction of three auxiliary functions. $\hat{C}_i(x)$ is the echelon cost at stage i when the local inventory level equals x, including the inventory holding cost for the subtree rooted at stage i and the backorder cost in the subtree. $C_i(y)$ represents the expected echelon inventory cost at stage i when its echelon inventory transit position equals y. $\underline{C}_i(v)$ is defined as the expected cost when the inventory shortage at $\mathcal{P}(i)$ due to stage i is v. $\underline{C}_i(v)$ is defined differently from the classical literature to facilitate the computation here.

For serial systems, the classical algorithm (Clark and Scarf [16]) start from the location facing direct outside customer demand, and moving upward to its suppliers. It recursively finds the optimal base-stock levels for each stage by optimizing covex single-variable functions. Similarly, we apply this idea to distribution systems under supply disruptions.

The Recursive Optimization Heuristic

1. For leaf nodes $i \in L$, we have

$$\begin{aligned}
\hat{C}_i(x) &= h_i x + (H_i + b_i)x^- \\
C_i(y) &= E_{W_{\mathcal{P}(i),i}} E_{D_i}[\hat{C}_i(y - D_i - W_{\mathcal{P}(i),i})] \\
S_i &= \arg\min_y C_i(y) \\
\underline{C}_i(v) &= C_i(S_i - v)
\end{aligned}$$

2. For non leaf nodes $i \in V \setminus L$, we have

$$\begin{aligned}
\hat{C}_i(x, \mathbf{B}_i) &= h_i x + \sum_{j\in\mathcal{S}(i)} \underline{C}_j(B_{i,j}) \\
C_i(y|S) &= E_{\mathbf{B}_i(y)} E_{W_{\mathcal{P}(i),i}} E_{D_i}[\hat{C}_i(y - D_i - W_{\mathcal{P}(i),i}, \mathbf{B}_i(y))] \\
S_i &= \arg\min_y C_i(y)
\end{aligned}$$

$$\underline{C}_i(v) = C_i(S_i - v)$$

Here $\mathbf{B}_i = \{B_{i,j}\}_{j \in \mathcal{S}(i)}$ denotes the random vector of inventory shortage at location $\mathcal{S}(i)$ due to stockouts at stage i. Similar to the algorithm introduced by Clark & Scarf [16] for serial systems, this algorithm also works in a bottom-up fashion. It first starts from every leaf node. For any non-leaf node i, once the base-stock levels for all its successors have been determined, the algorithm finds the base-stock level for i by finding y that minimizes $C_i(y)$.

This recursive optimization heuristic does not guarante an optimal solution for the distribution system, neither for the leaf nodes, nor for the non-leaf nodes. This algorithm assumes that the leaf nodes are completely independent from each other, while in fact a shortage at one leaf node i might require more inventory from its predecessor $\mathcal{P}(i)$, thus $\mathcal{P}(i)$'s other successors might be affected. Although our heuristic algorithm does not find the optimal base-stock levels necessarily, it still yields good performance according to our numerical studies.

3.5 Numerical Experiments

In this section, we test how well our recursive optimization algorithm performs, and compare it with the solution found by coordinate descent method with bisection search.

We test our algorithms on several different distribution system structures. System 1 that we consider is shown in Figure 3.4.

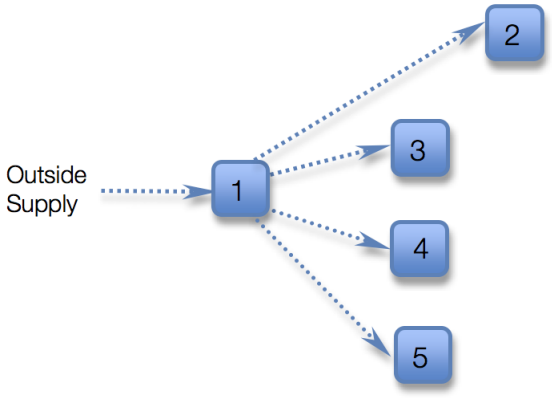

Figure 3.4: One-Warehouse, Four-Retailer System

This is a one-warehouse, four-retailer system, with retailer 2, 3, 4 and 5 ordering from the warehouse, stage 1. All stages are under supply risk. The leadtime is $[l_1, l_2, l_3, l_4, l_5] = [2, 3, 2, 2, 2]$, and the penalty cost is $p = 100$ per unit per time for stages 2, 3, 4, and 5. The demand rate is $[\lambda_2, \lambda_3, \lambda_4, \lambda_5] = [2.5, 7.5, 3.5, 4.5]$. In this system, we use different combinations of local inventory holding cost, and disruptions parameters. We test all 48 combinations of the holding cost and disruption parameters are shown in Table 3.1.

$(h_1, h_2, h_3, h_4, h_5)$	Average disruption length	Average time between disruption
(1, 10, 5, 5, 5)	(0.4, 0.667, 0.667, 0.667, 0.667)	(4, 2, 2, 2, 2)
(1, 5, 10, 5, 5)	(0.667, 0.4, 0.667, 0.667, 0.667)	(2, 4, 2, 2, 2)
(1, 5, 5, 10, 5)	(0.667, 0.667, 0.4, 0.667, 0.667)	(2, 2, 4, 2, 2)
	(0.667, 0.667, 0.667, 0.4, 0.667)	(2, 2, 2, 4, 2)

Table 3.1: System 1 parameters for performance test

For this instance, we calculate the base-stock levels with the recursive optimization algorithm. The computation time for this algorithm is within 5 seconds on average, relatively fast. But to evaluate the solution quality is relatively hard. First, there is no way to calculate the exact expected inventory cost per time unit, even given the base-stock levels for each stage. Instead we test the solution's performance by simulation. We simulate this five stage distribution system in C++ with the given parameter settings, generating random samples for both demand and disruption. The simulation runs for 5000 time units, and this simulation process takes a relatively longer time. Second, there is no way to calculate the optimal base-stock levels for this system directly. For small sized problems, it is possible to use exhaustive search over all possible base-stock level vectors. However, this is too cumbersome even for a five stage system, as it would take hundreds of hours to find the optimal base-stock levels for merely one instance. Instead, we use coordinate descent search for finding base-stock levels for comparison. For each line search, we use bisection search to find the minimum for each coordinate.

We refer to the parameter settings using the notation (c_1, c_2, c_3), which means this instance has the c_1-th holding cost, c_2-th average disruption length, and c_3-th average time between disruptions from Table 3.1. For these 48 different instances, we obtain the computational results in the following table.

Instance No.	Algorithm	Coordinate Descent	Gap	Computation Time
(1,1,1)	333.768	329.499	0.012956033	4.496026
(1,1,2)	333.009	329.851	0.009574020	4.366212
(1,1,3)	308.458	306.550	0.006224107	4.365593
(1,1,4)	330.603	327.463	0.009588870	4.301088
(1,2,1)	337.905	333.360	0.013633909	4.305687
(1,2,2)	355.055	347.097	0.022927309	4.216075
(1,2,3)	329.633	319.244	0.032542507	4.246899
(1,2,4)	346.966	339.083	0.023247995	4.222778
(1,3,1)	307.282	302.340	0.016345836	4.276114
(1,3,2)	322.625	310.933	0.037602956	4.207488
(1,3,3)	318.835	307.254	0.037691942	4.248709
(1,3,4)	319.746	308.142	0.037657963	4.230181
(1,4,1)	335.449	329.306	0.018654382	4.257256
(1,4,2)	347.150	337.525	0.028516406	4.205890
(1,4,3)	326.937	314.893	0.038247913	4.225795
(1,4,4)	350.924	341.150	0.028650154	4.218673
(2,1,1)	342.002	336.730	0.015656461	4.296909
(2,1,2)	342.261	338.022	0.012540604	4.252085
(2,1,3)	315.098	312.570	0.008087788	4.316965
(2,1,4)	339.198	335.299	0.011628427	4.242342
(2,2,1)	347.866	342.257	0.016388270	4.290649
(2,2,2)	366.298	357.876	0.023533291	4.229326
(2,2,3)	337.434	328.568	0.026983760	4.275360
(2,2,4)	358.139	349.842	0.023716421	4.256779
(2,3,1)	312.373	308.201	0.013536621	4.331191
(2,3,2)	329.810	319.203	0.033229638	4.301468
(2,3,3)	327.840	314.592	0.042111688	4.284425
(2,3,4)	325.333	316.049	0.029375192	4.259284

(2,4,1)	344.516	337.509	0.020760928	4.313255
(2,4,2)	359.321	347.992	0.032555346	4.297210
(2,4,3)	334.140	323.564	0.032685960	4.278498
(2,4,4)	361.498	351.254	0.029164081	4.214616
(3,1,1)	334.754	329.820	0.014959675	4.330365
(3,1,2)	334.232	330.598	0.010992202	4.285378
(3,1,3)	309.052	306.962	0.006808660	4.290658
(3,1,4)	330.307	327.441	0.008752722	4.293266
(3,2,1)	339.091	334.174	0.014713892	4.337701
(3,2,2)	355.292	348.496	0.019500941	4.276365
(3,2,3)	329.798	320.623	0.028616163	4.280471
(3,2,4)	347.332	339.982	0.021618792	4.247018
(3,3,1)	307.573	302.460	0.016904715	4.378230
(3,3,2)	323.790	312.020	0.037721941	4.354901
(3,3,3)	318.402	307.977	0.033849930	4.303414
(3,3,4)	319.513	308.387	0.036078045	4.269379
(3,4,1)	332.715	328.757	0.012039287	4.323005
(3,4,2)	345.317	337.847	0.022110600	4.245490
(3,4,3)	323.507	314.854	0.027482579	4.314054
(3,4,4)	347.390	341.030	0.018649386	4.356956
Max	366.298	357.876	0.042111688	4.496026
Min	307.282	302.34	0.006224107	4.205890
Average	334.282	326.972	0.022433673	4.285781

Table 3.2: System 1 test results

Based on our simulation for these 48 different instances, the recursive optimization algorithm finds a set of base-stock levels that yield a performance with an average 2.24% gap, and 4.21% maximum gap, as well as 0.62% minimum gap.

We also tested our algorithm on a three-echelon system, as shown in Figure 3.5.

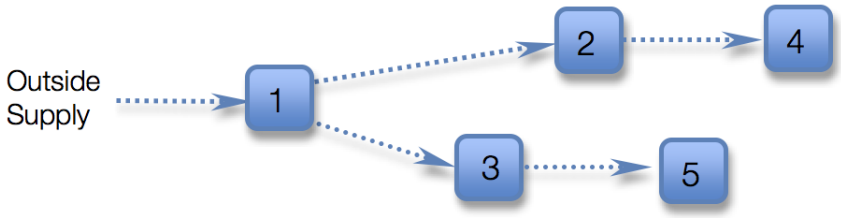

Figure 3.5: Three-Echelon Distribution System

This is a three-echelon system, with retailer 4 and 5 ordering from warehouse stages 2 and 3 repectively, and warehouses 2 and 3 replenishing from factory 1. All stages are under supply risk. The leadtime is $[l_1, l_2, l_3, l_4, l_5] = [2, 3, 2, 2, 2]$, and the penalty cost is $p = 100$ per unit per time for stages 2, 3, 4, and 5. The demand rate is $[\lambda_4, \lambda_5] = [3.5, 4.5]$. In this system, we use different combinations of local inventory holding cost and disruptions parameters, as well. We test all 64 combinations of the holding cost and disruption parameters shown in Table 3.3.

$(h_1, h_2, h_3, h_4, h_5)$	Average disruption length	Average time between disruption
(1, 2, 1.75, 3.5, 4.5)	(0.4, 0.667, 0.667, 0.667, 0.667)	(4, 2, 2, 2, 2)
(1, 1.75, 2, 4.5, 3.5)	(0.667, 0.4, 0.667, 0.667, 0.667)	(2, 4, 2, 2, 2)
(1, 2, 1.75, 4.5, 3.5)	(0.667, 0.667, 0.4, 0.667, 0.667)	(2, 2, 4, 2, 2)
(1, 1.75, 2, 3.5, 4.5)	(0.667, 0.667, 0.667, 0.4, 0.667)	(2, 2, 2, 4, 2)

Table 3.3: System 1 parameters for performance test

As before we calculate the base-stock levels with the recursive optimization algorithm for this system, and we use coordinate descent search for finding the optimal base-stock levels. We compare the simulated average inventory cost using base-stock levels from our heuristic and from the coordinate descent method. For these 64 different instances, we obtain computational results in the following table in the same manner.

Instance No.	Algorithm	Coordinate Descent	Gap	Computation Time
(1,1,1)	308.396	302.036	0.021057093	1.481445
(1,1,2)	304.971	299.720	0.017519685	1.486538
(1,1,3)	303.963	298.081	0.019732891	1.452269
(1,1,4)	294.931	286.088	0.030910070	1.443500
(1,2,1)	306.048	300.147	0.019660366	1.440851

(1,2,2)	312.724	303.027	0.032000449	1.465480
(1,2,3)	309.259	299.317	0.033215621	1.448981
(1,2,4)	296.830	287.158	0.033681806	1.442139
(1,3,1)	303.164	293.958	0.031317399	1.433473
(1,3,2)	307.494	294.734	0.043293275	1.448619
(1,3,3)	311.498	296.015	0.052304782	1.489318
(1,3,4)	296.352	280.741	0.055606413	1.446018
(1,4,1)	286.437	277.243	0.033162244	1.425333
(1,4,2)	288.599	279.009	0.034371651	1.439938
(1,4,3)	288.602	276.773	0.042738995	1.428596
(1,4,4)	290.254	278.865	0.040840550	1.439884
(2,1,1)	307.625	301.618	0.019915920	1.420893
(2,1,2)	303.804	299.125	0.015642290	1.428191
(2,1,3)	301.719	298.728	0.010012453	1.438345
(2,1,4)	290.480	286.358	0.014394569	1.464393
(2,2,1)	304.051	299.472	0.015290244	1.443602
(2,2,2)	310.703	302.720	0.026370904	1.453025
(2,2,3)	305.646	299.822	0.019424859	1.444528
(2,2,4)	294.778	287.044	0.026943604	1.460868
(2,3,1)	299.936	295.313	0.015654577	1.431490
(2,3,2)	304.683	296.120	0.028917331	1.470644
(2,3,3)	307.555	298.763	0.029428008	1.465821
(2,3,4)	289.819	282.747	0.025011760	1.450274
(2,4,1)	281.925	277.775	0.014940149	1.429031
(2,4,2)	285.395	279.086	0.022605935	1.437430
(2,4,3)	282.934	277.970	0.017858042	1.452474
(2,4,4)	288.605	279.692	0.031867197	1.449941
(3,1,1)	308.962	300.307	0.028820507	1.440473
(3,1,2)	304.020	297.244	0.022796087	1.428477

(3,1,3)	304.209	297.598	0.022214531	1.439822
(3,1,4)	291.834	284.895	0.024356342	1.439082
(3,2,1)	304.126	298.004	0.020543348	1.424192
(3,2,2)	309.830	300.876	0.029759768	1.462393
(3,2,3)	306.835	297.962	0.029778965	1.448338
(3,2,4)	293.952	284.981	0.031479292	1.458598
(3,3,1)	303.446	294.137	0.031648518	1.449693
(3,3,2)	307.043	294.369	0.043054805	1.473040
(3,3,3)	309.401	295.798	0.045987464	1.454740
(3,3,4)	293.276	280.149	0.046857208	1.472656
(3,4,1)	283.271	276.152	0.025779281	1.454977
(3,4,2)	285.619	277.041	0.030962926	1.468801
(3,4,3)	285.372	275.561	0.035603732	1.601772
(3,4,4)	290.929	277.052	0.050088070	1.574551
(4,1,1)	308.950	303.869	0.016721021	1.569934
(4,1,2)	306.557	301.975	0.015173442	1.700061
(4,1,3)	303.354	300.262	0.010297673	1.610240
(4,1,4)	293.577	288.850	0.016364895	1.751267
(4,2,1)	307.762	302.271	0.018165818	1.567337
(4,2,2)	315.225	305.199	0.032850697	1.809944
(4,2,3)	309.815	302.248	0.025035732	1.689135
(4,2,4)	297.656	289.494	0.028194021	1.642737
(4,3,1)	301.523	296.104	0.018301002	1.691942
(4,3,2)	306.878	297.534	0.031404814	1.773961
(4,3,3)	309.065	298.952	0.033828173	1.454972
(4,3,4)	292.896	283.678	0.032494589	1.672950
(4,4,1)	285.092	280.100	0.017822206	1.432178
(4,4,2)	288.374	281.460	0.024564769	1.573679
(4,4,3)	286.164	280.129	0.021543646	1.511542

(4,4,4)	288.932	281.758	0.025461566	1.604674
Max	315.225	305.199	0.055606413	1.809944
Min	281.925	275.561	0.010012453	1.420893
Average	299.2675781	291.2699063	0.027556969	1.498460781

Table 3.4: System 2 test results

Based on our simulation for these 64 different instances, the recursive optimization algorithm finds a set of base-stock levels that yield a performance with an average 2.76% gap, and 5.56% maximum gap, as well as 1.00% minimum gap.

From these numerical experiments on these two different network structures, we can see that our recursive heuristic algorithm does provide relatively fast solutions to problems with different parameter settings, normally within seconds. It reaches average gaps of 2.24% and 2.74% respectively, which means it find solutions with good quality for both network structures.

3.6 Conclusion and Future Work

In this work, we introduce the recursive optimization heuristic to calculate the base-stock levels for a distribution inventory system under supply disruption risk. The main idea is to modify the classical recursive algorithm for serial systems to adapt to our distribution system under supply risk. We analyze the effects of supply disruptions on the inventory shortage for each stage first, then incorporate the effects into our recursive optimization algorithm. Our recursive algorithm works in a "bottom-up" way, which starts from leaf nodes, then moves upstream toward the root node. Whenever it calculates the base-stock level for a stage, it takes the inventory cost at all its successors into consideration.

We assess the performance of the heuristic by comparing the results from it and the results obtained by implementing a coordinate descent search. Our heuristic not only provides results very close to optimal, but also yields solution relatively fast. Finding the exact base-stock levels for multi-echelon systems takes days or more of computational time, while

our recursive algorithm finds solutions within seconds.

For research directions, we plan to investigate whether the Decomposition-Aggregation heuristic (Rong et al. (2012)) can be applied to distribution systems with supply risk. It works very well for the case without supply risk. We can examine whether the effects of disruptions can be incorporated into their heuristics or not, and if so, how well it would perform. Our heuristics could possibly be applied to more generalized settings. Other possible extensions we plan to work on are fixed ordering costs, as well as stochastic leadtimes.

Chapter 4

A Bilevel Model for Retail Electricity Pricing with Flexible Loads

4.1 Introduction

In the USA, traditional electric utility and transmition systems were designed in an integrated way to serve customers. This worked extremely well for decades until the deregulation of electricity markets. As the economy grows, demand and supply vary continuously. However, electricity is hard to store by nature, and it has to be available on demand. Balancing this supply-demand relationship requires a large number of financial transactions. Consequently there is a physical requirement for a controlling agency to coordinate the dispatch of electricity units to match actual demand across the power grid. Independent System Operators (ISO) and Regional Transmission Organizations (RTO) were established to handle the vastly increased number of transactions that take place in a competitive environment. The day-ahead energy market emerged after the deregulation of electricity. It is a financial market where participants purchase and sell electricity for the following day under binding day-ahead prices. If there is a difference between the amount purchased from the day-ahead market and the amount required, electricity can also be purchased from

the real-time market. The prices in the real-time market are determined by the location marginal price (LMP) algorithm to balance electricity demand from available generation units. Real-time prices are often higher than those in the day-ahead market.

Electricity service providers (utilities, electricity retailers, etc.) provide electricity to users. One of their main tasks is to satisfy all demand from consumers. But aggregated daily consumer demand profiles typically exhibit significant some peaks, which would incur large costs for grid operators, as a great amount of generation capacity is reserved but rarely used. Traditionally consumers receive a flat rate electricity price all day long. This makes consumers indifferent to the time of day for their electricity consumption. For many utilities, peak demand occurs during afternoon and early evening as a result of flat prices. A huge amount of resources have to be allocated to prepare for peak hour usage, and the cost to satisfy peak demand is also tremendous.

If the service provider is allowed to fluctuate its prices, and the consumer is capable of responding to price fluctuations, marginal-cost based prices will encourage more efficient energy usage, and thus help to mitigate these problems. Therefore, service providers use demand response programs to encourage consumers to shift demand from peak hours to off-peak hours, or ideally to generate demand that matches the available supply. Typically demand response programs aim at some of the following objectives: reducing peak hour consumption, shifting loads to adjacent non-peak hours, and decrease peak-to-average ratios in load demand. It is believed that demand response may have a huge impact as smart grids are put into practice.

There are two common types of demand response programs. One is direct load control. In this approach, the electricity service provider can remotely control the operations of certain appliances, based on an agreement between the service provider and the consumer. For example, it may adjust the operations of thermal comfort equipment, refrigerators, etc., to accommodate its other concerns. However, this raises great concerns on users' privacy when it comes to residential customers, and its implementation is limited. The alternative approach is smart pricing, where consumers are allowed and encouraged to manage their own loads. Service providers normally set the price high during the peak hours, and set the price low in the off-peak hours. So there are incentives for consumers to reduce short-term

demand when capacity is tight, increase usage when capacity is abundant, or even delay non-urgent electricity demand from peak hours to non-peak hours.

Real-time pricing (RTP), critical-peak pricing (CPP), and time-of-use (TOU) pricing are among the most popular pricing schemes. For a typical real-time pricing scheme, the electricity prices vary at different times of the day. They would normally be higher in peak hours like afternoon and early evening. The consumer is expected to respond to the time-differentiated prices by shifting their own loads from the high price hours to the low price hours.

To better estimate potential peak hour demand and shift demand from peak hours to non-peak hours, utilities need to understand demand as well as its patterns responding to price variation among different hours of the day. For TOU pricing, electricity prices are announced for each specific time period in a day in advance. Customers know this pricing information in advance, which allows them to decide their usage in response to price variations throughout a day. Typically demand peaks form during afternoon and early evening under flat rate pricing. Under TOU pricing, while the price effects of hourly pricing might discourage customers from concentrating consumption during peak hours, it might inadvertently encourage customers to delay demand to adjacent hours, thus forming a new "rebound" peak and leaving the service provider no better off even than flat rates. As a result of day ahead pricing, the possibility of a rebound peak could increase since price rate adjustment is not permitted after the price has been announced in advance.

CPP are very similar to TOU pricing, except for certain peak days, when prices could reflect the actual generation or purchasing cost in the wholesale market. CPP has the same problem as TOU pricing, in that a rebound peak might form during adjacent hours to peak hours. RTP, also known as dynamic pricing, allows electricity prices to vary as often as hourly. Customers receive price signals on an advanced basis, which reflect the utility's generation or purchasing cost in the wholesale market. Taylor et al. [81] showed hourly RTP tariffs have been implemented successfully for large industrial and commercial firms. We investigate a special type of RTP, where the price is dependant on the actual electricity consumption quantity. Higher electricity costs occur as a result of more consumption. This pricing mechanism discourages consumers from concentrating their load at a specific time,

and encourage them to evenly spread out their consumptions.

There has be various demand side management work focusing on the interaction between the utility company and end users, by implementing different pricing schemes. Typically, each user is expected to respond to the time-differentiated price set by the utility company by shifting its loads, and the utility company tries to either minimize total cost or maximize total social wellfare. We provide a brief review on this literature in section 4.2.1. In contrast, we consider the interaction between the utility and the consumer from a new perspective.

In this work, we consider a service provider and a large electricity consumption entity. The service provider's objective is to encourage this entity to generate a demand profile that the service provider has identified a priori as desirable. The service provider may also take deviation from the desired load profile into consideration. If the actual demand is greater than the desired amount, this demand surplus has to been satisfied by purchasing more electricity from the real-time market, which results in increased costs. If the actual demand is less than the desired amount, some generating unit has to be put back into spinning reserve, which also results in some extra cost or underutilization that would diminish system performance. Meanwhile, the electricity consumer's objective is to determine when to operate its own devices to minimize the total cost, accounting for the exogenous electricity cost as given by the service provider, as well as the inconvenience of delaying certain devices' operation when needed. We consider both deterministic and stochastic demand in this work.

This problem can be naturally modeled as a bilevel programming problem: the service provider (the leader) sets prices, and the consumer (the follower) makes his own consumption decisions based on the electricity prices. Normally bilevel programming problems cannot be solved directly. In our work, we investigate the properties of the lower level problem. For the deterministic demand case, we analyze the optimality conditions. These conditions allow us to write the lower level problem in terms of the optimality conditions, and the whole problem can be reduced into a single level problem. This newly obtained single level problem can be solved easily in CPLEX. For the stochastic demand case, we investigate the best response function of the follower, and approximate its best response function to facilitate calculating the leader's optimal strategy.

4.2 Literature Review

Our work studies the interaction between the electricity service provider and electricity consumer. There are several research fields related to our topic.

4.2.1 Demand Side Management

Demand-side management (DSM) is fundamental in upgrading the aging power grid into a more reliable and economically operated smart grid. It has a much needed positive impact in smoothing peak hour demand, lowering generation cost, and so on. Many works modeling demand side management have already been undertaken. Hobbs et al. [37] present a bilevel model where the electric utility at the upper level tries to minimize costs or maximize benefits while controlling electric rates and subsidizing energy conservation programs, while the lower level maximizes consumers' net benefits by consuming electricity and investing in conservation. It also considers factors such as free riders and rebound effect. Hamalainen et al. [35] examines a model in which a coalition of consumers can be set up in a hierarchical framework. The coalition buys electricity from a wholesale market, and its objective is to optimize the total welfare of the coalition. The individual consumers optimize dynamically their own consumption based on the price. A Time-of-Use pricing program is considered for this model. A simulation is used to find the optimal Time-of-Use price based on the optimal marginal price. Lavigne et al. [44] study the electricity market under different pricing mechanisms. They present a heuristic decomposition technique to solve pure competition, regulated and Stackelberg-type (tempered monopoly) equilibria. Energy sectors are modeled through activity analysis models. An application to the province of Quebec is presented.

Mohsenian-Rad et al.[55] studies demand-side management programs focusing on an energy consumption scheduling game, where the users play the game by scheduling their household appliances and load. The utility company can adopt pricing tariffs at different times and energy levels. They include the interactions among the users. A distributed algorithm is proposed to find the Nash Equilibrium. Their results show the users can maintain privacy and they would have an incentive to participate in the energy consumption

scheduling game. Samadi [68] consider a real-time pricing algorithm to encourage energy consumption interaction among subscribers and the energy provider. The authors formulate the real-time pricing problem as an optimization problem to maximize the aggregated users' utility, which is each user's preferences and engergy consumption patterns. Tarasak [80] extends their work to include the effect of load uncertainty. The authors derive the optimal prices under three types of load uncertainty: bounded uncertainty model, Gaussian model and unknown distribution model. They also show its influences on power consumption and generating capacity. Samadi [70][69] further extend this work by proposing a Vickrey-Clarke-Groves (VCG) mechanism which aims to maximize the aggregate utility functions of all users minus the total energy cost. This mechanism requires each user to provide its demand information, while the energy provider will determine each user's electricity bill payment. The authors study some properties of this mechanism and show it could benefit both users and utility companies.

Zhu et al. [87] study an electricity consumption scheduling mechanism for residential load management using linear integer programming. The authors minimize peak hourly load to achieve a balanced daily load schedule. Huang et al. [38] study an operation scheme that allows the utility to perform demand response and power procurement jointly, to maximize social welfare. The authors develop a low complexity algorithm called the welfare maximization algorithm to perform power procurement and dynamic pricing jointly. Their results show this algorithm reaches a close to optimal utility. Bu et al. [11] consider a problem in which multiple electricity retailers can co-exist in the system, and they can compete or cooperate with each other to reach the highest revenue, either individually or together. The authors consider traditional fixed rate users and opportunistic electricity users who can change their own demand, or even turn to another retailer. They present two game formulations, and some results to show the effectiveness of the proposed real-time pricing scheme. Bu et al. [10] study the interaction between the retailer and electricity customers as a four-stage Stackelberg game. Simulation results show the effectiveness of the proposed scheme, as well as the effects of system parameters on procurement and price decisions. Li et al. [49] study the effect of demand side mangement in reducing peaks and adapting elastic demand to fluctuating generation. The authors show that time-varying prices can align

individual optimality with social optimality, which means when each household optimizes its own benefits, social welfare is maximized automatically. They propose a distributed algorithm to jointly compute the optimal prices and demand schedules.

Gatsis and Giannakis [26] study the problem of minimizing the electricity provider cost plus the total user dissatisfaction, subject to the individual load constraints. The authors solve this problem by a distributed subgradient method with Lagrange multipliers. Lee et al. [48] present a design of an electricity consumption scheduler for a demand respsonse system. The authors also evaluate its performance in smart grid homes or buildings, aiming at reducing the peak load in them and system wide networks. Qian et al. [63] investigate a real time pricing scheme to reduce the peak-to-average ratio through demand response by solving a two stage optimization problem. The user maximizes the gap betwen its quality-of-usage and payment, while the retailer maiximizes its profits. The authors develop a simulated-annealing-based price control algorithm to solve the optimization problem.

Kishore and Snyder [41] consider the electricity conumption both within a home and across several within a neighborhood. They provide an optimizaiton model to schedule appliance to take advantage of lower electricity rates at off peak hours, but simulation indicates resulting solution might be more picky thereby negating the benefits. Then they provides a distributed scheduling mechanism to shape peak load. And they also introduce an EMC optimizaiton model that account for electricity capacity constraints based on dynamic programming.

Palensky and Dietrich [58] present an overview for demand side management, and analyze several types of demand side managment. We refer to this paper for more details on demand-side management.

4.2.2 Mathematical Programs with Equilibrium Constraints (MPEC)

A large body of literature considers the interaction between the electricity service provider and electricity consumers. This interaction can be captured using game theory, and formulated as a bilevel programming problem. This overlaps with many works mentioned previously, and some existing work formulate this intrinsically hard problem as an MPEC. They are closely related to our work, and thus we provide a brief review o them.

Hobbs et al. [36] works on an oligopolistic market with several dominant generating firms in an electric power network. They model a single firm as a MPEC while the lower level problem of maximizing social welfare is replaced by its KKT conditions. A penalty interior point method is used to solve the single firm problem. The paper also includes numerical examples for multi-firm Nash equilibria in which each player solves an MPEC of the single-firm type.

Bjørndal and Jørnsten [8] study a Stackelberg game for a network-constrained energy market. They reformulate equilibrium conditions of the associated MPEC into disjunctive constraints, and nonlinear terms are linearized, so that the MPEC can be reformulated as a mixed-integer linear program.

Chen et al. [13] proposes a Real-Time Pricing based power scheduling scheme as demand response for residential power usage. This scheme also considers the Stackelberg game between service provider and enery management controller. The equilibrium is obtained by information exchange between them. The simulation indicates that the scheme could both save cost for consumer, and reduce peakload and variance between demand and supply.

Saad et al. [67] provide a review of game theory applied to smart grid. This paper summarizes the potential of applying game theory for addressing relevant and timely open problems, including micro-grid systems, demand-side management, and communications. We refer to this paper for more details, especially on game theory applications for demand-side management. Dempe et al. [19] summarize the main directions of research and applications of bilevel programming. This paper provides a good review of MPECs, especially useful for solving other MPEC problems.

Most modeling studies focus on maximizing total social welfare or minimizing total cost for the game leader. In our work, we consider something similar but different. In a deregulated electricity market, an electricity generation and consumption plan is determined in a day-ahead market. Any new demand must be satisfied by purchasing electricity from the real-time market at an extra cost. Consumers who produce with less demand than expected would also probably incur extra costs for the service provider as some generators would be sent to spinning reserve. The service provider must satisfy all the demand from the consumer, but it would also prefer this consumption profile to be as close as possible

to the predesigned demand profile. We are more interested in how close the real demand profile can get to the predetermined demand profile using the pricing scheme to encourage the consumer to shift loads.

4.3 Stackelberg Game Model

Consider a smart power system with a huge demand customer. Assume it has enough demands in each time period so that the demand in each period can be considered to be infinitely divisible. We also assume that the customer is equipped with a smart meter, which can receive real-time pricing signals and schedule the devices' operations. As there are so many demands, we only consider load scheduling in the aggregate scale.

This work considers a time horizon of T periods. In period t, the consumer has an original unshifted deterministic demand D_t, and the service provider has a desired load $O_t > 0$. We relax te deterministic demand assumption in section 4.6. The amount of demand that occurs at time i, that is shifted to time t, is denoted as x_i^t. Assume there is a constant inconvenience cost C per unit per period for delaying demand.

We assume teprice charged by the service provider is a quadratic function of the load, equal to $a_t x^2 + b_t x$, wher x is the load. The service provider can determine the electricity price scheme by picking pricing parameters a_t and b_t. If the service provider sees a total load of $\sum_{i \le t} x_i^t$ from all customers at time t, then the total electricity cost charged at time t is:

$$a_t \left(\sum_{i \le t} x_i^t \right)^2 + b_t \sum_{i \le t} x_i^t.$$

We assume a_t and b_t can be chosen within the intervals $[a^L, a^U]$ and $[b^L, b^U]$, respectively. We use this pricing scheme to discourage large concentrations of electricity consumption in any specific period. The electricity unit price increases simultanously as the consumption increases due to the second order term in this pricing function. We use A_t^+, A_t^- to denote the surplus and deficit respectively from the desired load profile at time t.

This problem can be formulated as a bilevel programming problem:

$$\min_{a_t, b_t, A_t^+, A_t^-, \bar{x}^t} \sum_t \left(A_t^+ + A_t^-\right) \tag{4.1}$$

$$\text{s.t.} \quad A_t^+ - A_t^- = \sum_i \bar{x}_i^t - O_t, \quad \forall t \tag{4.2}$$

$$A_t^+, A_t^- \geq 0, \quad \forall t \tag{4.3}$$

$$a^L \leq a_t \leq a^U, \quad \forall t \tag{4.4}$$

$$b^L \leq b_t \leq b^U, \quad \forall t \tag{4.5}$$

$$\begin{array}{ll} \bar{x}_i^t = \arg\min_{x_i^t} & \sum_t a_t \left(\sum_i x_i^t\right)^2 + \sum_t b_t \sum_i x_i^t + \sum_{t,i} C(t-i) x_i^t \\ \text{s.t.} & \sum_{t \geq i} x_i^t = D_i, \quad \forall i \\ & x_i^t = 0, \quad \forall i > t \\ & x_i^t \geq 0, \quad \forall i, t \end{array} \tag{4.6}$$

The upper level problem objective function (4.1) says that the service provider minimizes the total deviation from the desired load profile throughout the time horizon, by deciding the pricing parameters. Constraint (4.2) requires that A_t^+ or A_t^- equals the deviation of the scheduled demand from the desired load profile at time t. The objective function of the lower level problem (4.6) minimizes the total cost accounting for the electricity cost as well as the inconvenience cost, given a fixed pricing scheme. The first constraints for the lower level problem require all demand occurring at time i to be satisfied, and the demand can only be shifted forward. The second constraints of the lower level problem indicate that the load cannot be shifted backward.

This is a very straightforward model of the problem introduced before. The leader makes a decision on price, and the follower reacts by scheduling demands based on the price to minimize cost. It is a concise and compact formulation for this problem. But it is inherently a hard problem to solve, simply because it is a nonlinear bilevel programming problem. If the demand is stochastic in each period, it would only make the problem even worse. In that case, the lower level problem itself is hard to solve. Dynamic programming would be necessary in order to solve it. But it would become computationally prohibitive. So at first we only consider fixed demand D_t in each period t to simplify the problem.

4.4 Optimality Conditions for Lower Level Problem

One way to deal with a bilevel programming problem is to reduce the lower level problem into some constraints, and then put these constraints back into the original problem, so that it becomes a single level problem which might be solvable. KKT conditions are a natural choice for this purpose. However, if there are some special properties of the lower level problem, it might be worthwhile to explore them, as they might give better constraints than the KKT conditions.

4.4.1 Properties of Optimal Solution for Lower Level Problem

Now we consider the lower level problem. First we introduce the notion of *marginal price* P_i^t:

$$\begin{aligned} P_i^t =& \frac{d}{dx_i^t}\left[a_t\left(\sum_{j\le t} x_j^t\right)^2 + b_t\sum_{j\le t} x_j^t + C(t-i)x_i^t\right] \\ =& 2a_t\sum_{j\le t} x_j^t + b_t + C(t-i) \end{aligned}$$

If we increase (or decrease) x_i^t by a very small amount ϵ, then the total cost would increase (or decrease) by $P_i^t\epsilon$. Demand occurring at time i can be scheduled in periods $i+1, i+2, \cdots, t_i^*$, where $t_i^* = \text{argmax}\{t|x_i^t > 0, x_i^{t+1} = 0\}$. t_i^* is an interesting time point for P_i^t. We will abbreviate t_i^* as t^* since i is fixed. There are some nice properties for $P_i^t, \forall t \le t^*$, and for $P_i^t, \forall t > t^*$:

Proposition 4.1. (Optimality conditions for lower level problem): Let $t^* = \text{argmax}\ \{t|x_i^t > 0, x_i^{t+1} = 0\}$. Then the optimal solution for the lower level problem satisfies

$$P_i^i = P_i^{i+1} = \cdots = P_i^{t^*-1} = P_i^{t^*}$$

and

$$P_i^t \ge P_i^{t^*}, \quad \forall t > t^*.$$

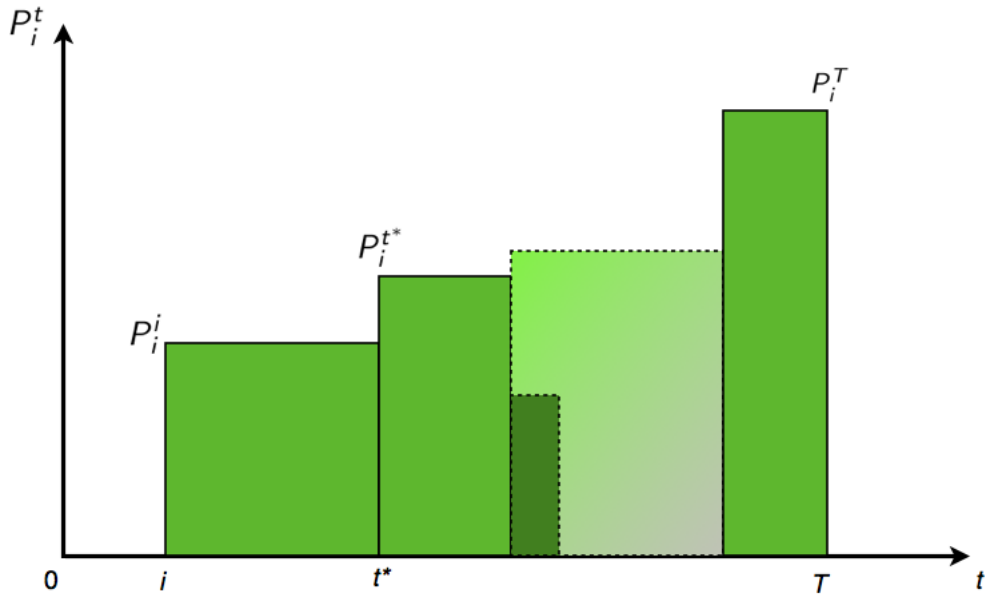

Figure 4.1: Segment of time horizon based on P_i^t

This Proposition states that the marginal price remains constant between i and t^* (see Figure 3.1.). We don't know how it behave for $t > t^*$, as in the light shaded area. But we do know $P_i^t \geq P_i^{t^*}, \forall t > t^*$ for certain. The demand requested at time i can be scheduled after t^*.

Proof. Suppose $\{x_i^t\}$ is the optimal solution to the lower-level problem and $\{P_i^t\}$ is the corresponding marginal price. We first prove that, for all $i \leq t^*$, $P_i^{t^*-1} = P_i^{t*}$. As $O_t > 0$, it is reasonable for us to assume, for an optimal solution, that the load allocated to period t should be close to O_t, or at least that there is load assigned to period t, i.e. $\sum_i x_i^{t^*-1} > 0$. Then if $P_i^{t^*} > P_i^{t^*-1}$, the cost would decrease if we shift a small load ϵ from t^* to $t^* - 1$, contradicting our assumption that x is optimal. Similarly, if $P_i^{t^*} < P_i^{t^*-1}$, the cost decreases if we shift ϵ from t^*-1 to t^*. Thus, $P_i^{t^*} = P_i^{t^*-1}$. We can generalize this to $t^*-1, t^*-2, \cdots, i$. Then we have:

$$P_i^i = P_i^{i+1} = \cdots = P_i^{t^*-1} = P_i^{t^*}.$$

If there is any $t > t^*$ which satisfies $P_i^t \leq P_i^{t^*}$, as in the darker block in the light shaded area of Fig. 3.1., the cost can be reduced by shifting load from t^* to t, which contradicts

our assumption that x is optimal. So

$$P_i^t \geq P_i^{t^*}, \quad \forall t > t^*.$$

□

This Proposition divides the time horizon based on x_i^t into two parts, $t \leq t^*$ and $t > t^*$. We don't know how it behaves for $t > t^*$, as in the light shaded area, but we know enough to proe this result. Now we are going to generalize the result to x_i^t for all i. But first we need another concept, *period-wise marginal price* p_t:

$$P^t = 2a_t \sum_{i \leq t} x_i^t + b_t + Ct.$$

P^t represents the marginal price for increasing electricity consumption at time t, and is independent of i as there is no term $-Ci$. As the next Proposition shows, based on P_t, the time horizon is divided into $n+1$ parts, and each part has equal marginal electricity price.

Proposition 4.2. There exists a set of n numbers $i_1, \cdots, i_n$, where $1 \leq n < T$, such that $1 < i_1 < i_2 < \cdots < i_n \leq T$, and the optimal solutions for the lower level problem satisfy:

$$\begin{aligned}
&P^1 = P^2 = \ldots = P^{i_1 - 1} \\
\leq &P^{i_1} = P^{i_1+1} = \ldots = P^{i_2-1} \\
\leq &P_{i_2} = P_{i_2+1} = \cdots = P_{i_3-1} \\
\leq &\ldots \\
\leq &P^{i_n} = P^{i_n+1} = \ldots = P^T
\end{aligned}$$

Moreoever, $\forall i, t$ where $i < t$, if there exists any i_k among these n numbers $i_1, \cdots, i_n$, which satisfy $i < i_k < t$, then

$$x_i^t = 0.$$

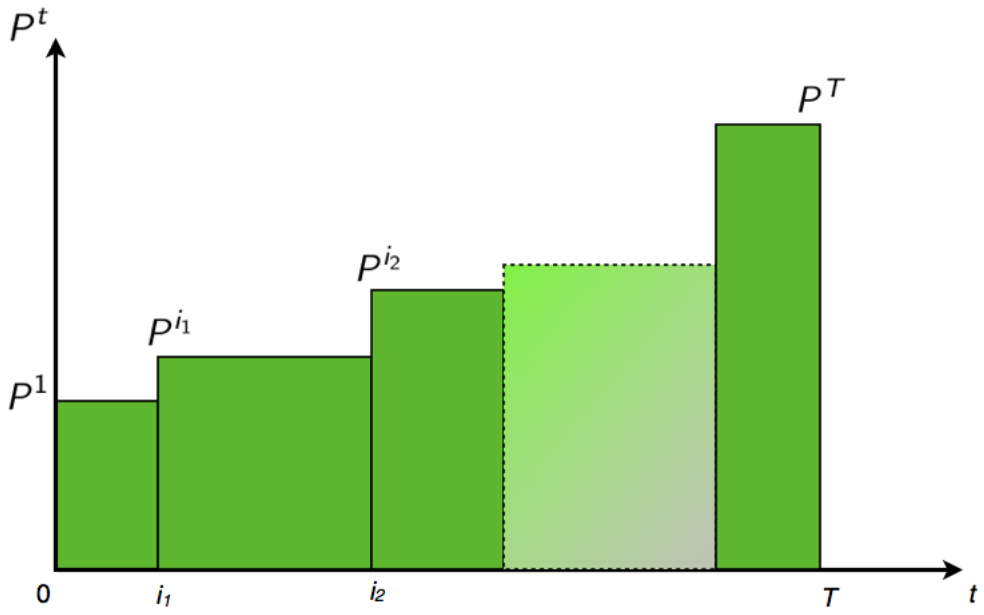

Figure 4.2: Segment of time horizon based on P^t

This Proposition indicates that P^t increases in stages. It stays constant for several periods, then increases. And it repeats this pattern. Demand requested within each stage can only be assigned within its stage (see figure 4.2).

Proof. First we will prove that for optimal solutions, $P^1 \leq P^2 \leq \cdots \leq P^T$. Based on Proposition 4.1, we have

$$P_1^1 = P^1 - C \leq P_1^2 = P^2 - C.$$

Obviously, $P^1 \leq P^2$ is true. Similarly we have $P^i \leq P^{i+1}, \forall i = 2, \cdots, T-1$. So

$$P^1 \leq P^2 \leq \cdots \leq P^T.$$

Before proving the second part of the Proposition, we need to introduce a set of binary variables z_t and break points $\{i_1, i_2, \cdots, i_n\}$. Starting from $t = 2$, if $P^{t-1} < P^t$, we assign $z_t = 1$ and set one of the break points to t, otherwise $z_t = 0$ and we do nothing with the break points. Repeating this until the end of the horizon, we get a set of $z_1, z_2, \cdots, z_T$, which satisfies $z_{i_1} = z_{i_2} = \cdots = z_{i_n} = 1$ and other $z_i = 0$, where $n < T$ and $1 \leq i_1 < i_2 <$

$\cdots < i_n \leq T$. Most importantly:

$$\begin{aligned} & P^1 = P^2 = \ldots = P^{i_1 - 1} \\ \leq & P^{i_1} = P^{i_1+1} = \ldots = P^{i_2-1} \\ \leq & P_{i_2} = P_{i_2+1} = \cdots = P_{i_3-1} \\ \leq & \ldots \\ \leq & P^{i_n} = P^{i_n+1} = \ldots = P^T \end{aligned}$$

To prove the second part of the Proposition, it is necessary to show that $x_i^t = 0, \quad \forall i < i_k < t$. Here we still denote $t_i^* = \text{argmax } \{t | x_i^t > 0, x_i^{t+1} = 0\}$. Then

$$x_i^t = 0, \quad \forall t_i^* < t.$$

So if $t_i^* < i_k$ for any i_k that is greater than i, then the claim is true. For simplicity, assume there is only one i_r which satisfies $i < i_r < t_i^*$. Then by the assumption of i_r, we know that $z_{i_r} = 1$ and $P^{i_r - 1} < P^{i_r}$. Because $i < i_r < t_i^*$, we can shift some load ϵ from $x_i^{t_i^*}$ to $x_i^{i_r - 1}$, and the total cost would be reduced by $\epsilon(P^{t_i^*} - P^{i_r - 1}) = \epsilon(P^{i_r} - P^{i_r - 1}) > 0$. So there can't be any i_r which satisfies $i < i_r < t_i^*$. □

This Proposition gives a better idea how the marginal price should behave for the optimal solution. It shows that the total time horizon can be divided into several sections such that the marginal price stays the same within each section, and it only increases overall. This provides better insight into the optimal solutions, which gives us a better way to formulate the lower level problem.

4.4.2 Formulating Optimality Conditions as Constraints

Our original problem is a bilevel problem, which is hard to compute directly. We would like to reformulate it into a single level problem, which would require reformulating the lower level problem by its optimality conditions. That's our purpose for Proposition 4.1 and Proposition 4.2. Next, we are going to formulate the optimality conditions in Proposition

4.2 as constraints.

Following the previous argument in Proposition 4.2, the lower level problem has to meet two conditions. The first condition suggests the whole time horizon is divided into sections such that the marginal price stays the same within each section, and it only increases overall. The first constraint is

$$P^t - P^{t-1} \geq 0, \quad \forall t > 1.$$

This constraint ensures that the marginal price is non-decreasing along the time horizon. The second constraint is:

$$P^t - P^{t-1} \leq M z_t, \quad \forall t > 1$$

for large M, where z_t is a binary variable that equals 1 if period t is a breakpoint, as in Proposition 4.2. This constraint makes sure that the marginal price remains the same within each section, and it can only increase at the breakpoints.

The second condition in Proposition 4.2 gives constraints on x_i^t, making sure that the demand incurred within a time horizon section is only scheduled within that time horizon section. It is equivalent to the following condition:

$$x_i^t \leq M(1 - z_j), \quad \forall i < j \leq t.$$

These constraints require x_i^t to be 0 if there is a break point between i and t. If there is not a break point between i and t, it means t lies within the time horizon section, so part of the load can be assigned to time t, and there is no constraint on it.

Based on Proposition 4.1 and Proposition 4.2, we have the following formulation for the lower level problem:

$$\begin{aligned}
\min_{x_i^t, P_t, z_t} \quad & 1 \\
\text{s.t.} \quad & \sum_{t \geq i} x_i^t = D_i, && \forall i && (4.7) \\
& x_i^t \leq M(1 - z_j), && \forall i < j \leq t && (4.8) \\
& P^t - P^{t-1} \leq M z_t, && \forall t > 1 && (4.9)
\end{aligned}$$

$$P^t - P^{t-1} \geq 0, \qquad \forall t > 1, \tag{4.10}$$

$$P^t = 2a_t \sum_j x_j^t + b_t + Ct, \qquad \forall t \tag{4.11}$$

$$x_i^t = 0, \qquad \forall i > t \tag{4.12}$$

$$x_i^t, P^t \geq 0, \qquad \forall i, t \tag{4.13}$$

$$z_t \in \{0, 1\}, \qquad \forall t \tag{4.14}$$

But Proposition 4.1 and Proposition 4.2 provide necessary conditions for an optimal solution. In order to replace the lower level problem with this new formulation, we also need to show there are sufficient conditions for optimality. In order to prove this, we need to introduce the KKT conditions for the lower level problem:

$$\sum_t x_i^t = D_i, \qquad \forall i \tag{4.15}$$

$$x_i^t \geq 0, \qquad \forall i, t \tag{4.16}$$

$$\mu_i^t \geq 0, \qquad \forall i, t \tag{4.17}$$

$$2a_t \sum_j x_j^t + b_t + C(t - i) - \mu_i^t + \lambda_i = 0, \qquad \forall i, t \tag{4.18}$$

$$\mu_i^t x_i^t = 0, \qquad \forall i, t \tag{4.19}$$

Proposition 4.3 Any solution that satisfies the new formulation also satisfies the KKT conditions for the lower level problem.

Proof. The first two KKT conditions (4.15) and (4.16) are obvious, as they are still required in (4.7) and (4.13).

The fourth KKT condition (4.18) is equivalent to $P_i^t + \lambda_i = \mu_i^t$. The KKT conditions require $\mu_i^t \geq 0$ and $\mu_i^t x_i^t = 0, \forall i, t$. These conditions would be true if we can show that both of the following two are true:

$$P_i^t + \lambda_i \geq 0$$

and

$$(P_i^t + \lambda_i) x_i^t = 0.$$

$P_i^t + \lambda_i \geq 0$ is obvious as Proposition 4.1 states P_i^t is increasing in t. Complementary slackness requires that if $x_i^t > 0$, then $P_i^t + \lambda_i = 0$; and if $x_i^t = 0$, then $P_i^t + \lambda_i \geq 0$. The second half is true as just mentioned. The first half is true because the condition in Proposition 4.2 indicates that $P_i^i = P_i^{i+1} = \cdots = P_i^{i_k - 1}$ and $x_i^i, x_i^{i+1}, \cdots, x_i^{i_k - 1} \geq 0$, and $x_i^{i_k - 1} = \cdots = x_i^T = 0$, where i_k is the smallest among all $i_j, j = 1, 2, \cdots, n$, such that $i_j > i$. So if $x_i^t > 0$, then $P_i^t = -\lambda_i$, where $\lambda_i = -P_i^i$. □

As argued in the proof, the new formulation provides sufficient conditions for the KKT conditions, so it also provides sufficient conditions for optimality. In conclusion, our new formulation is equivalent to the lower level problem.

Incorporating this new formulation of the lower level problem back into the original problem, the bilevel programming problem can be formulated as:

$$
\begin{aligned}
\min_{a_t, b_t, A_t^+, A_t^-, x_i^t, z_t, P_t} \quad & \sum_t A_t^+ + A_t^- & \\
\text{s.t.} \quad & A_t^+ - A_t^- = \sum_i x_i^t - O_t, & \forall t \\
& \sum_{t \geq i} x_i^t = D_i, & \forall i \\
& x_i^t \leq M(1 - z_j), & \forall i < j \leq t \\
& P^t - P^{t-1} \leq M z_t, & \forall t > 1 \\
& P^t - P^{t-1} \geq 0, & \forall t > 1, \\
& P^t = 2a_t \sum_j x_j^t + b_t + Ct, & \forall i, t \\
& x_i^t = 0, & \forall i > t \\
& x_i^t, P^t \geq 0, & \forall i, t \\
& z_t \in \{0, 1\}, & \forall t \\
& A_t^+, A_t^- \geq 0, & \forall t \\
& a^L \leq a_t \leq a^U, & \forall t \\
& b^L \leq b_t \leq b^U, & \forall t
\end{aligned}
$$

4.4.3 Linearizing Nonlinear Constraints

The newly formulated problem is a mixed integer problem with nonlinear constraints. The nonlinearity makes the problem very hard to solve. From our preliminary numerical experiments, it can be solved within minutes for a small scale problem with $T = 20$ using the solver Couenne. But once the problem size goes up to a reasonable size such as $T = 240$, it becomes very slow, and the computation time would exceed the time limit easily. Eliminating the nonlinearity would have a profound role in simplifying the problem and speeding up computation.

Carefully examing the problem, there is only one nonlinear constraint:

$$P^t = 2a_t \sum_j x_j^t + b_t + Ct, \quad \forall t. \tag{4.20}$$

Luckily, this is the only nonlinear constraint. The nonlinear component comes from the product of a_t and $\sum_j x_j^t$. Notice that the variables a_t only appear in this constraint. There are not any additional constraints for a_t. In other words, a_t is determined uniquely by this constraint. If the values of the other variables P^t, x_j^t, b_t are obtained from solving this model, then there would be a unique solution for a_t. And since there are upper and lower bounds for a_t, these can be incorporated into the constraints, replacing the nonlinear constraints by linear constraints.

As the upper bound (or lower bound) for a_t is a^U (or a_L), we have the following conditions:

$$P^t \leq 2a^U \sum_j x_j^t + b_t + Ct, \quad \forall t \tag{4.21}$$

and

$$P^t \geq 2a^L \sum_j x_j^t + b_t + Ct, \quad \forall t. \tag{4.22}$$

Replacing the nonlinear constraint (4.20) with these two new linear constraints (4.21) and (4.22), we have the following problem:

$$\min_{b_t, A_t^+, A_t^-, x_i^t, z_t, P_t} \text{cost} = \sum_t A_t^+ + A_t^-$$

$$
\begin{aligned}
\text{s.t.} A_t^+ - A_t^- &= \sum_i x_i^t - O_t, && \forall t \\
\sum_{t \geq i} x_i^t &= D_i, && \forall i \\
x_i^t &\leq M(1 - z_j), && \forall i < j \leq t \\
P^t - P^{t-1} &\leq M z_t, && \forall t > 1 \\
P^t - P^{t-1} &\geq 0, && \forall t > 1, \\
P^t &\leq 2a^U \sum_j x_j^t + b_t + Ct, && \forall t \\
P^t &\geq 2a^L \sum_j x_j^t + b_t + Ct, && \forall t \\
x_i^t &= 0, && \forall i > t \\
x_i^t &\geq 0, && \forall i, t \\
z_t &\in \{0, 1\}, && \forall t \\
A_t^+, A_t^-, P^t &\geq 0, && \forall t \\
b^L &\leq b_t \leq b^U, && \forall t
\end{aligned}
$$

After solving the whole model, a_t can be computed easily as:

$$a_t = \frac{P^t - b_t - Ct}{\sum_j x_j^t}.$$

This new formulation solves the same Stackelberg problem, based on the assumption that the inconvenience cost is the same for every demand, and the price parameter a_t is bounded. This work reduces a bilevel programming problem with nonlinear constraints into a single level linear mixed integer problem. The original problem is very hard to solve, and it cannot be coded as an input for any solver. The final problem is a regular MIP with $O(T)$ binary variables and $O(T^2)$ continuous variables. It can be solved within seconds for a moderate sized problem.

4.5 Consumer with Local Storage

This section deal with the case consumer in which the integrates local storage into his system. That is, the customer has one load that is a storage unit which can be used to store energy when prices are low and to draw power locally during those times when prices may be high, etc. Intuitively, similar results should still hold.

We still use same notations under the same assumptions. Now assume there is a storage unit for the customer. Denote the amount of electricity available at the storage unit as s_t, where $\underline{s} \leq s_t \leq \bar{s}$. And the amount of energy variation in period t is y_t. If $y_t > 0$, electricity is bought and injected into the storage unit, and if $y_t < 0$, electricity is withdrawn from the storage unit and consumed.

The service provider can determine the electricity pricin scheme by picking pricing parameters a_t and b_t. If the service provider sees a total load of $\sum_{i \leq t} x_i^t$ from all customers at time t, then the total electricity cost charged at time t is:

$$a_t \left(\sum_{i \leq t} x_i^t + y_t \right)^2 + b_t \left(\sum_{i \leq t} x_i^t + y_t \right).$$

So the whole problem can be formulated as a bilevel programming problem:

$$\begin{aligned}
\min_{a_t, b_t, A_t^+, A_t^-, \bar{x}^t} \quad & \sum_t \left(A_t^+ + A_t^- \right) \\
\text{s.t.} \quad & A_t^+ - A_t^- = \sum_i \bar{x}_i^t - O_t, \quad \forall t \\
& A_t^+, A_t^- \geq 0, \quad \forall t \\
& a^L \leq a_t \leq a^U, \quad \forall t \\
& b^L \leq b_t \leq b^U, \quad \forall t
\end{aligned}$$

$$
\begin{aligned}
\bar{x}_i^t = \arg\min_{x_i^t, s_t, y_t} \quad & \textstyle\sum_t a_t \left(\sum_i x_i^t + y_t\right)^2 + \sum_t b_t \left(\sum_i x_i^t + y_t\right) + \sum_{t,i} C(t-i) x_i^t \\
\text{s.t.} \quad & \textstyle\sum_{t \geq i} x_i^t = D_i, \quad \forall i \\
& x_i^t = 0, \quad \forall i > t \\
& x_i^t \geq 0, \quad \forall i, t \\
& s_t = s_{t-1} + y_t, \quad \forall t \\
& \underline{s} \leq s_t \leq \bar{s}, \quad \forall t
\end{aligned}
$$

Everything is the same as the original model, except the cost function modification for the game follower and storage transition function. If we denote $P^t = 2a_t \left(\sum_j x_j^t + y\right) + b_t + ct$, then we have similar results as the case without local storage.

Proposition 4.3. There exists a set of n numbers $i_1, \cdots, i_n$, where $1 \leq n < T$, such that $1 < i_1 < i_2 < \cdots < i_n \leq T$, and the optimal solutions for the lower level problem satisfy:

$$
\begin{aligned}
& P^1 = P^2 = \ldots = P^{i_1 - 1} \\
\leq & P^{i_1} = P^{i_1+1} = \ldots = P^{i_2-1} \\
\leq & P^{i_2} = P^{i_2+1} = \cdots = P^{i_3-1} \\
\leq & \ldots \\
\leq & P^{i_n} = P^{i_n+1} = \ldots = P^T
\end{aligned}
$$

Moreoever, $\forall i, t$ where $i < t$, if there exists any i_k among these n numbers $i_1, \cdots, i_n$, which satisfy $i < i_k < t$, then:

$$x_i^t = 0.$$

Formulating the Optimality Conditions

Proposition 4.3 is still constructed in the same way using two constraints. The first one is:

$$P^t - P^{t-1} \leq M z_t^0, \quad \forall t > 1.$$

The second condition is:

$$x_i^t \leq M(1 - z_j^0), \quad \forall i < j \leq t.$$

Repeat the work we have done for the case without local storage, we have the following formulation for the lower level problem:

$$\min_{x_i^t, P_t, \bar{P}_t, z_t^i, s_t, y_t} \quad 1$$

$$\text{s.t.} \sum_{t \geq i} x_i^t = D_i, \qquad \forall i \qquad (4.23)$$

$$\bar{P}^t = 2a_t \left(\sum_j x_j^t + y_t \right) + b_t, \qquad \forall t \qquad (4.24)$$

$$P^t = \bar{P}^t + Ct, \qquad \forall t \qquad (4.25)$$

$$P^t - P^{t-1} \geq 0, \qquad \forall t > 1, \qquad (4.26)$$

$$\bar{P}_t - \bar{P}_{t+1} = \mu_1^t - \mu_2^t, \qquad \forall t < T \qquad (4.27)$$

$$\bar{P}_T = \mu_1^T - \mu_2^T \qquad (4.28)$$

$$x_i^t \leq M(1 - z_j^1), \qquad \forall i < j \leq t \qquad (4.29)$$

$$P^t - P^{t-1} \leq M z_t^1, \qquad \forall t > 1 \qquad (4.30)$$

$$\mu_1^t \leq M z_t^2, \qquad \forall t \qquad (4.31)$$

$$s_t - \underline{s} \leq M(1 - z_t^2), \qquad \forall t \qquad (4.32)$$

$$\mu_2^t \leq M z_t^3, \qquad \forall t \qquad (4.33)$$

$$\bar{s} - s_t \leq M(1 - z_t^3), \qquad \forall t \qquad (4.34)$$

$$z_t^2 + z_t^3 \leq 1, \qquad \forall t \qquad (4.35)$$

$$s_t = s_{t-1} + y_t, \qquad \forall t \qquad (4.36)$$

$$\underline{s} \leq s_t \leq \bar{s}, \qquad \forall t \qquad (4.37)$$

$$\mu_1^t, \mu_2^t, x_i^t, P^t \geq 0, \qquad \forall i, t \qquad (4.38)$$

$$x_i^t = 0, \qquad \forall i > t \qquad (4.39)$$

$$z_t^i \in \{0, 1\}, \qquad \forall i = 1, 2, 3, \forall t \qquad (4.40)$$

Asymmetrical Information Regaring Storage

The case when the utility does not know whether the customer has local storage is captured in the original model. Intuitively, if the customer will not sell any electricity back to the utility, the utility would reach a lower deviation. Typically, the utility would set the price high if the demand is higher than the desired load, and set the price low if the demand is lower than the desired load. The storage unit would give the customer a chance to require less electricity when the price is high, and request more electricity when the price is low. That is same as saying, if the customer has a storage unit, and the utility thinks the customer doesn't, this will not hurt the utility.

4.6 Random Demand

The model we have so far solves the case with deterministic demand. However, the demand is never deterministic in reality. It flunctuates from day to day, hour to hour, minute to minute. It is vital to study how the model might be modified to handle stochastic demand. In this section, we add stochastic loads to our model for pricing and demand response. Suppose the total load in period t has some probability distribution function $f_t(\cdot)$ whose parameters (mean, variance, etc.) may be different in each period. Demands are independent across periods. It would be very hard to formulate and solve the leader's problem accurately while accounting for this stochasticity. Therefore, we focus on an approach in which the follower solves a more accurate problem (e.g., accounting for stochasticity) while the leader uses a more naive approach (e.g., ignoring stochasticity and assuming the follower does too).

4.6.1 Folding Horizon

In this section, we suppose that the follower solves his problem on a folding horizon basis. That is, in period 1 he solves the problem for periods $1, \cdots, T$. In period 2 he solves the problem for periods $2, \cdots, T$, updating the parameters to reflect what happened in period 1. In general, in period t he solves the problem for periods $t, \cdots, T$ updating the parameters to reflect what happened in periods $1, \cdots, t-1$. The ending period stays the same. The

leader, on the other hand, solves the problem once, in period 1, for the horizon $1, \cdots, T$. This seems plausible since the service provider must announce tomorrow's prices for the entire day and can't change the prices once announced, whereas the follower can modify his consumption decisions based on how the stochastic loads are realized over time. Assume that the follower realizes the load in period t before he has to make consumption decisions; that is, when he solves the problem in period t, D_t is deterministic but D_s is stochastic for $s = t+1, \cdots, T$.

4.6.2 Consumer's Policy

The consumer needs to solve the following problem for every period n.

$$\begin{aligned}
\min_{x_i^{nt}} & \sum_{t \geq n} a_t \left(\sum_i x_i^{nt} \right)^2 + \sum_{t,i \geq n} b_t x_i^{nt} + \sum_{t,i \geq n} C(t-i) x_i^{nt} \\
\text{s.t.} & \sum_{t \geq i} x_i^{nt} = D_i, \quad \forall i > n \\
& \sum_{t \geq n} x_n^{nt} = \bar{D}_n + \sum_{t \geq n} x_{n-1}^{n-1,t} \\
& x_i^{nt} = 0, \quad \forall i > t \\
& x_i^{nt} \geq 0, \quad \forall i, t
\end{aligned}$$

Here $\sum_{t \geq n} x_{n-1}^{n-1,t}$ is the delayed unmet demand from previous period $t < n$. It is constant since n is fixed. $\bar{D}_n$ is the actual demand in period n, which is also a constant. We can ignore $\sum_{t \geq n} x_{n-1}^{n-1,t}$, and include it in $\bar{D}_n$, as the demand to be met in period n. For period $t > n$, D_t represents the random demand in the future. This problem is equivalent to the following dynamic programming (DP) problem for each period n:

$$f_n(y) = \min_x \; a_n x^2 + b_n x + C(y-x) + \bar{f}_{n+1}(y-x) \tag{4.41}$$

$$\bar{f}_n(x) = E_{D_n}\left(f_{n+1}(x + D_{n+1}) \right) \tag{4.42}$$

For each period, the solution to $f_n(y)$ can be obtained with exact dynamic programming. It follows a policy that is very similar to a base-stock policy. Figure 4.3 shows a plot of

$x_n(y)$ v.s. y, where $x_n(y)$ is the minimizer in (4.41) and (4.42).

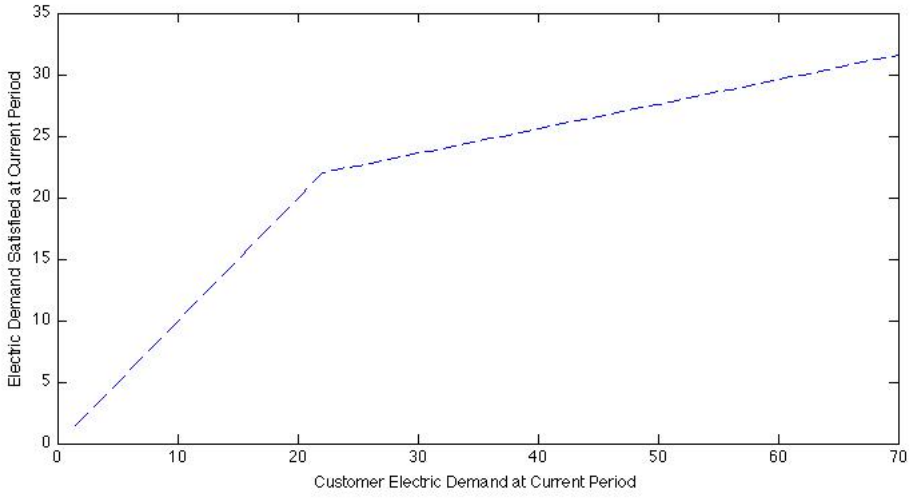

Figure 4.3: Consumer's best response function

This plot indicates how the consumer allocates his demand during a given period. If the total demand y, including delayed unmet demand from the previous period, as well as new demand in period t, is less than s_t, it is all satisfied in the current period. If the total demand y is greater than s_t, then only a part of it is satisfied, while the rest is delayed until the next period.

4.6.3 Difficulty with the Optimal Response Function

The optimal response function for the electricity consumer can be formulated as

$$g(x) = \begin{cases} x & x \leq s_t \\ \alpha x + (1-\alpha)s_t & x > s_t. \end{cases}$$

Here α is the slope at which x increases when $x > s_t$. For this policy, there are two parameters to be found: s_t and α.

Here we explore the optimality condition for minimizing the expected deviation. If $s_t > O_t$, the total expected deviation is:

$$F(s_t, \alpha) = \int_0^{O_t} (O_t - x) f_t(x) dx + \int_{O_t}^{s_t} (x - O_t) f(x) dx + \int_{s_t}^{+\infty} (\alpha x + (1-\alpha)s_t - O_t)\, f_t(x) dx$$

where $y = \frac{O_t}{\alpha} + \frac{\alpha-1}{\alpha} s_t$, which satisfies $\alpha y + (1-\alpha)s_t = O_t$. If $s_t \leq O_t$, the total deviation is:

$$\begin{aligned} F(s_t, \alpha) = &\int_0^{s_t} (O_t - x) f_t(x) dx + \int_{s_t}^{y} \left(O_t - \alpha x - (1-\alpha)s_t\right) f_t(x) dx \\ &+ \int_y^{+\infty} \left(\alpha x + (1-\alpha)s_t - O_t\right) f_t(x) dx \end{aligned}$$

For the case $s_t > O_t$, taking the partial derivatives with respect to s_t and α, we have:

$$\begin{aligned} \frac{\partial F(s_t, \alpha)}{\partial s_t} =& (s_t - O_t) f(s_t) + \int_{s_t}^{+\infty} (1-\alpha) f_t(x) dx - \left(\alpha s_t + (1-\alpha)s_t - O_t\right) f_t(s_t) \\ =& (1-\alpha) \int_{s_t}^{+\infty} f_t(x) dx \\ \frac{\partial F(s_t, \alpha)}{\partial \alpha} =& \int_{s_t}^{+\infty} (x - s_t) f_t(x) dx \end{aligned}$$

So obviously $\frac{\partial F(s_t,\alpha)}{\partial s_t} > 0$, and $\frac{\partial F(s_t,\alpha)}{\partial \alpha} > 0$. In order to let the deviation be as small as possible, s_t and α should be as small as possible, if $s_t > O_t$. This is equivalent to saying s_t should be less than or equal to O_t.

For the case $s_t \leq O_t$, again taking the partial derivatives with respect to s_t and α, we have:

$$\begin{aligned} \frac{\partial F(s_t, \alpha)}{\partial s_t} =& (O_t - s_t) f(s_t) + \int_{s_t}^{y} (\alpha - 1) f_t(x) dx - \left(O_t - \alpha s_t - (1-\alpha)s_t\right) f(s_t) \\ &+ \int_y^{+\infty} (1-\alpha) f_t(x) dx \\ =& -\int_{s_t}^{y} (1-\alpha) f_t(x) dx + \int_y^{+\infty} (1-\alpha) f_t(x) dx \\ \frac{\partial F(s_t, \alpha)}{\partial \alpha} =& -\int_{s_t}^{y} (x - s_t) f_t(x) dx + \int_y^{+\infty} (x - s_t) f_t(x) dx \end{aligned}$$

The necessary condition for the deviation to reach its minimum is $\frac{\partial F(s_t,\alpha)}{\partial s_t} = 0$ and $\frac{\partial F(s_t,\alpha)}{\partial \alpha} = 0$, which are equivalent to:

$$\int_{s_t}^{s_t + \frac{O_t - s_t}{\alpha}} f_t(x) dx = \int_{s_t + \frac{O_t - s_t}{\alpha}}^{+\infty} f_t(x) dx$$

and

$$\int_{s_t}^{s_t+\frac{O_t-s_t}{\alpha}}(x-s_t)f_t(x)dx=\int_{s_t+\frac{O_t-s_t}{\alpha}}^{+\infty}(x-s_t)f_t(x)dx.$$

But

$$\int_{s_t}^{s_t+\frac{O_t-s_t}{\alpha}}(x-s_t)f_t(x)dx<\int_{s_t}^{s_t+\frac{O_t-s_t}{\alpha}}\frac{O_t-s_t}{\alpha}f_t(x)dx$$

and

$$\int_{s_t+\frac{O_t-s_t}{\alpha}}^{+\infty}\frac{O_t-s_t}{\alpha}f_t(x)dx<\int_{s_t+\frac{O_t-s_t}{\alpha}}^{+\infty}(x-s_t)f_t(x)dx.$$

This means

$$\frac{O_t-s_t}{\alpha}\frac{\partial F(s_t,\alpha)}{\partial s_t}<\frac{\partial F(s_t,\alpha)}{\partial \alpha}.$$

This equality shows that $\frac{\partial F(s_t,\alpha)}{\partial s_t}$ and $\frac{\partial F(s_t,\alpha)}{\partial \alpha}$ can not equal 0 at the same time. Even if they could, the equations $\frac{\partial F(s_t,\alpha)}{\partial s_t}=0$ and $\frac{\partial F(s_t,\alpha)}{\partial \alpha}=0$ are very hard to solve, with s_t and α in the integration upper bound and lower bound.

4.6.4 Approximation of Optimal Response Function

Due to the difficulty of computing the original optimal response function for the electricity consumer, we need to find an approximation for this best response function, which is easy to use but also yields good results.

We propose to approximate the electricity consumer's response with

$$g(x)=\min\{x,s_t\}.$$

This response function chooses to satisify all the load request if the total load request is less than s_t, or satisify exactly s_t if the total load request is at least s_t. This functions very similarly to a base-stock policy.

Choice of s_t

In this model, the electricity service provider still prefers to have as little deviation as possible. If the game leader could force the follower to choose any s_t, a proper s_t which minimizes the total weighted deviation would be his target. Here we explore which s_t would

minimize the weighted deviation. The only thing that can be changed is s_t. The game leader considers two problems:

$$\begin{aligned} \min_{s_t} \quad & \lambda^- \int_0^{s_t} (O_t - x) f_t(x) dx + \lambda^+ \int_{s_t}^{+\infty} (O_t - s_t) f_t(x) dx \\ \text{s.t.} \quad & s_t \leq O_t \end{aligned}$$

and

$$\begin{aligned} \min_{s_t} \quad & \lambda^- \int_0^{O_t} (O_t - x) f_t(x) dx + \lambda^+ \int_{O_t}^{s_t} (x - O_t) f(x) dx + \lambda^+ \int_{s_t}^{+\infty} (s_t - O_t) f_t(x) dx \\ \text{s.t.} \quad & s_t > O_t \end{aligned}$$

In this expected deviation, both O_t and f_t are given in advance. The partial derivative with respect to s_t of the objective function in the first problem is:

$$-\lambda^- \int_{s_t}^{+\infty} f_t(x) dx \quad \text{for} \quad s_t \leq O_t$$

And the partial derivatives with respect to s_t of the objective function in the second problem is:

$$\lambda^+ \int_{s_t}^{+\infty} f_t(x) dx \quad \text{for} \quad s_t > O_t$$

This indicates that the deviation is decreasing for $s_t \leq O_t$, and increasing for $s_t > O_t$. The minimum is obtained when $s_t = O_t$. So ideally, the game leader would set a_t, b_t in a way such that the follower would set $s_t = O_t$.

Merits of Response Function Approximation

This approximation of the optimal response function greatly simplifies the problem. First, originally we have two parameters to decide: the slope and intercept. Now we have only one parameter to calcualte. Second, this one parameter has a very simple target, just to match s_t with O_t. It is simple, intuitive, and straightforward. Third, with only one parameter to manipulate, the electricity provider has a better chance to acheive its goal, i.e., to set

a_t, b_t such that $s_t = O_t$. Because higher electricity cost discourages demand and vice versa, s_t is a monotonically decreasing function of a_t and b_t. We can use bisection search to find apporpriate values of a_t, b_t such that $s_t = O_t$.

4.6.5 Algorithm for Finding a_t and b_t

So far, our analysis provides a goal for the electricity service provider in determining its pricing parameters a_t and b_t. The optimal a_t and b_t are the ones that make $s_t = O_t$.

If we consider s_t as a function of a_t and b_t, then s_t is a monotonely decreasing function of a_t and b_t. An increase in a_t and b_t would raise the electricity cost in period t, thus encouraging more demand to be delayed into the next period, which means s_t would decrease. This monotonicity allows an easier search for good a_t and b_t. In our numerical experiments, we fix a_t first, then search over all b_t. If the resulting s_t is smaller than O_t, b_t needs to be decreased, and vice versa. Bisection search is utilized to speed up this search. If no $b_t > 0$ can reach the point where $s_t = O_t$, that means $s_t < O_t$, and the price should be reduced further. In this case we decrease a_t by 10%, and search for b_t which yields $s_t = O_t$. We repeat this process iteratively - fixing a_t, solving for b_t, decreasing a_t if no good b_t are available, until we finally find a pair of (a_t, b_t) with the resulting s_t equal to O_t.

Another related issue is how to find s_t given a_t and b_t. s_t is one parameter of the electricity consumer's best response function

$$g(x) = \begin{cases} x & x \leq s_t \\ \alpha x + (1-\alpha)s_t & x > s_t. \end{cases}$$

This best response function is obtained by solving electricity consumer's problem in period t:

$$f_n(y) = \min_x \; a_n x^2 + b_n x + C(y-x) + \bar{f}_{n+1}(y-x) \tag{4.43}$$

$$\bar{f}_n(x) = E_{D_n}\left(f_{n+1}(x + D_{n+1})\right) \tag{4.44}$$

Thus we solve this problem starting from the last period T, and moving backward to period

$T-1, \cdots, 1$.

The algorithm we used to search for (a_t, b_t) is summarized in the following pseudocode in Table 4.1.

Steps	
1:	Initialization:
	set $a_t = \bar{a}, b_t^1 = 2\bar{b}, b_t^2 = \bar{b}, b_t^3 = \bar{b}/2$
	calculate s_t^1 with b_t^1, s_t^2 with b_t^2, s_t^3 with b_t^3 by solving (4.43) (4.44)
2:	**for** $t = T-1$ to 1 **do**
3:	**while** $\lvert s_t^2 - O_t \rvert > \epsilon_1$ and $b_t^2 > \epsilon_2$ **do**
4:	**if** $O_t \leq s_t^1$ **do**
	set $b_t^3 = b_t^1, b_t^1 = 2b_t^1, b_t^2 = (b_t^1 + b_t^3)/2$
5:	**else if** $s_t^1 < O_t \leq s_t^2$ **do**
	set $b_t^3 = b_t^2, b_t^2 = (b_t^1 + b_t^3)/2$
6:	**else if** $s_t^2 < O_t \leq s_t^3$ **do**
	set $b_t^3 = b_t^2, b_t^2 = (b_t^1 + b_t^3)/2$
7:	**else if** $s_t^3 < O_t$ **do**
	set $b_t^1 = b_t^3, b_t^3 = b_t^3/2, b_t^2 = (b_t^1 + b_t^3)/2$
8:	**end if**
	update s_t^1, s_t^2 and s_t^3 correspondingly by solving (4.43) (4.44)
9:	**end while**
10:	**if** $b_t^2 < \epsilon_2$ **do**
11:	set $a_t = 0.9a_t, b_t^1 = 2\bar{b}, b_t^2 = \bar{b}, b_t^3 = \bar{b}/2$
12:	calculate s_t^1 with b_t^1, s_t^2 with b_t^2, s_t^3 with b_t^3
13:	go to 3
14:	**end if**
15:	set $a_t = a_t^2, b_t = b_t^2$
16:	**end for**

Table 4.1: Flow chart of algorithm: search for a_t and b_t

4.7 Numerical Experiments

In this section, we test how our models handle the interaction between the electricity supplier and electricity consumer. The orginal problem is very hard to solve. So one important thing is how hard it is to solve our models. Furthermore, our model provides a pricing scheme for the electricity provider which incentives the consumer to react in an expected way. The effectiveness of this pricing scheme is another thing we test in this section.

We have built models for the case where the demand is deterministic, the case where there is electricity storage in the system with deterministic demand, as well as the case

where demand is stochastic. Since the case with electricity storage is very similar to the case without it, we only test the case with deterministic demand, and the case with stochastic demand in our numerical studies.

4.7.1 Deterministic Demand Model

For the model with only deterministic demand and no electricity storage, the final model is a linear mixed integer programming problem, which can be solved using any commercial solver. We implement this MIP problem in AMPL, and solves it using CPLEX 12.2.0.0.

First we test how fast this model can be solved. In these numerical experiments, demand in each period is a uniformly distributed random variable on [0, 100]. The predetermined load profile is set to be the average, i.e., 50 in every period. The problem size is determined by one factor - the total number of periods. We test this model for different sizes as listed in Table 4.2, which also lists the CPU time. For moderate sized problems, the problem can

Number of Periods	100	150	200
Running time (seconds)	14.8	27.1	122.2

Table 4.2: Running time for deterministic demand model

be solved within a couple of minutes. The running time is acceptable. But as the number of periods increases, the problem becomes increasingly hard to solve. On the other hand, T values larger than 150 or so are less likely in practice, since service provider paln prices for one day at a time.

Next we test how well the pricing scheme works. For each given demand and predetermined load profile, there is a minimum possible deviation. Note that the predetermined load profile might be greater than the actual load, in which case there is no way to delay the actual load to match the predetermined load profile; or the actual load could be too big, in which case when it comes to the last period, it has to be satisfied, thus bringing the allocated amount much higher than the desired amount. Thus for any given pair of demand and predetermined load profile, there is a minimum possible deviation. Our objective is to get as close as possible to this minimum deviation.

We pick a random set of data with a total of 50 periods. Demand D_t is uniformly

distributed from [0, 100], and the objective O_t is uniformly distributed between 30 and 70. For this set of data, the minimum deviation is 547. We test how the feasible region for (a_t, b_t, C) affects the deviation.

b_t	C	a_t	Deviation
$500 \leq b_t \leq 600$	50	1	822
$500 \leq b_t \leq 600$	25	1	694.67
$500 \leq b_t \leq 600$	5	1	590.5
$500 \leq b_t \leq 600$	5	$1 \leq a_t \leq 2$	554.2
$400 \leq b_t \leq 600$	5	$1 \leq a_t \leq 2$	547
$500 \leq b_t \leq 600$	25	$1 \leq a_t \leq 2$	651
$500 \leq b_t \leq 600$	25	$5 \leq a_t \leq 7.5$	636
$500 \leq b_t \leq 600$	5	$5 \leq a_t \leq 7.5$	593
$400 \leq b_t \leq 600$	5	$5 \leq a_t \leq 7.5$	571.4

Table 4.3: Deterministic model experiments with different parameter settings

Based on our experiments, only one set of bounds on the parameters, $(1 \leq a_t \leq 2, 400 \leq b_t \leq 600, C = 5)$, achieves this minimum deviation. It has the largest feasible region among all given combinations of parameter bounds. This set of parameter bounds has more pricing power, specifically because $C = 5$, which puts more weight on a_t and b_t, or in other words, gives a_t and b_t more choices to achieve the minimum deviation. The proper choice of (a_t, b_t, c_t) depends on the preference of the consumer in delaying load, as well as the electricity service provider's limits on determing the electricity price.

4.7.2 Stochastic Demand Model

For the stochastic demand case, our analysis does not reduce the original bilevel problem into a single level problem directly. It provides a strategy that is easy to implement for the electricity service provider to determine its pricing schemes. The algorithm is described in Table 4.1, and we implement this algorithm in Matlab. The problem we study in this chapter is about the utility company and electricity consumer. Thus, we obtain our numerical experiement data from NYISO to make it more realistic. NYISO provides ISO load forecast data, and integrated real time actual load from item P-7 and P-58C at `mis.nyiso.com/public/`, which is available publicly. Since the ISO makes its plan according to its load forecast, we assume this forecast is the desired load level O_t in our model.

And the integrated real time actual load is D_t used in our model. NYISO makes a forecast for each hour for the coming few day, and it provides real time actual load for every 15 minute interval. We use integrated hourly demand to be consistent with the prediction data. NYISO divides New York state into 11 regions. We use the integrated actual load data and forecast data for 01.25.2014 through 02.01.2014 for the Millward region to test our model.

As in section 4.7.1, we test how fast can each set of data can be solved with the stochastic demand model, and we test how well this pricing scheme works. For each set of data, we first solve the problem for the electricity service provider based on the demand forecast, calculating the pricing parameters for it, and obtaining the consumer's best response function in the meantime. With the consumer's best response function, and actual consumer demand data, we can calculate how the consumer would allocate his load according to the pricing information using the best response function. Thus we can get the actual deviation for this set of forecast-demand information. We report the running time, minimum deviation, and actual deviation for each set of data in the Table 4.4.

Date	Computational Time (seconds)	Minimum Deviation	Actual Deviation	Actual/Minimum Deviation Ratio
Jan 25	603.76	642.80	643.70	1.0014
Jan 26	610.88	1590.50	1609.70	1.0121
Jan 27	563.27	750.10	760.30	1.0136
Jan 28	729.27	6399.40	6399.40	1.0000
Jan 29	578.83	2589.10	2589.10	1.0000
Jan 30	600.45	578.90	579.80	1.0016
Jan 31	369.92	712.50	714.00	1.0021
Feb 01	315.02	628.20	629.90	1.0027

Table 4.4: Performance of stochastic demand model

Our numerical experiments yield very good results, attaining errors of less than 2% verse the minimum deviation. On one hand, this indicates that this pricing scheme is working properly: second order pricing has great power in discouraging consumption from concentrating in peak hours, and encouraging consumption in off-peak hours. On the other hand, it shows there is great flexibility in this pricing scheme, as it works well for every single day we test. For some day, it is easier to provide incentives to the consumer such that the actual load matches the predetermined load, and for other days it is harder to acheive.

However, according to our experiments, there always exists a set of price parameters for each day to match the actual load with predetermined load exactly. Our search algorithm is always capable of finding such a_t and b_t.

Choice of Pricing Parameters

Our pricing scheme is very effective in providing incentive to consumers, and it also provides a lot of flexibility to the service provider. Picking a good pricing scheme among many candidates is an interesting problem. According to our price search algorithm, a_t always increases with t. For example, Figure 4.3 plots a_t v.s. t for our experiment on Jan 30th's demand and load data.

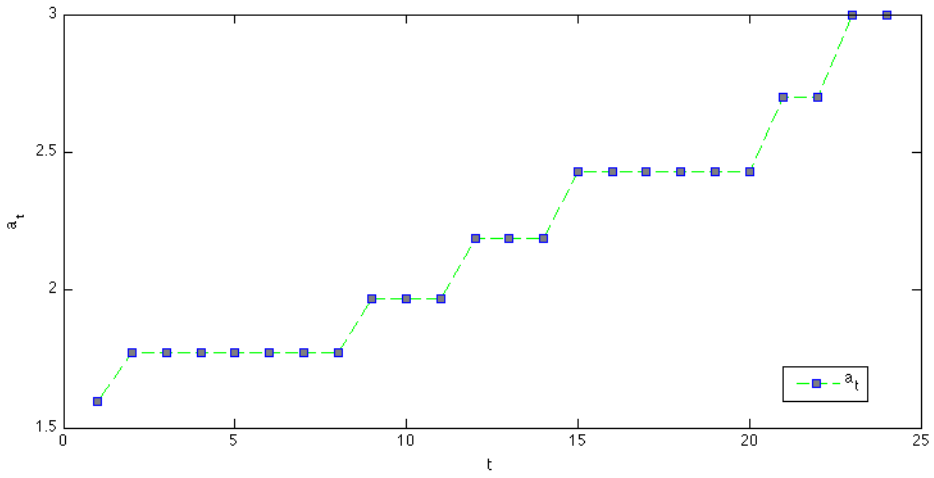

Figure 4.4: Price Parameter a_t for Jan 30th

Due to our search strategy, we start with the last period, then move backward. So if we have decreased a_t for the current period t, then it would be harder to push demand to period t when deciding a_{t-1}. That is why a_{t-1} is smaller than a_t, and thus this trend of increasing a_t appears. However, there is no such general trend for b_t. Because a_t has a much larger impact on the total cost, especially when the amount of load is high, it plays a more dominant role in affecting the consumer's behavior. b_t provides less incentive, and its effect comes second to a_t's. Figure 4.4 plot b_t v.s. t from our experiment on Jan 30th's demand and load data.

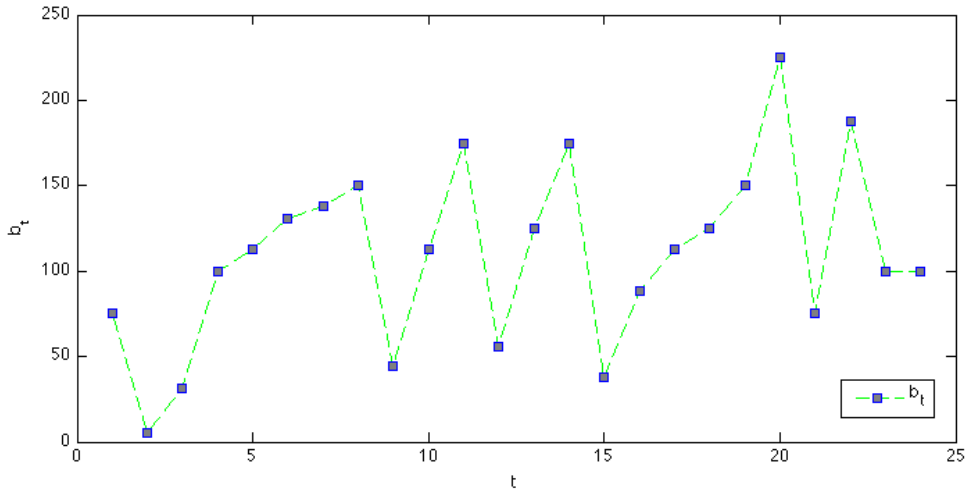

Figure 4.5: Price Parameter b_t for Jan 30th

As can be seen from this graph, b_t fluctuates from the first hour to the last hour of Jan 30th. It appears to be varying randomly. If we take a closer look, we can observe that when a_t stays at a constant level, b_t increases as t increase. This is for the same reason that a_t increases over time. When a_t increases, b_t decreases. This is a direct result of our search algorithm. As we move from the last hour to the first hour, whenever b_t gets too small, we decrease a_t to acheive our goal. Thus we get this pattern of b_t's behavior.

4.8 Conclusion and Future Work

In this chapter, we study the interaction between an electricity service provider and a large electricity consumer. We introduce a second order pricing scheme for the service provider to discourage consumption concentration. The consumer reacts to this pricing scheme by allocating its own demand. We study both the deterministic demand case and the stochastic demand case. For the deterministic demand case, the optimality condition for the consumer's problem is developed, and brought back into the original problem. The resulting problem is a single level linear MIP, which can be solved using exisiting commercial software. For the stochastic demand case, the consumer's optimal response function is analyzed and approximated, such that the service provider's approximate strategy can be easily calculated. The numerical experiments show the effectiveness of our pricing schemes,

and tractability of our models.

Our model studies the most important question of how to decide the pricing parameters such that the consumer would react in a prefered way. There are still several other interesting questions to be explored. The first question concerns what happens if the consumer has different preferences for various loads. The demand for lighting is obviously more important than the demand for the dish-washer. How to integrate this demand difference into our model is one interesting problem. The second question is what pricing parameters might be best. As shown in the numerical experiments, it seems there is more than just one set of pricing parameters that can attain the minimum deviation. Which are more practical and which are easier to implement? That is another topic that deserves further research.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Multi-echelon inventory systems are hard systems to analyze when compared to single echelon systems. The optimal inventory policy, and the allocation policy if required, are still unknown for many different types of multi-echelon inventory systems. Our work contributes to the research on designing a reliable and robust supply chain to prepare for possible supply disruptions. Among the three most fundamental inventory network structures, we studied both the assembly system and the distribution system. For the assembly system, we investigated the reason why an assembly sytem can not be reduced to a serial system in the presence of disruptions. The long-run balance property that leads to the reduction of an assembly system without supply risk is now interfered with by supply risk. We discovered that even though this long-run balance property does not hold, a similar concept called generalized item-specific long-run balance holds for assembly systems under supply risks when operated optimally. This result allowed us to present a method for reducing an assembly system with disruptions to an "almost serial" system with some extra stages, and propose an inventory policy based on this reduction. Even though the "almost serial" system obtained is not equivalent to the original one, it provides us with a method of approximating the optimal policy for the original system. We proposed a recursive algorithm to facilitate the computation of the inventory policy. This allows for easier comparison of the impact of different disruptions across systems with different structures. Significant savings were

acheived by our inventory policy with base-stock levels from our recursive algorithm, when compared with the best known policy and algorithm.

Following our work on assembly systems, we moved on to distribution systems under supply risk. As another important type of inventory system, a distribution system is also difficult to optimize its inventory policy. We considered a general distribution model under continuous review where a first-come, first-served allocation policy is implemented. We proposed a heuristic procedure to approximate the base-stock levels of distribution system under supply risk. The main idea behind this heuristic is to analyze the effect of supply disruptions and stockouts at a stage on inventory shortages at the stage's successors. We incorporated this information into a classical algorithm for serial systems, which computes the base-stock levels in a bottom-up way. Our numerical experiments demonstrated the effectiveness of this approach by reaching a 3% gap on average when compared to a bisection search based coordinate descent search.

Our last chapter shed light on a different topic - smart grid. We focus more on the interaction between an electricity service provider and an electricity consumer. We investigated the problem of how the service provider should determine the pricing scheme such that the electricity consumer reacts to this pricing scheme so that the shifted actual load is as close to a predetermined load profile as possible. This problem was formulated as a nonlinear bilevel problem. We studied the lower level problem, for both the deterministic demand case and the stochastic demand case. For the deterministic demand case, the lower level problem can be reformulated in terms of its optimality conditions. This allows us to bring it back to the upper level problem, and the whole problem was reduced into a linear single level problem which makes it solvable. For the stochastic demand case, we studied the best response function for the consumer, and approximated this best response function to facilitate calculating the electricity service provider's optimal strategy. The numerical experiments validate our analysis, as well as the tractability of our models.

5.2 Future Work

Inventory mangement is an area with plenty of questions to be answered. We studied two basic forms of inventory systems: assembly systems and distribution systems with supply risks. In reality, however, an inventory system is typically a combination of assembly systems, distribution systems, and serial systems. How to generalize our research to a more generic inventory network structure is one topic worth exploring.

For assembly systems, our work showed that the reduced "almost serial" system is not equivalent to the original system. This is related to the situation in which where a stage should keep shipping orders to its successor when a supply disruption is present in the system. To answer this question better, we need to explore further the conditions under which continuing shipping yields better results. And if continuing shipping is favored, is this the reduced system equivalent to the original system when operated optimally?

For distribution systems, we are interested in investigating how to apply the Decomposition-Aggregation heuristic (Rong et al. (2012)) to distribution systems with supply risk. This heuristic performed very well for the case without supply risk. We would like to know how this heuristic perform if it can be applied to the case with supply disruptions. In addition, more general settings, such as fixed ordering costs, and stochastic leadtimes, are some possible extensions worth considering.

Our work on smart grid provides a prototype for further understanding the interaction between service suppliers and electricity consumers. We work under the assumption of uniform load delaying cost, where in fact this cost might be different for different consumers. Integrating heterogeneous delaying cost is an interesting question. Our model also provides an approach for studying other objectives for service providers on the basis of interactions between it and the consumer. It would be helpful to understand how the service provider and consumer would behave if the service provider uses other objectives. Finally, what kind of pricing schemes are most effective? We only studied a second order pricing scheme. Picking a good set of parameters for it is a hard question. How can this be done optimally? Do there exist some other pricing schemes that are more powerful in affecting load allocation? These are interesting questions to be studied in future research.

Bibliography

[1] Jonas Andersson, Sven Axsäter, and Johan Marklund. Decentralized multiechelon inventory control. *Production and Operations Management*, 7(4):370–386, 1998.

[2] Antonio Arreola-Risa and Gregory A DeCroix. Inventory management under random supply disruptions and partial backorders. *Naval Research Logistics*, 45(7):687–703, 1998.

[3] Sven Axsäter. Simple solution procedures for a class of two-echelon inventory problems. *Operations Research*, 38(1):64–69, 1990.

[4] Sven Axsäter. Exact analysis of continuous review (R, Q) policies in two-echelon inventory systems with compound Poisson demand. *Operations Research*, 48(5):686–696, 2000.

[5] Sven Axsäter and Lars Juntti. Comparison of echelon stock and installation stock policies for two-level inventory systems. *International Journal of Production Economics*, 45(1):303–310, 1996.

[6] Sven Axsäter, Johan Marklund, and Edward A Silver. Heuristic Methods for Centralized Control of One-Warehouse, N-Retailer Inventory Systems. *Manufacturing & Service Operations Management*, 4(1), December 2002.

[7] Sven Axsäter and Kaj Rosling. Notes: Installation vs. echelon stock policies for multilevel inventory control. *Management Science*, 39(10):1274–1280, 1993.

[8] Mette Bjørndal and Kurt Jørnsten. The Deregulated Electricity Market Viewed as a Bilevel Programming Problem. *Journal of Global Optimization*, 33(3):465–475, November 2005.

[9] R Bollapragada, U Rao, and J Zhang. Managing inventory and supply performance in assembly systems with random supply capacity and demand. *Management Science*, 2004.

[10] Shengrong Bu, F Richard Yu, and Peter X Liu. A game-theoretical decision-making scheme for electricity retailers in the smart grid with demand-side management. pages 387–391, 2011.

[11] Shengrong Bu, F Richard Yu, and Peter X Liu. Dynamic pricing for demand-side management in the smart grid. pages 47–51, 2011.

[12] Gérard P Cachon. Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review. *Operations Research*, 49(1):79–98, 2001.

[13] Chen Chen, Shalinee Kishore, and Lawrence V Snyder. An innovative RTP-based residential power scheduling scheme for smart grids. pages 5956–5959, 2011.

[14] F Chen and YS Zheng. Lower bounds for multi-echelon stochastic inventory systems. *Management Science*, pages 1426–1443, 1994.

[15] S Chopra, G Reinhardt, and U Mohan. The importance of decoupling recurrent and disruption risks in a supply chain. *Naval Research Logistics*, 2007.

[16] A.J Clark and H Scarf. Optimal policies for a multi-echelon inventory problem. *Management Science*, pages 475–490, 1960.

[17] M Dada, N C Petruzzi, and L B Schwarz. A Newsvendor's Procurement Problem when Suppliers Are Unreliable. *Manufacturing & Service Operations Management*, 9(1):9–32, 2007.

[18] Gregory A DeCroix. Inventory Management for an Assembly System Subject to Supply Disruptions. *Management Science*, 59(9):2079–2092, September 2013.

[19] Stephan Dempe. Annotated Bibliography on Bilevel Programming and Mathematical Programs with Equilibrium Constraints, 2003.

[20] Gary Eppen and Linus Schrage. Centralized ordering policies in a multi-warehouse system with lead times and random demand. *Multi-level production/inventory control systems: Theory and practice*, 16:51–67, 1981.

[21] Nesim Erkip, Warren H Hausman, and Steven Nahmias. Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Science*, 36(3):381–392, 1990.

[22] Markus Ettl, Gerald E Feigin, Grace Y Lin, and David D Yao. A supply network model with base-stock control and service requirements. *Operations Research*, 48(2):216–232, 2000.

[23] Awi Federgruen and Paul Zipkin. Approximations of dynamic, multilocation production and inventory problems. *Management Science*, 30(1):69–84, 1984.

[24] Awi Federgruen and Paul Zipkin. Computational issues in an infinite-horizon, multiechelon inventory model. *Operations Research*, 32(4):818–836, 1984.

[25] G Gallego, O Ozer, and P Zipkin. Bounds, Heuristics, and Approximations for Distribution Systems. *Operations Research*, 55(3):503–517, May 2007.

[26] Nikolaos Gatsis and Georgios B Giannakis. Residential load control: Distributed scheduling and convergence with lost AMI messages. *Smart Grid, IEEE Transactions on*, 3(2):770–786, 2012.

[27] Paul Glasserman and Sridhar Tayur. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Science*, 41(2):263–281, 1995.

[28] Stephen C Graves. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science*, 31(10):1247–1256, 1985.

[29] Stephen C Graves. A multiechelon inventory model with fixed replenishment intervals. *Management Science*, 42(1):1–18, 1996.

[30] Stephen C Graves and Sean P Willems. Supply chain design: safety stock placement and supply chain configuration. *Handbooks in operations research and management science*, 11:95–132, 2003.

[31] Refik Güllu, Ebru Önol, and Nesim Erkip. Analysis of a deterministic demand production/inventory system under nonstationary supply uncertainty. *IIE transactions*, 29(8):703–709, 1997.

[32] Diwakar Gupta. The (q, r) inventory system with an unreliable supplier. *Infor*, 34(2):59–76, 1996.

[33] Mustafa Cagri Gurbuz. Coordinated Replenishment Strategies in Multi-item Inventory/distribution Systems, 2006.

[34] Mustafa Çagri Gürbüz, Kamran Moinzadeh, and Yong-Pin Zhou. Coordinated replenishment strategies in inventory/distribution systems. *Management Science*, 53(2):293–307, 2007.

[35] Raimo P Hämäläinen, Juha Mäntysaari, Jukka Ruusunen, and Pierre-Olivier Pineau. Cooperative consumers in a deregulated electricity market—dynamic consumption strategies and price coordination. *Energy*, 25(9):857–875, 2000.

[36] Benjamin F Hobbs, Carolyn B Metzler, and J-S Pang. Strategic gaming analysis for electric power systems: An MPEC approach. *Power Systems, IEEE Transactions on*, 15(2):638–645, 2000.

[37] Benjamin F Hobbs and Sushil K Nelson. A nonlinear bilevel model for analysis of electric utility demand-side planning issues. *Annals of Operations Research*, 34(1):255–274, 1992.

[38] Longbo Huang, Jean Walrand, and Kannan Ramchandran. Optimal Power Procurement and Demand Response with Quality-of-Usage Guarantees. *arXiv.org*, math.OC, December 2011.

[39] Karl Inderfurth and Stefan Minner. Safety stocks in multi-stage inventory systems under different service measures. *European Journal of Operational Research*, 106(1):57–73, 1998.

[40] Peter L Jackson. Stock allocation in a two-echelon distribution system or "what to do until your ship comes in". *Management Science*, 34(7):880–895, 1988.

[41] Shalinee Kishore and Lawrence V Snyder. Control mechanisms for residential electricity demand in smartgrids. pages 443–448, 2010.

[42] W Karl Kruse and Alan J Kaplan. Technical Note—On a Paper by Simon. *Operations Research*, 21(6):1318–1322, 1973.

[43] Mark LaPedus. Nikon, tel impacted by quake, 2011.

[44] Denis Lavigne, Richard Loulou, and Gilles Savard. Pure competition, regulated and Stackelberg equilibria: Application to the energy system of Quebec. *European Journal of Operational Research*, 125(1):1–17, 2000.

[45] Hau L Lee and Corey Billington. Material management in decentralized supply chains. *Operations Research*, 41(5):835–847, 1993.

[46] Hau L Lee and Kamran Moinzadeh. Operating characteristics of a two-echelon inventory system for repairable and consumable items under batch ordering and shipment policy. *Naval Research Logistics (NRL)*, 34(3):365–380, 1987.

[47] Hau L Lee and Kamran Moinzadeh. Two-parameter approximations for multi-echelon repairable inventory models with batch ordering policy. *IIE transactions*, 19(2):140–149, 1987.

[48] Junghoon Lee, Hye-Jin Kim, Gyung-Leen Park, and Mikyung Kang. Energy consumption scheduler for demand response systems in the smart grid. *Journal of Information Science and Engineering*, 28(5):955–969, 2012.

[49] Na Li, Lijun Chen, and Steven H Low. Optimal demand response based on utility maximization in power networks. pages 1–8, 2011.

[50] Zhaolin Li, Susan H Xu, and Jack Hayya. A periodic-review inventory system with supply interruptions. *Probability in the Engineering and Informational Sciences*, 18(1):33–53, 2004.

[51] Bin Liu and Jinhua Cao. Analysis of a production–inventory system with machine breakdowns and shutdowns. *Computers and Operations Research*, 26(1):73–91, 1999.

[52] Edward J McGavin, Leroy B Schwarz, and James E Ward. Two-interval inventory-allocation policies in a one-warehouse N-identical-retailer distribution system. *Management Science*, 39(9):1092–1107, 1993.

[53] RR Meyer, MH Rothkopf, and SA Smith. Reliability and inventory in a production-storage system. *Management Science*, pages 799–807, 1979.

[54] Esmail Mohebbi. Supply interruptions in a lost-sales inventory system with random lead time. *Computers & Operations Research*, 30(3):411–426, 2003.

[55] Amir-Hamed Mohsenian-Rad, Vincent W S Wong, Juri Jatskevich, Robert Schober, and Alberto Leon-Garcia. Autonomous Demand-Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid. *IEEE Transactions on Smart Grid*, 1(3):320–331, December 2010.

[56] Kamran Moinzadeh and Prabhu K Aggarwal. An Information Based Multiechelon Inventory System with Emergency Orders. *Operations Research*, 45(5):694–701, 1997.

[57] Suleyman Özekici and Mahmut Parlar. Inventory models with unreliable suppliersin a random environment. *Annals of Operations Research*, 91:123–136, 1999.

[58] Peter Palensky and Dietmar Dietrich. Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads. *IEEE Transactions on Industrial Informatics*, 7(3):381–388.

[59] M Parlar and D Perry. Inventory models of future supply uncertainty with single and multiple suppliers. *Naval Research Logistics*, 1996.

[60] Mahmut Parlar. Continuous-review inventory problem with random supply interruptions. *European Journal of Operational Research*, 99(2):366–385, 1997.

[61] Mahmut Parlar, Yunzeng Wang, and Yigal Gerchak. A periodic review inventory model with markovian supply availability. *International Journal of Production Economics*, 42(2):131–136, 1995.

[62] MJM Posner and M Berg. Analysis of a production-inventory system with unreliable production facility. *Operations Research Letters*, 8(6):339–345, 1989.

[63] Li Ping Qian, Ying Jun Angela Zhang, Jianwei Huang, and Yuan Wu. Demand Response Management via Real-Time Electricity Price Control in Smart Grids. *IEEE Journal on Selected Areas in Communications*, 31(7):1268–1280.

[64] Ying Rong, Zumbul Atan, and Larry V. Snyder. Heuristics for base-stock levels in multi-echelon distribution networks with first-come first-served policies. *Under revision*, 2013.

[65] K Rosling. Optimal inventory policies for assembly systems under random demands. *Operations Research*, 1989.

[66] Andrew M Ross, Ying Rong, and Lawrence V Snyder. Supply disruptions with time-dependent parameters. *Computers and Operation Research*, 35(11):3504–3529, 2008.

[67] Walid Saad, Zhu Han, and H Vincent Poor. Game Theoretic Methods for the Smart Grid. *arXiv.org*, cs.IT, February 2012.

[68] Pedram Samadi, A Mohsenian-Rad, Robert Schober, Vincent WS Wong, and Juri Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. pages 415–420, 2010.

[69] Pedram Samadi, Hamed Mohsenian-Rad, Robert Schober, and Vincent W S Wong. Advanced Demand Side Management for the Future Smart Grid Using Mechanism Design. *IEEE Transactions on Smart Grid*, 3(3):1170–1180, February 2012.

[70] Pedram Samadi, Robert Schober, and Vincent WS Wong. Optimal energy consumption scheduling using mechanism design for the future smart grid. pages 369–374, 2011.

[71] Charles P Schmidt and Steven Nahmias. Optimal policy for a two-stage assembly system under random demand. *Operations Research*, 33(5):1130–1145, 1985.

[72] Amanda J Schmitt and Lawrence Snyder. Infinite-Horizon Models for Inventory Control under Yield Uncertainty and Disruptions. pages 1–35, February 2009.

[73] Kevin H Shang and Jing-Sheng Song. Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Science*, 49(5):618–638, 2003.

[74] Craig C Sherbrooke. METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16(1):122–141, 1968.

[75] R M Simon and RAND CORP SANTA MONICA CALIF. Stationary Properties of a Two-echlon Inventory Model for Low Demand Items, 1969.

[76] Lawrence Snyder, Atan Zumbul, Peng Peng, Rong Ying, Amanda Schmitt, and Burcu Sinsoysal. OR/MS models for supply chain disruptions: A review. *Under revision*, 2012.

[77] Simon Stafford. Japanese earthquake & tsunami: Nikon begin the recovery process update, 2011.

[78] Antony Svoronos and Paul Zipkin. Estimating the performance of multi-level inventory systems. *Operations Research*, 36(1):57–72, 1988.

[79] Antony Svoronos and Paul Zipkin. Evaluation of one-for-one replenishment policies for multiechelon inventory systems. *Management Science*, 37(1):68–83, 1991.

[80] Poramate Tarasak. Optimal real-time pricing under load uncertainty based on utility maximization for smart grid. pages 321–326, 2011.

[81] Thomas N Taylor, Peter M Schwarz, and James E Cochell. 24/7 hourly response to electricity real-time pricing with up to eight summers of experience. *Journal of regulatory economics*, 27(3):235–262, 2005.

[82] Brian Tomlin. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science*, 52(5):639, 2006.

[83] Brian Tomlin and Yimin Wang. On the Value of Mix Flexibility and Dual Sourcing in Unreliable Newsvendor Networks. *Manufacturing & Service Operations Management*, 7(1):37–57, January 2005.

[84] Geert-Jan Van Houtum. Multi-echelon Production/inventory Systems: Optimal Politics, Heuristics, and Algorithms. 2006.

[85] JHCM Verrijdt and AG De Kok. Distribution planning for a divergent depotless two-echelon network under service constraints. *European Journal of Operational Research*, 89(2):341–354, 1996.

[86] Yao Zhao. Evaluation and Optimization of Installation Base-Stock Policies in Supply Chains with Compound Poisson Demand. *Operations Research*, 56(2):437–452, April 2008.

[87] Ziming Zhu, Jie Tang, Sangarapillai Lambotharan, Woon Hau Chin, and Zhong Fan. An integer linear programming based optimization for home demand-side management in smart grid. pages 1–5, 2011.

[88] Paul Herbert Zipkin. *Foundations of inventory management*, volume 2. McGraw-Hill New York, 2000.

Biography

Lin He joined the Department of Industrial and Systems Engineering at Lehigh University in 2008. He is currently a Ph.D. candidate working on supply chain management and smart grid under the supervision of Dr. Lawrence Snyder. He received his M.Sc. and B. Sc. in Mathematics from Beijing Normal University, Beijing, China in 2008 and 2005 respectively. He is a member of INFORMS. Lin joined the Nomis Solutions in San Bruno, CA, as an optimization engineer in Feb 2014.