HYPERCONNECTED FULFILLMENT AND INVENTORY ALLOCATION AND DEPLOYMENT MODELS

A Dissertation
Presented to
The Academic Faculty

By

NaYeon Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology

May 2021

HYPERCONNECTED FULFILLMENT AND INVENTORY ALLOCATION AND DEPLOYMENT MODELS

Thesis committee:

Professor Benoit Montreuil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Alan Erera
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Walid Klibi
Supply Chain Center of Excellence
KEDGE Business School

Professor Eric Ballot
Centre de Gestion Scientifique
MINES-ParisTech - PSL

Professor Alejandro Toriello
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date approved: April 20, 2021

To my family

Seungtae, Eunkyung, and Yoonsung

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisors, Dr. Benoit Monreuil and Dr. Walid Klibi, for their guidance and support. Not only I learned a lot, but I am also very sure I could not have completed Ph.D. without them. I enjoyed our discussions on papers and research topics, conferences, and projects with our industry partners.

I also would like to thank my committee members and collaborators, Dr. Alejandro Toriello, Dr. Alan Erera, Dr. Eric Ballot, and Dr. Mohamed Zied Babai, for their advice on improving the thesis work. Special thanks to Dr. Shabbir Ahmed, who inspired my third chapter a lot. I also thank many other professors in Georgia Tech for their amazing classes.

Besides academic faculties, I would like to thank Claude Bernier from South Shore Furniture, for his support for research collaborations and sharing industry's view and insights which was a great inspiration.

Special thanks to my friends of all time, Sona Kang, Yeonsu Kim, Sara Kaboudvand, Kumbit Hwang, Dan Tu, and Timur Tankayev. I would like to thank many others I met in ISyE and Kedge although I cannot list all here. It would have been a lot harder without them.

Most importantly, I cannot thank enough for my family. Especially my dear parents, Eunkyung Park and Seungtae Kim. I would not have been able to achieve any of this without their support. They have always been the best teachers and friends for me. My parents and my younger brother, Yoonsung, are the best things in my life. I am lucky to have such an amazing family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Consumption patterns have been changed dramatically over the past decades, notably by the growth of e-commerce. With the prevalence of e-commerce and home delivery, customer expectations for a faster, punctual, and cheap delivery are increasing. In fact, many customers are expecting for same-day or x-hour deliveries now and offering fast delivery becomes more and more critical for e-retailers to survive in a fierce market competition. However, many companies are simply lacking financial, physical, and/or operational resources to increase their responsiveness.

Focusing on solving the challenges in the perspective of fulfillment and inventory, we aim to find a breakthrough from a recently emerging logistics innovation movement induced by the introduction of the Physical Internet (PI). PI can potentially enable responsive yet affordable fulfillment for companies of any size through open asset utilization and multi-player operations. The key of PI innovation is transforming asset-driven logistics operations to service-driven logistics operations. This thesis provides an academic foundation for hyperconnected fulfillment to effectively satisfy the growing customer expectations on responsive deliveries. We first present a comprehensive design and evaluation of a hyperconnected fulfillment system. Then, we focus on providing inventory operations models, inventory allocation and deployment respectively, which maximally utilize the key features of hyperconnected fulfillment system: connectivity, flexibility, and decentralization.

In Chapter 2, a hyperconnected fulfillment and delivery system is designed in the context of the last-mile operations in urban areas. A comprehensive system and decision architecture of the hyperconnected system is modeled. We carefully design the scenarios to show a gradual transformation from dedicated to hyperconnected system in each thread of delivery and fulfillment so as to reveal the marginal impact of each step of transformation. We conduct a scenario analysis using a simulation platform built upon the system and decision architecture where autonomous agents are optimizing their decisions and interact with the environment. The experimental results clearly demonstrate the potential benefit of hyperconnected urban fulfillment and delivery system by concurrently improving often opposing performance criteria: economic efficiency, service capability and sustainability.

Chapter 3 tackles an optimal inventory allocation problem among multiple sales outlets. Specifically, we analyze a case where a dropshipper allocates availability to multiple e-retailers via availability promising e-contracts (APCs). Under the APC, the e-retailers do not observe actual availability and this information asymmetry leads them to pose a promised availability threshold (PAT). PAT is a threshold on remaining promised availability set by an e-retailer for a product of a dropshipper, below which the e-retailer unlists the product and thus does not accept any more orders from customers, until the promised availability is climbed above the threshold by the dropshipper. The dropshipper's APC problem with PAT is modeled as 2-stage stochastic program with two stochastic parameters: demand and PAT. We design and evaluate three contract policies differentiated by the allowance level for overpromising: guaranteed fulfillment, controlled fillrate, and penalty-driven fillrate policies. We also present a modeling approach to convert the endogenous demands, per-retailer-distribution of which is affected by the APCs, to exogenous demands with linear substitution constraints. The numerical results show the penalty-driven fillrate policy is the dominating strategy for dropshippers especially under a lean availability.

Chapter 4 tackles an inventory deployment problem under the context of open asset utilization and responsive fulfillment. When it comes to very responsive deliveries, such as X-hour deliveries, the physical availability of inventories near the delivery locations becomes necessary, which requires a broad and dense fulfillment network. The open asset utilization and service-driven fulfillment operations of the PI can enable affordable access to such decentralized fulfillment network comprised of the open fulfillment centers. We evaluate the benefit of such decentralized fulfillment network for a responsive fulfillment and develop an appropriate inventory deployment model, which possesses a partially pooled demand and inventory structure induced by responsiveness requirements, as a variant of Newsvendor. We derive a pragmatic heuristic inventory solution, W-solution, and present an efficient binary search based solution heuristic, W-heuristic. Then, via numerical experiments over both theoretical and empirical demand distributions, we demonstrate the advantage of decentralized network and w-solution over centralized network and allocation-based inventory model, pre-allocation model, respectively. We also report rather counter-intuitive observations that the w-solution which accounts for pooling leads to more inventory than pre-allocation model which does not account for pooling under low sales margin.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Recent delivery and fulfillment challenges

Consumption patterns have been changed dramatically over the past decades, notably by the growth of e-commerce. Reports of Statista ([1]) indicate that the total e-commerce sales in 2020 reached 4.3 trillion dollars, which is almost doubled compared to the amount 3 years ago (2.4 trillion dollars in 2017). In 2019, the e-commerce sales account for 14.1% of total retail sales while it accounted for only 8.6% in 2016 ([2]). The number of digital buyers is estimated to exceed 2 billion worldwide by 2021 ([3]). In US, the e-commerce sales revenue reached 432 billion dollars in 2020 and projected to be 549 billion dollars in 2024 ([4]). The growth is not only coming from pure e-commerce companies, but also coming from a growing number of omnichannel players. The growth of e-retail has been even more accelerated by COVID pandemic in 2020 due to the changed customer behavior ([5]). Forbes ([6]) reported that e-commerce recorded 129% of annual sales growth and 146% growth in terms of the number of orders. It is a remarkable growth considering that the overall retail market size decreased in 2020. In fact, in 2020, e-commerce sales accounted for 21% of total retail sales ([7]).

Along with the prevalence of the e-commerce and home deliveries, customer expectations for a faster, punctual, and cheap delivery are increasing. The recent survey of Invesp ([8]) reported that 56% of young online customers expect to have same-day deliveries and 61% of customers are willing to pay extra for the same-day deliveries. Moreover, it said that fast delivery meant same-day delivery for 96% of customers and more than half of customers want even faster delivery, such as 1-hour or 3-hour deliveries. Considering the 2-day delivery offered by Amazon Prime was a groundbreaking offer a few years ago, it can be seen that the customer expectations are rapidly growing. At the moment, many e-commerce companies are already offering very fast deliveries. For example, Amazon now offers a same-day Prime delivery for selected items and a 2-hour delivery with Prime Now. Walmart offers 2-hour Express Delivery and Target offers same-day delivery. Also, many start-ups, such as Flexe.com, Darkstore, and Deliv (acquired by Target), have been launched in offering fulfillment or distribution resources or services to support very-fast delivery. In fact, offering fast deliveries becomes more and more critical for e-retailers to survive in a fierce market competition.

These massive and irreversible wave of changes in retail, especially in e-commerce, poses immense logistics challenges. For example, the short delivery lead time, to be referred as responsiveness here, requires a physical availability of inventory near delivery locations, a fast pick-up, and an immediate transportation. Moreover, e-retailers tend to offer more variety of products (heavy tailed product portfolio) and e-orders tend to be smaller and more customized, which are all contributing factors for the operational complexity and higher logistics cost. Given the thin margin of logistics operations, it requires even more efficient operational schemes than currently available. This thesis targets to tackle this logistics challenge, especially focusing on inventory management and fulfillment operations.

1.2 Hyperconnected fulfillment system

Facing the fulfillment challenges, numerous studies have been conducted. Agatz, Fleischmann, and Van Nunen [9] provides an extensive review for scholarly articles in E-fulfillment and multi-channel distribution. As addressed, the critical logistics challenge to focus here is to meet the customer expectations on the tight responsiveness while remaining profitable. Notably, to satisfy such tight responsiveness requirements such as same-day or x-hour delivery, physical availability of inventory within certain distance to delivery locations is necessary. That means inventories must be highly distributed over a broad and dense network of fulfillment facilities. However, even though needed, most of the companies simply do not have financial capability to operate such broad and dense fulfillment network.

This work aims to find a breakthrough from a recently emerging logistics innovation movement induced by the introduction of the Physical Internet ([10]). The Physical Internet (PI) is defined as a ”hyperconnected global logistics system enabling seamless open asset sharing and flow consolidation through standardized encapsulation, modularization, protocols and interfaces to improve the efficiency and sustainability of fulfilling humanity's demand for physical object services.” ([10]). As can be seen from the definition, PI is a comprehensive logistics system covering from manufacturing to delivery and usage design. The focus of this work is a distribution web among the five PI logistics webs presented by Montreuil, Meller, and Ballot [11].

Among the distinctive features of the PI, two are the most critical to enable the responsive fulfillment and the efficient inventory management over a decentralized fulfillment network: open asset utilization and multi-player operations. The former brings economies of scope and the latter brings economies of scale. Open asset utilization transforms a fulfillment network model from an asset-driven model to a service-driven model. In other words, retailers who did not have financial capability to build a sufficiently large and dense fulfillment network on their own can now deploy inventories over a broad and dense fulfillment network provided by the fulfillment service providers. This changes the cost structure as well. When operating their own network, the costs comprise of investment cost, facility maintenance cost and operational costs. When using an open fulfillment network, they pay for the service. That is, no huge capital investment cost is needed and the service cost only occurs when they used the service. On the other hand, the service provider can reduce its cost by gaining economies of scale from multiple users. The multi-player operations are realized through those service providers. In fact, the existence of the service providers enables multi-player operations without each user's knowledge of or interaction with other users. Some early business models of the service providers can be found from *Fulfillment* *by Amazon*, *ES3* ([12]), *Flexe.com*, and *Darkstore (www.darkstore.com)*.

Note that the open asset sharing and multi-player operation in the PI is beyond the traditional collaborations ([13]). In the perspective of players, the differences are the existence of exclusive collaboration partners and fulfillment service providers. It enables to gain from multi-player operations without negotiating collaboration agreements. Also, the usage of certain fulfillment services can be controlled more dynamically by paying for the fulfillment service on-demand. In the perspective of business models and operations, it changes the focus of retailer's or supplier's operation from how to build and manage its own fulfillment network to how to utilize the open fulfillment network and how to design service offers to customers (e.g. pricing, delivery lead time options). Also, the size and density of the open fulfillment network is expected to exceed those of the collaborative network on average and it is same to all users of the service regardless of their own size and asset status. Figure 1.1 illustrates and contrasts the dedicated, collaborative and hyperconnected fulfillment conceptually. The dedicated and collaborative systems are exemplified with two players, marked in blue and red respectively. The open fulfillment facilities operated by one or more service providers are marked in black, and the other players using the service is marked in grey which can represent more than one players.

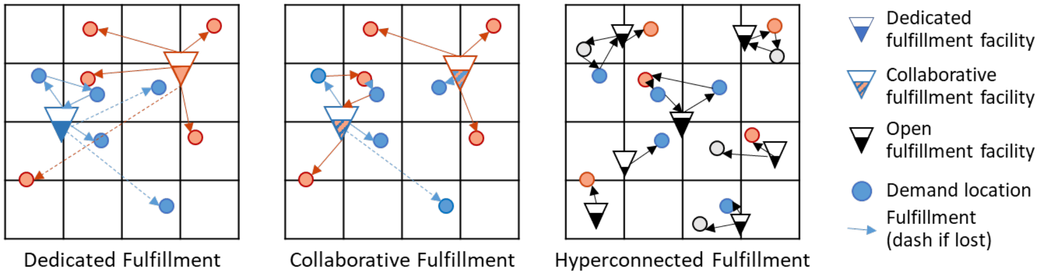

Dedicated Fulfillment
Collaborative Fulfillment
Hyperconnected Fulfillment
Dedicated fulfillment facility
Collaborative fulfillment facility
Open fulfillment facility
Demand location
Fulfillment (dash if lost)

Figure 1.1: Illustration of dedicated, collaborative and hyperconnected fulfillments

However, the access to a broad and dense fulfillment network is not sufficient to resolve the challenge. For efficient fulfillment operations, a proper inventory model adapted to the new context is needed to maximize the benefit of the open fulfillment network. In fact, decentralization has been widely considered to increase cost due to reduced pooling, since the famous work of Eppen [14]. Therefore, it must be answered if a decentralized network can be more profitable than a centralized network given the setting. There are several reasons why such contrast between the centralized and decentralized network and new inventory model are needed. Firstly, under the tight responsiveness constraints, centralization may lose demands due to its limited capacity to satisfy tight delivery lead time, trading off the pooling benefits. Secondly, a decentralized inventory network with responsiveness requirements results in a complex pooling structure, which is named as partial pooling here. Because the traditional inventory deployment models do not reflect such complex structure induced by the responsiveness requirements, they may not lead to optimal results in this new setting. Thirdly, besides the physical allocation of inventory over the fulfillment network, often times inventories also need to be allocated to multiple sales channels in e-commerce environment. The sales channels can be a mixture of a company's own e-retail site and other e-retailer's site or can be multiple e-retailers it supplies. Both problems have partial inventory and demand pooling structure in common. Accordingly, new inventory allocation and deployment models are needed that exploit the partial pooling and maximally utilize the hyperconnected fulfillment.

1.3 Contributions and summary

The objective is to build a framework for hyperconnected fulfillment, and inventory deployment and allocation under it, so as to evaluate the potential of the new fulfillment scheme in terms of operational efficiency, profitability, service capability and sustainability facing the growing needs for tighter responsiveness. The context is closely related to e-commerce while the core of results and insights is not limited to e-commerce operations. Accordingly, we present a modeling framework and solution methodology to evaluate the benefit of hyperconnected fulfillment and to optimally allocate inventories both digitally and physically under hyperconnected system. We also provide the managerial insights on potential business opportunities and caveats gained through analyzing the results of numerical experiments. This thesis is structured as follows. Firstly, in Chapter 2, we build a comprehensive system and decision architecture for the proposed hyperconnected fulfillment system and demonstrate the potential benefits of it. Then, in Chapters 3 and 4, we tackle the operational models for digital and physical allocations of inventory, which we refer as inventory allocation and deployment models respectively. We first present an inventory allocation model to multiple sales outlets through a form of an availability promising e-contract. Then, a network inventory deployment model is presented and the benefit of decentralized fulfillment network operations under tight responsiveness requirements is assessed. In the rest of the current section, we summarize the overview and contributions of each chapter.

In Chapter 2, a hyperconnected fulfillment and delivery system is designed in the context of the last-mile operations in urban areas. A comprehensive system and decision architecture is modeled that supports the hyperconnected system. Then, multiple scenarios for gradual transformation from dedicated system to hyperconnected system in each thread of delivery and fulfillment are carefully designed. The scenario analysis shows the marginal impact of each step of transformation. For the experiment, an optimization-supported agent-based simulation platform is built that enables the evaluation of potential benefits of hyperconnected system via scenario analysis. The computational experiment is conducted in the context of the last-mile delivery of large items. The experimental results clearly demonstrate the potential of hyperconnected urban fulfillment and delivery system by concurrently improving often opposing measures: economic efficiency, service capability and sustainability. In short, utilizing open peri-urban fulfillment centers and a network of open urban logistic hubs lead to a reduction of operational costs by 24% with minimal delivery lateness, and reduction of undesirable atmospheric emissions by 45$\sim$50%. Furthermore, although the proposed design framework and results are from the specific context, the essence of the design and insights can be scaled to a hyperconnected fulfillment system in larger areas such as a country or a continent.

Chapter 3 tackles an optimal inventory allocation problem in the context where a dropship manufacturer allocates its availability to multiple e-retailers via daily availability promising e-contracts (APCs). Availability, e.g. available-to-promise quantity, is used instead of inventory because the digital allocation contract does not involve any physical transaction. This means that inventories in transit or in production process can be promised to fulfill orders as long as delivery lead time can be met. Moreover, it allows the dropshipper to promise more than available quantity so that it can maximally utilize the fact that its inventory is pooled until actual fulfillment at a potential penalty cost. The problem belongs to a family of inventory allocation problems but we are the first to study the problem where promised availability threshold (PAT) is imposed by e-retailers. PAT is a threshold on remaining promised availability internally set by an e-retailer for a product of a dropshipper, below which the e-retailer unlists the product and thus does not accept any more orders from customers, until the promised availability is climbed above the threshold by the dropshipper. PAT value is not shared with the dropshipper and we proved that when PAT value is shared it is equivalent to have no PAT. The dropshipper's APC problem under PAT is modeled with 2-stage stochastic optimization with two stochastic parameters: demand and PAT. Due to the existence of PAT and the nature of e-shopping where customers explore multiple websites at the same time, demand becomes an endogenous parameter. We circumvent modeling endogenous parameter by modeling customer's shopping behavior with linear constraints. We also design and evaluate three contract policies by varying the allowance level of overpromising: Guaranteed fulfillment, controlled fillrate, and penalty-driven fillrate policies. The results show the penalty-driven fillrate policy is the dominating strategy for dropshippers especially under lean availability.

In Chapter 4, we solve an inventory deployment problem under the context of tight responsiveness requirements and open asset utilization. The problem context is closely related to the current market trend where the customer expectations for responsive fulfillment are growing with prevalent home delivery practices. Tight responsiveness requirements, such as x-hour deliveries, make the physical availability of inventories near demand locations necessary which requires a broad and dense fulfillment network. However, it is often infeasible or financially unviable to build such a decentralized network. The open asset utilization and service-driven fulfillment operations of the PI can enable affordable access to such decentralized fulfillment network comprised of the open fulfillment centers. In this chapter, we first evaluate the benefit of a decentralized network under tight responsiveness requirements and second to develop an inventory deployment model appropriate to the given setting. We build a Newsvendor-based inventory model with a partial risk and inventory pooling induced by responsiveness requirements, followed by a pragmatic and easy-to-compute heuristic solution, W-solution. We also designed a binary search based heuristic algorithm, W-heuristic, to calculate the W-solution. Then, via numerical experiments over theoretical demand distributions and empirical case study, we demonstrate the advantage of decentralized networks and W-solution over centralized networks and allocation-based inventory model, pre-allocation model, respectively. We also report rather counter-intuitive observations that the W-solution which accounts for pooling leads to more inventory than the pre-allocation model which does not account for pooling under low sales margin, while W-solution always costs less then pre-allocation model.

In short, the thesis provides academic foundations for hyperconnected fulfillment to effectively satisfy growing customer expectations on responsive fulfillment. The key system and decision architecture of hyperconnected fulfillment is designed and rigorous yet practical methodologies are presented that support key operational decisions related to inventory: inventory allocation and deployment. Besides methodological contributions, we offer tools to be used to evaluate the new business opportunities, for potential open fulfillment service providers and users. Also, the proposed pragmatic models are designed to support the operational decisions in industry with minimal customization.

Remarks

The works presented in this thesis involves collaborations with three faculty members, Dr. Benoit Montreuil, Dr. Walid Klibi, and Dr. Mohamed Zied Babai. Furthermore, the environmental impact analysis, estimating the emission levels of greenhouse gases, $PM_{2.5}$ and fuel consumption, in Chapter 2 was conducted by one of our collaborators, Nitish Kholgade, who was a master student at Georgia Tech.

Chapters 3 and 4 involve the collaboration work with our industry partner, South Shore Furniture. Thanks to the partnership, the works were able to be motivated by and linked to the real-world problems faced in business practices and we could present the interesting case studies.

CHAPTER 2
HYPERCONNECTED URBAN FULFILLMENT AND DELIVERY

2.1 Introduction

Logistics operations in urban areas, notably urban fulfillment and last-mile delivery, tend to be the most expensive of the entire logistic process while being the most critical in shaping customer experience. There are two key trends which constantly challenge urban logistics: ever-growing urbanization and customer expectations for faster and more precise, punctual and affordable delivery. High urban population density creates economic, environmental, and social issues such as logistics-induced road congestion and air pollution ([15]).

Growing demand for more convenient delivery services from time-sensitive customers is an important challenge for logistics service providers and retailers. For example, customers used to be satisfied with X-day delivery, while many are now expecting same-day, Y-hour order-to-delivery time. The trend has certainly been accelerated with the growth of e-commerce. In addition, there are emerging needs for precise and punctual delivery, with customers requiring service within a precise delivery time window in line with their availability and a precise delivery location: their home, office, car, smart locker, or store ([16]). Although urban logistics has received attention from many scholars and practitioners in the last decade ([17]; [18]; [19]; [20]), emerging urban challenges still remain. Fast, precise, and punctual urban logistic services are difficult and sometimes beyond the capability of current dedicated and fragmented logistic systems. When achieved, it is usually at high cost and high environmental impacts. The major drawback of the current fragmented systems is that they quickly lose efficiency and sustainability as logistics services become customized and individualized. To overcome the limitations of current systems, fundamental paradigm shifts are needed. Savelsbergh and Woensel [16] have identified opportunities for addressing these challenges, and smart urban logistics and the physical internet (PI) initiatives are one of them.

Through its novel perspectives and principles, the PI ([21]; [22]) offers the potential to enable breakthrough solutions for current problems. Montreuil [10] defines the PI as a ”hyperconnected global logistics system enabling seamless open asset sharing and flow consolidation through standardized encapsulation, modularization, protocols and interfaces to improve the efficiency and sustainability of fulfilling humanity’s demand for physical object services.” As also implied in [16], the notion of PI lies in systematic change enabling open asset utilization, which is beyond the concept of collaboration. Multi-player operations using open assets in a hyperconnected logistics system are achieved as business models change from being assets-driven (e.g., fulfillment centers or fleets) to becoming service-driven (e.g., fulfillment services), without requiring collaboration agreements such as strategic alliances and partnerships. *Fulfillment by Amazon*, *ES3* ([12]), *Flexe.com*, and *Darkstore (www.darkstore.com)* provide early examples of such business models. These service businesses act as trusted intermediaries, enabling open asset sharing beyond the usual group-limited collaborative agreements, and offering collaborative benefits while circumventing collaboration conflicts. Hereafter, open logistics resources (e.g., facilities or vehicles) refer to PI assets whose services can be used by any player.

A conceptual framework for hyperconnected urban logistics -the application of PI to smart urban logistics- was first introduced by Crainic and Montreuil [23]. The framework notably emphasizes interconnecting the urban logistics networks of different transportation modes, scopes and/or functionalities via a web of urban logistics centers. Figure 2.1 illustrates the transformation from a current dedicated system (left side) toward a hyperconnected system (right side), as investigated in this chapter. The model city is conceptually described as a grid on which essential logistics facilities and operations are shown using four color-coded retailers serving the city.

The typical current system is described on the left side of Figure 2.1, where last-mile

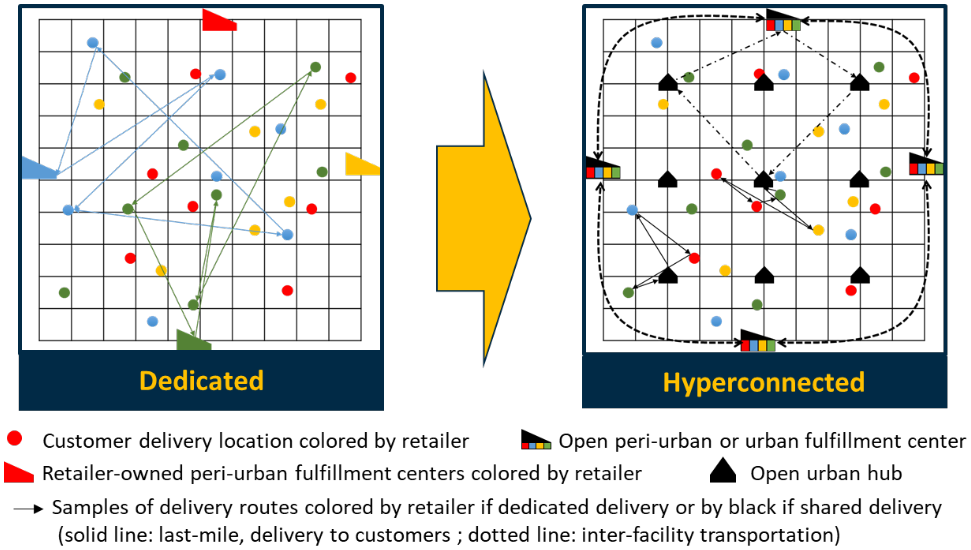

Dedicated
Hyperconnected
Customer delivery location colored by retailer
Open peri-urban or urban fulfillment center
Retailer-owned peri-urban fulfillment centers colored by retailer
Open urban hub
Samples of delivery routes colored by retailer if dedicated delivery or by black if shared delivery
(solid line: last-mile, delivery to customers ; dotted line: inter-facility transportation)

Figure 2.1: From dedicated to hyperconnected urban logistic system, according to a physical internet induced transformation (PI transformation)

fulfillment and delivery operations are dedicated to individual retailers using their own resources, such as dedicated peri-urban fulfillment centers. Generally speaking, in current systems, each retailer owns and/or operates independently its digital/physical sales channel(s) and fulfillment center(s) (FCs) and owns or uses an exclusively contracted fleet of vehicles and delivery/installation personnel. Due to their fragmented nature, the reliance on current systems has limited potential to improve urban fulfillment and last-mile delivery.

The right side of Figure 2.1 shows a proposed hyperconnected urban logistics system for last-mile fulfillment and delivery. Here, fulfillment is carried out through open FCs, which handle a multi-retailer inventory mix instead of dedicated FCs. The retailers can distribute and redeploy their inventories among the open FCs, notably aiming to improve proximity to customers through the diversified storage locations. Multi-retailer operations at FCs can facilitate delivery consolidation at origin. Open urban hubs (OUHs) further improve operations by enabling two-tier delivery, for example first-tier delivery from FCs to OUHs (dotted line) in the early morning and second-tier delivery from the OUHs to customers during the day (solid line). Tier-1 delivery enables further consolidation at OUHs, and tier-2 delivery enables the use of smaller, efficient and environmentally friendly last-mile vehicles. Hyperconnected delivery can potentially increase vehicle utilization, lower travel miles, and lower greenhouse gas (GHG) emissions. Hereafter, we refer to this transformation from the dedicated system to the hyperconnected system described in Figure 2.1 as a PI transformation.

The main goal of this study is to investigate at a strategic level the potential of this PI transformation toward hyperconnected fulfillment and delivery in the last mile in urban areas. The contribution of this chapter can be summarized into four points. First, we provide a decision and system architecture of the proposed hyperconnected urban system. Open facilities in the proposed system is characterized by conceptualizing their types and functionalities. Then, the system architecture defines key players in the market and the logistics network with open facilities and the decision architecture defines key operational components of the system such as inventory, routing, and scheduling. Second, a comprehensive set of scenarios is designed to demonstrate the gradual transformation toward a hyperconnected system in each stream of fulfillment and delivery, which can provide better insight on the implementation of a hyperconnected system. Third, an optimization supported agent-based simulation platform is built according to the system architecture. The simulation platform enables assessing the performance and dynamics of alternative urban logistics systems under a stochastic environment, capturing the interactions between players. In a nutshell, our experimental results, which have been built upon the case of large item delivery in urban area, indicate significant improvements by evolving from the dedicated system to the hyperconnected system. The overall operation costs are reduced by 24% with minimal delivery lateness, while reducing atmospheric emissions by about 50%. As a fourth contribution, this chapter fills a literature gap by addressing a challenging case study of large-item delivery logistics when conducting a simulation experiment. This business case has unique features of routing and scheduling due to the considerable delivery and install time caused by large item sizes and white glove services.

The chapter is structured as follows. Section 2 reviews the related literature and positions the contribution of the chapter. Section 3 describes the decision and system architecture. Section 4 presents the simulation structure, the scenarios, and the key performance measures. Section 5 reports the experimental results for the main scenarios as well as key results from sensitivity analyses. Section 6 concludes the chapter with insights, limitations and avenues for future research.

2.2 Literature Review

General trends, challenges, and modeling approaches in urban logistics can be found in Taniguchi, Thompson, and Yamada [18], Savelsbergh and Woensel [16], and Bektaş, Crainic, and Woensel [17]. In this section, we focus on reviewing three key literature threads related to this chapter. First, we review the most relevant contributions on urban logistics with a focus on urban multi-party consolidation and collaborative systems. We notably highlight past work related to urban consolidation centers (UCCs), which is widely studied and practiced in the context of multi-player consolidation for urban delivery. Second, we review recent trends in literature on two-tier and hyperconnected urban logistic systems that match with PI principles. Third, we review literature on urban logistics simulation with an emphasis on simulation-based methodologies incorporating decision-making processes among players. This chapter develops from early works of Goyal, Cook, Kim, Montreuil, and Lafrance [24], Kim and Montreuil [25] and Kim, Kholgade, and Montreuil [26] who have reported exploratory investigations with interesting preliminary results for PI driven large-item delivery systems.

2.2.1 Urban Consolidation and Collaborative Systems

UCCs can be found with varying names, such as city/urban distribution centers (CDCs/UDC) and freight consolidation centers (FCCs). UCCs have been tested in many cities since the new millennium as extensively summarized in BESTUFS [27]. Also, it has been widely studied academically in various settings ([28]; [29]; [30]; [31]). For example, Duin, Quak, and Muñuzuri [30] investigated the success factor for the UCC in the city of Hague and Faure, Montreuil, Burlat, and Marqués [28] analyzed the UCC operations in the context of Physical Internet. Some researchers specifically focus on the routing problem from UCC (Heeswijk, Mes, and Schutten [32]). Although there are a few cases of UCC that serve a single property ([29]), the UCCs serving the entire city are closer to the hyperconnected system proposed in this chapter. The main benefit of UCCs is stemming from a multiplayer delivery consolidation to increase last-mile vehicle fillrate and the potential use of more efficient and environmentally friendly transportation means in congested city areas, although the functionality of UCCs can be extended to other logistics or retail activities, such as temporary storage ([29]; [33]; [31]). Despite myriad attention to their potential, most UCC operations fail to sustain financially once the initial public subsidies cease ([34]; [29]; [31]). Quak and Tavasszy [33] and Heeswijk, Larsen, and Larsen [35] pointed out that it is critical for UCCs to secure enough flow to be profitable. Up to now, studies on financial viability of UCCs lead to one key factor: the need for an efficient operation ensuring the benefit of better consolidation to outweigh the cost of extra handling ([36]; [34]; [37]). However, the UCC practices have not been able to replace FC operations due to limited or lack of storage functionality, which could contributed to extra inbound shipping and handling cost. Here, we expand the scope to fulfillment operations and propose a more comprehensive system which covers both fulfillment and delivery operations.

Multi-player shipment consolidation, which is one of the key enablers for efficient and sustainable urban logistics typically achieved via UCCs or two-tier delivery systems, is in general aimed to be achieved through collaboration. Collaboration in logistics has been studied and practiced historically in diverse contexts such as transportation operations of carriers ([38]; [39]; [40]) and some of them have reported significant savings from multiparty shipment consolidation ([41]). Cleophas, Cottrill, Ehmke, and Tierney [42] provide an extensive summary of collaborative transportation examples, specifically in urban settings. Multiparty storage and fulfillment are less studied compared to multi-party delivery.

Collaborative logistics has significant potential to achieve economies of scale and scope, yet it suffers from three issues: it is generally slow to implement, tough to adapt, and hard to scale ([10]), because it requires executive approval from each party, the collaborative solutions are designed to address a common specific problem in specific context, and these solutions are conceived to benefit the group of collaborators, not to be expanded widely. PI based hyperconnected logistic systems addressed next are aiming to get away from such limitations and to facilitate seamless large-scale multi-player operation. Cleophas, Cottrill, Ehmke, and Tierney [42], who categorize urban collaborative transportation into vertical and horizontal collaboration, point out that PI-enabled transportation achieves the combination of both vertical and horizontal coalitions.

2.2.2 Two-tier and Hyperconnected Systems

In order to overcome the financial viability operational efficiency issues of UCCs, Crainic, Ricciardi, and Storchi [43] proposed a shift from single-tier systems to two-tier systems, exploiting a two-tier modeling framework. This approach has been used in tactical planning models ([44]) and operational transportation models ([44]; [45]; [46]). Some researchers expanded such two-tier systems to further incorporate other variations, such as collection-and-delivery points ([47]). The emphasis is on the optimization of loads of different shippers and carriers when they coordinate their transportation activities within a city ([18]; [48]; [20]). These models have roots in the classical location-allocation models ([49]) and/or location-routing models ([50]). Several papers ([17]; [51]) discussed the fact that due to the combinatorial complexity of the produced models, aggregation and approximation techniques are still necessary to solve them, which may lead to dismissing details highly valuable for solution use. The two-tier hierarchical urban delivery system uses a network of satellite centers ([52]), as a precursor step toward a hyperconnected delivery system with open hubs. This study builds on a two-tier hierarchical urban delivery system to model open hub operations while expanding the scope for fulfillment and inventory operations. This is also a distinguishing feature of our proposed system versus relying exclusively on UCCs where storage is only a minor function or non-existent. In this chapter, we investigate the tiered delivery and fulfillment model at a fine level of granularity through simulation implementation, for example, congestion-affected travel speed, stochastic delivery time, and varied hub usage and storage cost by location.

The PI introduced in Montreuil [53] and Montreuil [21] spans entire worldwide supply chains and logistic systems. It emphasizes open asset sharing and flow consolidation at any phase of supply chains, notably including both delivery and fulfillment. Pan, Ballot, Huang, and Montreuil [54] summarize papers in PI. In the context of urban logistics, Crainic and Montreuil [23] and Bektaş, Crainic, and Woensel [17] design and address the potential of hyperconnected urban logistics. Crainic and Montreuil [23] have extended the two-tier hierarchical urban delivery system of Crainic, Ricciardi, and Storchi [52] to a multi-tier multi-party delivery system enabled through exploiting a network of various types of open logistic hubs and open transportation/delivery service providers. A hyperconnected system not only restructures the logistics system with open logistic resources, but also redesigns the use of logistic resources by enabling dynamic multiplayer operation. Such multiplayer operations are not limited to coalition, rather leveraging standardized protocols and interfaces, which mainly distinguish hyperconnected systems from collaborative systems. The open resources are typically, but not necessarily, operated by third-party logistics providers (3PL) and offered as a service. Examples of close-to-PI models in fulfillment are *ES3, Flexe.com, Fulfillment by Amazon, Warehouse Anywhere*, and *Darkstore*. Also, close-to-PI examples in transportation include *Uber Freight* and *Convoy*. These service providers enable users to bypass collaboration contracts and extra investment to connect physical and digital systems of participants for multi-player resource use and operation. In other words, finding collaborators ([55]) or potential competition between collaborators ([56]) is not a barrier for the PI system.

Up to now, the early modeling and optimization oriented papers on two-tier and hyperconnected urban logistics ([57]; [52]) still mainly focus on transportation/delivery systems, with fulfillment systems being kept out of the scope of study. This chapter aims to fill the gap by encompassing both transportation and fulfillment systems. Also, the optimization-based approach has limitations in capturing the multidimensional nature of the system, including its operational efficiency, cost, profitability of each player, service, and sustainability. Sustainability encompasses the economic, societal and environmental performance facets, and is often related to the three Ps: planet, people, and profit (e.g., [58]). Improving logistics sustainability is at the fundamental core of the PI ([21]). In the context of city logistics study, achieving sustainability and improving quality of life for residents by reducing congestion and air pollution, while providing better service level, is important ([16]; [58]), which is one of the main goals pursued by a hyperconnected logistic systems ([23]). The chapter can also fill the gap in the literature by revealing the dynamics of such various measures through a simulation-based approach.

2.2.3 Simulation in Urban Logistics

Simulation is a powerful tool to model complexity and dynamics of urban logistic systems with a level of fidelity that can hardly be captured with a traditional optimization approach. Taniguchi, Thompson, and Yamada [59] reviewed the use of optimization and simulation techniques in urban logistics studies. One must notice that most of the related works resort to optimization or to a combination of techniques ([60]; [61]) and that only a few resort to simulation to tackle the problem. However, simulation enables capturing detailed factors such as congestion, and the evolution of a wide set of interlaced key performance indicators ([61]). Duin, Kortmann, and Boogaard [62], in line with previous work by Verbraeck [63], address five conditions that a simulation model benefits. Among these conditions, three are relevant to the research in this chapter. (1) The problem is complex in such a way that the outcomes are not simple and one-sided. (2) Because the new concept we want to study does not exist yet, it is too expensive and time-consuming to experiment with a real-world model. (3) There is no simple, analytical solution to the mathematical models of the system. Agent-based simulation (ABS) has an additional advantage in contexts such as in this chapter: it enables investigating complex interactions between multiple decision-making players, as well as the relative impact of different scenarios on each player. Although less common than optimization models, simulation has been used for one or more of the described reasons in the area of urban logistics and PI study. Concerning urban logistics study, Heeswijk, Albertus, Mes, Schutten, and Zijm [64], Wangapisit, Taniguchi, Teo, and Qureshi [65], Duin, Kolck, Anand, Taniguchi, *et al.* [66], and Roorda, Cavalcante, McCabe, and Kwan [67] used multi-agent simulation approach, some specifically to study UCC. In area of PI, Sarraj, Ballot, Pan, Hakimi, and Montreuil [68] and Pan, Nigrelli, Ballot, Sarraj, and Yang [69] used ABS to respectively study efficiency and sustainability of logistic networks and optimal inventory policies. Notably, Heeswijk, Albertus, Mes, Schutten, and Zijm [64], Holmgren, Davidsson, Persson, and Ramstedt [70] and Roorda, Cavalcante, McCabe, and Kwan [67] provided a great simulation framework for modeling behaviors and interactions of urban actors on policy and contract decisions. Our framework focuses more on operational decisions and interactions, incorporating dynamics such as congestion and stochastic delivery times. It also models inventory management in detail while the previous literature focuses on transportation.

2.3 Decision and System Architecture

In this section, we describe the decision and system architecture modeling. The key logistic resources and key players are defined and key operational decisions are modeled for supporting alternative fulfillment and transportation systems. System and network architecture is first described to clarify the scope of the business context, such as urban logistic resources and key players. Following is the key decision architecture including an inventory policy and routing method. The decision and system architecture forms the blueprint for the simulation platform.

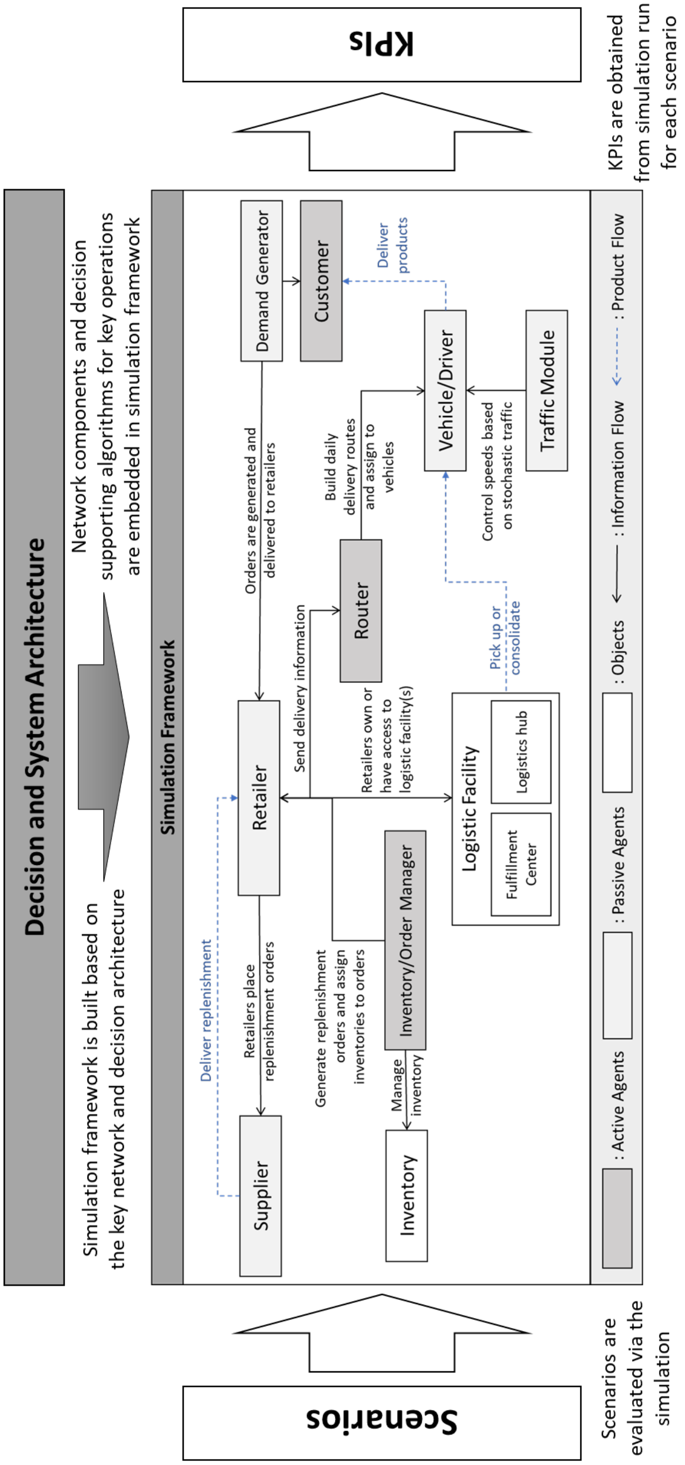

Scenarios
Scenarios are evaluated via the simulation
Decision and System Architecture
Simulation framework is built based on the key network and decision architecture
Network components and decision supporting algorithms for key operations are embedded in simulation framework
Simulation Framework
Supplier
Deliver replenishment
Retailers place replenishment orders
Retailer
Orders are generated and delivered to retailers
Demand Generator
Customer
Send delivery information
Router
Build daily delivery routes and assign to vehicles
Vehicle/Driver
Deliver products
Inventory
Manage inventory
Inventory/Order Manager
Generate replenishment orders and assign inventories to orders
Retailers own or have access to logistic facility(s)
Logistic Facility
Fulfillment Center
Logistics hub
Pick up or consolidate
Control speeds based on stochastic traffic
Traffic Module
: Active Agents
: Passive Agents
: Objects
: Information Flow
: Product Flow
KPIs
KPIs are obtained from simulation run for each scenario

Figure 2.2: Schematic description of modeling and experiment structure

Figure 2.2 shows the overall experimental design through the link between the simulation platform and network and decision architecture design, scenarios, and key performance indicators (KPIs), which are described in detail in Sections 3 and 4. The decision and system architecture affects the simulation framework structure. All operation policies and associated optimization algorithms are embedded in the simulation. It can be seen that the simulation is modeling key players, such as retailers, customers, and suppliers, and that each key player has key agents performing key operations, such as an inventory manager and a router. Each component in the simulation framework is described in detail in this section. Then, different scenarios are evaluated via the simulation using appropriate KPIs. The experimental design is described in detail in Section 4.

The section begin with a comprehensive list of notations to be used in the following sections.

2.3.1 Notations

- F: Fulfillment center (DPF: Dedicated Peri-urban FC, OUF: Open Urban FC)
  - ic: Unit daily inventory holding cost
- H: Hub (OUH: Open Urban Hub)
  - hc: Unit hub usage cost
- R: Set of retailers. For $r \in R$:
  - F_r: Set of hubs r has access to
  - P_r: Product portfolio of r
  - $I_r = \{I_{rp} = \{I_{rpf} \forall f \in F_r, I^i_{rp}\} \forall p \in P_r\}$: Inventory of r for all products in P_r, I_{rp} which comprised of on-hand inventory in each FC (I_{rpf}) and in-transit inventory (I^i_{rp})
  - C_r: Set of customers to serve
- Q: Set of Suppliers. For $q \in Q$:

  - P_q: Product portfolio of supplier q

- P: Set of Products. For $p \in P$:
  - q_p: Supplier of p (unique)
  - R_p: Retailer selling p
  - l_p: Stochastic replenishment leadtime of p
  - ξ_p: Stochastic delivery time of p
  - ω_p: Stochastic install time of p

- C: Customer
  - C^t: Set of customers to make delivery to on day t
  - r_c: Retailer received order of a customer c
  - p_c: Product that c orders
  - (x_c, y_c): Latitude and longitude of delivery location for c
  - $TW_c = (ED_c, LD_c)$: Delivery time window for c defined as the earliest and latest delivery start time (ED_c, LD_c)
  - f_c: Fulfillment center from which c is fulfilled
  - h_c: Hub from which the delivery to c is made directly

- O_t: Routes to be served on day t. For a route o in O_t:
  - f_o or h_o: A fulfillment center or a hub that is the depot for route o
  - T_o: Departure time at the depot
  - $\{\overrightarrow{C_o}\}$: Sequence of customers to visit
  - $d(a, b)$: Distance between a and b where a, b are customer delivery location, hub, or FC

  - w_h: Weight for distance from hub h calculated as the hub h's usage cost (hc_h) divided by usage cost of hubs in suburban area (hc_{suburb})

- v: Vehicle
  - $type_v$: Type of vehicle v
  - V_v, W_v: Volume and weight capacity of V
  - o_v: Route currently assigned to v to serve

2.3.2 Logistic Network and Resources

The main geographical area of interest is a single city. The two important physical elements are logistics facilities and delivery vehicles. There are two main types of logistics facilities modeled in this chapter: fulfillment centers and logistic hubs.

Model City

A grid city model shown in Figure 2.1 is used for a generic representation. It consists of districts and a main road network, which are represented as basic square areas and edges of the grid respectively. The main road network represents highways or boulevards. Vehicles move along the main road network when moving between districts by entering to and exiting from the closest point on the network. When moving within a district, vehicles are assumed to move along an implicit small road network modeled as a straight line travel. The average speed on the inter-district road network is higher than the speed on intra-district roads. However, the actual speed is a stochastic variable varying by hour and day on each road. Such a road network structure represents the essence of the common structure of large cities.

As pointed, the geographically simplified grid city model provides flexibility to simulate different cities by varying parameters such as demand density distribution, population, or size. Its simplicity also enables us to focus on the logistic transformations and their impact on the key performance objectives.

Urban Logistic Facilities (F,H)

There are two fundamental types of urban facilities in the study: fulfillment centers (F) and logistic hubs (H). Fulfillment centers are the main storage facility to fulfill customer orders in the city. Hubs enable consolidation, crossdocking and transshipment of goods. Both are further differentiated according to their location, as peri-urban (P) or urban (U) and/or according to whether they are open (O) to all retailers or dedicated (D) to a specific retailer or an exclusive group of retailers. For example, OUH stands for an Open Urban Hub and a PF stands for a Peri-urban Fulfillment center. In general, PFs have larger capacity than UFs due to lower space cost and availability, so only a limited set of top-selling products will be stored in UFs. The location of each facility is defined as an (x,y) coordinate in the grid representing the city.

The costs associated with the facility use are the inventory holding cost (ic) and the hub usage cost (hc). Each cost is measured in $/item,day and $/item, respectively. The costs used in the experiment are determined exploiting representative market prices in *Flexe.com*. Both costs differ only by the location of the facilities. For example, the usage cost of urban facilities in the city center is higher than the cost of using facilities in peri-urban areas, but the inventory holding cost is the same in dedicated peri-urban FCs and open peri-urban FCs. Because the focus is on the operational efficiency rather than on network optimization, facility opening cost is not considered.

Each fulfillment center f and hub h has the following attributes:

$$f = \{Operation(\text{D/O}), Area(\text{P/U}), location = (x,y), ic_f, R_f\}$$

$$h = \{Operation(\text{D/O}), Area(\text{P/I}), location = (x,y), hc_h\}$$

where R_f is the set of retailers who has access to the fulfillment center f. In this study, all the hubs are assumed to be open facilities.

Delivery Vehicles (v)

Different types of vehicles are used for each route type in the study. Last-mile routes from FCs to customers are served by $17'$ trucks. Smaller trucks -$10'$ trucks- are used for tier-2 routes which tend to be shorter and carry fewer items. Last, larger trucks -$26'$ trucks- are used for tier-1 routes. Vehicles are assumed to be sourced on demand from a 3PL company.

Given a route and schedule assignment, vehicles travel at speeds according to actual traffic of the time of day. Because of traffic fluctuations, the travel time is stochastic. The times to deliver and install the goods are also stochastic. It means that delivery trucks may arrive at each customer location earlier or later than scheduled. This, in turn, can cause delivery lateness failing to arrive within the customer time window. Late delivery may cause delivery failure and delivery rescheduling to other days. Here, we assume that late deliveries are still made on the same day, causing customer dissatisfaction. When delivery persons arrive earlier than the earliest delivery time specified at the time window, they have to wait to start delivery. Waiting does not affect customer satisfaction, but it decreases operational efficiency and causes road congestion. Each vehicle has the following attributes: $v = \{type_v, V_v, W_v, o_v\}$, where $type$ is vehicle type, V_v and W_v are volume and weight capacity, and o is an assigned route.

2.3.3 Market

Three key players form the market: retailers (R), customers (C), and suppliers (Q). All instances of the three key players are modeled as independent agents in the simulation with their operations, behavior, and interactions implemented as rule-based policies or heuristics.

Note that although municipalities often serve a significant role in city logistics initiatives, they are not included in the key players here because the proposed hyperconnected system is designed to be market-driven. However, municipalities' objectives and constraints are captured implicitly through social and ecological requirements.

Retailers (R)

Retailers (R) are at the center of the market in the study. They receive and fulfill customer demand while managing inventory. Each retailer r has its own product portfolio of products (P_r) and dynamically manages its inventory ($I_r = \{I_{rp}, \forall p \in P_r\}$) over the fulfillment network of accessible fulfillment centers (F_r). C_r is the set of customers for r to serve. A retailer r has the following attributes:

$$r = \{F_r, P_r, I_r, C_r\}$$

I_r is composed of the on-hand inventory of each product $p \in P_r$ in each FC $f \in F_r$ (I_{rpf}), and the in-transit inventory of each product I^i_{rp}. Each retailer r manages its own inventory independently. Inventory management includes replenishment, distribution, and fulfillment. The inventory management scheme is described in detail in subsubsection 2.3.4.

Note that, because the retailers are not manufacturers, some products are sold by multiple retailers ($P_r \cap P_{r'} \neq \emptyset$) and customers are assumed not to have a preference among retailers. This affects customer behavior, as explained in subsubsection 2.3.3.

Suppliers (Q) and Products (P)

Suppliers (Q) receive replenishment orders from one or more retailers and ship their products to them. Each supplier (q) produces or supplies an exclusive set of products, P_q. That is, a product p is single sourced from a supplier q but can be sold by multiple retailers. The set of retailers offering p is denoted as R_p. The stochastic lead time l_p of each product p combines a production lead time and delivery lead time. Production lead time is assumed to be zero for make-to-stock (MTS) products and is a positive stochastic variable for make-to-order (MTO) products. Delivery lead time is affected by proximity of the supplier's facility. It is assumed suppliers always have available inventory.

There are a delivery time (ξ_p) and an install time (ω_p) associated with the delivery of p, and both are stochastic variables. In the experiments, the delivery and install times are assumed to follow $Uniform[\underline{\xi_p}, \overline{\xi_p}]$ and $Uniform[\underline{\omega_p}, \overline{\omega_p}]$, respectively. Delivery time is always positive but installation time can be zero for some products, such as a small chair. A product p and supplier q are defined with their attributes as follows:

$$p = \{q_p, R_p, l_p, \xi_p, \omega_p\}$$

$$q = \{type(MTS/MTO), P_q\}$$

Customer (C)

Each customer (c) represents a delivery order. The customer order arrival process is modeled as a Poisson process by product (p), district (d), and retailer (r) with a constant arrival rate λ_{dpr}, meaning there is no seasonality over the study period. Some customers may find the product they are looking for is not available at the first retailer. If the same product is available at other retailer(s), the customers will order from any of the retailers with available stock (retailer substitution). If no other retailer has available stock, customers will wait or leave with probability $u(w)$ or $1 - u(w)$ respectively. The probability $u(w)$ is a function of expected waiting time (w). C^t is the final set of customers to whom the delivery must be made on day t.

The customer arrival rate varies by district in the study. In the main experiment, the arrival rate is assumed to be higher in the center of the city. Such demand distribution is commonly found in the large cities where population and businesses are concentrated in the city center. However, different demand distribution is investigated via sensitivity analysis as presented in subsubsection 2.5.2.

Each customer is assumed to order one product (p) to retailer (r). The customer requires a delivery to location (x, y) and also will choose the delivery time window $TW = (ED, LD)$ within the carrier's available work hours. The delivery time window (TW) defines the earliest and the latest time (ED and LD) when the delivery process must start. The length of a time window is $|TW| = LD - ED \in [2, 4]$ hours where $ED \in [9AM, 6PM]$ and $LD \in [11AM, 8PM]$. Next-day delivery is assumed for all customers. When multiple FCs are available, customers are assigned to a FC f from which to be fulfilled by an inventory manager of the corresponding retailer (see subsubsection 2.3.4) and assigned to hub h if hubs are used by a router (see subsubsection 2.3.4).

In summary, each customer order c has the following attributes:

$$c = \{r_c, p_c, (x_c, y_c), TW_c = (ED_c, LD_c), f_c, h_c\}$$

where (x_c, y_c) represents the delivery location and f_c and h_c represent assigned FC and hub from which the customer c is served.

2.3.4 Last-Mile Fulfillment and Delivery

Inventory management and delivery routing under the hyperconnected fulfillment and transportation system are different from those under the dedicated system and, therefore, must be well designed to enable proper comparison of the two systems. Inventory management includes inventory tracking and replenishment as well as deployment and order allocation when multiple FCs are available. The delivery routes and schedules must cover all customers, and aim to respect the delivery time window while minimizing total labor hours. The inventory management scheme and routing algorithm are described in detail in this section.

Inventory Management

As mentioned in the previous section, inventory is managed by each retailer independently, even in the hyperconnected settings, because it is at the core of retailers' business. Among the plethora of inventory policies available, here we assume that inventory management is based on a service-level driven (s_{rp}, S_{rp}) policy for each product $p \in P_r$. The reorder point (s_{rp}) and order-up-to level (S_{rp}) are calculated using a three-standard-deviation (3σ) service level rule, with a normal approximated demand distribution.

When multiple open FCs are available with hyperconnected fulfillment, retailers can deploy inventories among the open FCs. In those scenarios, inventory deployment and allocation to customers are key dynamic decisions. The deployment among multiple FCs must be done in a way to balance inventories with respect to demand in the network, which means having the right amount of stock in each FC to fulfill future demand rather than having equal stock level in each FC. Inventory can be balanced by smartly redeploying inventories between FCs or deploying new stock among FCs. We assume suppliers stop and unload the new stock at multiple open FCs at the request of retailers to balance inventory, assuming close cooperation between retailers and suppliers. Inventory allocation to customers requires deciding from which FC to fulfill each customer. One way is to fulfill a customer order is to fulfill from the nearest available FC, which is the nearest to the customer location among those that have available inventory of the ordered product. This allocation decision will be the input to delivery routing method.

Note that the increase in the number of FCs does not increase the required inventory level because delivery zones are not preassigned to each FC. In other words, customer orders are allowed to be fulfilled from any FC so that there is complete inventory pooling among exploited FCs. Because the number of exploited FCs does not change demand of the city, the overall required inventory level and (s_{rp}, S_{rp}) level remain the same.

Routing

A router (O) generates daily delivery routes based on order information such as delivery location and time windows (TWs), inventory location assignments, expected traffic pattern in terms of travel speed, and expected delivery and installation time. A router represents the transportation department of a retailer or a 3PL service provider. It can belong to one retailer in scenarios with dedicated transportation, but with hyperconnected transportation, a router is modeled as an independent service provider who manages routing for all retailers simultaneously. Daily routes are constructed in the early morning every day and inventory location assignment made by an inventory manager of each retailer is forwarded to the router.

When no PI hub is used, all routes depart from an FC and all customers in the route are fulfilled by the FC. When hubs exist, two-tier routes are constructed. Last-mile routes (tier-2 routes) from hubs to customers are built by first assigning customers to the nearest hubs based on weighted distance ($w_h d(h, c_i)$). Then the customer-hub assignments are used as an input to build routes from fulfillment centers to the hubs (tier-1 routes). The weight for hub h $w_h = hc_h / hc_{suburb}$ is calculated by dividing the per-item usage cost of hub h (hc_h) by the per-item usage cost of hubs in suburban area (hc_{suburb}). Because the usage cost of a hub located in the city center is more expensive than in suburban area, hubs in the city center are penalized.

Each route (o) consists of an origin facility (FC (f_o) or hub (h_o)), a sequence of customers to visit ($\{\overrightarrow{C_o}\}$), and the departure time at the depot (T_o):

$$o = \{f_o \text{ or } h_o, T_o, \{\overrightarrow{C_o}\}\}$$

Note that when the sequence of customers to visit is known, a sequence of delivery locations and delivering items at each location is also known, as well as the expected delivery schedule referring to the starting time of the route.

The vehicle routing problem with time windows (VRPTW) is a well-known problem and there is a large variety of heuristics proposed in literature. For example, Solomon [71] proposed several construction heuristics, such as savings heuristics, nearest-neighbor heuristics, insertion heuristics, and sweep heuristics. The most successful heuristic was an insertion heuristic working as follows: at any position of a given route, insert if feasible a new customer among unrouted customers that maximizes distance and time savings as opposed to direct serving of the chosen customer. Initial routes can be constructed, for instance, by choosing an unrouted customer with the latest feasible time window. Campbell and Savelsbergh [72] concisely summarize information to be maintained, route updates, and route finalization for VRPTW with insertion heuristics. They suggest that pushing delivery schedules to the latest possible time minimizes waiting time of any feasible route. The VRPTW variant addressed by Liu, Guo, Yu, and Zhou [73], who consider stochastic assembly time in addition to a deterministic travel time, is the closest to what we propose in this chapter. They first modeled the problem as stochastic optimization with chance constraint, derived an equivalent deterministic program, and presented heuristics in sequence. The approach in this chapter is similar to the time-sliding mechanism and saving algorithm in Liu, Guo, Yu, and Zhou [73]. In the context of a large-item delivery, Weigel and Cao [74] present a routing-and-scheduling algorithm that involves worker and service request assignment.

The route construction heuristic with time window (TW) used in this chapter for last-mile routing is a variation of the best insertion heuristics described in Solomon [71] reflecting the observations of Campbell and Savelsbergh [72] on minimum waiting time. For tier-1 routing, which is a lot simpler due to a small number of delivery points and no customer TW constraint, the CW construction heuristic ([75]) is used. Insertion based heuristics and CW heuristics are still commonly used as a part of advanced routing algorithms these days ([76]; [77]). Although more advanced and sophisticated routing algorithms are available, a simple construction heuristic that can repeatedly generate good feasible routes in a short time satisfies the strategic goal of the chapter.

Algorithm 1: Routing Heuristic

```
    Input: Customer set $C_t$, Inventory status $I(r,p,f)\ \forall p \in P_r, f \in F_r, r \in R$
    Output: Tier-1 routes and tier-2 routes
 1: for all $c \in C_t$ do
 2:     $f_c = argmin_{f \in F}\{d(f,c) | I(r_c, p_c, f) > 0\}$
 3:     $h_c = argmin_{h \in H}\{w_h d(h,c)\}$
 4: end for
 5: Set $Unrouted = C_t$
 6: Set $O_t = \emptyset$
 7: while $|Unrouted| > 0$ do
 8:     $i = argmax_{i \in Unrouted}\{LD_i\}$
 9:     $o = \{h_o = h_i, T_o = LD_i - \tau(h_o, i | LD_i), \overrightarrow{C_o} = C_i\}$
10:     $b_o = LD_i$
11:     $t(o) = \tau(h_o, i|b_o) + E[ST_i] + \tau(h_o, i|b_o + ST_i)$
12:     $Unrouted \leftarrow Unrouted \setminus \{i\}$
13:     while $|Unrouted| > 0$ do
14:         $j = argmin_{j \in Unrouted}\{\Delta t(o, o_{+j}) | t(o_{+j}) \leq T_{labor}, v(o_{+j}) \leq V_{v_2}, w(o_{+j}) \leq W_{v_2}, h_j = h_o,$
                $b_{o_{+j}} = min(LD_j, b_o - \tau(j, i|b_o) - E[ST_{p_j}]) \geq ED_j\}$
        where $o_{+j} = \{h_o, T_{o_{+j}} = T_o - \Delta t(o, o_{+j}), \overrightarrow{C_{o_{+j}}} = \{j, \overrightarrow{C_o}\}\}$
15:         if j not NULL then
16:             $o \leftarrow o_{+j}$
17:         else
18:             $O_t \leftarrow \{O_t, o\}$ and go to line 7
19:         end if
20:     end while
21: end while
22: Set $Unrouted1 = C_t$
23: $O'_t = \{o_i \forall i \in C_t\}$ where $o_i = \{f_{o_i} = f_i, T_{o_i} = T_i, \overrightarrow{C_{o_i}} = \{i\}\}$
24: while $|Unrouted1| > 0$ do
25:     for $o_i \in O'_t$ do
26:         $o_{j(i)} = argmax_{o_j \in O'_t \setminus o_i}\{t(o_i) + t(o_j) - t(o_{i,j}) | t(o_i) + t(o_j) - t(o_{i,j}) > 0,$
                $t(o_{i,j}) \leq T_{labor,1}, v(o_{i,j}) \leq V_{v_1}, w(o_{i,j}) \leq W_{v_1}, f_i = f_j\}$
        where $o_{i,j} = \{h_i, T_1, \overrightarrow{C_{o_{i,j}}} = \{\overrightarrow{C_{o_i}, C_{o_j}}\}$
27:         $savings_i = t(o_i) + t(o_j) - t(o_{i,j})$
28:     end for
29:     Select $o_{i*} = argmax_{o_i \in O'_t}\{savings_i\}$
30:     if $o_{i*}, o_{j(i*)}$ not NULL then
31:         $O'_t \cup o_{i*,j(i*)} \setminus \{o_{i*}, o_{j(i*)}\}$
32:     else
33:         Go to line 36
34:     end if
35: end while
36: return tier-2 routes $O_t$ and tier-1 routes $O'_t$
```

The routing heuristic for two-tier routing with hubs is described in Algorithm algorithm 1. Routes are generated everyday using the set of customers to serve on day t (C_t) and the current inventory as an input. Each customer c is served from the nearest hub (h_c) using weighted distance and from the nearest available FC (f_c) (line 1 to 4). Lines 5 to 21 describe the tier-2 route construction algorithm. The routes are constructed sequentially and greedily and added to the set of tier-2 routes of day t (O_t). A route o begins to be shaped by choosing a seed customer i among unrouted customers who has the latest delivery time window (lines 8 to 12). b_o is the time of service start time at the first customer of route o. $t(o)$ is the expected total duration of route o and it consists of expected travel and service times. Travel and service times are stochastic, so expected values are used for routing. Expected travel time between a and b at time of day dt is calculated based on expected traffic of the day in time dt ($\tau(a, b|dt)$). Service time at customer j (ST_j) consists of delivery time (xi_{p_j}) and install time (ω_{p_j}). Iteratively, a new, not yet routed customer j who increases the total route duration the least is selected. Then, feasibility of adding customer j before the first customer of route o is checked as in line 14. Routing is constrained by maximum labor hour (T_{labor}), volume and weight capacity of each vehicle v (V_v, W_v), and time windows ($TW = (ED, LD)$). Different vehicle types are used for no-hub routes, tier-1 routes, and tier-2 routes (v_0, v_1, v_2). If feasible, route o is updated; otherwise, route o is finalized and added to O_t (lines 15 to 19). The tier-1 route construction algorithm is described in lines 22 to 35. The time window constraints are not applied to tier-1 routes delivering to hubs. As mentioned, a simple CW construction heuristic is used for tier 1. Because the early-morning pre-delivery to hubs cannot take full regular labor hours, shortened maximum labor hours ($T_{labor,1}$) are used for tier 1. An initial set of tier-1 routes of day t (O'_t) consists of single-stop routes to all customers. These routes are merged iteratively by taking two routes that maximizes time saving (line 29) until no merge is feasible and desirable.

Routing without hubs is the same as tier-2 routing, only it uses larger vehicles, and FCs rather than hubs. To be specific, line 3 and lines 22 to 35 are deleted and in lines 9 and 14, f_o,f_i replace h_o,h_i and v_0 replaces v_2. Finally, in line 36, it returns routes for each tier O_t and O'_t.

Urban Traffic Pattern

Traffic in each city area has a distinctive pattern, such as peak-hour distribution, and it has a significant impact on travel time for delivery. To approximately model such a traffic congestion pattern, the traffic in the city is stochastically set as an average travel speed for each hour of each day along each road type in the city. Even though it is possible to model with fine granularity vehicles moving across a city and thus explicitly model traffic signal and congestion for urban logistics study ([78]; [79]; [80]), this approach is highly resource-intensive. For our strategic modeling purposes, we have selected to do as Taniguchi and Der Heijden [81] have shown where a central traffic module sets average road-based and day-and-time-based travel speeds dynamically.

Traffic varies by road type, hour of the day, and by weekday. For example, Monday to Friday have heavy traffic during rush hours but traffic does not vary in such a way during weekends. Traffic on small roads varies by district but is assumed to be identical within a district. For example, small road traffic tends to be worse in the districts in the center of the city. Traffic on main roads is differentiated only by city center and perimeter: whether a path of a vehicle includes the city center or not. Average travel speed data by hour and weekday is collected between different pairs of points in the real world map which represent different segments: main roads (highways/boulevards) passing through the city center, main roads not passing through the city center, small roads at the city center, and small roads at the perimeter in South, West, North, or East of the city. Figure 2.3 shows typical hourly traffic patterns for weekdays and weekends.

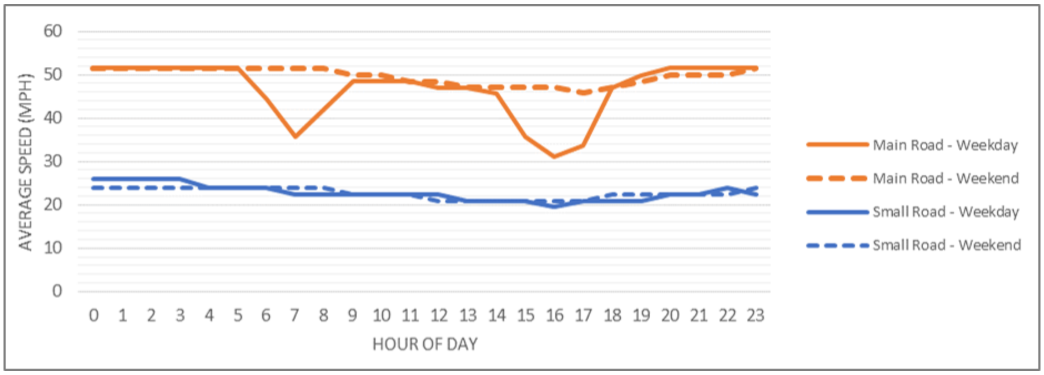


Figure 2.3: Typical traffic pattern in average speed by hour of day by road type in a weekday or in weekend

2.4 Simulation-based Experiment Design

To investigate the potential impact of the hyperconnected fulfillment and transportation system on last-mile delivery of large items as shown in Figure 2.1, scenario analysis via an optimization-supported simulation platform is used. Simulation-based scenario analysis is well proven to be an effective tool to compare different systems at a strategic level by evaluating overall performances, especially when the system is very complex involving dynamicity, interactions, and stochasticity. In this section, we start with scenarios designed to represent gradual PI transformations in each thread of fulfillment and transportation. Followed are KPIs carefully chosen to evaluate the scenarios in three dimensions: economic, service level, and environmental aspects. Last, we describe the optimization supported simulation platform built to evaluate scenarios based on the logistics system and decision architecture defined in the previous section.

2.4.1 Scenario Design

As mentioned, last-mile logistics in a city area can be divided into two main logistic operations: fulfillment and delivery. From a facility's point of view, hyperconnected fulfillment enables storage spaces such as FCs to be used on demand by any company. From a network's point of view, hyperconnected fulfillment enables any company to access and exploit a large, broad fulfillment network regardless of the company's size. Hyperconnected delivery involves a dynamic and broad range of flow consolidation and modularization of delivery activities by region and functional layers. It is often implemented through a network of various types of open logistic (PI) hubs. The modularization of routes enables dynamic flow consolidation and route specialization. For example, when long-haul routes are divided into short routes connected to each other at PI hubs, shipments from different long-haul routes can be merged fast and efficiently in some of the short routes. In the context of urban delivery, the multi-tier delivery system described by Crainic and Montreuil

[23] is one of the ways to implement hyperconnected delivery. The multi-tier system allows the last-tier routes, which are shorter and carry fewer items compared to original routes, to be operated with smaller greener vehicles in a multimodal approach including, for example, bicycles and electrical scooters.

Scenarios are constructed by transforming the current urban large-item logistics system into a hyperconnected system, first along each thread of fulfillment or delivery, and then by combining both transformations. Figure 2.4 illustrates urban logistics network topology and operation of each investigated scenario, representing the transformation into the hyperconnected system along the thread of fulfillment and delivery. Similar to Figure 2.1, the operations of four retailers and their customers are described in different colors. Each scenario in Figure 2.4 is described in detail hereafter.

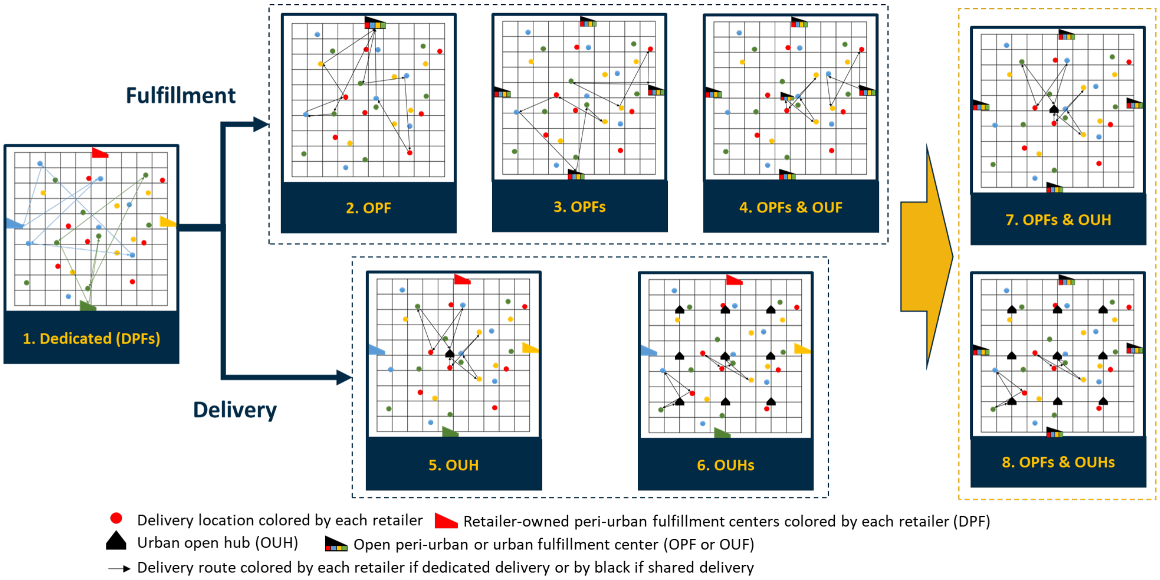

Fulfillment
Delivery
1. Dedicated (DPFs)
2. OPF
3. OPFs
4. OPFs & OUF
5. OUH
6. OUHs
7. OPFs & OUH
8. OPFs & OUHs
Delivery location colored by each retailer
Retailer-owned peri-urban fulfillment centers colored by each retailer (DPF)
Urban open hub (OUH)
Open peri-urban or urban fulfillment center (OPF or OUF)
Delivery route colored by each retailer if dedicated delivery or by black if shared delivery

Figure 2.4: Sample scenarios illustrating the gradual transformation from dedicated to hyperconnected logistic structure and routes with delivery time windows

Scenario 1 represents typical current dedicated operations. Each retailer operates its own fulfillment center, typically located in a peri-urban area (DPF, dedicated peri-urban fulfillment center), and each retailer optimizes its own delivery routes served by its own or exclusively contracted fleet of vehicles and delivery team. Scenario 1 forms the baseline to which other scenarios are compared.

Scenarios 2 to 4 are constructed along the thread of hyperconnected fulfillment. In these scenarios, customers of any retailer fulfilled from the same FC are served in multi-retailer routes. The scenario mimics the use of a third-party logistics orchestrator ([82]) contracted by all retailers to consolidate orders out of an FC to ensure efficient routes to customers, while the actual drivers and trucks may be from various businesses contracting to perform specific routes. Scenario 2 exploits a single open peri-urban FC (OPF) in which all retailers store multi-product inventory and from which all customers are served. Note that when delivery routes are dedicated to each retailer, delivery related metrics remain the same as scenario 1 given symmetric demand distribution and network structure.

Scenario 3 exploits multiple OPFs. Each retailer smartly deploys its inventory in any of the OPFs to ensure inventory availability at the nearest FC to future customers. Customers are served from the nearest FC that has available inventory. Proximity to customers is increased for all retailers. This scenario can be interpreted as a scenario in which existing DPFs of the retailers start to be open, operated by a third-party logistics orchestrator. When the current FCs are the most spread-out, as assumed in the scenario, the proximity to customers can be increased the most.

Scenario 4 introduces an additional open FC located at the city center (OUF) to scenario 3. The capacity of the OUF is limited and unit storage cost at the OUF is higher compared to that of OPFs due to the limited land and high rent at the city center.

Scenarios 5 and 6 experiment further hyperconnected delivery, beyond the multi-retailer delivery in scenarios 2-4, by exploiting open urban hub(s) (OUH(s)). FCs are retailer-dedicated in these scenarios, but products are first shipped to OUHs from FCs in the early morning on a large truck and shipped to customers from OUHs during the day on a small truck. The former routes are referred as tier-1 routes and the latter routes are referred as tier-2 routes. Scenario 5 exploits a single OUH at the city center and scenario 6 exploits a network of OUHs spread over the city.

Scenarios 7-8 represent combined hyperconnected fulfillment and delivery systems. Both of them exploit multiple OPFs as in scenario 3. Scenario 7 exploits a single OUH and scenario 8 exploits a network of OUHs. Scenario 7 can be seen as a combination of scenarios 3 and 5 and scenario 8 can be seen as a combination of scenarios 3 and 6.

2.4.2 Key Performance Indicators

It is critical to use proper KPIs to evaluate scenarios from multiple perspectives. We categorize the KPIs into three major groups by the type of impacts they measure: economic, environmental and social impacts.

All indicator values are measured as a daily expected value because the last-mile operations including customer arrivals are on a daily basis. That is, when a simulation runs for T days, T observations of each indicator value are obtained and the expected value is calculated based on these observations.

Measurement of Economic Impacts

The most straightforward KPIs to use are travel distance and several cost components. Cost components related to delivery are labor, fuel, and vehicle rental cost. Cost components related to facility cost are hub usage and inventory holding cost. The sum of all cost components provides the total induced cost. In this experiment, the unit of all costs is $/day.

To define the KPIs, first let the distance in miles between locations i and j be d_{ij}. O_t is a set of routes to serve on day t, $d(o) = \sum\sum_{i,j\in\{depot,\overrightarrow{C}_o,depot\}} d_{ij}$ is the travel distance of route o. The depot would be an FC f_o or a hub h_o. V_t is a set of vehicles used on day t, and $lc(v), fc(v), vc(v)$ are labor cost, fuel cost, and vehicle cost associated with each vehicle $v \in V_t$ respectively. Different hourly rates are applied to each personnel based on his or her role; therefore, the labor cost $lc(v)$ is calculated differently by the type of routes assigned to each vehicle v. For example, installation staff requires higher hourly wages than delivery staff. To serve a deliver-and-install route, one installation staff and one delivery staff form a team. Alternatively, to serve a tier-1 route, only one driver is needed. Vehicle cost can be the cost of daily/hourly vehicle rental or vehicle maintenance cost depending on context. In this chapter, vehicles are assumed to be rented by the hour in all scenarios for consistency, so the rental cost is proportional to usage time.

The cost of using a hub is charged by item going through each hub. hc_h is the cost of using hub h per item, and it is more expensive for a hub in the city center. $TH_t(h)$ is the throughput at hub h on day t. ic_f and $IO_t(f)$ are unit inventory holding costs ($/item,day) and on-hand inventory level on day t at FC f, respectively. Similar to hub costs, unit holding costs are more expensive for a FC at the city center. In addition to costs, we present travel miles to provide better insight because delivery costs are not always proportional to travel miles, especially with hub operations.

The economic KPIs are expressed as follows:

- (KPI1a) Expected daily travel miles: $E[TM] = \frac{1}{T}\sum_{t=1}^{T}\sum_{o\in O_t} d(o)$
- (KPI1b) Expected daily labor cost: $E[LC] = \frac{1}{T}\sum_{t=1}^{T}\sum_{v\in V_t} lc(v)$
- (KPI1c) Expected daily fuel cost: $E[FC] = \frac{1}{T}\sum_{t=1}^{T}\sum_{v\in V_t} fc(v)$
- (KPI1d) Expected daily vehicle rental cost: $E[VC] = \frac{1}{T}\sum_{t=1}^{T}\sum_{v\in V_t} vc(v)$
- (KPI1e) Expected hub usage cost: $E[HC] = \frac{1}{T}\sum_{t=1}^{T}\sum_{h\in H} hc_h * TH_t(h)$
- (KPI1f) Expected inventory holding cost: $E[IC] = \frac{1}{T}\sum_{t=1}^{T}\sum_{f\in F} ic_f * IO_t(f)$
- (KPI1g) Expected total delivery cost: $E[TTC] = E[LC] + E[FC] + E[VC]$
- (KPI1h) Expected total facility cost: $E[TFC] = E[HC] + E[IC]$
- (KPI1i) Expected total induced cost: $E[TC] = E[LC] + E[FC] + E[VC] + E[HC] + E[IC]$

Measurement of Service Level Impacts

In this chapter, service capability is used to measure social impacts, which can be potentially improved by a hyperconnected urban logistics system. When a delivery time window (TW) is required by a customer, lateness plays a critical role in the customer's experience, which can result in delivery postponement and rescheduling or even lost sales. As direct measures, expected lateness and conditional expected lateness time are calculated.

- (KPI2a) Expected lateness per order: $E[lateness] = \frac{\sum_{t=1}^{T}\sum_{o\in O_t}\sum_{c\in o}[DT_c-LD_c]_+}{\sum_{t=1}^{T}\sum_{o\in O_t}|o|}$
- (KPI2b) Conditional expected lateness per order: $E[lateness|lateness > 0] = \frac{\sum_{t=1}^{T}\sum_{o\in O_t}\sum_{c\in o}[DT_c-LD_c]_+}{\sum_{t=1}^{T}\sum_{o\in O_t}|o_+|}$

where $|o|$ is the number of customers in route o, o_+ is customers in route o whose delivery was late, DT_c is delivery time to customer c, and LD_c is the latest delivery time of time window of customer c. Note that the delay is calculated as $[DT_c - LD_c]_+$. Also, the (KPI2b) will be simply zero when there is no late delivery.

Montreuil, Labarthe, and Cloutier [83] modeled various types of TW requirements and satisfaction curves and showed the impact of delays and early delivery in a make-to-order environment. Similarly, customer satisfaction level is also measured using three different types of utility function.

- (KPI2c-k) Expected customer satisfaction level with utility function g_k, $k = 1, 2, 3$: $E[Satisfaction|g_k] = \frac{\sum_{t=1}^{T}\sum_{o\in O_t}\sum_{c\in o} f_k(DT_c,LD_c)}{\sum_{t=1}^{T}\sum_{o\in O_t}|o|}$ where
  - $g_1(DT_c, LD_c) = \frac{\rho-[DT_c-LD_c]_+}{\rho}$: linear utility function
  - $g_2(DT_c, LD_c) = \frac{1}{1+[DT_c-LD_c]_+}$: Inverse utility function
  - $g_3(DT_c, LD_c) = exp^{-[DT_c-LD_c]_+}$: Exponential utility function

Note that all the utility functions have a satisfaction level of 100% when the delay is 0 and a satisfaction level converging on 0 as delay time increases. For all calculations of the utility functions, DT_c and LD_c are measured in hourly units. In g_1, the constant $\rho = 8$ is used because the regular working hour is 8 hours a day.

Measurement of Environmental Impacts

From an environmental perspective, GHG or toxic gas emissions are important because delivery vehicles are one of the most critical sources of GHG emissions and air pollution in urban areas. We measure carbon dioxide (CO_2) emission, the most critical GHG, emitted by the combustion of fossil fuels in motor vehicles. GHGs are linked to global warming, and there is a strong global push to curb CO_2 emissions. In addition to CO_2, vehicular emission also leads to the release of toxic gases such as sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) into the atmosphere. The emission of fine particles, especially the atmospheric particulate matter ($PM_{2.5}$), is another important measure of environmental impact. $PM_{2.5}$ can penetrate into the lungs during inhalation and can lead to severe respiratory illnesses ([84]). In addition to the atmospheric emission, fuel consumption of each vehicle v, $fuel(v)$, is measured.

To estimate the cumulative emissions of CO_2, SO_2, NO_2, and $PM_{2.5}$, we used the Motor Vehicle Emission Simulator (MOVES) software produced by the US Environmental Protection Agency (EPA). The software estimates the emissions based on the number of vehicles, vehicle types, road types, travel miles, and average travel speed. The input parameters are generated from the output of each scenario-based simulation run using our urban sandbox simulator. We use Fulton County, Atlanta, GA as an example city for MOVES. Even though this approach is sufficient for our purposes, it is possible to get more accurate measures by using additional simulators for GHG emissions, such as that proposed by Osorio and Nanduri [80] who combine a traffic simulator and fuel consumption simulator for accurate fuel consumption estimation.

The environmental KPIs are calculated as follows:

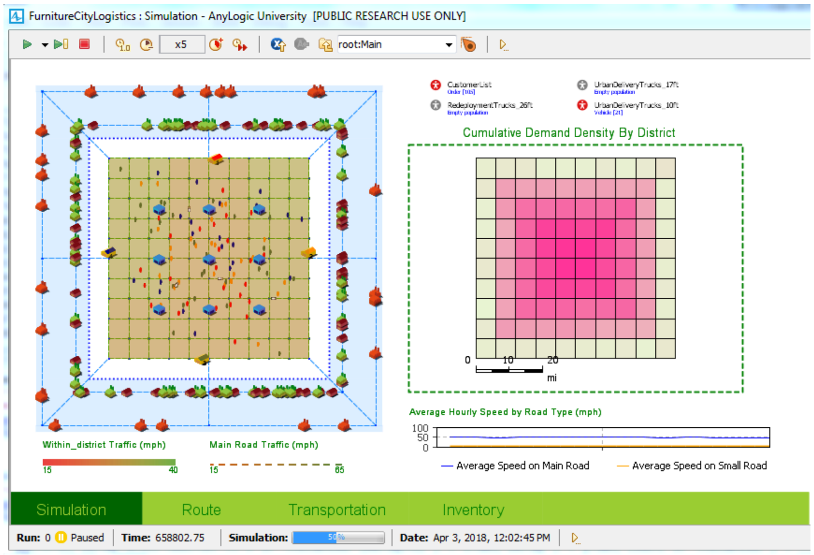
- (KPI3a) Expected daily CO_2 emissions: $E[CO_2] = \frac{1}{T}\sum_{t=1}^{T}\sum_{v\in V_t} CO_2(v)$

- (KPI3b) Expected daily SO_2 emissions: $E[SO_2] = \frac{1}{T} \sum_{t=1}^{T} \sum_{v \in V_t} SO_2(v)$
- (KPI3c) Expected daily NO_2 emissions: $E[NO_2] = \frac{1}{T} \sum_{t=1}^{T} \sum_{v \in V_t} NO_2(v)$
- (KPI3d) Expected daily $PM_{2.5}$ emissions: $E[PM_{2.5}] = \frac{1}{T} \sum_{t=1}^{T} \sum_{v \in V_t} PM_{2.5}(v)$
- (KPI3e) Expected daily fuel consumption: $E[Fuel] = \frac{1}{T} \sum_{t=1}^{T} \sum_{v \in V_t} fuel(v)$

2.4.3 Optimization-supported Simulation Platform

Here, we describe the optimization-supported simulation platform built on the network and decision architecture described in the previous section. An agent-based discrete-events simulation is exploited in this study. Figure 2.2 in section 3 shows the structure of the simulation using a conceptual diagram illustrating the key agents, objects, and interactions between them described as information and physical flows in the middle. The key players and logistics resources are modeled as agents. For example, retailers and customers are modeled as individual agents as seen in Figure 2.2. The optimization algorithms for key decisions are embedded in active agents. For example, routing and scheduling algorithms and inventory algorithms are respectively embedded in the router agent inventory manager agent. Customers are classified as active agents due to the retailer substitution process described in subsubsection 2.3.3. Figure 2.5 shows a snapshot of the simulation used in the study.

Along with a simulation platform, one of the critical components of a simulation experiment is data. The experiment in this chapter is designed to explore novel systems without using industry data or existing benchmark data. Therefore, the market data, including product specifications, sales, and retailer portfolios, are designed to be as realistic as possible referring to information available on the web, for example, exploiting information on the website of an e-commerce retailer selling furniture and large appliances. A total of 440 products are modeled, each belongs to one of 13 product categories. The product categories define general types of products such as a desk, chair, and refrigerator. The relative



FurnitureCityLogistics : Simulation - AnyLogic University [PUBLIC RESEARCH USE ONLY]

Run: 0 Paused | Time: 658802.75 | Simulation: 50% | Date: Apr 3, 2018, 12:02:45 PM

Figure 2.5: Model city and simulation snapshot

sales volume is designed to follow a Pareto curve, both in total and within each category, as seen in Figure 2.6.

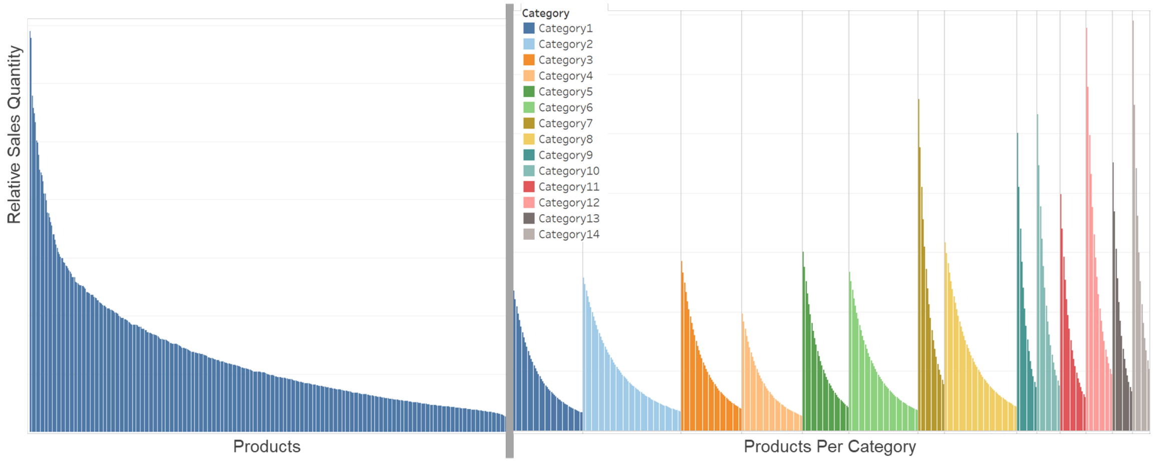


Figure 2.6: Pareto sales: total and per category

The product portfolio of each retailer is carefully designed as well. The most critical characteristic of retailers to model is that retailers are selling products produced by different manufacturers. In other words, based on the size or strategy of the manufacturer, the same products can be sold by a single retailer (exclusive contract), by only a few retailers, or by all retailers. The product portfolio of the four base retailers, used for the main experiments, is designed accordingly. Figure 2.7 shows the number of products sold by the four

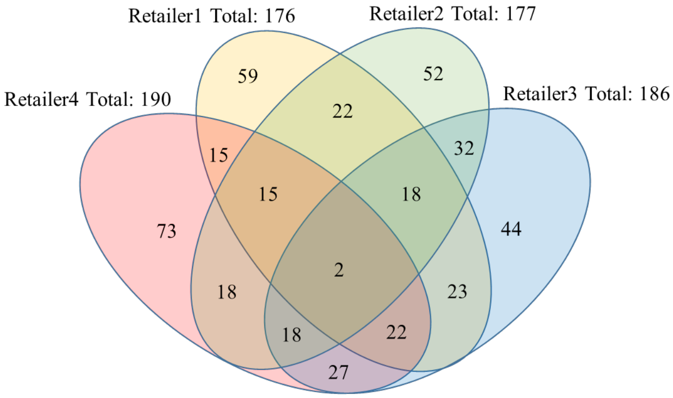


Figure 2.7: Number of products sold by retailers

base retailers. The numbers in overlapped areas represent the number of products sold by multiple retailers.

2.5 Experimental Results

The decision and system architecture and scenarios constructed in previous sections are built into a simulation using Java and AnyLogic 8.2.3 for computational experiments. Each scenario is ran for 2.5 years where the first 0.5 year is used as a warm-up period. From the results, the KPIs proposed in the subsection 2.4.2 are calculated based on which scenarios are evaluated and compared. Results of the eight main scenarios are hereafter analyzed, followed by a sensitivity analysis.

2.5.1 Comparative Results

Experimental results for each of the eight scenarios described in section 4.1. are presented here. The different combinations of dedicated or hyperconnected fulfillment and delivery options have led to highly distinctive routing and scheduling patterns. With hyperconnected fulfillment, routes can be shortened because customers can be served from nearer FCs and deliveries of different retailers can be consolidated as shown in scenarios 2-4 in Figure 2.4. With hyperconnected delivery, as in scenarios 5 and 6 in Figure 2.4, customers are served with very short routes from OUHs, where environmentally friendly vehicles can be used.

Short tier-2 routes can reduce road congestion in the city during the day, and tier-1 delivery can be done in the early morning when traffic is not heavy. Combined, tier-1 routes can also be efficient because deliveries to hubs can be served from nearer FCs and consolidated among retailers while maintaining the benefits of multi-tier delivery, as shown in scenarios 7 and 8 in Figure 2.4.

Let us concentrate first on economic performance. The total induced cost (KPI1i) can summarize the economic efficiency of each scenario most effectively. Figure 2.8 shows the total induced cost along with its elements, fulfillment, and delivery costs of each scenario. On top of each hyperconnected scenario, the total cost reduction rate with respect to scenario 1 is shown.

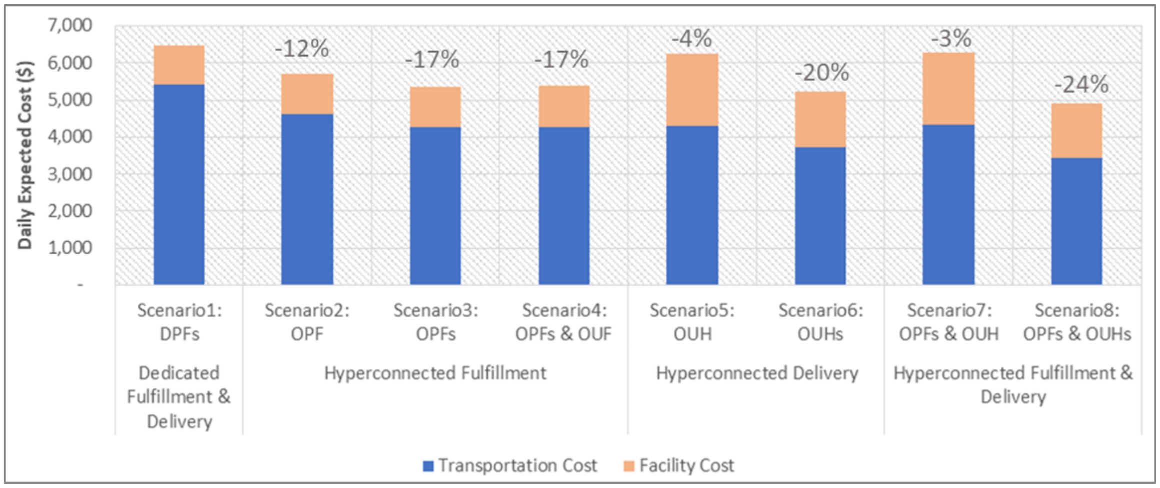


Figure 2.8: Total induced cost consists of fulfillment and delivery cost by scenario with total cost reduction rate with respect to scenario 1

With open FCs as in scenarios 2-4, delivery costs are reduced because many customers can be served from closer FCs. Facility costs remain almost unchanged throughout scenarios 1-3 because the inventory holding cost in peri-urban FCs is set to be the same. Even when a more expensive urban FC is used in scenario 4, it remains almost the same because only a small amount of inventory is stored in the urban FCs due to limited capacity. However, with open hub(s), as in scenarios 5 and 6, delivery costs are further reduced with the two-tier routing scheme but facility costs are increased due to the added hub costs. Compared to scenarios with a single hub at the expensive city center, hub costs are saved by using hubs in the suburban area in scenarios with multiple hubs.

Cost reduction rates of scenarios 2-8 in comparison to scenario 1 are shown in Figure 2.8. Each cost component and the total cost are shown in Table A.1 in Appendix A. From scenarios 2 and 3, it can be seen that open PF(s) can reduce total costs by 12% to 17% depending on the number and location of the OPF(s). However, by comparing scenarios 3 and 4, it can be inferred that an open UF (OUF) hardly brings more economic gain. Scenario 3 is dominant among the scenarios in hyperconnected fulfillment.

The total cost was only reduced by 3∼4% by using a single OUH regardless of fulfillment operations, as shown in scenarios 5 and 7. However, significant cost reduction in scenario 6 shows that having multiple OUHs, especially those not in the city center where the hub cost is a lot cheaper, can be beneficial. The cost achieved with hyperconnected delivery in scenario 6 is about 3% smaller than the cost achieved with the hyperconnected fulfillment in scenario 3 because the savings from efficiency of tiered delivery outweighs the cost of open hub use.

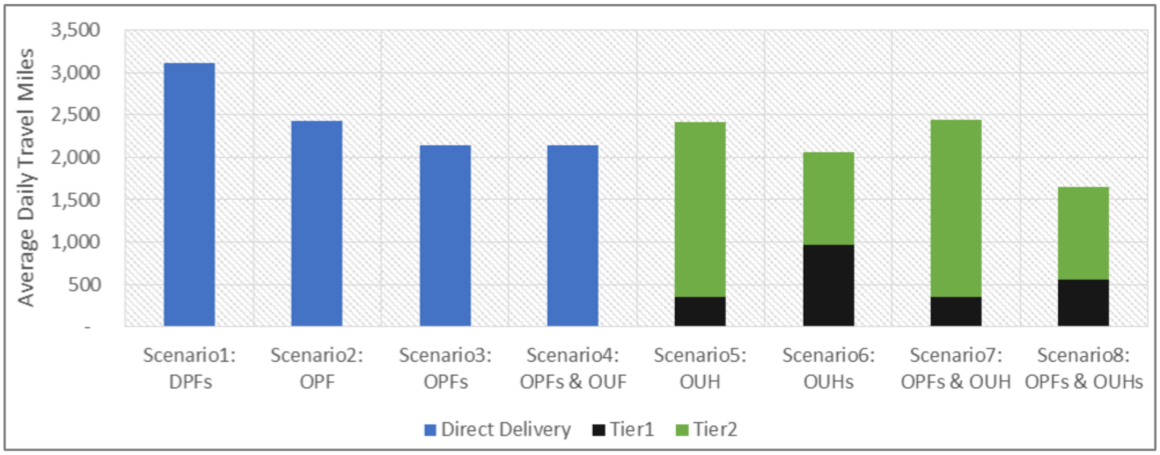


Figure 2.9: Average daily travel miles by delivery tier by scenario

As mentioned earlier, the cost of scenario 7 is not significantly different from the cost of scenario 5, varying by about only 1%. However, the cost of scenario 8 is significantly smaller than the cost of scenario 6, differing by 4%. That is, under the operation of the multiple OUHs, combining hyperconnected fulfillment can further reduce the cost mainly by improving tier-1 operations. From Figure 2.9 which presents average daily travel miles (KPI1a) by delivery type, it can be inferred that by distributing inventories in OPFs, tier-1 operations are improved significantly in scenario 8 compared to scenario 6 where each retailer must supply all hubs from a single PF independently. Such a benefit does not exist with a single hub operation where all tier-1 routes are simply single-stop round trips. Scenario 8 reduces the total cost the most, by 24%, and later on it will be shown that other KPIs have similar patterns. Therefore, we can say that scenario 8 dominates over all the other scenarios. However, it must be noted that either hyperconnected fulfillment or delivery can solely achieve about three quarters of the cost reduction found by the ultimate hyperconnected fulfillment and delivery scenario. In practice, this implies that implementing either one of the threads can be a good first step toward a hyperconnected system. Meanwhile, the marginal gain from transforming the remaining thread when one of the threads is transformed first is still significant.

Figure 2.10 illustrates the service level improvement through the reduction rate of overall and conditional delivery lateness level per order (KPI2a and KPI2b) with respect to baseline scenario 1. It can be seen that the expected lateness per order can be reduced by $\sim$40% and the conditional lateness by $\sim$30% with hyperconnected delivery. Using multiple hubs, hyperconnected delivery can reduce these measures by up to $\sim$95% and $\sim$50%, respectively. With hyperconnected delivery, the difference between overall and conditional lateness increases, meaning that the lateness probability is decreased significantly. Note that under hyperconnected delivery, having multiple hubs can reduce delivery lateness more compared to the reduction achieved by a single hub, as shown by smaller reduction rates in scenarios 5 and 7 compared to the reduction rates in scenarios 6 and 8.

Customer satisfaction level improvement by different utility function types $-$linear (KPI2c-1), inverse (KPI2c-b), and exponential (KPI2c-3)$-$ is described in Figure 2.11. The impact of lateness reduction on customer satisfaction level depends on the shape of

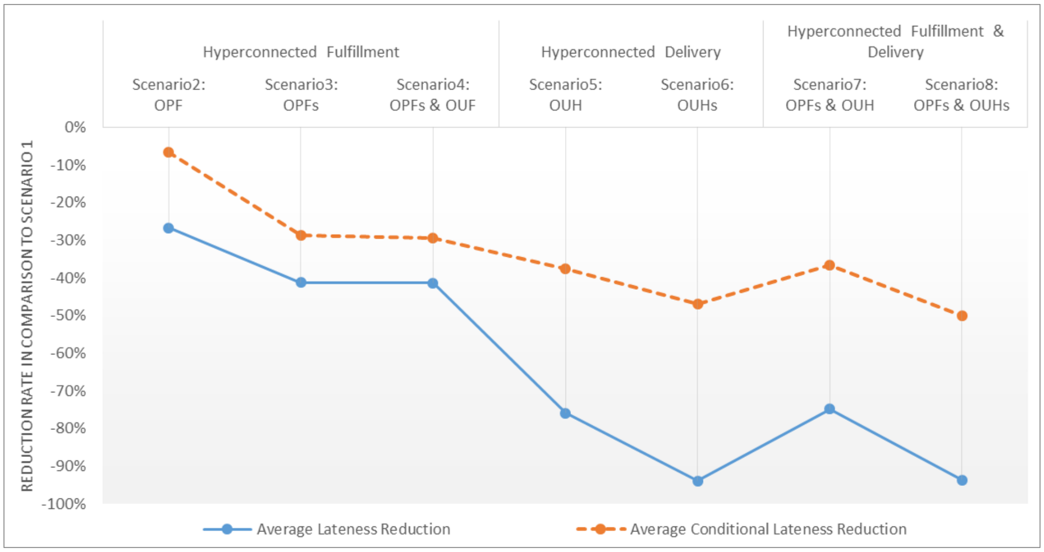


Figure 2.10: Lateness reduction rate by scenario

customer utility function. With a linear utility function, there is only very marginal improvement. However, with inverse or exponential utility function, meaningful improvement in customer satisfaction level is observed. This is because the satisfaction level decreases faster with inverse and exponential functions than with a linear function.

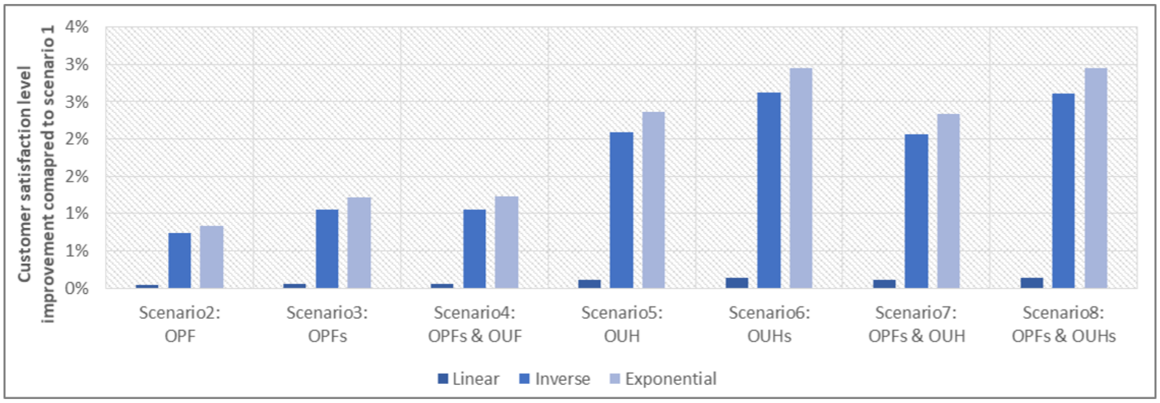
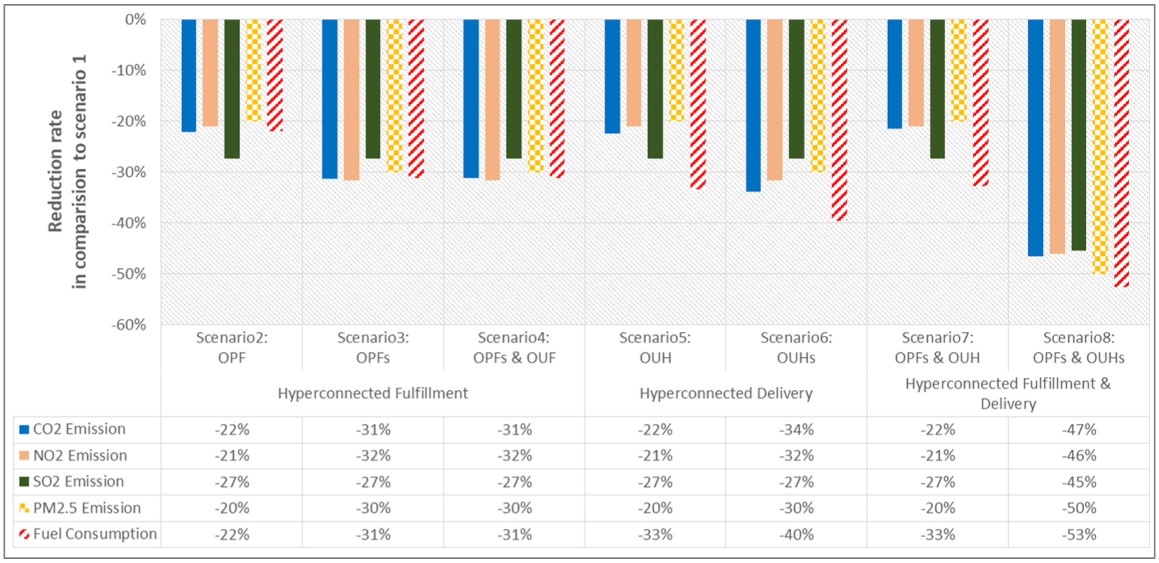


Figure 2.11: Customer satisfaction level improvement by utility function type by scenario

Environmental KPI improvement is illustrated in Figure 2.12. With little variation, all four environmental KPIs follow a similar pattern. As in the previous performance perspectives, scenarios 3, 6, and 8 are the best scenarios with respect to environmental KPIs from their corresponding thread of transformation. Similar to the insights from economic KPIs, OPFs do not make any difference in cases of single hub operations (comparing scenario 5 and 7), whereas OPFs have significant impact when multiple hubs are used (comparing 6 and 8). In all scenarios except scenario 8, the reduction rate only reaches 20∼30%. However, in scenario 8, all emissions are reduced by almost half. This is an inspiring result for companies with environmental goals or city governments needing to meet GHG emission limits. Moreover, such significant reduction affects air quality in the city, which in turn can improve quality of life of people living and/or working in the city. Also, the reduction in fuel consumption can help companies to save operation costs.



	Scenario2: OPF	Scenario3: OPFs	Scenario4: OPFs & OUF	Scenario5: OUH	Scenario6: OUHs	Scenario7: OPFs & OUH	Scenario8: OPFs & OUHs
	Hyperconnected Fulfillment			Hyperconnected Delivery		Hyperconnected Fulfillment & Delivery	
■ CO2 Emission	-22%	-31%	-31%	-22%	-34%	-22%	-47%
■ NO2 Emission	-21%	-32%	-32%	-21%	-32%	-21%	-46%
■ SO2 Emission	-27%	-27%	-27%	-27%	-27%	-27%	-45%
▨ PM2.5 Emission	-20%	-30%	-30%	-20%	-30%	-20%	-50%
▨ Fuel Consumption	-22%	-31%	-31%	-33%	-40%	-33%	-53%

Figure 2.12: Environmental KPI improvement by scenario

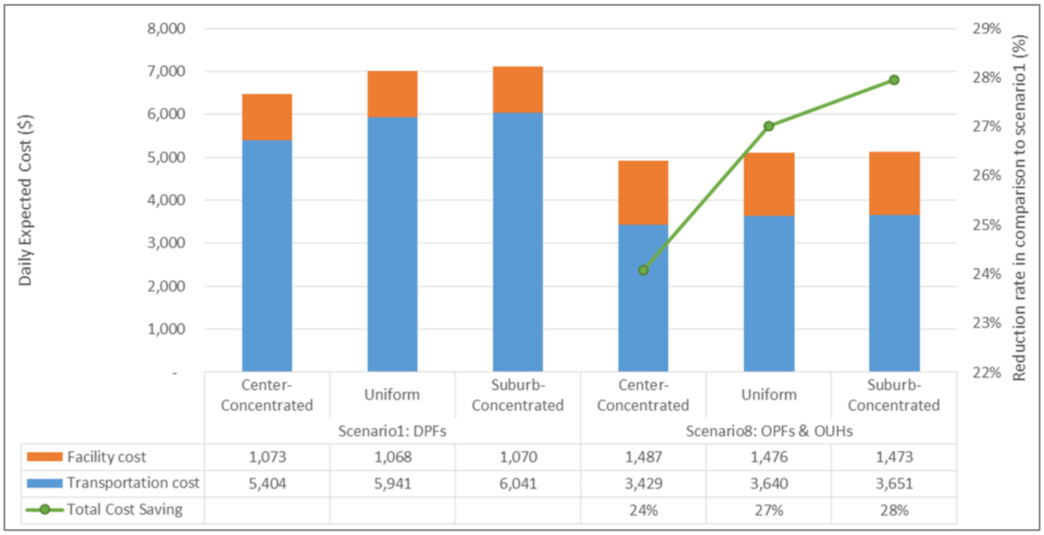
In summary, hyperconnected fulfillment and delivery scenario 8 shows improvement in all KPIs - 24% cost reduction, more than 90% reduction in delivery lateness, and about 50% reduction in GHG emission - in comparison to base scenario 1, exploiting dedicated fulfillment and delivery. Moreover, as can be seen from scenarios 3 and 6, hyperconnected fulfillment or delivery alone can achieve about 60% of the KPI improvements in scenario 8.

2.5.2 Sensitivity Analysis

In this section, we perform a sensitivity analysis of the robustness of the previous results with respect to key assumptions: (1) demand distribution, (2) demand size, (3) the number of retailers, (4) storage capacity limits, and (5) usage cost of open facilities. Each variation is explained and its impacts are shown with experimental results for selected scenarios.

Demand Distribution

Each city has its own shape and population distribution. Such attributes of the city determine the distribution of delivery locations. In the main experiment, city-center-concentrated demand distribution is assumed, where demand share is diminishing with distance from the city center. However, demand may not be concentrated in the city center in some cities. For example, it can be rather concentrated in the suburbs where residential complexes are located. To examine the impact of demand distribution, the baseline scenario 1 and the dominant hyperconnected scenario 8 are simulated under geo-uniform demand and suburb-concentrated demand distributions. The resulting expected daily cost is shown in Figure 2.13 and each cost component is shown in Table A.2 in the Appendix.



	Center-Concentrated	Uniform	Suburb-Concentrated	Center-Concentrated	Uniform	Suburb-Concentrated
	Scenario1: DPFs			Scenario8: OPFs & OUHs		
🟧 Facility cost	1,073	1,068	1,070	1,487	1,476	1,473
🟦 Transportation cost	5,404	5,941	6,041	3,429	3,640	3,651
—●— Total Cost Saving				24%	27%	28%

Figure 2.13: Expected daily cost of scenario 1 and 8 by demand distribution and cost reduction rate of scenario 8 in comparison to baseline scenario 1

For all demand distributions, scenario 8 saves about a quarter of the total induced daily cost. This suggests that PI transformation can bring significant benefit to any city regardless of population or demand distribution. Cost savings are the smallest at 24% under center-concentrated demand distribution and increased to 27% and 28% under uniform and suburb-concentrated demand distribution, respectively. This is because the more demand is spread over a large area, the more gains in delivery efficiency can be achieved with the hyperconnected system. It can be seen that most of the additional savings come from the delivery cost. In fact, the savings in delivery cost with scenario 8 under center-concentrated, uniform and suburb-concentrated distributions are 37%, 39%, and 40%, respectively.

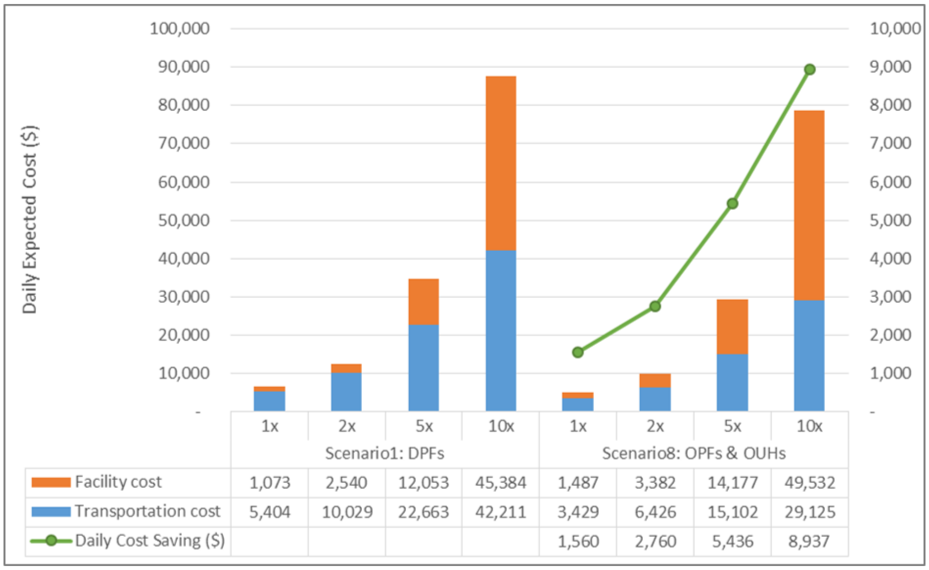
Demand Size

Demand size, measured as the average number of customers to be served a day, also affects the hyperconnected logistics system. For a main scenario comparison, each retailer serves 20 customers per day on average. Demand size is here increased by 2, 5, and 10 times so that each retailer is to serve 40, 100, and 200 customers every day on average. Scenarios 1 and 8 are evaluated at the increased demand size. Resulting expected daily costs under each demand size are shown in Figure 2.14.

Total daily cost savings increase with the size of demand, up to $8,937 with 10x demand size from $1,560 with 1x demand size. In fact, the cost savings are increasing slightly less than linearly relative to demand size. This does not mean that there will be no benefit from hyperconnected logistics when the size of each retailer is large enough to have its own economies of scale. It should rather be interpreted that the cost savings would converge, reaching a maximum cost savings with respect to demand size. Also, in all experiments, the number of retailers was fixed. When more retailers get involved as the demand size increases, the total cost savings can be changed, and it is expected to climb because there are more retailers for a given urban demand size.

Number of Retailers

The baseline case used for the main computational experiment offers a business context in which four large item retailers operate in the city. In this subsection, instances with an



	Scenario1: DPFs				Scenario8: OPFs & OUHs			
	1x	2x	5x	10x	1x	2x	5x	10x
🟧 Facility cost	1,073	2,540	12,053	45,384	1,487	3,382	14,177	49,532
🟦 Transportation cost	5,404	10,029	22,663	42,211	3,429	6,426	15,102	29,125
● Daily Cost Saving ($)					1,560	2,760	5,436	8,937

Figure 2.14: Expected daily cost of scenario 1 and 8 by demand size and cost reduction rate of scenario 8 in comparison to baseline scenario 1

expanded number of retailers are tested. This is done by (1) fixing the total market size or (2) proportionally increasing the total market size while the relative sizes of retailers are set to be the same. The results of the sensitivity analysis with 8 and 12 retailers are presented in Table 2.1 in terms of savings in total costs and in transportation costs in scenario 8 with respect to scenario 1. We notice that the network of open facilities for scenario 8 is fixed as in the main experiment and that the additional experiments were run on an updated version of the simulation platform Anylogic 8.5.1.

The incremental percentage cost savings, summarized in Table 2.1, is given by number of retailers and market size. These complementary results underline the robustness of our previous findings because the percentage savings have not changed significantly by the number of retailers regardless of total market size. In other words, the percentage savings for each retailer remains almost the same, regardless of the total number of retailers or the size of each retailers. It also means the percentage savings are not benefited by economies of scale which is in line with the results in the sensitivity analysis with respect to demand size. The number of retailers increasing while the total market size is fixed indicates reduced size of each retailer. This brings the additional cost savings to near zero. The additional cost savings were slightly negative in the case of eight retailers due to increased safety stock to cover relatively higher demand variation with respect to expected demand size. The major insight here is that the benefits of the hyperconnected system came not only from the size of total market size or the number of participants, but also from the structural change of PI transformation. In other words, because the multiplayer operations are offered and used as a service, a hyperconnected system is not constraining the number of players using the service, in contrast to a collaborative system.

Table 2.1: Incremental percentage savings transforming from scenario 1 to scenario 8 with respect to the four-retailer case as the number of retailers increases

Number of Retailers	**8**	**12**	**8**	**12**
Relative Market Size	**1x**	**1x**	**2x**	**3x**
Δ % Saving in Total Cost	-0.5%	0.2%	1.8%	2.9%
Δ % Saving in Transportation Cost	0.1%	2.1%	2.3%	3.6%

Storage Capacity Limits

All scenarios have been run assuming a very large/unlimited storage capacity of all FCs. However, in several practical contexts, each FC could have a tight capacity limit that can be an obstacle toward deploying inventory at a desired location. For scenarios with dedicated FCs, capacity constraints can be relaxed without loss of generality if we assume that each FC has been designed to handle the demand variability of the owning retailer. Therefore, capacity constraint is applied only to dominant hyperconnected scenario 8 and compared to scenario 1 and 8 without capacity constraints. Results of scenario 1 are used to determine storage capacity of each peri-urban FC measured as the number of products. Note that no more capacity should be needed for hyperconnected fulfillment because inventory levels in the network remain the same due to a demand-driven inventory policy.

The expected daily cost for scenarios 1 and 8 with no capacity constraint as experimented for the base scenario analysis and the expected daily cost for scenario 8 with limited storage capacity are shown in Table 2.2. The storage capacity constraints only slightly increase daily costs. This is because the impact of sub-optimal demand allocation to FCs with capacity constraints is mitigated with 2-tier delivery. This result strengthens the scenario analysis results in this chapter, which are done without capacity constraints. To summarize, benefits of the hyperconnected system can be obtained without additional storage capacity.

Table 2.2: Expected daily cost in scenarios 1 and 8 with no capacity constraint and scenario 8 with capacity constraint

Scenario ID	**Capacity Constraint at Each Fulfillment Center**	**Average Daily Cost ($)**
1	No	6,477
8	No	4,917
8	Yes	4,959

Usage Cost of Open Facilities

The usage cost of open facilities, the open FCs, and open hubs is determined based on the closest current market practice. However, as a new operational scheme, the usage cost of open facilities is rather uncertain. Therefore, the sensitivity analysis is conducted to measure the impact of usage cost of the open facilities on the cost reduction rate of a hyperconnected fulfillment and delivery system (scenario 8) with respect to a dedicated system (scenario 1). Figure 2.15 shows the expected cost reduction rate changes over a range of unit open hub usage costs and unit inventory holding costs at an open FC. The x-axis shows the size of each cost as a multiplication of the base cost. Point A corresponds to the base cost results shown in the main computational analysis in the chapter. Point B (2.5x) and C (4.7x) are the points at which the operation costs of the hyperconnected system become equal to that of the dedicated system in terms of each cost, respectively. The results show that even though the holding costs at open FCs is twice more as expensive as than the holding costs at dedicated FCs, hyperconnected systems still result in $\sim$10% cost reduction. Also, even though the hub usage cost is tripled, we still can get more than a 10% overall cost reduction.

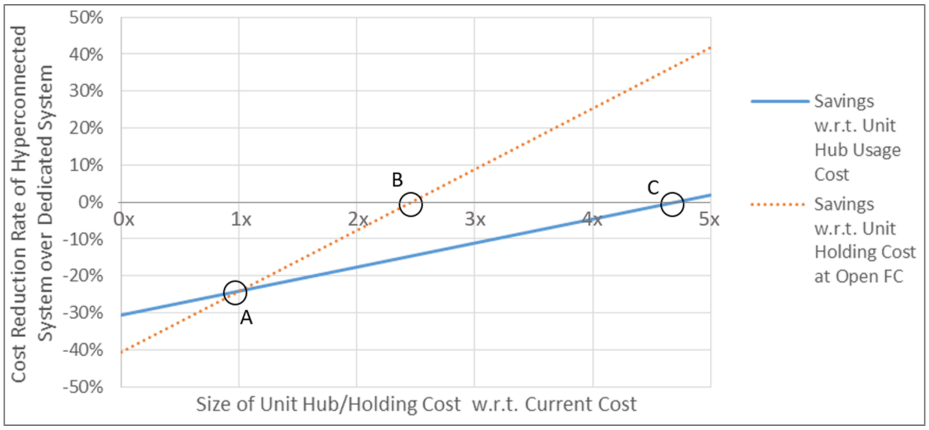


Figure 2.15: Sensitivity analysis with respect to open facility usage cost

2.6 Conclusion

In this chapter, a novel urban logistics system in line with Physical Internet concepts, a hyperconnected urban logistics system, is designed for last-mile delivery and fulfillment. A comprehensive decision and system architecture for the proposed system is presented and the gradual transformation from a current dedicated system towards a proposed hyperconnected system is illustrated through eight scenarios. Along with the comprehensive modeling of system and decision architecture, the scenario-based analysis provides a potential guideline to implement such PI transformations in practice. The potential benefits of the proposed system are examined from economic, service, and environmental perspectives through simulation-based scenario analysis on a simulation platform built upon the proposed architecture. A case of large-item delivery is used for experimental purposes. Results show the marginal impact of each transformation step. In short, compared to a dedicated system, a hyperconnected fulfillment and delivery system through open OPFs and OUHs could operate with 24% lower cost concurrent to reducing GHG and atmospheric particulate matter ($PM_{2.5}$) emission by about 50% while enhancing customer satisfaction by reducing delivery lateness. Hyperconnected fulfillment or delivery alone could reduce the cost by 17% and 20%, respectively with about 30% reduction in GHG and $PM_{2.5}$ emission while offering better delivery punctuality.

Under the pressure of customer expectations for better service, low operation budgets due to market competition, and environmental goals to reduce pollution and emissions in the city, diverse solutions have already been proposed and implemented. However, the issues of urban logistics are not completely resolved. The encouraging results in this chapter indicate that the PI enabled hyperconnected logistics systems for urban fulfillment and last-mile delivery can be significant contributors to the efficiency and sustainability of smart cities and retailers serving these cities. However, the assumption of having readily available open FCs or hubs is not realistic in current cities. Also, the main difference between hyperconnected and collaborative systems is their business structure. For example, any company in need can purchase and use the fulfillment service at open FCs, whereas collaborative FCs can be accessed only by exclusive members; the use of such collaborative fulfillment is not flexible because actors typically rely on mid- or long-term agreements. However, although still in an embryonic phase, there are emerging businesses providing such open logistics services. In the fulfillment thread, especially, *ES3, Flexe.com, Fulfillment by Amazon, Warehouse Anywhere* and *Darkstore* are good examples of successful business models. The on-demand delivery or transportation services, which are implicitly assumed in the modeling, can be seen in the offerings of businesses such as *Uber Freight* and *Convoy*. The PI initiative is driven by market, but policies can be made to urge such systems, given the significant potential of the hyperconnected urban delivery systems to reduce GHG emissions and congestion. This study can provides preliminary analysis on the hyperconnected system to offer insights to potential business models and city logistics systems.

There is a rich potential for future research on hyperconnected urban logistics to address the limitations of this study. First, the impact of demand seasonality on the performance of hyperconnected systems can be studied, especially how the synchronization of seasonality between retailers affects operations, for example in terms of storage capacity requirements. Future studies may investigate the impact of seasonality on the general efficiency of hyperconnected systems. A smart way to use the open logistics resources to handle demand peaks on special days such as Thanksgiving and pre-Holidays weeks can also be explored. Second, the study can be extended to model the open fulfillment and delivery service providers' business including pricing and network optimization. The results of sensitivity analysis on total market size and the number of retailers indicate that the impact of network configuration and structural changes on cost savings is significant. In fact, the fulfillment network optimization, which is highly dependent on the characteristics of the city, has not been considered in the study as the goal is to conduct a strategic analysis for general case. However, analyzing the impact of the number of open FCs and locations of them is one of the interesting future research topics. Moreover, the fulfillment network optimization will be required to implement the hyperconnected fulfillment network in a specific city. Third, as an extension of the second point, hyperconnected systems can be contrasted to collaborative systems of various degrees and orientations. The thorough comparison can further emphasize the difference between hyperconnected and collaborative systems. Note that for this contrast, it is essential to model a large number of retailers to capture the open-to-market aspects of hyperconnected systems and to explore various collaboration possibilities. Fourth, delivery failure and second-time delivery can be considered for more accurate reflection of practices. Although it is assumed in the chapter that customers are always present at the first delivery regardless of delivery lateness, it is common in home delivery practice to visit customers multiple times due to the absence of customers at delivery. Fifth, the impact of routing algorithms could be investigated. The selection of routing algorithm could be subject to experiment, as the potential and impact of improving routing efficiency are well recognized with ample scientific literature on routing with diverse variations, yet with limited investigation of routing in hyperconnected systems. In addition, the potential opportunity of faster delivery options such as same-day and X-hour delivery of large items can be explored, which would require online dynamic routing algorithms to be run numerous times each day.

Despite the limitations of the study, the results presented in the chapter demonstrate the potential of designing, implementing and operating a hyperconnected urban logistics system, as a future urban logistics system alternative for urban fulfillment and last-mile delivery, notably for large items. It is expected that the insights obtained from this study can be among the building blocks enabling to shape an efficient, capable, and environmentally-friendly future urban logistics system.

CHAPTER 3
OPTIMAL ALLOCATION UNDER AVAILABILITY PROMISING E-CONTRACT IN DROPSHIP OPERATIONS

3.1 Introduction

Dropshipping, where e-retailers advertise products and transfer received customer orders to suppliers while suppliers (dropshippers) manage their inventories and fulfill the orders, is often adopted in e-commerce practice. This is especially beneficial when an e-retailer offers a large variety of products, and those from a supplier have relatively small sales volume at each retailer and/or require special treatment. This enables both e-retailers and suppliers to focus on their own expertise ([85]). It can reduce inventory assets at the e-retailer and improve flow and inventory consolidation at the dropshipper. Capitalizing on financial and operational advantages, many e-retailers and manufacturers have adopted dropshipping as their joint mode of operation ([86]). However, the lack of control on inventory and order fulfillment poses potential risks for the e-retailers ([87]; [88]). In fact, Cheong, Goh, and Song [89] showed that accurate inventory information in dropship operations plays a critical role not only for e-retailers but also for dropshippers. To reduce the risk, some e-retailers take a dual fulfillment strategy using both internal stock and dropshipping ([90]; [91]) or make risk-reducing contracts with dropshippers, notably with service level agreement terms ([92]).

In this chapter, we focus on an alternative approach based on availability promising contracts (APCs). APC is driven by dropshippers unlike the typical service level agreements which mainly impose e-retailer's requirements to dropshippers ([92]). Through an APC with each e-retailer, the dropshipper promises a certain quantity of each product to be available for fulfilling the e-retailer's customers. While the APC is enforced, it limits the maximum number of products ordered by each e-retailer to the promised quantity stated in the APC. In practice, it is possible to renew APCs on a daily basis or even more often, allowing to adapt promises based on demand and supply fluctuations. On one side, the dropshipper can promise through APCs more than what is actually available (overpromising) as these e-contracts do not involve immediate physical transactions. Although the total promised quantity across all e-retailers may exceed the available quantity with overpromising, APCs can be robustly upheld for each e-retailer by leveraging demand uncertainty pooling. On the other side, each e-retailer can enforce a promised availability threshold (PAT). An e-retailer may unlist any product from its online platform or mark it as 'out-of-stock', when the remaining promised quantity falls below this PAT. Each e-retailer decides whether to enforce PATs for the contracted products to protect their service level against potential fulfillment failure due to overpromising. The PAT value is typically not shared with dropshippers.

The optimal allocation problem under APC addressed in this chapter is to determine the quantity of each product to be promised to each e-retailer so as to maximize expected profit of the dropshipper, given its product availability, uncertain demand and unknown PATs. In other words, setting the APC quantity for each product is an 'ex-ante' activity performed when demand is not yet realized, unlike the typical inventory allocation that is modeled as an 'ex-post' activity performed once demand is realized. The structure of the APC design problem fits with two-stage stochastic programming framework ([93]; [94]), as the daily renewing APC is made in the morning once a day (first stage) and orders are received throughout the day and fulfillment decisions are made at the end of the day according to the realization of demand and threshold which were unknown in the morning (second stage). Our study stems from a partnership with a dropship furniture manufacturer in the North American market who is continually facing such an availability promising contract design problem with some of its e-retailers exploiting hidden PATs. The schematic description of the APC problem is described in Figure 3.1. We focus on a daily-renewing APC problem under the existence of PATs. The problem is naturally decomposed into independent subproblems by product as each PAT is product-specific and the dropshippper's products are unsubstitutable.

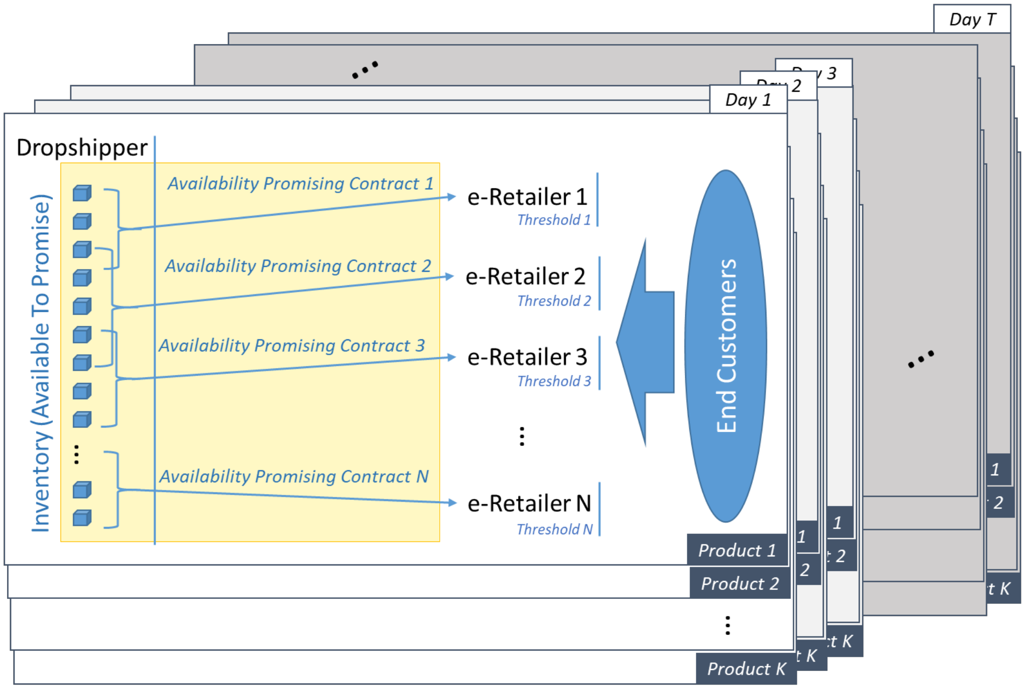

Day T
Day 3
Day 2
Day 1
Dropshipper
Inventory (Available To Promise)
Availability Promising Contract 1
Availability Promising Contract 2
Availability Promising Contract 3
Availability Promising Contract N
e-Retailer 1
Threshold 1
e-Retailer 2
Threshold 2
e-Retailer 3
Threshold 3
e-Retailer N
Threshold N
End Customers
Product 1
Product 2
Product K

Figure 3.1: Illustration of daily availability promising contract

The essence of the problem is here described using a simple example. Assume a dropship manufacturer has 100 units of a type-x table on hand when setting its APC for two e-retailers A and B at the beginning of a day. The two APCs respectively promise 80 units to e-retailer A and 30 units to e-retailer B. The type-x tables are sold throughout the day and at 3pm, A sells its 60^{th} table and suddenly the product is marked 'out-of-stock' on e-retailer A's site. No more customer order is accepted on that day by A although there still remains 20 units promised to A. Meanwhile, the manufacturer sold 15 units via B and the product has remained available until the end of the day. Observing the patterns over time, the dropship manufacturer can guess that the e-retailer A has a PAT of value between 15 and 22 for the type-x tables, which may be dynamically changing daily, and that e-retailer B has no PAT. Now assume that the actual demand at A and B are 75 and 15 units respectively that day, and the PAT of A that day is known to be 20. If the demand and PAT have been known in advance, the manufacturer could have promised 100 units to A instead of 80 units knowing that the last 20 units won't be sold. Such APC would lead to the maximum sales, 75 units and 15 units via A and B respectively. Yet in reality, the manufacturer usually does not know with certainty either the demand or PAT of each e-retailer, making the APC decisions complex.

Accordingly, the contributions of the chapter are (1) to investigate a new type of contract (APC), faced by a dropshipper under the existence of unknown retailer-specific thresholds (PAT), (2) to provide 2-stage stochastic optimization models which build on a linear formulation of endogenous uncertainty from substitution behavior and on closed-form stockout probability bounds and (3) to deliver managerial insights on the performance of three contract policies, guaranteed fulfillment, controlled fillrate and penalty-driven fillrate policies, and the impact of PAT on both dropshipper and e-retailer.

The chapter is structured as follows. In section 3.2, relevant literature is analyzed so as to clearly position the original contributions of the chapter. Then the problem structure and components are described in section 3.3, followed by section 3.4 which describes the modeling and solution approaches. Then, computational results from a case study and sensitivity analysis are presented in section 3.5. Lastly, the chapter concludes with section 3.6. Throughout the chapter, we use the term retailer instead of e-retailer for simplicity.

3.2 Literature Review

Dropshipping became a common practice with the growth of e-commerce. It has been studied for more than a decade ([95]; [85]) and is still an ongoing research topic. One of the main research streams in recent years is to model the retailer's order fulfillment planning leveraging the dropshipping option ([96]; [97]), and to complementarily model the retailer's inventory planning problem when dropshipping is an option in addition to in-house inventory ([98]; [87]; [99]; [100]). There are relatively less studies focusing on the dropshipper's perspective. Peinkofer, Esper, Smith, and Williams [101] and Dennis, Cheong, and Sun [102] studied the supplier's capability and profitability of expanding to dropship operations. The availability promising contract (APC) is a new problem that is dropshipper focused. APC is distinguished from previously studied dropship service level agreements ([87]). Broadly, the APC problem belongs to a category of inventory allocation problems, which will be reviewed first in this section. The information asymmetry in the dropship setting allows overpromising and this, in turn, makes some retailers adopt promised availability thresholds (PATs). Therefore, the relative literature for each aspect will be reviewed, following the inventory allocation literature review. Lastly, we review recent pertinent literature on 2-stage stochastic optimization.

The inventory allocation problem, which is part of a broader category of resource allocation problem, arises when an inventory owner deals with multiple retailers or sales channels to whom it needs to allocate available inventory, which is to bound sales through each retailer or channel. This is well-studied problem, but the term 'inventory allocation' is used for different classes of problems. It may refer to an inventory positioning problem, notably in a multi-echelon setting ([103]). It can refer to an order allocation problem which matches inventories and capacities to realized demand ([104]) or received orders from downstream players which may have been exaggerated to gain more allocation ([105]), which can also happen in a dropshipping context ([106]). When demand is not observed directly by a supplier, it is important to understand the value of information shared between retailer and supplier ([107]). Inventory rationing is one of the closely related inventory allocation problems, occurring when there are multiple customer classes and limited inventory availability ([108]; [92]; [109]).

Typical objectives of inventory allocation problems are related to maximizing profit, notably through minimizing lost sales ([110]) or backorder costs ([111]). The retailers or channels may have different preferences or margins ([112]), or require different service levels ([113]). In a multi-period setting, order acceptance decisions in the current period are affected by future orders and availability ([114]). The APC problem is defined as a single-period expected profit maximization problem. It deals with multiple retailers and allocates inventory to them optimally to maximize expected profit as in most inventory allocation problems. However, the combination of relying on dropship operations and contracting on promised availability prior to demand instantiation, instead of committing inventory to realized demand, creates critical differences between APC and traditional inventory allocation problems. As promising availability does not involve physical transactions, a dropship manufacturer can promise more than what is actually available (overpromising). This, combined with the facts that APC is dropshipper-driven and that there is competition between retailers, leads some retailers to adopt PAT. Also, unlike most of inventory allocation where allocation decision is made to fulfill realized demand ('ex-post' activity), APC problem promises inventory before demand realization ('ex-ante' activity). Such problem characteristics require optimization under uncertainty, which is commonly tackled using stochastic programming ([94]).

Stochastic optimization is a well-established methodology and widely used to solve supply chain and logistics problems ([93], [94]). Pre-allocating resources in the first stage and optimally using them in the second stage upon realization of demand is one of the most common problems formulated as a stochastic optimization model ([115], [116]). Another common first-stage decision is resource planning such as production capacity ([117]). Stochastic optimization is also largely employed in network design ([118], [119]), distribution network deployment ([120], [121]), or asset allocation ([122]) problems. The quality of the solutions produced by stochastic location-allocation models compared to their deterministic counterparts is discussed in Klibi, Martel, and Guitouni [123]. One of the most common approaches to solve a stochastic model is the sample average approximation (SAA) ([124]; [125]). SAA uses N scenarios, each corresponds to a specific realizations of the stochastic parameters, with probability $\frac{1}{N}$, which converts the original stochastic problem to an equivalent deterministic problem. Choosing proper N to balance solution optimality and computational effort is important. Kleywegt, Shapiro, and Homem-de-Mello [125] presented an algorithm to calculate the performance gap for a SAA solution and to determine a proper N. Dupačová, Gröwe-Kuska, and Römisch [126] demonstrated how to reduce the optimality gap by sampling a good quality of scenarios. As indicated earlier, stochastic optimization fits well with the APC design problem structure and inherent uncertainty, justifying its use in this chapter.

As stressed earlier, APC is a virtual commitment of availability. One of the major advantages of dropshipping and APC on inventory management is that the physical inventories remain pooled until the point of order fulfillment ([91]). Such a setting creates a main differentiation in optimal promising strategy, as it enables overpromising. In fact, overpromising (related to overallocation or overbooking) is commonly found in revenue management. For example, overbooking has been practiced in airline and lodging industry where firms try to maximize expected revenue considering the potential no-shows and overselling that respectively lead to empty seats and boarding denial ([127]; [128]). An extension of such problems includes minimizing penalty via substitutable resources ([129]). The problem of overbooking has also been applied to patient scheduling ([130]). Such overbooking operations can increase potential profit but can also result in order decline causing explicit penalties charged by retailers or indirect penalties induced from disappointed customers ([131]). Although APC shares the idea of overallocation with these cited problems, the risk structure is different. Overbooking hedges against potential order cancellation while overpromising hedges against retail demand and retailer PAT uncertainty. The promised quantity, combined with retailer substitution behavior of customers, affects the demand distribution by limiting the maximum number of product units to be ordered. Also, while the overbooking decision is made sequentially as customers make reservations in traditional booking models, the APC decision is made once and only once to all retailers for the target sales period. Lastly, APC does not change product margin or price, while dynamic pricing is often used along with overbooking in revenue management ([132]).

Another unique aspect of APC is the product availability thresholds (PATs) set by retailers for each product. PAT is a new concept and has not been presented in the existing literature at the best of the author's knowledge. Accurate inventory information is important to both retailers and supplier in dropship operations ([89]), but more critical to retailers as inventory is only indirectly observed by them via APCs which may have been overpromised. Typical retailer strategies for protecting their service level consist of carrying some owned inventory in addition to dropshipping ([90]; [91]), contracting to multiple dropshippers ([114]), or penalizing dropshipping company for poor fulfillment performance. However, the PAT strategy has never been studied. In this chapter, we firstly aim to design optimal APCs under PATs. Secondly, we aim to evaluate the effectiveness of PAT to retailers through dropship manufacturer's APC decisions under varying PATs.

3.3 Problem Description and Modeling

As described, the dropshipper promises available inventory quantities to retailers through daily availability promising contracts (APCs) knowing that some retailers have promised availability thresholds (PATs) that, when reached, stop these retailers from accepting further customer orders. It is assumed that the location of inventory is not part of the contract. In short, the APC problem is a single-period (day, hour, etc.), single-product, and single-location problem. The dropshipper makes the decision to maximize expected profit based on demand forecast and PAT estimation.

Figure 3.2 shows the structure of the APC design problem by describing information and decision flows between the three main players: dropshipper, e-retailers and customers. The forecast and estimation of uncertain information at the point of dropshipper's APC decision are shown in blue. Information flow between players are numbered by the chronological order of events. The modeling of each important element of the problem shown in Figure 3.2.

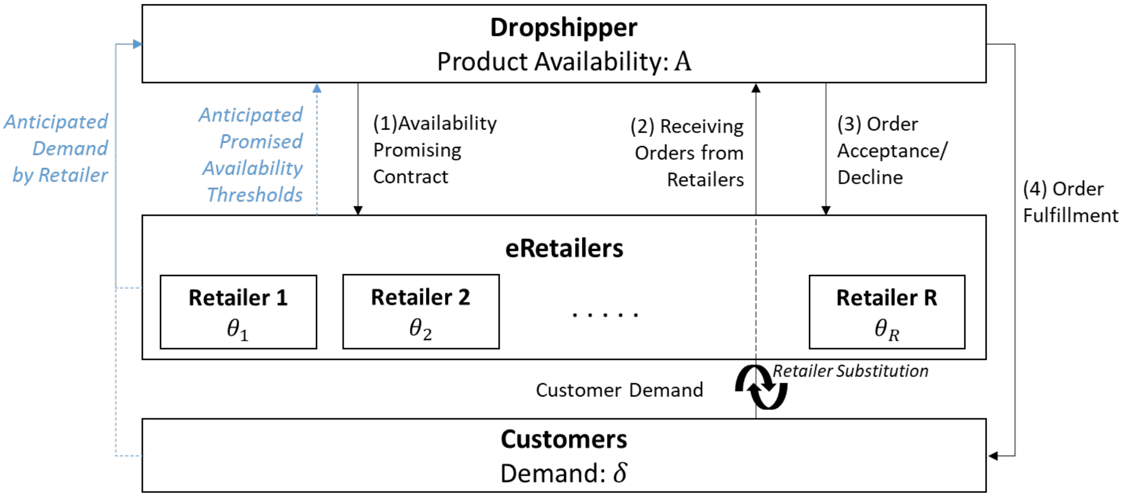


Figure 3.2: Structure of the inventory promising contract problem

3.3.1 Availability promising contract (APC)

The availability promising contract (APC) is made on a periodic basis (e.g. daily) in anticipation of unknown demand and promised availability thresholds (PATs) for the period. Note that APC is unilateral: based on its forecasts, the dropshipper decides the contract quantity ($z_r\ \forall r$) and transmits the contract to the retailers. Dropshippers exploiting APCs have to face customer demand rather closely as compared to traditional setting where vendors allocate available inventory based on retailer orders which may have been exaggerated. In other words, the dropshipper observe true customer orders, not the potentially exaggerated orders of retailers who try to gain priority over other retailers. Since online customers typically visit multiple websites before purchasing a product while exhibiting some preferences among the online retailers, the dropshipper also needs to understand how its APC decisions affect customer behavior and its sales. Under dropshipping and APC, retailers only received the digitally contracted quantity and do not observe the dropshipper's availability directly. Such information asymmetry allows the dropshipper to pool availability for multiple retailers via overpromising. Overpromising has the potential to increase profit as sales are limited by the promised quantities, not necessarily by actual available inventory quantities. On the other hand, this is the reason why some retailers set PATs to protect themselves against potential order fulfillment failure caused by overpromising. Order rejection or expedited production, in case of orders exceeding availability due to overpromising, create extra cost for the dropshipper and also harm the relationship with retailers and its reputation to customers. This indicates that setting an optimal contract policy is neither trivial nor insignificant.

In addition to profit maximization, the dropshipper also needs to take its relationship with retailers into consideration when designing APCs. For example, the dropshipper may promise availability only to retailer 1, and none to retailer 2, so as to maximize the expected profit in the current period. This may harm its long-term profit potential with retailer 2, as this retailer may switch to consider the dropshipper as uncommitted and/or unreliable. Such myopic decision may put the profitability of the dropshipper at risk, as well as its collaboration with retailers and/or channel diversification strategies. To prevent this, a dropshipper may set an internal policy forcing minimum and maximum contract quantities or minimum and maximum fractions of the total available inventory (l_r, u_r) to be promised to each retailer r.

The three most critical inputs for APC design are the available-to-promise (ATP) quantity, the retailer specific PAT estimates, and the overall and retailer-specific demand forecasts. The ATP level is deterministic data when the dropshipper relies on on-hand inventory, yet may have some degree of uncertainty when relying on current production and/or in-transit supplies. On the other hand, retailer thresholds and demand are in most cases unknown and the accuracy of their estimation is critical to APC design. Each input is described in detail here, assuming a single-day period for descriptive simplicity.

Product Availability

The APC is made based on a given availability level (or available-to-promise level) at the start of the day (A) ([133]; [134]). In general, the availability level is considered low or high with respect to replenishment lead time. However, for a daily APC, we categorize availability levels with respect to the daily demand forecast. For example, F, $\overline{F}$, $\underline{F}$, $\overline{\overline{F}}$ and $\underline{\underline{F}}$ in Figure 3.3 can be the most likely demand forecast, $F+\sigma$, $F-\sigma$, $F+3\sigma$, and $F-3\sigma$ respectively and availability zones are defined accordingly. The exact criteria values or number of zones can vary based on the estimated probability distribution or company policy. Five zones are used for the categorization as shown in Figure 3.3: scarce, low, lean, high, and overfull.

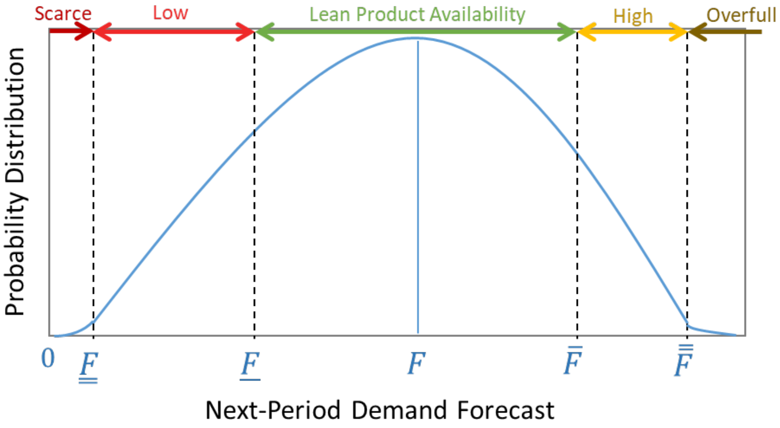

Scarce
Low
Lean Product Availability
High
Overfull
Probability Distribution
0
$\underline{\underline{F}}$
$\underline{F}$
F
$\overline{F}$
$\overline{\overline{F}}$
Next-Period Demand Forecast

Figure 3.3: Illustration of availability zone definition with respect to daily demand forecast

As mentioned earlier, the dropshipper has the ability to pool via overpromising. However, the impact of overpromising may depend on the daily availability level, which to be shown via experimental results.

Promised Availability Threshold (PAT)

Let θ_r denote the promised availability threshold (PAT) of retailer r. To protect their service level and potential stockout under such unilateral contracts with no information on true availability levels, some retailers set internal threshold, PAT, (θ_r). There are two intuitive ways to model PAT: absolute value and fraction of promised quantity, z_r. For example, in the first case, the retailer cut off sales once remaining promised quantity reaches θ_r. In the second case, the retailer cut off sales once $1-\theta_r$ of promised quantity is sold, which makes the absolute cut-off threshold level endogenous to allocation, $(1-\theta_r)z_r$. In the context our study, the first PAT definition is chosen. Moreover, when a large (small) amount is promised it is more likely for the dropshipper's availability is in high or overfull (low or scarce) zone, which makes the first definition more reasonable than the second. Yet more complex models can be used such as defining PAT as a step-wise stochastic function of promised quantity. However, because the focus of this chapter is on designing optimal allocation for a dropshipper under APC with PAT, we assumed the PAT will be determined as a simple stochastic value in a given period. Another assumption on the characteristics of PAT made in this study is that PAT value of a given period is only determined by historical APCs and actual fulfillment records. For example, if an retailer observes that the dropshipper tended to decline 10 to 15 customer orders per period, it may set its PAT between 10 to 20 for the given period regardless of how many units are promised at the moment. For the case study presented in this chapter, we estimated retailer PATs from historical APC and sales data and a few direct observations on the availability on retailer's website. Note that we do not have direct observations or deterministic information of retailer PAT for the case study.

Given the definition, when remaining promised availability level falls below the PAT for a product, they remove the product from their online sales platform or mark the product out-of-stock. In case of positive PAT, the effective promised quantity ($z_r - \theta_r$) will be smaller than nominal promised quantity (z_r). The PAT values (θ_r) are not shared with the dropshipper and potentially changing everyday, so the dropshipper needs to calibrate its APC based on the uncertain threshold. It is critical for the dropshipper to understand the PAT because it affects the maximum sales directly. Also, when the PAT is estimated accurately, the dropshipper can overpromise by PAT quantity without risk of stockout knowing that the additional quantity will never be sold (i.e. overpromising at risk of thresholds). However, since the PAT value is not shared with the dropshipper, in practice the dropshipper typically has to rely on its estimation of a plausible range of the PAT ($\underline{\theta_r}, \overline{\theta_r}$) based on its historical observations.

An important observation about PAT is that when the PAT θ_r is known by the dropshipper, there is an optimal promising quantity corresponding to the optimal quantity when $\theta_r = 0$. Proposition 1 presents it analytically and demonstrates it.

Proposition 1. *If θ_r is known, then the optimal promising quantity is $z_r^* = z_r^0 + \theta_r$ where z_r^0 is an optimal promising quantity when $\theta_r = 0$.*

Proof. Let the demand to retailer r is δ_r and let y_r^0 be the number of orders received through r when z_r^0 is promised and $\theta_r = 0$. Then, $y_r^0 = min(z_r^0, \delta_r)$. Now, assume the threshold of the retailer r, θ_r, is known to the dropshipper and the corresponding promising quantity is $z_r^* = z_r^0 + \theta_r$. That is, the effective promising quantity is $z_r^* - \theta_r = z_r^0$ and resulting number of orders is $y_r^* = min(z_r^* - \theta_r, \delta_r) = min(z_r^0, \delta_r) = y_r^0$. Note that the optimal profit with a threshold cannot be better than the optimal profit when $\theta_r = 0 \; \forall r \in R$ (no threshold). Therefore, $z_r^* = z_r^0 + \theta_r$ is the optimal promising quantity under known threshold θ_r. □

Customer Demand

Demand forecast is one of the critical stochastic inputs to APC along with retailer PAT estimation. Recall the two dotted blue arrows in Figure 3.2. Demand forecasting is a common practice in industry and various forecasting tools are used, ranging from simple extrapolation and exponential smoothing methods to state-of-the-art machine learning models, at varying time horizon and aggregation level ([135]). Most of the supply chain, production and inventory related forecasts are mid- or long-term demand forecasts, e.g. over a lead time, quarterly, or yearly. For daily APC, next-day demand forecasts are required. Another complexity is to estimate demand including lost sales which is not observable. Simply estimating lost demand due to stockout is already a significant challenge in demand forecasting as Ferreira, Lee, and Simchi-Levi [136] stressed. Estimating lost demand under the existence of unknown retailer PAT and partially pooled demand adds extra complexity. Appropriate forecasting method must be selected that performs the best for the target case, allowing to provide point forecasts coupled with prediction intervals and probability distributions. The forecasting method used for our case study is described in subsection 3.4.3 in detail.

APC problem requires next day demand forecast δ_r for the contracted product at each retailer r. However, as mentioned, demands are partially pooled between retailers since customers visit multiple e-retailers at the same time, with almost no extra cost. In fact, demand dependency between channels ([137]) or products ([138]) is common, causing significant challenges in e-commerce demand forecasting and fulfillment modeling. Majority of customers also have a preferred retailer and only order from other retailers if there is no availability at preferred retailer. Such partially pooled demand structures impose significant complexity to the model by affecting customer choice of retailers with APC. In this chapter, we circumvent such endogenous uncertainty by modeling customer's retailer substitution behavior as described in subsubsection 3.3.3.

3.3.2 Contract Policy

We examine three contract policies differentiated by overpromising allowance: guaranteed fulfillment, controlled fillrate, and penalty-driven fillrate. The two extremes are not allowing overpromising at all (guaranteed fulfillment) and allowing full flexibility to overpromise but at stockout penalty (penalty-driven fillrate). Controlled fillrate policy, which limits maximum overpromising quantity based on retailer threshold uncertainty, is a policy designed specific to the business context.

Recall that there are two stochastic input parameters: demand (δ) and threshold (θ). That is, when overpromising is allowed to take advantage of inventory pooling effect, there are two sources of potential stockout: (1) The dropshipper overpromised assuming certain threshold levels may face stockout when actual thresholds were smaller than expected and, as a results, it received more orders through some retailers than expected maximum. (2) The dropshipper overpromised assuming the total maximum demand of the day (D) would be $D \leq A$ may face stockout if actual total demand was larger than availability A. This means that the dropshipper can control the probability of stockout with respect to threshold uncertainty, demand uncertainty, or both. The controlled fillrate policy focuses on limiting overpromising quantity to control the probability of stockout with respect to threshold uncertainty.

Let x_r be the internal pre-allocation quantity to retailer r where $\sum_{r \in R} x_r \leq A$ and o_r be overpromised quantity to retailer r which makes contracted quantity $z_r = x_r + o_r$. Then, along with demand δ_r, the effective contracted quantity ($[z_r - \theta_r]_+$) sets the upper bound of orders received via retailer r, y_r: $y_r = min([z_r - \theta_r]_+, \delta_r)$. Note that since $\sum_{r \in R} x_r \leq A$, the stockout can only occur if and only if there exists r overpromised quantity exceeds threshold level.

$$P[stockout] = P[\sum_{r \in R} y_r > A] \leq P[\max_r(o_r - \theta_r) > 0] \tag{3.1}$$

When distributional information on threshold is available, we can control the probability of stockout by setting upper bounds for o_r with respect to θ_r. Let $\Theta_{\alpha,r} = argmax\{\Theta : P[\theta_r < \Theta] \leq \alpha\}$. Constraining upper bound of o_r with $\Theta_{\alpha,r}$ would ensure stockout probability to be no more than α. Proposition 2 proves this formulation is correct and conservative.

Proposition 2. *If $o_r \leq \Theta_{\alpha,r}$ $\forall r \in R$, then $P[stockout] = P[\sum_{r \in R} y_r \geq A] \leq \alpha$.*

Proof. Assume a sufficiently large demand. That is, sales is bounded by allocation and inventory availability, not by demand. Then, the probability of stockout is bounded as shown in (Equation 3.1). If $o_r \leq \Theta_{\alpha,r}$,

$$\begin{aligned} P[stockout] &= P[\sum_{r \in R} y_r > A] \\ &\leq P[\max_r(o_r - \theta_r) > 0] \\ &\leq P[\theta_{r^*} < o_{r^*}] \\ &= P[\theta_{r^*} < o_{r^*} \leq \Theta_{\alpha,r^*}] \end{aligned}$$

$$\leq P[\theta_{r^*} < \Theta_{\alpha,r^*}]$$

$$\leq \alpha$$

where $r^* = argmax_r(o_r - \theta_r)$. Due to the assumption of large demand, this is a conservative upper bound for stockout probability. □

3.3.3 Order Characterization

It is important to understand online sales structure and customer behaviors to characterize the correlated demand. As described in the previous section, APC and PAT affect order quantities via availability. In the mean time, they also affect customer behaviors, especially supplier and retailer substitution. To properly structure the order capturing process in the given business context, the impact and interaction of these factors must be clarified. Figure 3.4 shows different stages from demand to sales with critical factors. This section characterizes order quantity in terms of demand, APC, and substitutions.

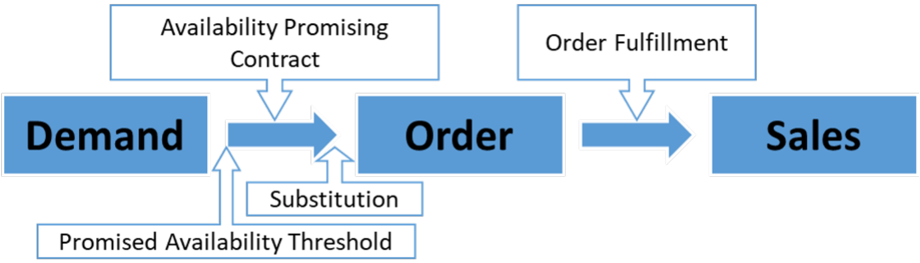

Availability Promising Contract
Order Fulfillment
Demand
Order
Sales
Substitution
Promised Availability Threshold

Figure 3.4: Order capturing and sales realization process

Note that sales realization from given order (order fulfillment) is relatively simple in a single period problem: sort orders by margin and fulfill up to available inventory. However, in a multi period setting, the fulfillment decision is not as simple since future profits and long term relationship with retailers must be taken into account at current order fulfillment decision.

Substitution

When there is no available inventory at their preferred or primary e-retailer, customers may buy a similar product of another dropshipper or supplier (supplier substitution), a similar product from the same dropshipper (product substitution), or the same product from another e-retailer with available inventory (retailer substitution). Customers may decide not to buy anything, but it is not different to supplier substitution for the dropshipper. Both are counted as lost opportunities. Moreover, since we focus on the single-product problem, product substitution is here counted as lost demand as well. This only leave retailer substitution or lost demand to the APC modeling.

Such substitution behavior can be explained with the brand power of the dropshipper, the product, and of the retailer. If the dropshipper's and/or the product's brand power is more critical, customers will be inclined to buy the product from any retailer who has available inventory because customers want 'the product'. On the other hand, if the retailer's brand power is more critical, then customers prefer to buy a similar product from the dropshipper or from other dropshippers available at their primary retailer.

When dropshipper brand power is dominant, the contract quantity becomes less critical as total profit will be determined mostly by total demand. This can happen when the dropshipper is actually a dropship manufacturer, not a vendor. When retailer brand power is dominant, the APC problem becomes simpler as demand at each retailer becomes rather independent. The APC problem becomes complex when both brand powers exist as it creates endogenous uncertainty: the demand is affected by contract quantity since customers may, but not always, switch to another retailer who has availability in case of unavailability at the primary retailer. It is known that stochastic optimization with endogenous uncertainty is very hard to solve ([139]).

Alternatively, we embed the endogenous uncertainty in the model by introducing retailer substitution ratio. In fact, modeling substitution with deterministic ratio is common approach ([87]). Note that, to model the endogeneity, the substitution ratio is applied to realized demand and inventory availability status, not to the demand rate. Substitution can be modeled as customer's have preference hierarchy from which purchase decision is made based on availability ([140]), but here we model the substitution as it happens with a certain probability when the product is unavailable at the most preferred retailer, yet only once: e.g. if there is no availability at the first substituted retailer, then customers will be lost without further substitution. Assume the demand created initially through retailer r, δ_r, cannot be fully captured by r due to limited contract quantity $z_r < \delta_r$. When the substitution ratio from r to r' is $\gamma_{rr'}$, we assume $\gamma_{rr'}$ fraction of unsatisfied demand, $\gamma_{rr'}(\delta_r - z_r)$, will be captured through retailer r'. γ_{rr} indicates the fraction lost. The substitution process is described in Figure 3.5 for 2-retailer case. Note that the substitution process is modeled as linear equations and it can be easily expanded to more than 2 retailer cases.

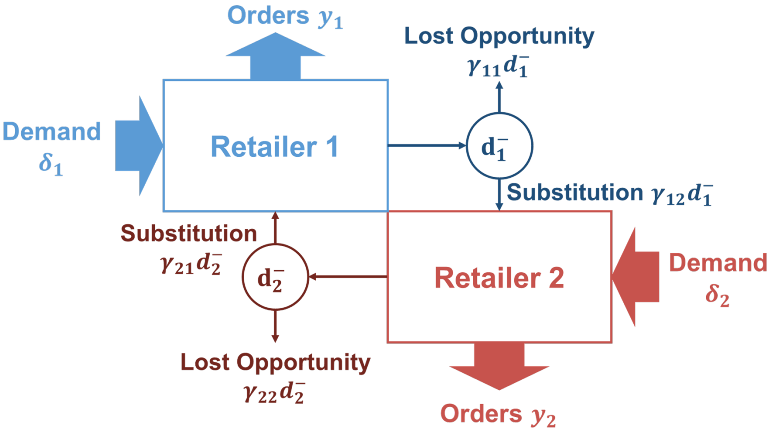


Figure 3.5: Description of demand and orders with retailer substitution

Order Characterization with Effective Contracted Quantity and Demand

Orders realized through retailer r, denoted by y_r, are limited by effective contracted quantity and demand. Effective contracted quantity is expressed as contracted quantity z_r subtracted by retailer threshold θ_r, $(z_r - \theta_r)$, as retailers do not accept any more orders once the remaining promised quantity reaches to the threshold. When customers are dedicated to a specific retailer (retailer brand power is dominant), order quantity can be expressed as below where δ_r represents the demand primarily captured by retailer r:

$$y_r = min\{\delta_r, [z_r - \theta_r]_+\} \tag{3.2}$$

With substitution, the demand is not independent by retailer anymore. Let effective demand be the sum of primary demand (δ_r) and substituted demand captured by a retailer r. Substituted demand is defined as a fraction of unsatisfied primary demand (d^-) as shown in Figure 3.5. Unsatisfied demand from a retailer r is expressed as below:

$$d_r^- = max\{0, \delta_r - [z_r - \theta_r]_+\} = [\delta_r - [z_r - \theta_r]_+]_+ \tag{3.3}$$

The substitution ratio $\gamma_{rr'}$ indicates the fraction of d_r^- switched to retailer r' where γ_{rr} is the fraction of d_r^- lost. Now, the order quantity can be expressed as the minimum between effective demand and contracted quantity:

$$y_r = min\{\delta_r + \sum_{r \neq r' \in R} \gamma_{r'r} d_{r'}^-, [z_r - \theta_r]_+\} \tag{3.4}$$

3.4 Models and Solution Approach

The ultimate goal of the chapter is to solve the dropshipper's APC problem addressing unknown PATs using a stochastic optimization that maximizes expected profit of a dropshipper based on demand forecast. In this section, we describe our model and solution approach. The overall scheme is described in Figure 3.6. As can be seen from the figure, there are three main dimensions for modeling and solution structure. First is alternative contract policy design which is closely related to managerial insight derivation. Second is uncertainty modeling. Scenarios are generated from it to solve the model using sample average approximation. Last is solution evaluation. The solution from stochastic optimization model is evaluated with respect to the solutions from comparison models. The section starts with the 2-stage stochastic optimization model and describes contract policy variations, followed by sample average approximation (SAA) approach. Then, uncertainty modeling and solution evaluation scheme are described.

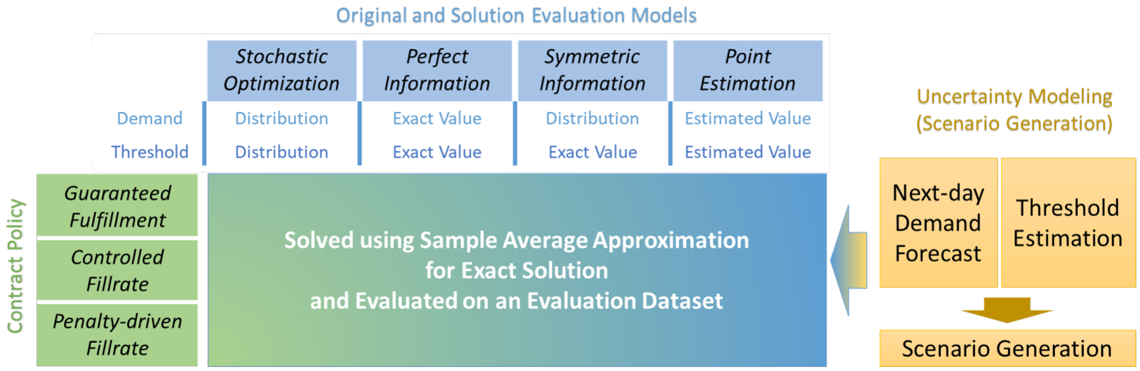


Figure 3.6: Model variation and solution structuring

3.4.1 Two-Stage Availability Promising Contract (APC) Model

The most general APC problem modeled as a two-stage stochastic problem is presented here which corresponds to penalty-driven fillrate contract policy. The APC is made in the first stage only knowing distribution of not-yet realized uncertain variables, so as to maximize expected profit. Retailer PAT values are revealed in the second stage. Demand is also realized in the second stage given the contract and PAT and received orders are accepted or declined so that the profit is maximized. This mimics the actual business procedure: APC is made in the beginning of the business day based on demand forecast and PAT estimation and the fulfillment decision is made at the end of business day given received orders. Notations used are listed below:

First stage variables

- z_r: Contracted(promised) quantity to retailer r via APC

Second stage variables

- y_r: Order quantity received through retailer r

- y_r^+: Accepted order quantity (sales) through retailer r
- y_r^-: Declined order quantity through retailer r due to shortage at dropshipper
- d_r^-: Demand lost by preferred retailer r due to unavailability at the retailer
- $d_{r,r'}^-$: Demand lost by retailer r and substituted with retailer r'; indicate lost demand without substitution when $r = r'$
- b_r: Binary variable indicating $d_r^- = 0$ or not ($b_r = 1$ if $\delta_r \leq z_r - \theta_r$ and $b_r = 0$ if $\delta_r > z_r - \theta_r$)
- v_r: Binary indicator variable ($v_r = 1$ if $z_r - \theta_r < 0$ and $v_r = 0$ otherwise)
- w_r: Binary variable indicating order quantity y_r is bounded by contracted quantity ($w_r = 0$) or by demand ($w_r = 1$)

Stochastic Parameters

- δ_r: Daily demand of product p with a preference in retailer r
  - $f_r(\delta)$: PMF/PDF of demand δ_r
  - $F_r(\delta)$: Distribution function of demand δ_r
- θ_r: Promised availability threshold of retailer r
  - $g_r(\theta_r)$: PMF/PDF of θ_r
  - $G_r(\theta_r)$: Distribution function of θ_r

Deterministic Parameters

- c_r: Unit sales margin of product p sold through retailer r
- π_r: Unit out-of-stock penalty associated to retailer r
- tl_r, tu_r: Estimated lower and upper bound of θ_r

- A: The available quantity of product p at dropshipper

- l_r, u_r: Minimum/maximum fraction of total available-to-promise that can be promised to retailer r (dropshipper's policy parameters)

- $\gamma_{rr'}$: Average fraction of customers originally reached through retailer r order through retailer r' when the product is not available at r

  - γ_{rr} represents the average fraction of customers lost when the product is not available at r, $\gamma_{rr} = 1 - \sum_{r \neq r' \in R} \gamma_{rr'}$ $\forall r \in R$

- M: Sufficiently large number (auxiliary parameter for model linearization)

General APC Model: Penalty-Driven Contract Policy

Here we present a general 2-stage optimization model under stochastic demand and threshold. Following Model (Equation 3.5) is a stochastic APC model with penalty-driven contract policy.

$$\max: \; E[Q(z, \delta, \theta)] \tag{3.5.1}$$

s.t. Min-Max Contract Quantities

$$z_r \leq u_r A \quad, \forall r \in R \tag{3.5.2}$$

$$z_r \geq l_r A \quad, \forall r \in R \tag{3.5.3}$$

$$z \in \mathbb{Z}_+ \tag{3.5.4}$$

$$where \quad Q(z, \delta, \theta) = \max \sum_{\forall r \in R} (c_r y_r^+ - \pi_r y_r^-) \tag{3.5.5}$$

s.t. Order Acceptance

$$\sum_{r \in R} y_r^+ \leq A \tag{3.5.6}$$

$$y_r = y_r^+ + y_r^- \quad, \forall r \in R \tag{3.5.7}$$

Order Characterization

$$y_r \leq z_r - \theta_r + Mv_r \quad ,\forall r \in R \tag{3.5.8}$$

$$y_r \leq \delta_r + \sum_{r \neq r' \in R} d^-_{r'r} \quad ,\forall r \in R \tag{3.5.9}$$

$$y_r \leq M(1 - v_r) \quad ,\forall r \in R \tag{3.5.10}$$

$$y_r \geq z_r - \theta_r - Mw_r - Mv_r \quad ,\forall r \in R \tag{3.5.11}$$

$$y_r \geq \delta_r + \sum_{r \neq r' \in R} d^-_{r'r} - M(1 - w_r) - Mv_r \quad ,\forall r \in R \tag{3.5.12}$$

Retailer Substitution

$$\sum_{r' \in R} d^-_{rr'} = d^-_r \quad ,\forall r \in R \tag{3.5.13}$$

$$d^-_{rr'} \leq \gamma_{rr'} d^-_r \quad ,\forall r, r' \in R, r \neq r' \tag{3.5.14}$$

$$d^-_{rr'} >= \gamma_{rr'} d^-_r - 1 + 1/M \quad ,\forall r, r' \in R, r \neq r' \tag{3.5.15}$$

Lost Customer

$$d^-_r \leq \delta_r + M(1 - v_r) \quad ,\forall r \in R \tag{3.5.16}$$

$$d^-_r \geq \delta_r - M(1 - v_r) \quad ,\forall r \in R \tag{3.5.17}$$

$$d^-_r \leq \delta_r - (z_r - \theta_r) + Mv_r + Mb_r \quad ,\forall r \in R \tag{3.5.18}$$

$$d^-_r \geq \delta_r - (z_r - \theta_r) - Mv_r - Mb_r \quad ,\forall r \in R \tag{3.5.19}$$

$$d^-_r \leq M(v_r + (1 - b_r)) \quad ,\forall r \in R \tag{3.5.20}$$

Control Binary Variables

$$b_r \geq \frac{(z_r - \theta_r) - \delta_r + 1}{M} \quad ,\forall r \in R \tag{3.5.21}$$

$$b_r \leq 1 + \frac{(z_r - \theta_r) - \delta_r}{M} \quad ,\forall r \in R \tag{3.5.22}$$

$$v_r \geq \frac{-(z_r - \theta_r)}{M} \quad ,\forall r \in R \tag{3.5.23}$$

$$(1 - v_r) \geq \frac{z_r - \theta_r}{M} \quad ,\forall r \in R \tag{3.5.24}$$

$$y, y^+, y^-, d^- \in \mathbb{Z}_+ \text{ and } b, v, w \in \{0, 1\} \tag{3.5.25}$$

First stage objective (Equation 3.5.1) is to make the APCs to maximize an expected margin with respect to δ and θ. Constraints (Equation 3.5.2) and (Equation 3.5.3) represent a dropshipper's policy on minimum and maximum promising quantity to each retailer. We will name them as 'business relationship constraints'. For general purpose, these constraints can be ignored. Promised quantities must be nonnegative (Equation 3.5.4).

Second stage objective (Equation 3.5.5) is to fulfill to maximize profit at a certain realization of demand and threshold. The profit is expressed as the sum of sales margin subtracted by stockout penalty.

Order acceptance constraints (Equation 3.5.6) and (Equation 3.5.7) models sales and order acceptance. The dropshipper can decline (y_r^-) or accept (y_r^+) customer orders (y_r) but cannot accept more than available inventory.

The total order quantity received through retailer r is

$$y_r = min\{\delta_r + \sum_{r \neq r' \in R} \gamma_{rr'} d_r^-, max\{0, z_r - \theta_r\}$$

which is ensured by order characterization constraints (Equation 3.5.8) - (Equation 3.5.12). Note that, in the model formulation, substituted demand $\gamma_{rr'} d_r^-$ is represented with an integer variable $d_{rr'}^-$ to ensure integer demand. As can be seen from constraints (Equation 3.5.13)-(Equation 3.5.15), substituted demand of retailer r by retailer r' is defined as $d_{rr'}^- = \lfloor \gamma_{rr'} d_r^- \rfloor$ and the rest d_{rr}^- is all lost. The max function for effective contract quantity ($max\{0, z_r - \theta_r\}$) is linearized via binary variables v_r. Constraints (Equation 3.5.11) and (Equation 3.5.12) explicitly ensures the equality of order characterization function at existence of positive penalty with binary variables w_r.

Following set of constraints (Equation 3.5.16) - (Equation 3.5.20) model lost customers at each retailer d_r^- (i.e. lost opportunity constraints). It is assumed that as long as there is remaining availability from effective contracted quantity, customers will order through their preferred retailer and the retailer will always accept the orders. The detailed description of linear modeling is in section B.1.

Lastly, constraints (Equation 3.5.21)- (Equation 3.5.22) and constraints (Equation 3.5.23)- (Equation 3.5.24) control binary variables b and v respectively (i.e. sales limit indicator constraints and effective contract quantity indicator constraints).

3.4.2 Models for Alternative Contract Policies

The general APC model corresponds to penalty-driven fillrate policy and serves as a basic backbone model. Recall the three contract policies from subsection 3.3.2: guaranteed fulfillment, controlled fillrate, and penalty-driven fillrate. Models (Equation 3.6) and (Equation 3.7) corresponds to the stochastic APC models with guaranteed fulfillment policy and controlled fillrate policy respectively.

Guaranteed fulfillment policy is the simplest and the most naive policy. The model is as follows:

$$\max E[Q(z, \delta, \theta)] \tag{3.6.1}$$

$$\text{s.t.} \sum_{r \in R} z_r \leq A \text{ (No Overpromising)} \tag{3.6.2}$$

$$(Equation\ 3.5.2) - (Equation\ 3.5.25) \tag{3.6.3}$$

An addition of maximum promising quantity constraint (Equation 3.6.2) to model (Equation 3.5) in the first stage is sufficient to model the guaranteed fulfillment policy, the model (Equation 3.6). Meanwhile, the second stage can further be simplified as all orders can be fulfilled ($y^- = 0$ and $y = y^+$) and constraint (Equation 3.5.6) is implied by the new first stage constraint. That said, the two second stage variables (y^-, y^+) and constraint (Equation 3.5.6) can be removed without losing model validity. In fact, under the guaranteed fulfillment policy, there is no pooling effect and it does not take advantage of dropship operation.

Controlled fillrate policy is designed to smartly incorporate PATs into APC decision as described in detail in subsection 3.3.2. To model controlled fillrate policy, another first-stage variable x should be added to the general model:

- x_r: Internal contract quantity to retailer r

According to the proposition 2, replacing $\Theta_{\alpha,r}$ with tl_r, a lower bound of θ_r, ensures zero stock out probability while preventing demand loss due to retailer thresholds. Model (Equation 3.7) solves the APC problem with controlled fillrate policy. Note that the constraint (Equation 3.7.3) controls stockout probability following Proposition 2.

$$\max E[Q(z, \delta, \theta)] \tag{3.7.1}$$

s.t. Controlled Overpromising

$$\sum_{r \in R} x_r \leq A \tag{3.7.2}$$

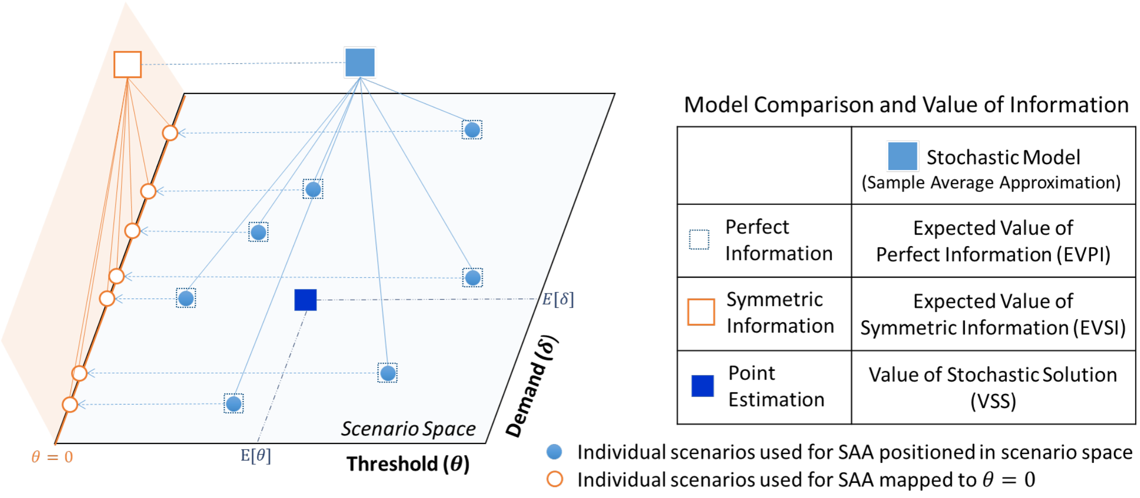
$$z_r \leq x_r + \Theta_{\alpha,r} \quad , \forall r \in R \tag{3.7.3}$$

$$x \in \mathbb{Z}_+ \tag{3.7.4}$$

$$(Equation\ 3.5.2) - (Equation\ 3.5.25) \tag{3.7.5}$$

3.4.3 Solution Generation and Evaluation

The modeling approaches described in Section 3 resulted in a 2-stage stochastic programming with linear constraints and objectives. That, in turn, enables the use of sample average approximation (SAA) to obtain solutions, which can be solved by using commercial solvers such as Gurobi or cplex. Such reduction to a simple solution approach also offers more practicality to model application in industry and better reflects the inherent generality of the problem where demand distribution varies by product, company, and/or season. Here, we describe the uncertainty modeling used for the case study, scenario generation for SAA, and evaluation models with respect to the two stochastic parameters in scenario space as illustrated in Figure 3.7.



Model Comparison and Value of Information

	■ Stochastic Model (Sample Average Approximation)
⬚ Perfect Information	Expected Value of Perfect Information (EVPI)
□ Symmetric Information	Expected Value of Symmetric Information (EVSI)
■ Point Estimation	Value of Stochastic Solution (VSS)

● Individual scenarios used for SAA positioned in scenario space
○ Individual scenarios used for SAA mapped to $\theta = 0$

Figure 3.7: Model positioning over scenario space and comparison

The scenarios are generated with respect to the two stochastic parameters: demand and threshold. As stressed earlier, the APC model is independent of the methods used to obtain the probability distributions, so any method can be used to model uncertainty which suits the best to the case. Firstly, to obtain a next-day demand forecast of the target product by retailer. We use a two-step top-down approach ([135]): (1) forecast the total market demand and (2) obtain retailer specific forecast by estimating the market share of each retailer. For part (1), we used dual-seasonality BM model ([141]), which is a variation of Holts-Winter's forecasting method. Time-series models are traditional and most commonly used for demand forecast ([142]), while new variations are still explored ([143]). The unobservable lost demand due to unavailability is captured by dynamic smoothing factor adjustment based on APC and sales. Simple exponential smoothing is used to forecast market share of retailers. The 2-step approach can intrinsically embed correlation between demands of multiple retailers. Intuitively, in the first step, positive correlation is included by estimating total demand and in the second step, negative correlation is included as the market shares must be summed up to 1. Along with point forecast, upper and lower bounds of forecast are estimated with confidence level 0.98.

Retailer threshold (PAT) is not directly observable from historical order and contract data as the dropshipper cannot distinguish if the leftover is due to lack of customer demand or to the threshold. It can be only observed via real-time monitoring of current sales status on all retailers' websites and sales status. Due to the difficulty of data acquisition, empirical PAT distributions based on business information and experience are used.

Demand and PAT scenarios are sampled from each distribution independently and combined to construct scenarios for SAA as illustrated in the 2-D space in Figure 3.7. Demand scenario is sampled in 2 steps: (1) total demand sample is obtained from demand distribution and (2) retailer market share is sampled from multinomial distribution using market share forecast as its parameter. Multiplying total demand from (1) and market share (2) gives demand by retailer. The main assumption underlying the 2-step sampling approach is total market size (demand) is independent of retailer market shares.

The solution generated from SAA with N training data is then evaluated over a separate scenario set, an evaluation data. The performance of solutions on evaluation set is then used to understand various policies and impact of stochastic parameters. The policy variation introduced earlier enables solution comparison with respect to contract policy and to find best policy. On the other hand, evaluating and comparing solutions at different levels of uncertainties also provide important insights. As described in Figure 3.6, three alternative models are used to evaluate value of information about the stochastic parameters. The first model is when both demand and threshold (PAT) is known deterministically (perfect information). This model provides objective upper bound for stochastic optimization model. The second model is when only PAT value is known with better intelligence of dropshipper or cooperative retailer (symmetric information). Note that such symmetric information does not enforce dropshipper to share availability status transparently with retailer. Comparison between the stochastic model and these two models gives the value of information: Expected value of perfect information (EVPI, or regret) and expected value of symmetric information (EVSI) respectively. The last is point estimation where APC is made only based on point estimation values of demand and PAT, not based on distributional knowledge. The value of stochastic solution (VSS) can be calculated by comparing this model with the stochastic model. Figure 3.7 shows the positioning of each model over the scenario space with respect to corresponding degrees of uncertainty and comparison tables between models. According to Proposition 1 the symmetric information model can be solved by projecting the 2-dimensional scenario space with demand and PAT to a one dimensional space where $\theta = 0$.

3.5 Computational Results

In this section, we present computational results from a case study and derive managerial insights from the results. Sensitivity analysis is also performed to better understand the impact of retailer substitution and retailer threshold.

3.5.1 Case Study

The case consists of 7 days and 4 retailers for a single product. The average of total market demand is forecasted to be 22 units per day. The case is built on the anonymized sample of our industry partner's data. The daily forecasts for the 7 days generate lower and upper bounds together with point forecasts as shown on the left of Figure 3.8. Triangular distribution is used with the point forecast and upper/lower bounds. On the right side of Figure 3.8, the four retailers are relatively positioned based on their threshold (PAT) uncertainty and market share. Simply, each can be categorized as large/small retailer with high/low PAT uncertainty. The sample size of 100 is chosen with a low statistical gap of 1.6% based on the algorithm presented in Kleywegt, Shapiro, and Homem-de-Mello [125]. For all retailers, the margin and penalty are set at 1 and no substitution is assumed. All models are solved to optimality using commercial solver Gurobi. Each APC problem is solved under five different availability levels: scarce, low, lean, high and overfull. Recall the five availability levels are defined w.r.t. daily forecast as illustrated in Figure 3.3.

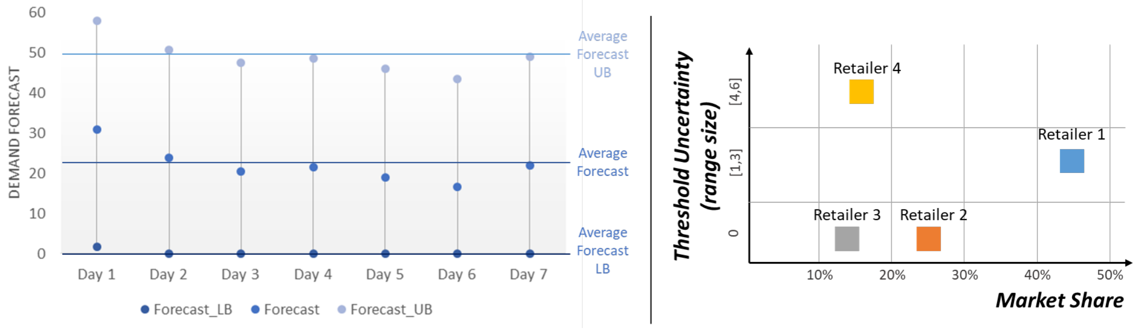
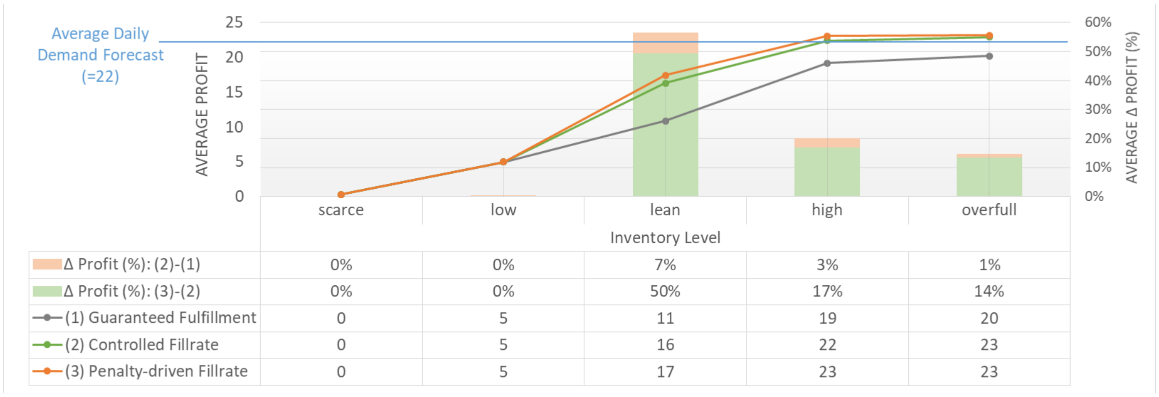


Figure 3.8: Daily forecast with lower and upper bound (left) and retailer positioning with respect to market share and threshold (right)

Figure 3.9 shows the average daily expected profit over 7 days by the three contract policies under varying availability level. The result of each day shows similar patterns although the profit size varies due to the demand size on each day. The daily results are attached in section B.2. At any availability level, a policy allowing more overpromising outperforms the others and the extra profit with overpromising flexibility is maximum at lean availability. Under scarce or low availability, the impact of contract policy is minor as the lack of availability is dominant constraint limiting sales. On the other hand, with high or overfull availability, the importance of contract policy decreases as there are enough products to offer to all retailers. That is, operating with smarter contract methodology becomes more critical when the availability is not too low or too high. Another insights from Figure 3.9 is that although penalty-driven fillrate policy always outperforms or equals controlled fillrate policy, most of the profit increase from guaranteed fulfillment which allows no overpromising can be achieved by controlled fillrate policy.

Figure 3.10 shows expected profit by model with different information as presented in Figure 3.7. Expected profit under perfect information model provides profit upper bound for corresponding policy and availability level combination. The difference between perfect information and stochastic models gives expected value of perfect information (EVPI) and the difference between symmetric information and stochastic models gives expected value of symmetric information (EVSI). The point estimation model cannot outperform



	scarce	low	lean	high	overfull
Δ Profit (%): (2)-(1)	0%	0%	7%	3%	1%
Δ Profit (%): (3)-(2)	0%	0%	50%	17%	14%
(1) Guaranteed Fulfillment	0	5	11	19	20
(2) Controlled Fillrate	0	5	16	22	23
(3) Penalty-driven Fillrate	0	5	17	23	23

Figure 3.9: Average daily expected profit by contract policy under varying availability over 7 days

stochastic model as the optimal solution of point estimation model is also a feasible solution for the stochastic model. The difference between these two models gives the value of stochastic solution (VSS). VSS tends to increase when availability increases because using point forecast values only underestimate demand although there is enough availability to satisfy it. However, under penalty-driven fillrate policy, VSS increases when availability level deviates from lean level in either direction. Because the point estimation ignores the potential of getting lower PAT value, point estimation model results in excessive overpromising which cannot be fulfilled under low or scarce availability.

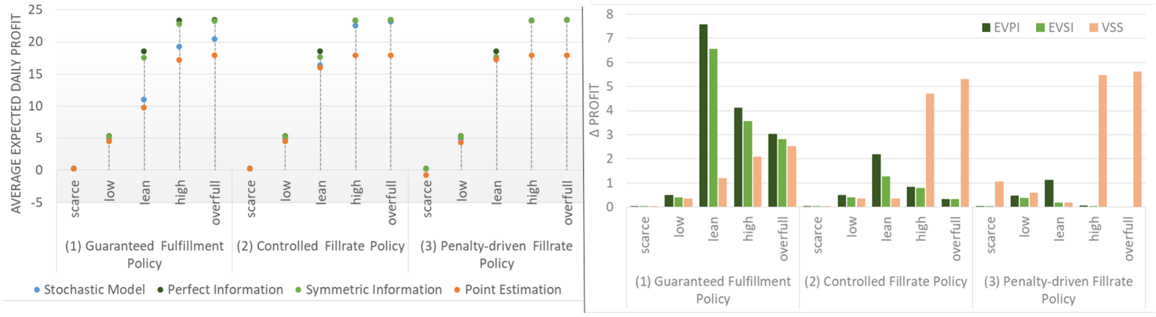


Figure 3.10: Average daily expected profit by varying level of information (left) over 7 days and average values of information (right)

Unlike VSS, it can be seen from Figure 3.10 that the value of information (EVPI, EVSI) is highest at lean availability and decreasing as the availability level deviates from it. This can be explained similar to the impact of smarter contract policy. When the availability level is too low, the information affects profit less as sales are dominantly bounded by availability. When availability level is very high, again, the information also affects profit less as enough availability can be provided to all retailers to cover most of demand and/or threshold variation. Comparing policies, it can be seen that knowing demand and/or threshold with higher certainty is more critical when overpromising is not allowed. This is because the APC is less flexible without overpromising. Another observation is that the difference between EVPI and EVSI is small compared to EVSI in most of the cases. That is, the uncertainty of thresholds has significant impact on the dropshipper's expected profit. In short, for APC which is made as an ex-ante decision, it is important to reflect the uncertainty into decision process and increase both accuracy and precision of the uncertain parameters, demand and threshold.

While the retailer threshold (PAT) certainly affects dropshipper's APC decisions and profit, it is questionable if the PAT is an effective strategy for retailers. Assuming the negotiation power of retailers to make the dropshipper assign the limited availability to them is equivalent, the threshold strategy benefits retailers when it leads the dropshipper to assign more availability to them. Figure 3.11 shows how many more or less quantities are promised to each retailer when some of the retailers adopt the threshold strategy. Recall that retailers 1 and 4 adopted threshold strategy while retailers 2 and 3 did not. When overpromising is allowed, retailers with PAT tend to get more availability assigned regardless of the overall availability level, which in some cases is taken from the retailers without PAT. When overpromising is not allowed and availability level is lean or lower, the dropshipper rather takes promised quantity away from the retailers with PAT and assign them to retailers without PAT. In other words, under the tight availability and inflexible policy, the dropshipper invests in more certain options than those with more risk. When the availability level is high, on the other hand, the dropshipper puts most of the availability to retailer 1 who has availability but has highest demand. Considering that retailer 1 has the largest market share, it can be interpreted that the best strategy for the dropshipper was to bet on the high-risk high-return option. In short, retailers can expect to benefit from having PAT with overpromising dropshippers, but it may not be an advantageous strategy with non-overpromising dropshippers, especially when their market share is not dominant.

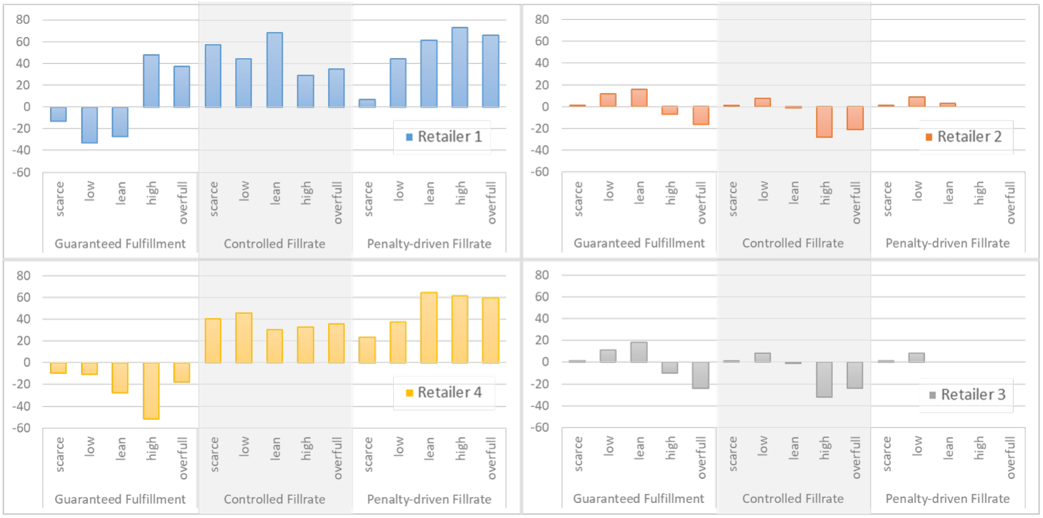


Figure 3.11: Total promised quantity over 7 days changes with threshold information by retailer (stochastic model - symmetric information)

3.5.2 Sensitivity Analysis of Retailer Substitution

For the purpose of sensitivity analysis in subsection 3.5.2 - subsection 3.5.4, the demand distribution of day 7 is used as it is the closest to the average demand distribution. Also, to focus on the sensitivity analysis, we limit the availability to only low, lean and high levels.

The impact of promised quantity on demand (endogenous uncertainty) induced by the demand pooling between retailers is a main reason which makes the APC and fulfillment problem nontrivial. The demand pooling impact can be modeled with retailer substitution as described in previous sections. The main results assume no substitution, which induce independent and exclusive demand by retailer, but in practice, customers check multiple websites before making their purchases although they usually have a preferred retailer.

In addition to the no-substitution scenario, two more substitution scenarios are assumed: equal substitution and proportional substitution. In equal substitution, customers will be lost or leave to any other retailers with equal probability in case of unavailability at their preferred retailer ($\gamma_{rr'} = \frac{1}{|R|}, \forall r, r' \in R$). In proportional substitution, the probability that the customer will leave to another retailer (or be lost if it is the preferred retailer) in case of unavailability at the preferred retailer is the same as the market share of the corresponding retailer ($\gamma_{rr'} = marketshare(r'), \forall r, r' \in R$).

As seen in Figure 3.12, positive substitution benefits the dropshipper in all policies and availability levels as demand is not just lost but may be captured by the dropshipper through another retailer. The percentage profit increase is the largest under the guaranteed fulfillment policy (no overpromising allowed) and lean availability level. This is firstly because there is no availability pooling through overpromising under this policy. Secondly, when the availability level is low or high, the impact of other factors, such as policy or substitution, is weakened as availability becomes a binding constraint or there is enough availability to absorb the variability.

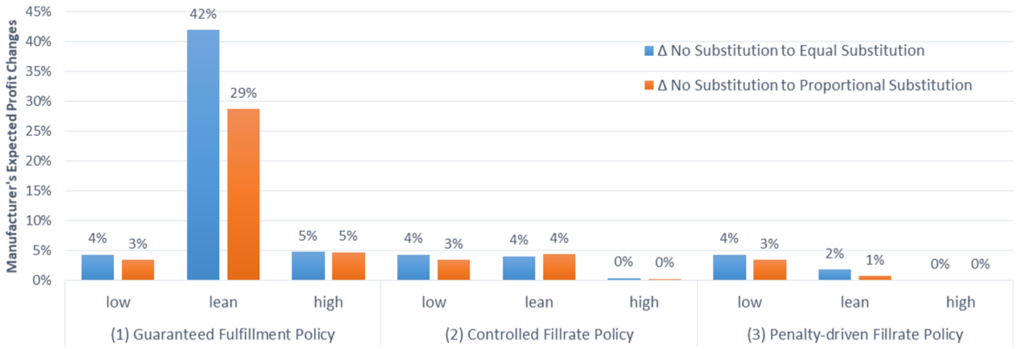


Figure 3.12: Impact of retailer substitution to dropshipper's expected profit

The promised quantities to the retailers change only marginally, except under guaranteed fulfillment policy with lean availability as shown in Figure 3.13. With a dropshipper who does not overpromise and has lean availability, positive retailer substitution makes the dropshipper remove promised quantity from a retailer with PAT (retailer 1) and assign it to retailers without PAT (retailers 2 and 3). This is because part of demand lost from retailer 1 can be captured via retailer 2 or 3 so the better strategy for the dropshipper is to reduce uncertainty from PAT and capture the deviated demand.

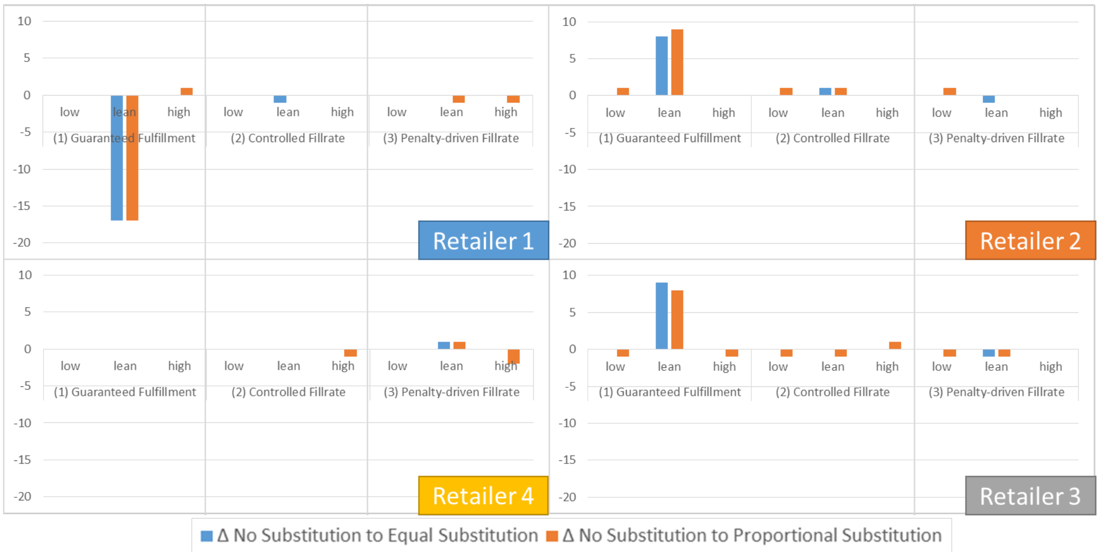


Figure 3.13: Promised quantity changes to each retailer by varying substitution scenarios with respect to no substitution scenario

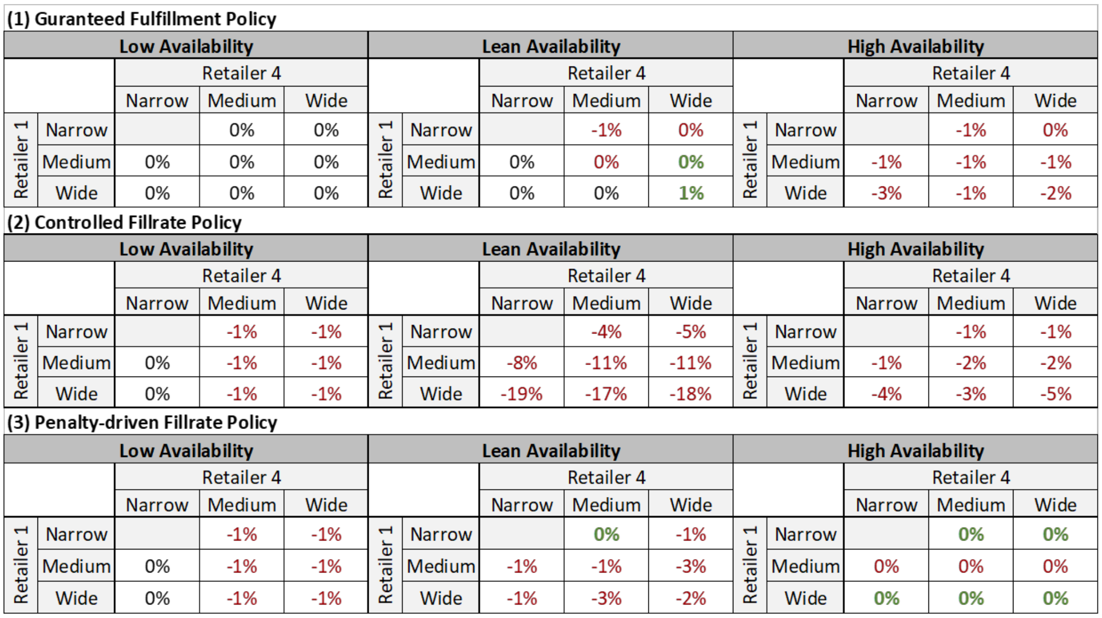
3.5.3 Sensitivity Analysis of PAT Uncertainty

Threshold (PAT) uncertainty can be represented as a standard deviation or range of a threshold distribution. To investigate the impact of PAT uncertainty, three threshold distributions are compared for retailers 1 and 4 respectively with respect to their expected demand $\hat{\delta}$: $Uniform[(1-\beta)\hat{\delta},(1+\beta)\hat{\delta}]\ where\ \beta = 0.1, 0.3, 0.5$. The three ranges will be referred as narrow, medium, and wide respectively. Recall that only retailers 1 and 4 have PAT.

Table 3.1 shows the dropshipper's expected profit change when a retailer's PAT uncertainty changes under each contract policy. PAT uncertainty is measured as the range of potential PAT values while the average remains unchanged. The results are highlighted by color green or red if the expected profit increases or decreases respectively. It can be seen that when the availability level is low or high, the PAT range has only a marginal impact. Also, the impact of PAT range is largest with overbook at risk of threshold policy. In general, the dropshipper's expected profit decreases as the threshold range, or the uncertainty of threshold, increases.

Figure 3.14 and Figure 3.15 show the promised quantity change induced by having narrower or wider range of PAT for each retailer 1 and retailer 4. When one retailer's range

Table 3.1: Dropshipper's expected profit change by threshold (PAT) range variation with respect to narrow PAT ranges



(1) Guranteed Fulfillment Policy

		Low Availability					Lean Availability					High Availability		
		Retailer 4					Retailer 4					Retailer 4		
		Narrow	Medium	Wide			Narrow	Medium	Wide			Narrow	Medium	Wide
Retailer 1	Narrow		0%	0%	Retailer 1	Narrow		-1%	0%	Retailer 1	Narrow		-1%	0%
	Medium	0%	0%	0%		Medium	0%	0%	0%		Medium	-1%	-1%	-1%
	Wide	0%	0%	0%		Wide	0%	0%	1%		Wide	-3%	-1%	-2%

(2) Controlled Fillrate Policy

		Low Availability					Lean Availability					High Availability		
		Retailer 4					Retailer 4					Retailer 4		
		Narrow	Medium	Wide			Narrow	Medium	Wide			Narrow	Medium	Wide
Retailer 1	Narrow		-1%	-1%	Retailer 1	Narrow		-4%	-5%	Retailer 1	Narrow		-1%	-1%
	Medium	0%	-1%	-1%		Medium	-8%	-11%	-11%		Medium	-1%	-2%	-2%
	Wide	0%	-1%	-1%		Wide	-19%	-17%	-18%		Wide	-4%	-3%	-5%

(3) Penalty-driven Fillrate Policy

		Low Availability					Lean Availability					High Availability		
		Retailer 4					Retailer 4					Retailer 4		
		Narrow	Medium	Wide			Narrow	Medium	Wide			Narrow	Medium	Wide
Retailer 1	Narrow		-1%	-1%	Retailer 1	Narrow		0%	-1%	Retailer 1	Narrow		0%	0%
	Medium	0%	-1%	-1%		Medium	-1%	-1%	-3%		Medium	0%	0%	0%
	Wide	0%	-1%	-1%		Wide	-1%	-3%	-2%		Wide	0%	0%	0%

changes, the other's PAT range is fixed at medium. In general, when overpromising is not allowed (or limited) and/or inventory level is low, increasing PAT range decreases promised quantity. This indicates that pressuring dropshippers by imposing high PAT uncertainty may not be a good strategy for retailers. Also, the pattern appears to both retailers whose market shares are different. Recall that retailers 1 and 4 have large and small market shares respectively.

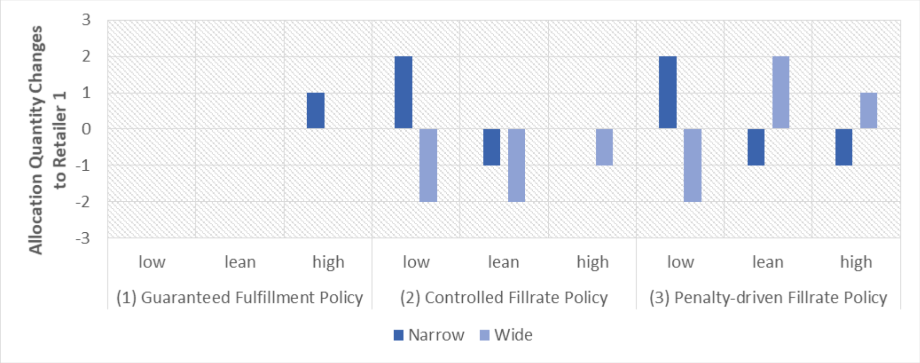


Figure 3.14: Changes in promised quantity to retailer 1 when its PAT range is narrowed/widened from medium range

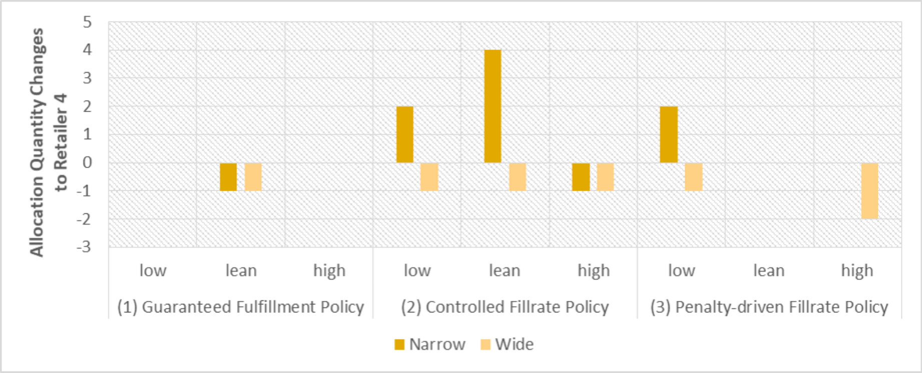


Figure 3.15: Changes in promised quantity to retailer 4 when its PAT range is narrowed/widened from medium range

3.5.4 Sensitivity Analysis of PAT Size

Threshold (PAT) size is defined as an expected value of PAT. We varied the PAT size from low to high with respect to demand size of the corresponding retailer. More specifically, three PAT sizes are studied: small, medium, and large, corresponding to x0.1, x0.5, and x1.0 of corresponding demand size. While varying the PAT size, the deviation and shape of PAT distribution remains unchanged. Also, since only the PAT size changes, the controlled fillrate policy is excluded in this sensitivity analysis.

Table 3.2 and Table 3.3 show the dropshipper's expected profit changes by the PAT sizes of retailers 1 and 4 as small thresholds for both retailers a baseline. With guaranteed fulfillment policy, expected profit decreases with larger thresholds, except under low availability level. Since no more quantity can be promised from availability, the dropshipper may not have enough overall availability to exceed retailer's PAT level with a larger PAT size. In the case of low availability, the APC or PAT affects less as the available quantity is a dominant constraint. In case of penalty-driven fillrate policy, on the other hand, the varying PAT size has no significant impact on the dropshipper's profit.

Table 3.2: Dropshipper's expected profit change by PAT size variation with respect to small PAT: Guaranteed fulfillment policy

		Low Availability			**Lean Availability**			**High Availability**		
		Retailer 4			Retailer 4			Retailer 4		
		Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
Retailer 1	Small		0%	0%		-3%	-3%		0%	-1%
	Medium	0%	0%	0%	-14%	-15%	-15%	-2%	-2%	-4%
	Large	0%	0%	0%	-25%	-26%	-28%	-6%	-7%	-10%

Table 3.3: Dropshipper's expected profit change by PAT size variation with respect to small PAT: Penalty-driven fillrate policy

		Low Availability			**Lean Availability**			**High Availability**		
		Retailer 4			Retailer 4			Retailer 4		
		Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
Retailer 1	Small		0%	0%		**0%**	**0%**		0%	**0%**
	Medium	0%	0%	0%	**1%**	**0%**	**0%**	**0%**	**0%**	**0%**
	Large	0%	0%	0%	**1%**	**0%**	**0%**	0%	**0%**	**0%**

3.6 Conclusion

This chapter studies the design of availability promising contracts (APCs), motivated from a partnership with a furniture dropship manufacturer. This dropshipper dynamically engages in APCs with multiple retailers, some of them having promised availability thresholds (PATs). When the remaining promised availability level falls below the retailer-specific PAT, no more orders can be received through the retailer. The dropshipper only knows the probability distribution of the threshold. As the contract does not involve physical transaction, the dropshipper may promise more quantities than available-to-promise quantity using the ignorance of retailers on actual availability level (overpromising). Three contract policies are investigated based on overpromising behavior: (1) Guaranteed fulfillment, (2) Controlled fillrate, and (3) Penalty-driven fillrate. The contract design optimization problem is modeled as a stochastic programming with two stochastic parameters: demand and threshold.

The model-driven decision support system can readily be used by companies with customized forecast and distribution assumptions. It can also generate customized managerial insights. The computational experiments presented in the chapter provide general yet critical managerial insights for APC and PAT operations. The results show that the overpromising policy always outperforms other policies and the benefit is maximized at lean availability level while the benefit is decreasing as the availability level decreases or increases. The retailer-specific PAT has significant impact on the dropshipper's expected profit while its benefit for retailers varies. When the dropshipper overpromises, adopting a PAT operation is beneficial to retailers as the dropshipper tends to commit more to compensate the extra uncertainty from thresholds. However, with a dropshipper who does not overpromise, having thresholds can rather reduce the quantity promised to the retailer unless the availability level of the dropshipper is high and the retailer has a dominant market share.

The chapter investigates a new type of contract for dropshipping operation, which opens a rich set of future research avenues. We here highlight three such avenues as an extension of current model and additional avenue as an exploration of the model in an different perspective. Firstly, the model can be extended to multi-period setting. With future sales included in the scope, the optimal strategy can be to promise less quantity today for securing future sales even though it may decrease today's profit. The extension covering the product's replenishment lead time can be easily linked to availability planning as well. A second potential avenue is a multi-location extension where the contracts are made by product and storage location. This has significant managerial impact on both dropshipper and e-retailers. E-retailers can offer faster delivery lead times and dropshippers can reduce delivery cost when fulfilling customers from closer locations. More specific demand forecasts by region is required for the multi-location extension. The model scope can be extended to a deployment problem as well. The third potential research avenue is to investigate and enable dynamical contract renewal. For example, more quantity can be assigned as the remaining quantity is getting close to the expected PAT or multiple contracts can be renewed to enable promised availability reallocation based on demand pattern observed during the day. The forth extension is to further investigate the point estimation model by finding the optimal quantile to be used instead of using the expected value. The point estimation value has not been optimized in the setting as it was not the focus of the sutdy. However, if the point estimation model with optimal quantile can generate near-optimal value compared to the stochastic models, the simpler deterministic model can be used to support more complex models such as multi-period extension models. Another future research avenue is to view the problem in a perspective of game theory by extending the problem to model retailer's PAT decision instead of treating PAT as an exogenous stochastic parameter. Such game theoretical model can deliver interesting insights on how effective the PAT strategy is and if the PAT and stockout penalty are complementary or supplementary.

CHAPTER 4

NETWORK INVENTORY DEPLOYMENT FOR RESPONSIVE FULFILLMENT

4.1 Introduction

With the growth of e-commerce and prevalent home delivery practices, expectations for shorter delivery lead times are increasing and meeting such expectations is critical for holding the lead in the competitive market. In 2020, the e-commerce sales revenue in the US exceeded 432B dollars and projected to be 549B dollars in 2024 ([4]). Moreover, a report of Invesp ([8]) indicated that 56% of young online customers expect same-day delivery and more than half of customers want even faster delivery, 1- to 3-hour-delivery. That is, customers are becoming more and more service conscious ([144]; [131]). Nowadays, Amazon offers a same-day Prime delivery for selected items and a 2-hour delivery with Prime Now. Walmart offers 2-hour Express Delivery and Target offers same-day delivery. Besides the big players, many start-ups, such as Flexe.com, Darkstore, and Deliv (acquired by Target), have been launched in offering delivery or distribution services to support very-fast delivery. Noticing the trend and growing customer expectations, existing e-commerce companies have started to explore new business opportunities and adjusting strategic plans to adopt to the market changes.

Our research is motivated by a collaborative project with an industry partner, a dropship manufacturer in the North American market offering furniture and home interior accessories that consist of over 2000 stock keeping units (SKUs). It ships per year almost a million items ordered online to more than 250 thousands customer locations by zip code. Currently, our industrial partner is operating with rather a centralized fulfillment/distribution network with three facilities in Utah, Texas, and Tennessee as shown in the left side of Figure 4.1. The locations are also used as mid- or long-term storage locations to hedge against demand uncertainty. However, capability to offer same-day or x-hour delivery is limited in such centralized fulfillment network. Accordingly, we investigated alternative fulfillment solutions in collaboration with our partner. First, the solution is to expand the fulfillment network by utilizing open fulfillment centers (FCs) through service providers spread all over the US as exemplified on the right hand side of Figure 4.1. Second, the solution aims to optimally deploy the fulfillment inventories in the decentralized network to better utilize the increased density and smartly leveraging the pooling structure than the current fulfillment scheme, based on a pre-allocation of demand zones to a specific FC (usually the nearest).

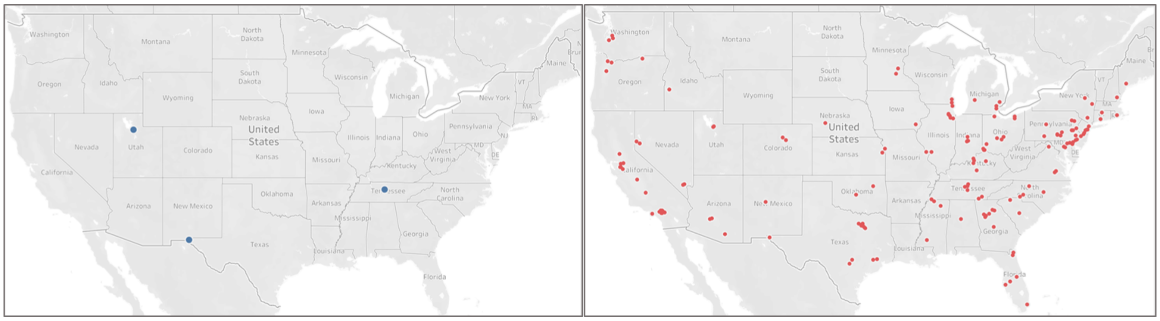


Figure 4.1: Current centralized FC network (left) and decentralized FC network (right) in Contiguous United States

As stressed, the main motivation of the study is the growing customer expectations on a very fast delivery. We will use the term ’responsiveness’ and ’responsiveness level’ to refer to how fast a delivery can be and delivery lead time respectively, following Snoeck and Winkenbach [145]. To meet such a tight responsiveness level requirement, products must be physically available within the lead time distance. This requires an access to a broad and dense network of FCs. In fact, when considering the US market, it is possible to exploit more expensive but fast transportation mode, such as air transportation, to cover the lack of physical availability if the shortest delivery lead time is next-day as pointed by Acimovic and Graves [146]. However, when it comes to same-day or X-hour delivery, the physical availability of items nearby becomes necessary as illustrated conceptually in Figure 4.2, which contrasts low and high responsiveness and centralized and decentralized network.

When higher responsiveness is required, the coverage area which is within the maximum distance to deliver to each customer in the required delivery time (shown in dotted line in Figure 4.2) shrinks. Naturally, a broader and denser FC network is required to maintain the same level of demand fill rate. Traditionally, however, although companies are aware of the benefit of a more decentralized network, often it is impossible or financially unviable as it implies a huge capital investment. We find that the Physical Internet (PI) initiative ([147]; [148]; [21]) can provide an alternative solution to overcome the hurdles. The PI builds a hyperconnected global logistics system enabling seamless asset sharing and flow consolidation by applying the mechanism of digital data operations in the web to physical object. We specifically focused on its emphasis on open asset utilization which can potentially enable affordable access to broad and dense fulfillment network by using services of open FCs on-demand instead of building own fulfillment network. For instance, the open FCs in Figure 4.1 can be a set of available FCs that a company can use by purchasing the service provided by one or more service providers. Several early business examples that offer fulfillment and storage services can be found in market already, such as the services provided by Flexe.com, ES3, WarehouseAnywhere, and Darkstore.

Yet the advantage of decentralized networks is not obvious as centralized inventory systems often outperformed decentralized systems due to maximized demand pooling as shown in many previous studies for decades under various settings ([14]; [149]; [150]). However, earlier research found it beneficial to rely on decentralized networks in other contexts such as supply chain disruption ([151]), omnichannel ([152]) and e-commerce ([146]). Therefore, it must be rigorously evaluated whether decentralization is beneficial and how pooling affects in our context where the fulfillment of online orders that require tight responsiveness. Responsiveness requirements results in a complex and interesting fulfillment capability structure that affects pooling structure in inventory management. Note that online customers do not differentiate where the products are shipped from, unlike customers at a store who can only buy products available at the store in traditional or om-

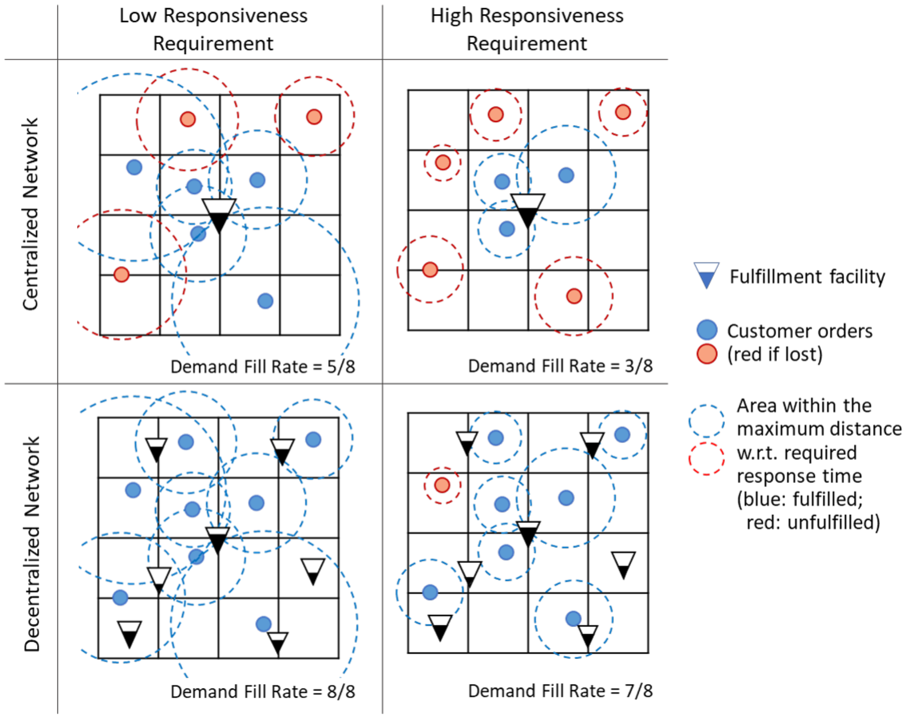

Low Responsiveness Requirement
High Responsiveness Requirement
Centralized Network
Decentralized Network
Demand Fill Rate = 5/8
Demand Fill Rate = 3/8
Demand Fill Rate = 8/8
Demand Fill Rate = 7/8
Fulfillment facility
Customer orders (red if lost)
Area within the maximum distance w.r.t. required response time (blue: fulfilled; red: unfulfilled)

Figure 4.2: Demand fulfillment from centralized and decentralized network under low and high responsiveness requirements

nichannel contexts ([152]). That is, online demands, that we are focusing on here, can be satisfied from any inventory location, e.g. FC, as long as the delivery lead time requirements are met. In the meantime, not all locations are capable of serving certain customer due to a tight responsiveness requirement. This is major difference to previous works that assumes the minimum delivery lead time is a day or more which can be covered by any inventory locations ([146]). As a results, demands become partially pooled, where each demand segment can only be served from a certain subset of inventory locations due to the responsiveness requirement (fulfillment capability). Such partial demand pooling leads to a very complex model as illustrated with a case of 3 inventory locations in Figure 4.3. The complexity tend to increase exponentially as the network size and the number of responsiveness options grows. Although there is a plethora of inventory and fulfillment studies considering the shipping cost and/or demand pooling, the context of such tight responsiveness level constraints has not been studied as much because it became critical only recently.

It is important to maximize the utilization of fulfillment flexibility and partial demand pooling.

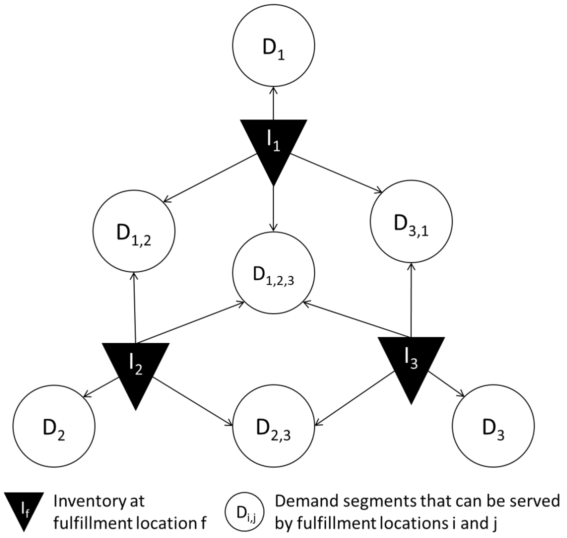

D_1
I_1
$D_{1,2}$
$D_{3,1}$
$D_{1,2,3}$
I_2
I_3
D_2
$D_{2,3}$
D_3
I_f Inventory at fulfillment location f
$D_{i,j}$ Demand segments that can be served by fulfillment locations i and j

Figure 4.3: Illustrative example with three inventory locations case

As the study focuses on online demand and home delivery, inventory locations are referred to as FCs throughout the manuscript.

This chapter builds and solves an inventory deployment model under tight responsiveness requirements for a company which needs to evaluate operations in a centralized fulfillment network given its resource and a decentralized fulfillment network accessed through a service provider. The chapter makes four major contributions.

- To the best of our knowledge, we are the first to introduce the notion of responsiveness to inventory modeling and formulate a Newsvendor model for a partially pooled network. We stress the complexity of the model as well as the intractability of the exact solution.

- We develop a pragmatic and easy-to-compute heuristic solution built upon the exact solution, a W-solution, which has an equal-fractile structure. We also present a binary search based heuristic to efficiently calculate the W-solution, referred to as a W-heuristic.

- We demonstrate the benefit of the decentralized network under tight responsiveness requirements over the centralized network and the pre-allocation-based inventory network. This is performed through numerical experiments on theoretical demand distributions. We also report rather counter-intuitive observations that solutions with pooling lead to more inventory than solutions without pooling under low sales margins.

- Using the case of a drop-ship manufacturer in the US market, we provide empirical evidence of the benefit of decentralization over the risk pooling when a responsive fulfilment is required. We also demonstrate the effectiveness of the W-solution over the empirical demand distributions and the practical-sized case.

The chapter is structured as follows. The section 3.2 reviews the relevant literature and positions the work. In section 4.3, the Newsvendor model with partial pooling induced by the responsiveness level requirements is built and a solution heuristic is presented. In section 4.4, computational experiments are first conducted over theoretical demand distributions. The experiment explores the benefit of distributed network over centralized network and the benefits of flexible fulfillment (pooling) over zone allocated fulfillment (no-pooling). Results of a case study are also presented using a real company data, evaluating the decentralization and the heuristic solution over an empirical distribution. Lastly, the chapter concludes in section 3.6 with a summary of the findings, contributions and limitations along with avenues for future research.

4.2 Literature Review

There are several streams of literature that are the most related to our study. The first refers to inventory models in decentralized single-echelon networks, especially those that consider constraints related to responsiveness. The second stream relates to fulfillment flexibility. These two streams are closely related to demand and inventory pooling. The third stream deals with the Newsvendor model which is the model used in this study among the various inventory models. The last is inventory modeling in the PI context.

Eppen [14] is one of the first who investigated the inventory pooling under centralized and decentralized networks. Following Eppen [14], a large body of literature studied the benefit of pooling in various settings ([150]; [153]; [149]; [154]; [155]; [156]). Among the large volume of research in this literature stream, some are more focused on comparing the trade-offs between centralization and decentralization. Schmitt, Sun, Snyder, and Shen [151] analyzed trade-offs between risk pooling and risk diversification under centralized and decentralized inventory systems and showed that decentralization is optimal under supply disruption, especially benefiting risk averse firms. Yang, Hu, and Zhou [157] compared both centralization/decentralization and pooling/no-pooling systems where the decentralized system without pooling can be represented as N independent Newsvendor models and the centralized system with pooling as a single Newsvendor model. Other implicit examples that provide good intuition are where demand or inventory can be pooled through transshipment ([158]; [159]; [160]). Govindarajan, Sinha, and Uichanco [152] presented decentralized and integrated inventory models under an omnichannel setting where the online demand is pooled among all stores and fulfillment centers whereas store demands are exclusive to each store. Alptekinoğlu and Tang [161] tackled the inventory ordering and allocation problem by decomposing the master problem into subproblems per inventory location by assigning demands and integrating them to find the optimal assignment. In our modeling framework, inventories and demands are partially pooled. Previously, partial pooling occurs as each company or inventory location decides to pool only a predetermined fraction of inventories ([162]; [160]) or inventories are physically distributed in a central (pooled) location and local (not pooled) locations ([163]). In our context, partial pooling arises due to the complex fulfillment capability structure induced by the responsiveness requirements.

One of the goals of our work is to find optimal inventory deployment over the network, as in Guo, Liu, and Wang [164], who examined the decisions for reserved inventories that hedges against sudden demand surge. Alptekinoğlu and Tang [161] study optimal allocation policy under the context of differentiated services. Govindarajan, Sinha, and Uichanco [152] is closely related to the study where it has a mixture of demands that can be exclusively served by each store and demands that are pooled. However, responsiveness and fulfillment capability induced by responsiveness requirements have not been modeled before. Such fulfillment capability structure with two levels, where some demand are exclusive to stores (as in omnichannel) or retailers and the others can be served by any locations, had been studied previously (e.g. [152]) but not in the context of responsiveness or further extended to more than two levels of pooling. Although responsiveness has started to gain attention, most of the literature focus on transportation ([145]; [165]; [166]) and/or network design ([167]). Some considers fast-ship options ([168]; [169]), which is not considered in our context. Note that the inventory allocation problem is often combined with the facility location problem ([170]; [171]), but in this study we focus on inventory order and allocation decisions for a given fulfillment network. Also, some inventory allocation literature considers service requirements ([172]; [104]), where the service level is measured as the minimum demand fill rate. However, in this study, we only consider the responsiveness requirements and the demand fill rate is determined as a result of cost minimization under a Newsvendor model. We contribute to the stream of inventory literature by first introducing the notion of responsiveness to inventory model and complex fulfillment structure induced by it.

Fulfillment flexibility is also a pertinent topic as it is not only related to responsiveness-induced fulfillment capability structure but also as it affects the inventory model. DeValve, Wei, Wu, and Yuan [173] emphasizes the fulfillment flexibility in an e-retail environment and many papers in the inventory literature with demand pooling address it directly or indirectly (e.g. [157]). However, in general, it is assumed that all demand segments can be served by any potential fulfillment location while it is not allowed due to pre-allocation (or pre-assignment) or fulfillment preference is set and controlled through the shipping cost. For example, in a pure e-commerce setting, Acimovic and Graves [146] modeled a base stock policy that minimizes the shipping cost by maximizing the inventory balance over the network and mitigating spillover. The underlying assumption is that even the shortest delivery lead time can be met by any fulfillment centers (FCs) utilizing a faster transportation mode such as air. Therefore, the total inventory level in the network can be set as a complete pooling inventory level and it is allocated to each FC. In a similar context, Acimovic and Graves [174] modeled a problem that consists to find from which FC online orders should be fulfilled to minimize the total discounted shipping cost. Process flexibility and production capacity in a context of manufacturing can be mapped to fulfillment flexibility and inventory in a context of demand fulfillment. Jordan and Graves [175] is the first that introduced the concept of chaining which is also relevant to our work. Chaining has been further studied and applied since Jordan and Graves [175] ([176]; [177]). In our model, the ’chain’ is already determined from responsiveness requirements. However, the concept of chaining gives a good intuition to understand the fulfillment flexibility. In our experiment, we compare our solution with the solution without fulfillment flexibility where demand segments are pre-allocated to specific fulfillment locations to further add on to the studies in fulfillment flexibility. Again, no previous study exists that considers a complex fulfillment capability structure as we present in the chapter. We allow a complete fulfillment flexibility given the fulfillment capability.

Here, we used the Newsvendor model to build the inventory ordering and allocation model. the Newsvendor model is one of the most well-known and well-studied inventory models. Khouja [178] provides an extensive review of traditional Newsvendor models. It is still one of the most actively used models until now with various extensions ([157]; [179]). Yet in most cases, the fulfillment capability is not tightly modeled as in our model even in multi-location models. For example, Govindarajan, Sinha, and Uichanco [179], who modeled a distribution-free multi-location Newsvendor and derived bounds, assumed that demand can be fulfilled by any locations while extra cost occurs when fulfilled from further locations. Our work is the most related to a variant: Newsvendor Network ([180]; [181]; [182]). The modeling and solution approach of the Newsvendor Network especially provided a great insight to tackle the new type of problems we present, although Newsvendor Network was not built in the fulfillment context. Although the Newsvendor model has a few limitations, especially due to the single-period nature, it is a basic component of many other multi-period models such as the basestock policy and it is still widely used. For example, Özer and Xiong [183] proposed a Newsvendor based heuristic to find the optimal basestock levels in a warehouse and multiple retailer distribution centers served by the warehouse with minimum fillrate requirements.

The Physical Internet (PI) is an innovative logistics transformation movements ([147]; [10]; [21]). Montreuil [21] defines PI as a ”hyperconnected global logistics system enabling seamless open asset sharing and flow consolidation through standardized encapsulation, modularization, protocols and interfaces to improve the efficiency and sustainability of fulfilling humanity’s demand for physical object services.”. Then, [147] summarizes recent progress of PI study and address future research avenues. Although the focus of the study is responsiveness, the sustainability PI system has also been shown in many literature ([21]; [148]). The open asset utilization invigorated by PI movement offers the potential to enable an affordable access to a decentralized fulfillment network through on-demand open fulfillment services for companies regardless of their market share. More importantly, here we aim to study on how to manage inventory efficiently over the open FC network. The hyperconnected distribution system, especially in comparison to a collaborative system, is modeled and evaluated by Sohrabi, Montreuil, and Klibi [13], Sohrabi, Montreuil, and Klibi [184] and Sohrabi, Montreuil, and Klibi [185]. Under the context of inventory management in the PI and hyperconnected fulfillment, Pan, Nigrelli, Ballot, Sarraj, and Yang [69] and Yang, Pan, and Ballot [186] studied the optimal (Q,R) inventory control policy in a multiechelon supply chain utilizing PI hubs in a hyperconnected network. Both reported reduction in cost and the global inventory level with PI. Ekren, Akpunar, and Mullaoglu [187] developed an optimal (s,S) policy with lateral transshipment in a two-echelon network under a PI environment. Yang, Pan, and Ballot [188] studied and demonstrated the positive impact of such PI inventory system on mitigating supply chain disruptions. Ji, Peng, and Luo [189] extend the scope and build an integrated production-to-distribution model. Although the literature on inventory models in PI is growing, up to now, most of the PI-based inventory policies emphasize the use of PI hubs instead of inventory policies in an open fulfillment network or responsiveness.

4.3 Newsvendor Model with Partial Pooling

Here, we develop the inventory model as a variant of the Newsvendor model. Firstly, we present a general linear programming model of the Newsvendor policy with responsiveness as shown in (Equation 4.1). FC, R and S are a set of available fulfillment centers (FCs), regions, and responsiveness levels respectively. The demand D_{rs}, from region $r \in R$ requiring a responsiveness level $s \in S$, is a stochastic parameter. Let $t(s)$ be the delivery lead time required for responsiveness level s and $t(i, r)$ be the time to deliver from FC i to region r. Then a subset of demand segments which can be fulfilled from $f \in FC$ is defined as $D_{rs(f)}$ where $rs(f) := \{(r, s) \in (R, S)|t(f, r) \leq t(s)\}$. Similarly, a subset of FCs that can serve demand segment D_{rs} is defined as $f(rs) := \{f \in FC|t(f, r) \leq t(s)\}$.

Three cost parameters are considered: the unit holding cost at FC i, h_i, the unit purchase cost of inventory c, as well as the unit sales profit when a unit $D_{r,s}$ is fulfilled from FC i, denoted by p_{irs}. Because the sales price of item, p, is in general the same regardless of the region, p_{irs} is defined as a unit sales price subtracted by the shipping costs. The unit overage and underage cost can be written as $c_o = (h_i + c)$ for each i and $c_u = p_{irs} - c$ for each i, r, s respectively. Since the goal of the study is to find the inventory policy that utilizes the partial pooling to fulfill demands with tight responsiveness, we assume that the shipping costs are the same, e.g. flat rate shipping. We also assume that the holding costs are the same across the network. In other words, as long as the responsiveness requirement is satisfied, there is no cost difference to serve a unit demand from a different FC.

The first stage decision variable is the inventory level at FC i, denoted by I_i, and there is a second stage variable x_{irs}, which is the sales quantity used to fulfill demand D_{rs} from FC i, such that $(r, s) \in rs(i)$. Note that the responsiveness requirements and partial pooling are already embedded in the model as a fulfillment feasibility.

$$\max_{I \in \mathbb{R}_+} E[Q(I)] - \sum_i (c + h_i) I_i \tag{4.1.1}$$

$$\text{where } Q(I) = \max \sum_{i,r,s} (p_{irs} + h_i) x_{irs} \tag{4.1.2}$$

$$\text{s.t.} \sum_{(r,s) \in rs(f)} x_{irs} \leq I_i \quad \forall i \in FC \tag{4.1.3}$$

$$\sum_{i \in f(rs)} x_{irs} \leq D_{rs} \quad \forall (r, s) \in \{R, S\} \tag{4.1.4}$$

$$X, I \in \mathbb{R}_+ \tag{4.1.5}$$

One way to solve the Newsvendor problem is to use dual values (shadow prices) corresponding to capacity constraints (Equation 4.1.3), following the method of Harrison and Van Mieghem [181] and Mieghem and Rudi [180]. More specifically, we are using Proposition 2 in Harrison and Van Mieghem [181]. As it was for the case of newsvendor network ([180]), the approach is very useful to gain insights and solve more complex version of newsvendor problem. When using the other approaches such as solving the expected profit equation directly as in Govindarajan, Sinha, and Uichanco [152], it becomes almost impossible to have the equation due to the flexible fulfillment options and feasibility in constraints (Equation 4.1.4) induced by service sensitive demand. One option is to assign priority to demand segments for each FC, which again becomes almost infeasible in practice when the system complexity increases with more number of FCs and service options.

To tackle the problem, here we start with a simple case with two FCs, which induces three demand segments: it can be served only by the first FC, only by the second FC, and by both FCs. We name it the W-case. Then extended models and their solutions are presented. The approximate solution of a general model is derived and named as the W-solution. Lastly, a heuristic to calculate the W-solution is presented, which will be referred to as the W-heuristic.

4.3.1 W-case

The assumption of a fixed shipping cost across the network simplifies the model but also adds more flexibility that may result in multiple optima. Therefore, we assigned a fulfillment priority to the closer fulfillment location to reduce the number of optimal solutions as shown in Figure 4.4 where the pooled demand is divided into D_a and D_b. These two demands are preferred to be served by I_1 and I_2 respectively. The goal of this section is to solve the case of Figure 4.4 and to use it as a base for a solution to a general problem.

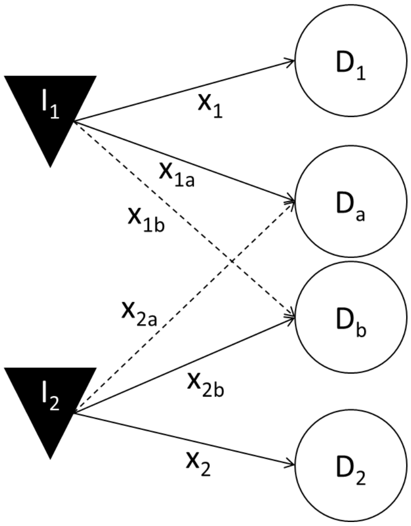


Figure 4.4: W-case with fulfillment preference

The linear relaxation of the primal distribution problem and the corresponding dual problem are described in Model Equation 4.2 and Model Equation 4.3. The primal corresponds to the second stage of Model Equation 4.1 where p is $p = p_{irs} + h_i$ for all i, r, s

combinations.

Primal

$$\max pX \tag{4.2.1}$$

$$\text{s.t. } A_1X \leq I \tag{4.2.2}$$

$$A_2X \leq D \tag{4.2.3}$$

$$X \in \mathbb{R}_+ \tag{4.2.4}$$

Dual

$$\min I\Lambda + D\mu \tag{4.3.1}$$

$$\text{s.t. } B_1\Lambda, B_2\mu \geq p \tag{4.3.2}$$

$$\Lambda, \mu \in \mathbb{R}_+ \tag{4.3.3}$$

Where $X = (x_1, x_{1a}, x_{1b}, x_2, x_{2a}, x_{2b})$, $A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$,

$A_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$, $I = (I_1, I_2)$, $D = (D_1, D_2, D_a, D_b)$, $\Lambda = (\lambda_1, \lambda_2)$,

$\mu = (\mu_1, \mu_2, \mu_a, \mu_b)$, $B_1 = A_1^T$, $B_2 = A_2^T$. The demand space partition is shown in Table 4.1 and Figure 4.5 gives more intuitive visualization of the demand space partition on a 2-D plot with I_1 and I_2 on each axis.

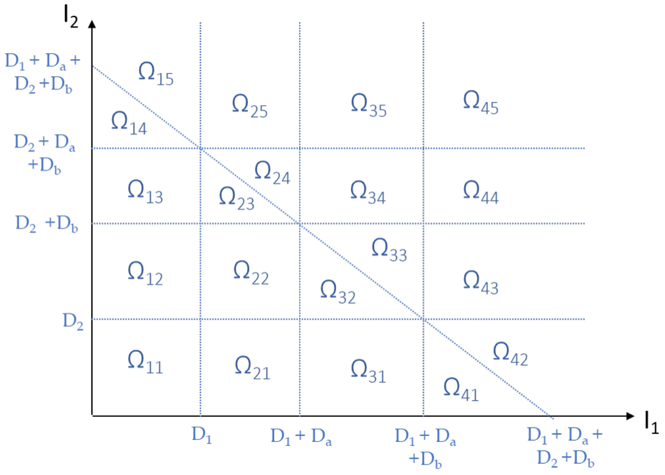


Figure 4.5: Demand space partition in 2-D plot

Table 4.1: Dual solutions by demand partitions, where $D_{ij} = D_i + D_j$

Demand Partition	I_1	I_2	I_1, I_2
Ω_{11}	$I_1 \leq D_1$	$I_2 \leq D_2$	-
Ω_{12}		$D_2 < I_2 \leq D_{2b}$	-
Ω_{13}		$D_{2b} < I_2 \leq D_{2ab}$	-
Ω_{14}		$D_{2ab} < I_2$	$I_1 + I_2 \leq D_{12ab}$
Ω_{15}		$D_{2ab} < I_2$	$I_1 + I_2 > D_{12ab}$
Ω_{21}	$D_1 < I_1 \leq D_{1a}$	$I_2 \leq D_2$	-
Ω_{22}		$D_2 < I_2 \leq D_{2b}$	-
Ω_{23}		$D_{2b} < I_2 \leq D_{2ab}$	$I_1 + I_2 \leq D_{12ab}$
Ω_{24}		$D_{2b} < I_2 \leq D_{2ab}$	$I_1 + I_2 > D_{12ab}$
Ω_{25}		$D_{2ab} < I_2$	-
Ω_{31}	$D_{1a} < I_1 \leq D_{1ab}$	$I_2 \leq D_2$	-
Ω_{32}		$D_2 < I_2 \leq D_{2b}$	$I_1 + I_2 \leq D_{12ab}$
Ω_{33}		$D_2 < I_2 \leq D_{2b}$	$I_1 + I_2 > D_{12ab}$
Ω_{34}		$D_{2b} < I_2 \leq D_{2ab}$	-
Ω_{35}		$D_{2ab} < I_2$	-
Ω_{41}	$D_{1ab} < I_1$	$I_2 \leq D_2$	$I_1 + I_2 \leq D_{12ab}$
Ω_{42}		$I_2 \leq D_2$	$I_1 + I_2 > D_{12ab}$
Ω_{43}		$D_2 < I_2 \leq D_{2b}$	-
Ω_{44}		$D_{2b} < I_2 \leq D_{2ab}$	-
Ω_{45}		$D_{2ab} < I_2$	-

The optimal dual solution corresponding to each demand partition is shown in Table 4.2. The primal solution value is shown in Table C.1 in Appendix. Also, the dual solution with extra shipping cost (s) for x_{1b}, x_{2a} is attached in Appendix: Table C.2, for additional reference.

To calculate the optimal solution, we need to find the gradient from the dual problem. In equation (Equation 4.4.1), the expected revenue, $E[Revenue]$, is the objective value of the primal model.

$$\begin{aligned}\frac{\partial}{\partial I_1} E[Revenue] &= \sum_{\forall k} \lambda^*_{1|\Omega_k} P[\Omega_k] \\ &= p\Big(P[I_1 \leq D_1] + P[D_1 < I_1 \leq D_1 + D_a \& I_1 + I_2 \leq D_{all}] \\ &\quad + P[D_1 + D_a < I_1 \leq D_1 + D_a + D_b \& I_1 + I_2 \leq D_{all}]\Big)\end{aligned} \tag{4.4.1}$$

The result is symmetric for I_2.

Table 4.2: Dual solution by demand partition

Demand Partition	λ_1	λ_2	μ_1	μ_2	μ_a	μ_b
Ω_{11}	p	p	0	0	0	0
Ω_{12}	p	p	0	0	0	0
Ω_{13}	p	p	0	0	0	0
Ω_{14}	p	0	0	p	p	p
Ω_{15}	p	0	0	p	p	p
Ω_{21}	p	p	0	0	0	0
Ω_{22}	p	p	0	0	0	0
Ω_{23}	p	p	0	0	0	0
Ω_{24}	0	0	p	p	p	p
Ω_{25}	0	0	p	p	p	p
Ω_{31}	p	p	0	0	0	0
Ω_{32}	p	p	0	0	0	0
Ω_{33}	0	0	p	p	p	p
Ω_{34}	0	0	p	p	p	p
Ω_{35}	0	0	p	p	p	p
Ω_{41}	0	p	p	0	p	p
Ω_{42}	0	p	p	0	p	p
Ω_{43}	0	0	p	p	p	p
Ω_{44}	0	0	p	p	p	p
Ω_{45}	0	0	p	p	p	p

By setting $\frac{\partial}{\partial I_1}E[Revenue] = c$, where the cost c is defined as $c = c + h_i$ for any location i, we can find the condition for the optimal solution:

$$\begin{aligned}
c &= \frac{\partial}{\partial I_1}E[Revenue] \qquad (4.5.1)\\
&= p\Big(P[I_1 \le D_1] + P[D_1 < I_1 \le D_1 + D_a \& I_1 + I_2 \le D_{all}] \\
&\qquad + P[D_1 + D_a < I_1 \le D_1 + D_a + D_b \& I_1 + I_2 \le D_{all}]\Big) \\
&= p\Big(P[I_1 \le D_1] + P[D_1 < I_1 \le D_1 + D_a + D_b \& I_1 + I_2 \le D_{all}]\Big) \\
&= p\Big(1 - P[D_1 < I_1] + P[D_1 < I_1 \& I_1 + I_2 \le D_{all}] \\
&\qquad - P[D_1 + D_a + D_b < I_1 \& I_1 + I_2 \le D_{all}]\Big) \\
&= p\Big(1 - P[D_1 < I_1 \& I_1 + I_2 > D_{all}] - P[D_1 + D_a + D_b < I_1 \& I_1 + I_2 \le D_{all}]\Big)
\end{aligned}$$

The above equation can be simplified as below.

$$\frac{p-c}{p} = P[D_1 < I_1 \& I_1 + I_2 > D_{all}] + P[D_1 + D_a + D_b < I_1 \& I_1 + I_2 \leq D_{all}] \tag{4.6.1}$$

$$= P[D_1 + D_a + D_b < I_1] + P[D_1 < I_1 \leq D_1 + D_a + D_b \& I_1 + I_2 > D_{all}] \tag{4.6.2}$$

In (Equation 4.6), the left hand side is Newsvendor ratio. In the right hand side, it is the probability of no stockout in the demands that can be fulfilled by location 1 ($D_{rs(1)}$). This can be seen clearer in Equation 4.6.2: The first term is the probability of $D_{rs(1)}$ completely fulfilled by I_1 alone. The second term is the probability of $D_{rs(1)}$ to be fulfilled by a combination of I_1 and I_2 where D_1 can only be served by I_1 exclusively. Although the expression seems not too complicated, the joint probability in the second term is actually adding a significant issue. We are to derive approximation that further simplifies the result to apply to more complicated cases as follows.

$$\frac{p-c}{p} \approx P[D_1 + D_a + D_b < I_1] + P[D_1 < I_1 \leq D_1 + D_a + D_b] P[I_1 + I_2 > D_{all}] \tag{4.7.1}$$

$$\approx P[D_1 + D_a + D_b < I_1] + (1 - P[D_1 + D_a + D_b < I_1]) P[I_1 + I_2 > D_{all}] \tag{4.7.2}$$

The approximation starts from Equation 4.6.2 and firstly approximate the second joint probability term by taking them independent as in (Equation 4.7.1). Then, assuming high enough fill rate is guaranteed and the demands that can be served by only one FC, D_1 and D_2, is not so significantly larger than pooled demand, $D_a + D_b$, that makes the pooling negligible, we can assume the probability of I_1 not enough to cover D_1 be very small, e.g. $P[D_1 < I_1] \approx 1$. Then we get the second approximation (Equation 4.7.2). Now we can see that there is no joint probability term that is extremely hard to solve and there are only three probabilities left with respect to the equation (Equation 4.7.2) for FC 1 and 2: $P[D_1 + D_a + D_b < I_1]$, $P[D_2 + D_a + D_b < I_2]$, and $P[I_1 + I_2 > D_{all}]$. Furthermore, $P[D_1 + D_a + D_b < I_1] = P[D_2 + D_a + D_b < I_2]$ and once the first two is solved, $P[I_1 + I_2 > D_{all}]$ is automatically set. In other words, finding solution is to find $\eta =$

$P[D_1+D_a+D_b < I_1] = P[D_2+D_a+D_b < I_2]$, that satisfies the approximation equation. It is interesting to find that the approximation is a family of equal fractile solutions which has been found by many inventory literature that embeds pooling in different contexts ([158]; [161]; [152]).

4.3.2 Case extensions

As stressed earlier, the main difference of our model compared to past works (e.g. [152]; [146]; [158]) is that there can be more than one layer of fulfillment feasibility induced by larger number of FCs and responsiveness requirement levels, as illustrated in Figure 4.3. Figure 4.6 shows a simpler demand fulfillment capability structure with 3 FCs.

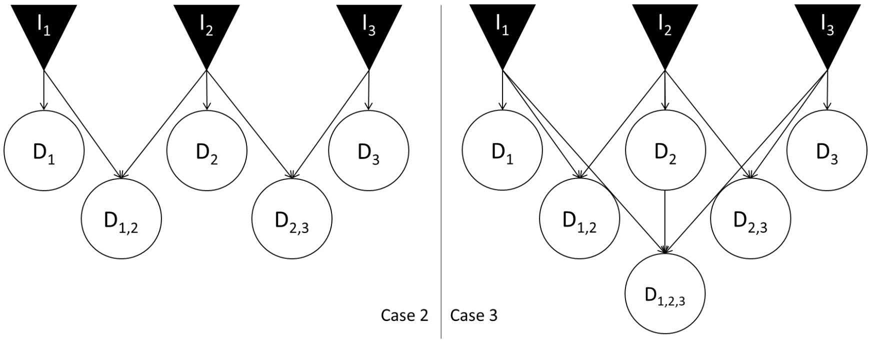


Figure 4.6: 3-FC Cases: I_f is an inventory level at FC f and D_u is demand segments that can be served by a subset of FCs, u

Recall (Equation 4.6), which show that for FC i, the optimal stock quantity is to satisfy $P[\text{no stockout in demand segments that can be fulfilled by } i, D_{rs(f)}] = \frac{p-c}{p}$. However, writing the optimality condition even for the simplest extension, such as case 2 in Figure 4.6, is not straightforward. For example, in case 2, there can be a chance to fill a unit demand in $D_{1,2}$ or in $D_{2,3}$ from I_2, which makes it impossible to write the equation although the it makes no difference for the objective value. That is, some priority rules must be made. Exclusive demand segments, D_f, or demand segments that can be served by a lower number of FCs, will be prioritized as we assumed in the W-case. Also, to write the optimality condition for I_1, we assume that I_2 will prioritize $D_{2,3}$ over $D_{1,2,3}$, which results in a more conservative I_1. With the priority rules, the optimality condition w.r.t. I_1 for case 2 can be written as Equation 4.8.1.

$$\begin{aligned}
\frac{p-c}{p} &= P[D_1 + D_{1,2} \leq I_1] \\
&\qquad + P\begin{bmatrix} D_1 \leq I_1 < D_1 + D_{1,2} \\ \& I_1 + [[I_2 - D_2]_+ - [D_{2,3} - [I_3 - D_3]_+]_+]_+ \geq D_1 + D_{1,2} \end{bmatrix} && (4.8.1) \\
&\leq P[D_1 + D_{1,2} \leq I_1] + P\begin{bmatrix} D_1 \leq I_1 < D_1 + D_{1,2} \\ \& I_1 + (I_2 + I_3 - D_2 - D_{2,3} - D_3) \geq D_1 + D_{1,2} \end{bmatrix} && (4.8.2) \\
&\approx P[D_1 + D_{1,2} \leq I_1] + P[D_1 \leq I_1 < D_1 + D_{1,2}]P\left[I_1 + I_2 + I_3 \geq D_{all}\right] && (4.8.3) \\
&\approx P[D_1 + D_{1,2} \leq I_1] + (1 - P[D_1 + D_{1,2} \leq I_1])P\left[I_1 + I_2 + I_3 \geq D_{all}\right] && (4.8.4)
\end{aligned}$$

Then, Equation 4.8.2 relaxed $[x]_+$ to x. In the W-case, it was not necessary because the joint probability guaranteed it implicitly due to the low complexity. Note that the assumption made here is similar to the ones made in the W-case. $P[I_f \geq D_f] \approx 1$ for each FC f and $P[I_2 + I_3 \geq D_2 + D_3 + D_{2,3}] \approx 1$. In other words, for any subset of demands and FCs with a higher fulfillment priority, there is an assumption for a sufficient inventory. We will refer to this assumption hereafter as the *'local sufficiency assumption'*. Following the approximation to Equation 4.8.4 is same to the W-case approximations.

Similar argument can be made for the optimality condition for I_1 in case 3, which results in (Equation 4.9). Step-by-step equations (Equation C.1) are included in section C.2 in the Appendix. Note that I_3 is symmetric to I_1 in case 2 and 3.

$$\frac{p-c}{p} \approx \begin{array}{l} P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] \\ +(1 - P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1])P[I_1 + I_2 + I_3 \geq D_{all}] \end{array} \qquad (4.9)$$

Solving for I_2 in case 2 also follows similar steps. For simplicity, here we denote demands that can be fulfilled by depot 2 as $D_{2*} = D_2 + D_{1,2} + D_{2,3}$.

$$\frac{p-c}{p} = P[D_{2*} \le I_2] + P\begin{bmatrix} D_2 \le I_2 < D_{2*} \\ \& I_2 + [I_1 - D_1]_+ \ge D_2 + D_{1,2} \\ \& I_2 + [I_3 - D_3]_+ \ge D_2 + D_{2,3} \\ \& I_2 + [I_1 - D_1]_+ + [I_3 - D_3]_+ \ge D_{2*} \end{bmatrix} \tag{4.10.1}$$

$$\le P[D_{2*} \le I_2] + P\begin{bmatrix} D_2 \le I_2 < D_{2*} \\ \& I_2 + I_1 - D_1 \ge D_2 + D_{1,2} \\ \& I_2 + I_3 - D_3 \ge D_2 + D_{2,3} \\ \& I_2 + I_1 - D_1 + I_3 - D_3 \ge D_{2*} \end{bmatrix} \tag{4.10.2}$$

$$\approx P[D_{2*} \le I_2] + P\begin{bmatrix} D_2 \le I_2 < D_{2*} \\ \& I_2 + I_1 + I_3 \ge D_{all} \end{bmatrix} \tag{4.10.3}$$

$$\approx P[D_{2*} \le I_2] + (1 - P[D_{2*} \le I_2])P\left[I_2 + I_1 + I_3 \ge D_{all}\right] \tag{4.10.4}$$

The first relaxations (Equation 4.10.2) and (Equation 4.10.3) are based on local sufficiency assumption. Then, the joint probability is relaxed as two independent probability. Equation 4.10 is directly applied to case 3 simply by setting $D_{2*} = D_2 + D_{1,2} + D_{2,3} + D_{1,2,3}$.

The W-case results and the case extensions results described in this section give intuitive and direct extension to general cases. Using the general notations introduced in section 4.3 and D_{all} to denote the sum of demand over all regions and responsiveness levels, the generalized W-solution for the general model, (Equation 4.1), can be written as follows.

$$\frac{p-c}{p} \approx P[D_{rs(f)} < I_f] + (1 - P[D_{rs(f)} < I_f])P[\sum_{i \in FC} I_i > D_{all}] \; \forall f \in FC \tag{4.11}$$

We will refer the solution in Equation 4.11 as **W-solution** throughout the chapter.

4.3.3 W-heuristic

Here, a heuristic to calculate the solution using the equations found in previous sections is presented. First, consider the functions g_f for all $f \in FC$ below where $I = (I_f \; \forall f \in FC)$, $I_{all} = \sum_{i \in F} I_i$, and f and F is probability and distribution function respectively. In fact, finding solution is equivalent to find solution that makes function $g = 0$.

$$\begin{aligned} g_f(I) &= P[D_{rs(f)} < I_f] + (1 - P[D_{rs(f)} < I_f])P[\sum_{i \in F} I_i > D_{all}] - \frac{p-c}{p} \\ &= F_{rs(f)}(I_f) + (1 - F_{rs(f)}(I_f))F_{all}(I_{all}) - \frac{p-c}{p} \end{aligned}$$

Note that for $f, q \in FC$ such that $f \neq q$,

$$\begin{aligned} \frac{\partial g_f}{\partial I_f} &= f_{rs(f)}(I_f)(1 - F_{all}(I_{all})) + f_{all}(I_{all})(1 - F_{rs(f)}(I_f)) \\ \frac{\partial g_q}{\partial I_f} &= f_{all}(I_{all})(1 - F_{rs(f)}(I_f)) \end{aligned}$$

Because $f \geq 0$ and $0 \leq F \leq 1$ due to the property of probability and distribution functions, g is monotonically increasing function of I_f. Also, since $\frac{\partial I_f}{\partial F_{rs(f)}} \geq 0$, g is monotonically increasing function of $F_{rs(f)}$ and the equal-fractile structure of the solution requires $F_{rs(f)}(I_f) = q$ for all f. Thanks to these properties, an efficient binary search using q as a search variable can find the solution as shown in algorithm 2. In algorithm 2, the overage and underage costs are used for generality. The line 5-line 8 keep the ratio of each inventory value obtained with the minimum f value and let the total inventory in the network to be determined as a single Newsvendor solution. This is needed when the demand distribution is bounded, as well as to compensate the practical limit of computational ability. Similarly, the line 9-line 12 scale the solution while keeping the ratio obtained with the maximum f when the upper bound of g is below zero. The line 31-line 35 is where the lower bound is forced that corresponds to the local sufficiency assumption made for the approximation.

However, this potentially adds significant computational complexity as the number of subsets grows exponentially to the number of FCs: $2^{|FC|}$. Therefore, in the line 29, we let the maximum size of checked subsets to be bounded with k, i.e., $|S_{FC}| \leq k|FC|^k$.

Algorithm 2: w-heuristic - binary search based heuristic for calculating w-solution in Equation 4.11

```
   Input : Demand distributions: PDFs (or PMFs) $f$ and CDFs (or CMFs) $F$
           Fulfillment centers FCs and fulfillment eligibility $rs(f)$
           Cost parameters (overage and underage cost): $c_o, c_u$
   output: $I$
1  Newsvendor Ratio $R = \frac{c_u}{c_o + c_u}$
2  Initialization: $I^l_f = F^{-1}_{rs(f)}(l), I^u_f = F^{-1}_{rs(f)}(u)\ \forall f \in FC$ where $l, u$ are lower and upper probability
     computation bounds
3  Calculate $g_l = l + (1-l)F_{all}(I^l_{all}) - R$ where $I^l_{all} = \sum_{f \in FC} I^l_f$
4           and $g_u = u + (1-u)F_{all}(I^u_{all}) - R$ where $I^u_{all} = \sum_{f \in FC} I^u_f$
5  if $g_l > 0$ then
6      $I_f \leftarrow F^{-1}_{all}(R)\frac{I^l_f}{I^l_{all}}\ \forall f \in FC$
7      Go to line line 31
8  end
9  if $g_u < 0$ then
10     $I_f \leftarrow F^{-1}_{all}(R)\frac{I^u_f}{I^u_{all}}\ \forall f \in FC$
11     Go to line line 31
12 end
13 while $|g_u - g_l| > 0$ do
14     Let $q = \frac{l+u}{2}$
15     $I^q_f = F^{-1}_{rs(f)}(q)\ \forall f \in FC$
16     Calculate $g_q = q + (1-q)F_{all}(I^q_{all}) - R$
17     if $g_q == 0$ then
18         $I_f \leftarrow I^q_f\ \forall f \in FC$
19         break and go to line line 31
20     else if $g_q < 0$ then
21         Update $I^l_f \leftarrow I^q_f\ \forall f \in FC$
22     else
23         Update $I^u_f \leftarrow I^q_f\ \forall f \in FC$
24     end
25 end
26 if $I$ is NULL then
27     $I_f \leftarrow \frac{I^l_f + I^u_f}{2}\ \forall f \in FC$
28 end
29 Define $S_{FC} = \{S \subset FC | \exists D_{rs}$ s.t. $f(rs) = S; |S| \leq k$; sorted in ascending order of $|S|$ $\}$
30       and $D_S = \sum D_{rs}$ s.t. $f(rs) \subseteq S$
31 for $S$ in $S_{FC}$ do
32     if $F_{D_S}(\sum_{f \in S} I_f) < R$ then
33         Update $I_f \leftarrow F^{-1}_{D_S}(R)\frac{I_f}{\sum_{f \in S} I_f}\ \forall f \in S$
34     end
35 end
36 Update and return $I = (I_f \leftarrow \lfloor I_f \rceil\ \forall f \in FC)$
```

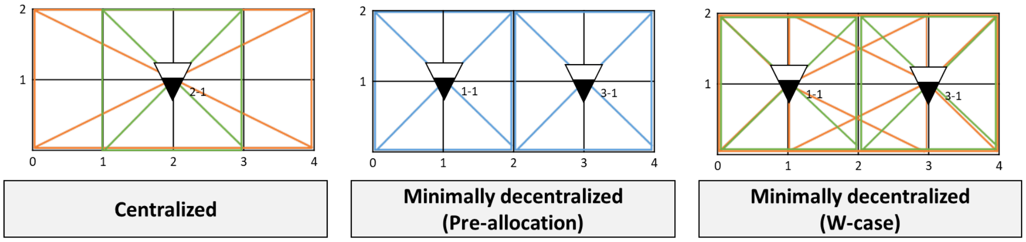
4.4 Numerical Experiment

The two main goals of the numerical investigation are: (1) to show the benefit of the decentralized network under high responsiveness and (2) to evaluate the effectiveness of the W-solution obtained in section 4.3 under partial pooling induced by responsiveness requirement. In other words, we are interested to investigate if the decentralized network is actually more profitable under tight responsiveness requirements, and how important is it to consider the partial pooling to deploy inventories over the network. For the second goal, the W-solution is compared to a simple pre-allocation model. In addition, we also aim to numerically evaluate the proposed heuristics. Therefore, in this section, we are to answer the questions via computational experiments using different settings of the demand distribution and costs. Normal distribution is used which is the most commonly used distribution for evaluating inventory models and due to its good theoretical properties, e.g. a convolution of independent normal distributions is also a normal distribution. Throughout the experiment in this section, the unit sales price and the purchase cost are set to 105 and 5 respectively and therefore the unit overage and underage cost are set to 5 and 100, unless set differently for a sensitivity analysis. The bound for the local sufficiency guarantee is set to $k = 1$ (refer line 29 in algorithm 2). Also, we used $l = 0.0001$ and $u = 0.9999$ as lower and upper bounds to calculate the probability and in line line 13, we replaced 0 with $1e^{-4}$ to ensure a reasonable convergence. Each obtained solution is evaluated for key performance indices (KPIs) such as the expected cost, the profit, and the fill rate via simulation over independently generated demand samples.

The section is organized as follows. Firstly, the proposed W-solution and W-heuristic (algorithm 2) are evaluated meticulously under the W-case (2-FC network) by varying the mean and variance of selected demand segments or randomizing demand distributions. Secondly, we investigate if the results of the W-case can be extended to more generalized cases such as a 10 by 10 grid model. Lastly, we present the case study results demonstrating the scalability of the results and the behavior over empirical demand distributions.

4.4.1 W-case experiment

We first start with tackling the W-case. For the experimental design, three coverage cases are considered as in Figure 4.7. There are eight regions and two responsiveness levels. Each region is defined, or named, by the lower left coordinate and each fulfillment center is named with its coordinate. For example, the region in the upper right corner is region 1-3, denoted as 'r1-3'. Similarly the central FC is FC2-1. Assume it takes one time period to move one region in the grid and travel time from a FC to a region is defined as the Manhattan distance from the FC to the center of the region. For instance, the travel time from FC1-1 to region r0-1 is $|1-0.5|+|1-1.5|=1$. Also, the responsiveness level s requires a delivery in s time period. In Figure 4.7, the coverage from each FC within time periods 1 and 2 are shown in green and orange respectively. Blue indicates pre-allocation of regions to the nearest FC. There can be one central FC or 2 FCs which is the minimum number of FCs to have a full coverage. With 2 FCs, the W-case considers partial pooling. That is, demand with responsiveness level 2 in regions r1-1, r2-1, r1-0 and r2-0 can be served by any FCs. On the other hand, the pre-allocation case exclusively assign regions to the nearest FCs and do not utilize the partial pooling. In this case, any demand satisfied is served with the minimum delivery distance.



Centralized

Minimally decentralized (Pre-allocation)

Minimally decentralized (W-case)

Figure 4.7: Three coverage cases for W-case experiment

Let D_{risj} be the demand from region i that require the responsiveness level j. In a link to Figure 4.4, D_1 in Figure 4.4 is equivalent to $D_{r0-1s1}+D_{r1-1s1}+D_{r0-0s1}+D_{r1-0s1}+$

$D_{r0-1s2} + D_{r0-0s2}$ and D_a is equivalent to $D_{r1-1s2} + D_{r1-0s2}$. Similarly, D_2 and D_b in Figure 4.4 are equivalent to $D_{r2-1s1} + D_{r3-1s1} + D_{r2-0s1} + D_{r3-0s1} + D_{r3-1s2} + D_{r3-0s2}$ and $D_{r2-1s2} + D_{r2-0s2}$ respectively. Additionally, for simpler notations, we denote all demands that can be covered by FC i as D_{iall}, e.g. $D_{1all} = D_1 + D_a + D_b$ using notations from Figure 4.4. Also, $D_{iassign}$ denotes all demands that are pre-allocated to FC i, e.g. $D_{1all} = D_1 + D_a$.

Note that the coverage cases above can be applied to inventory and fulfillment solutions separately. For example, applying pre-allocation in inventory solution means that demands are allocated to the FCs and independent Newsvendor problems are solved to determine inventory level at each FC. This does not necessarily require same pre-allocation coverage for fulfillment. That is, once demand is realized, demands are allowed to be fulfilled from any FC as long as responsiveness requirements are met (e.g. spillover). Accordingly, five solutions are investigated in the experiment.

The Table 4.3 compares the solutions with respect to FC network and coverages for inventory solution and fulfillment. The difference between the pre-allocation and pre-allocation with spillover is a fulfillment flexibility. In other words, both have same inventory level at each FC, calculated from pre-allocation that solves independent Newsvendor given the pre-allocation, but W-case allows spillover at fulfillment (flexible fulfillment) while pre-allocation only allows to serve demands that are allocated to each FC. W-solution is the solution calculated from W-heuristic, e.g. algorithm 2. W-solution without lower bound is the solution from algorithm 2 without line 31-line 35. This is a direct solution from Equation 4.7.2. Hereafter, subscripts c, a, aso, w, and $wnlb$ stand for centralization, pre-allocation, pre-allocation with spillover, W-solution, and W-solution with no lower bound respectively. For example, the total inventory in the network under pre-allocation will be $Itotal_a$, so does the total inventory under pre-allocation with spillover $Itotal_a = Itotal_{aso}$. The total cost under these pre-allocation and pre-allocation with spillover will be $cost_a$ and $cost_{aso}$ respectively. For extended cases, the pre-allocation and w-case for inventory solution coverage can be interpreted as no-pooling and partial pooling. Similarly, the pre-allocation and w-case for fulfillment coverage can be interpreted as inflexible fulfillment and flexible fulfillment respectively. The W-solution with and without lower bound with respect to varying demand distributions is shown in subsection C.3.1 in the Appendix.

Table 4.3: Solution comparison by fulfillment network and coverage for inventory solution and fulfillment for W-case experiment

	FC Network	Inventory Solution Coverage	Fulfillment Coverage
Centralization	Centralized	Centralized	Centralized
Pre-allocation	Decentralized	Pre-allocation	Pre-allocation
Pre-allocation with Spillover	Decentralized	Pre-allocation	W-case
W-solution	Decentralized	W-case	W-case
W-solution with no Lower Bound	Decentralized	W-case	W-case

To observe how the different coverage and solutions perform under certain distributions, three phases of experiments and sensitivity analysis on margin have been conducted. First is to vary the mean and the coefficient of variance (cov) of a demand segment that can only be served by FC 1 or central FC: D_{r1-1s1}. Second is to vary the mean and the variance of a demand segment that can be served by both FC 1 and FC2 or central FC: D_{r1-1s2}. Lastly, we randomly vary the mean and the variance of demand segments. The three cases are conducted with the default costs, $c_o = 5, c_u = 100$. Then, we repeat the first and the last experiments with $c_o = 100, c_u = 25$ as a sensitivity analysis. The default demand distribution has been set up as follows: $D_{r0-1s1}, D_{r1-1s1}, D_{r2-1s1}, D_{r3-1s1}, D_{r1-1s2}, D_{r2-1s2}$ follow Normal with mean of 50 and cov of 0.2. The others are set to zero. With the default distribution, $D_i, D_{iassign}, D_{iall}$ follow $N(100, \sqrt{200}), N(100, \sqrt{200}), N(150, \sqrt{300}), N(200, 20)$ respectively for both $i = 1, 2$. Demand under coverage of central DC follows $N(200, 20)$.

Figure 4.8 shows the total inventory in the network by solutions when varying the mean and the cov of D_{r1-1s1}. $Itotal_w$, $Itotal_wnlb$, $Itotal_c$, and $Itotal_a$ represents the total inventory in the network from W-solution, W-solution with no lower bound, centralization and preallocation respectively. Preallocation with spillover is omitted because $Itotal_{aso} = Itotal_a$. For all cases, it is shown that centralization has significantly a smaller inventory, which is not because of efficiency obtained with better pooling but rather because of losing demands requiring tight responsiveness from regions 1, 4, 5 and 8. Pre-allocation tends to require more inventory and this is because it is ignoring the pooling. Lastly, in most cases, the W-solution and the W-solution w/o LB result in the same inventory but it deviates when the mean of D_{r1-1s1} becomes much higher than the others.

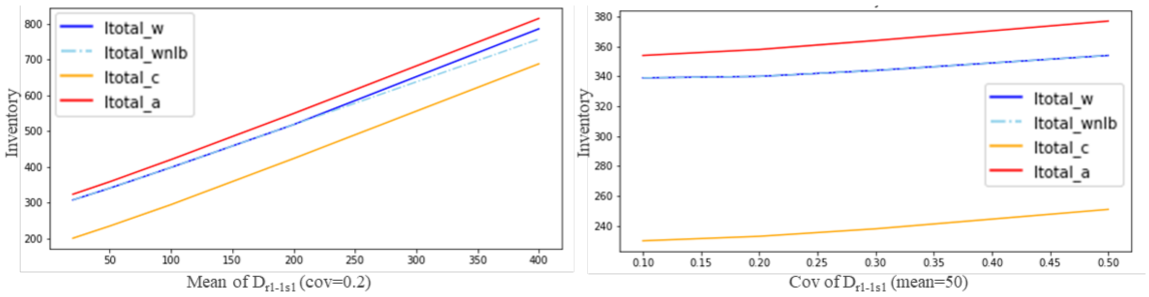


Figure 4.8: Total inventory by mean and cov of D_{r1-1s1}

Figure 4.9 plots the expected cost by solution when varying the mean and the cov of D_{r2s1}. $cost_w$, $cost_wnlb$, $cost_a$, and $cost_{aso}$ represents the total inventory in the network from W-solution, W-solution with no lower bound, preallocation and preallocation with spillover respectively. In the plot of the expected cost, the centralized coverage is excluded for better readability as it is significantly higher than the other solutions. As explained, because the centralized network is not capable of serving demands requiring tight responsiveness from regions 1,4,5 and 8, these lost demands incur high opportunity costs. In other words, due to the responsiveness requirements, the cost of lost demands surpass the benefit of extra pooling. In all cases, including both the centralized and decentralized cases, the W-solution outperforms the other solutions which demonstrates the importance of utilizing the partial pooling properly. Also, although the pre-allocation solution results in higher costs than the W-solution, it can be seen that allowing spillover can reduce costs significantly given the pre-allocation solution (e.g. $cost_a \geq cost_{aso}$). This support the benefit of fulfillment flexibility.

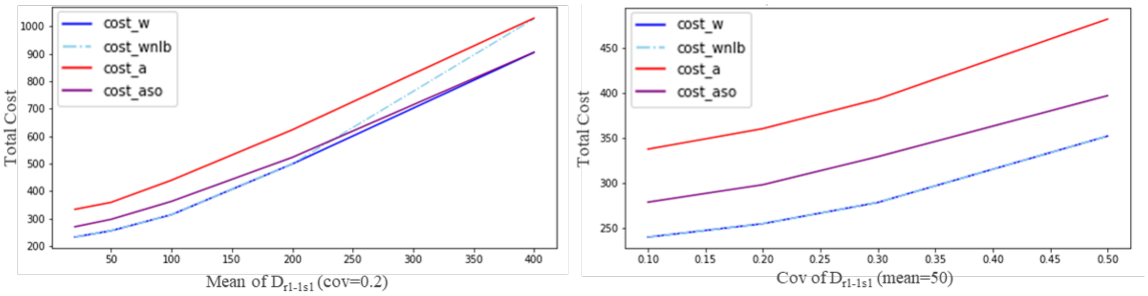


Figure 4.9: Expected cost by mean and cov of D_{r1-1s1}

Besides the economic performances, the demand fill rate is an interesting performance measure. Expected demand fill rate by mean of D_{r1-1s1} and fill rate per unit inventory by total inventory the network are plotted on each side of Figure 4.10 respectively. Recall that the fill rate is measured as the fraction of demands that is satisfied directly from stock. The left curve clearly shows the limitations of centralized network to capture demands by offering satisfying responsiveness, which results in a high opportunity cost. The efficiency curve also shows the value of the distributed network over the centralized one while there is no significant difference between the variations of the distributed network, in line with the results in Figure 4.8 and Figure 4.9.

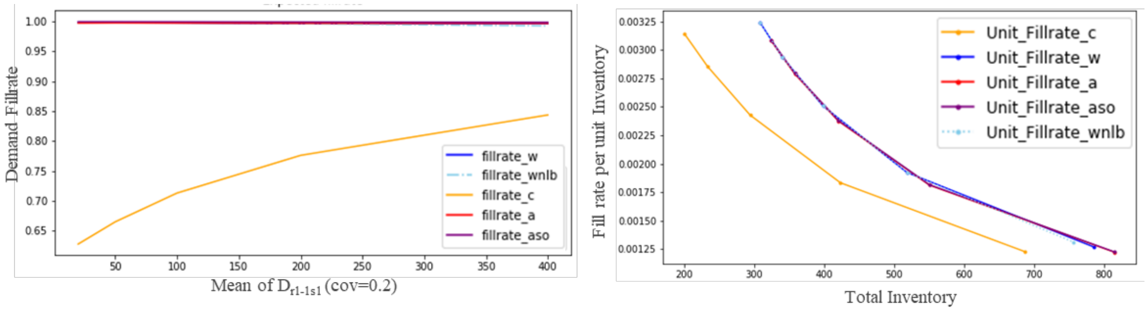


Figure 4.10: Expected fill rate by mean of D_{r1-1s1} (left) and the fill rate per unit inventory (right) when varying the mean of D_{r1-1s1}

The results of the second experiment, by varying mean and cov of D_{1-1s2} are similar to the results of the first experiment, as shown in Figure 4.11 and Figure 4.12. Figure 4.11 shows the total inventory in the network by varying mean and cov of D_{r1-1s2}. Figure 4.12

shows the expected cost by varying mean and cov of D_{r1-1s2} for distributed network. The results are very similar to the first experiment where D_{r1-1s1} was varied, except in this case, the difference between w-case and w-solutions to preallocation is seemed to be smaller and at all cases, w-solution and w-solution with LB resulted in the same inventory levels.

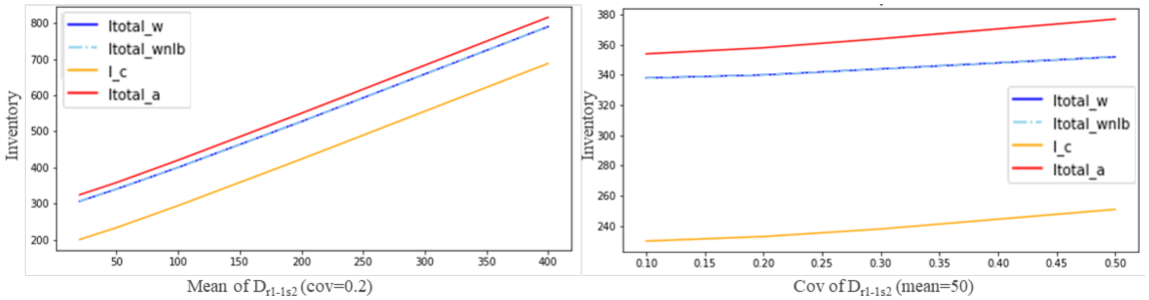


Figure 4.11: Total inventory by mean and cov of D_{r1-1s2}

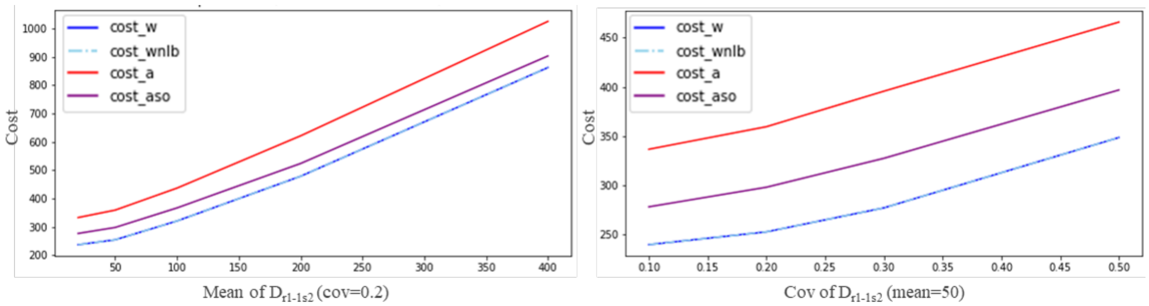


Figure 4.12: Expected cost by mean and cov of D_{r1-1s2}

The randomized experiment is done in the following manner. Starting from the default demand distribution, each demand segment, by region and responsiveness level, is revisited and changed with probability of 0.3. Once it is chosen for a random change, the Normal distribution with mean$\sim Uniform[20, 170]$ and cov$\sim Uniform[0.1, 0.4]$ is randomly assigned. The number of demand segments to get the normal distribution with random parameters are also chosen randomly to maximize the impact of randomization. The Table 4.4 summarizes the results as a percentage change of the total inventory and the expected cost, profit and fill rate with respect to W-solution. The results show that the centralized solution is the least profitable and the cost of lost demand is very high. From the expected fill rate, it can be seen that centralized solution is losing almost 30% of the demand than decentralized solutions due to its limited capability to offer responsiveness. As mentioned earlier this is due to the high margin and we are to make comparison to the low margin case through the sensitivity analysis. Under the decentralized network, the W-solution results in less inventory than the pre-allocation with a lower cost and slightly a higher profit. The W-solution w/o LB performs similar to the W-solution, but the W-solution brings extra cost saving.

Table 4.4: Relative KPIs w.r.t. W-solution from W-case random experiment

	Total Inventory	Cost	Profit	Fill rate
Centralized	-26.3 %	2930.3%	-28.9%	-29.1%
Pre-allocation	5.8%	33.5%	-0.3%	0.0%
Pre-allocation w/ Spillover	5.8%	11.1%	-0.1%	0.2%
W-solution w/o LB	0.0%	0.1%	0.0%	0.0%

The efficiency curve is shown in Figure 4.13. The curve for centralized solution is fluctuating due to the fluctuations in demands that cannot be served by the central location. However, the trend is clear that efficiency of centralized solution is inferior to others which results in higher opportunity cost.

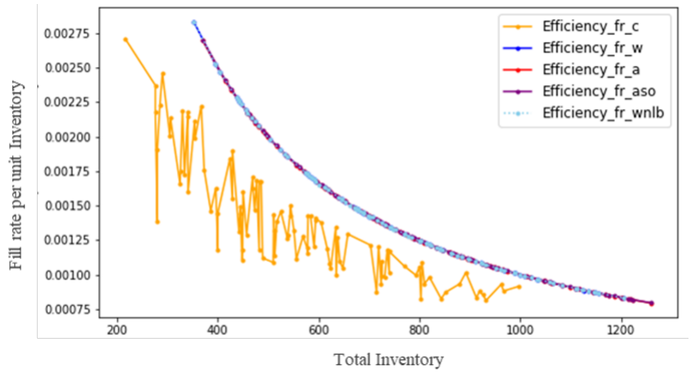


Figure 4.13: Efficiency curve with respective to fill rate per inventory and total inventory for W-case with random variation

The results of sensitivity analysis with respect to margin are presented from here. One of the following questions to answer is if w-solution works well when margin is not as high as the default. The sensitivity analysis has been conducted by repeating the first and third w-case experiments assuming low margin, where $c_o = 100, c_u = 25$. We begin with presenting the first experiment results.

Figure 4.14 shows the total inventory in the network by solution when mean and cov is varied on D_{r1-1s1}. Unlike the default case where margin is very high (refer Figure 4.8), w-solution results in higher inventory level in total at all cases.

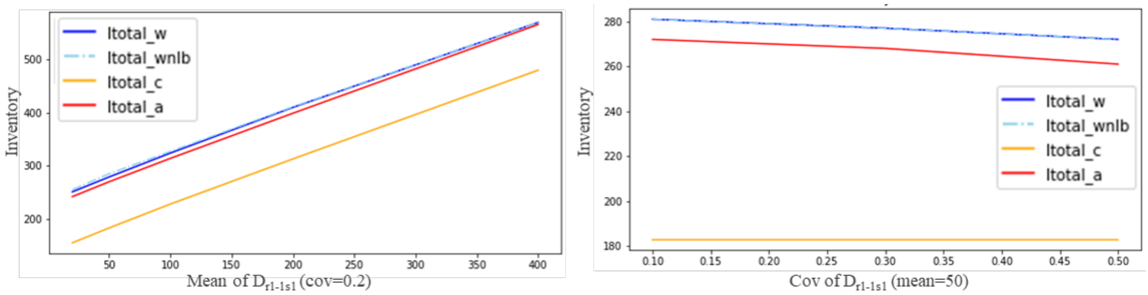


Figure 4.14: Total inventory by mean and cov of D_{r1-1s1} under low margin where $c_o = 100, c_u = 25$

However, w-solution remains to be the most profitable solution as shown in Figure 4.15. That is, preallocated model often overstock or understock by ignoring partial pooling effect. In other words, it often ignores the capability for the allocated demand to be partially served from other FCs, resulting in overstock, or ignores the capability for the FC to serve part of demands allocated to other FCs, resulting in understock. It is rather trivial that centralized model is losing demand due to limited capability to satisfy all responsiveness levels, as in the previous case.

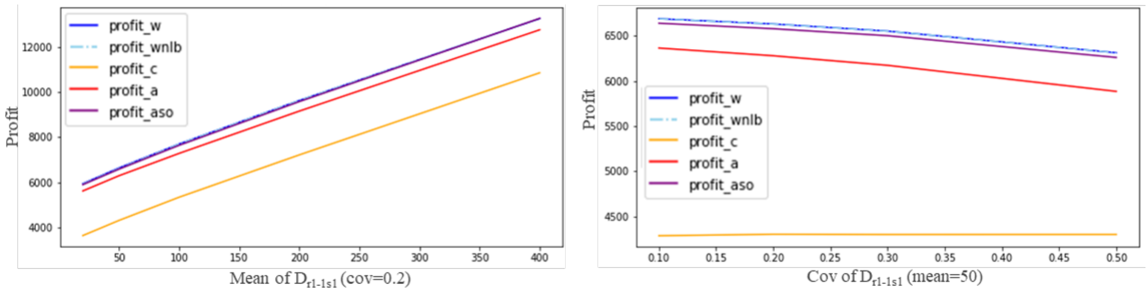


Figure 4.15: Expected profit by mean and cov of D_{r1-1s1} under low margin where $c_o = 100, c_u = 25$

Under the randomized experiment, w-solution tend to have more inventory in the network when the margin is low while resulting in lower cost and higher profit. The sensitivity results with randomized demand are summarized in Table 4.5. In short, the results show that the W-solution remains the most profitable and the least costly solution and the centralized network remains more costly than the decentralized network under a low margin as well. However, as can be seen by comparing Table 4.4 and Table 4.5, the cost gap between the centralized solution and the decentralized solutions is significantly smaller when the margin is low. Because in both cases, centralized solution is losing about 30% of the demand than W-solution, the gap decrease is due to the reduced cost of the lost demands under low margin. We also observe that, with a low margin, the W-solution results in a higher inventory level in the network than the pre-allocation solution as shown in Table 4.5. This observation will be discussed in more details in the next section.

Table 4.5: Relative KPIs w.r.t. W-solution from W-case random experiment under low margin where $c_o = 100, c_u = 25$

	Total Inventory	Cost	Profit	Fill rate
Centralized	-31.0%	206.7%	-32.6%	-31.9%
Pre-allocation	-2.8%	34.0%	-5.3%	-3.3%
Pre-allocation w/ Spillover	-2.8%	4.0%	-0.6%	-2.4%
W-solution w/o LB	0.0%	0.0%	0.0%	0.0%

In short, w-solution remains the most profitable and the least costly model under low margin as well. The cost gap between centralized solution and decentralized solution has been decreased significantly because the total cost of lost demand by centralized system is smaller when margin is low. We also observed that, with low margin, w-solution results in higher inventory level in the network than preallocation. This observation will be discussed more in detail in the next section.

4.4.2 Extended case experiment

The more generalized experiment is conducted on 10*10 grid network with decentralized network with 25 FCs as shown in Figure 4.16. Based on the results of previous experiments that demonstrate the benefit of having the lower bound for the W-solution, hereafter we omit the W-solution w/o LB. The definition of demand regions as well as travel time remains the same. Here, we assume 5 responsive levels, s1, s2, s3, s5 and s9, each requires delivery in 1, 2, 3, 5 and 9 time units, respectively. For instance, it can represent 4-hour, same-day, next-day, 2-day, and 5-day delivery options. The coverage from the FCs in the middle, FC5-5, is shown in color on the right side in Figure 4.16. The coloring scheme is from very prompt and urgent in red to no-rush in blue and it will be consistently used in the manuscript. In the grid experiment, we did not fix the location of the centralized FC. Instead, for each experiment one FC is chosen that can cover the most demand as a central FC.

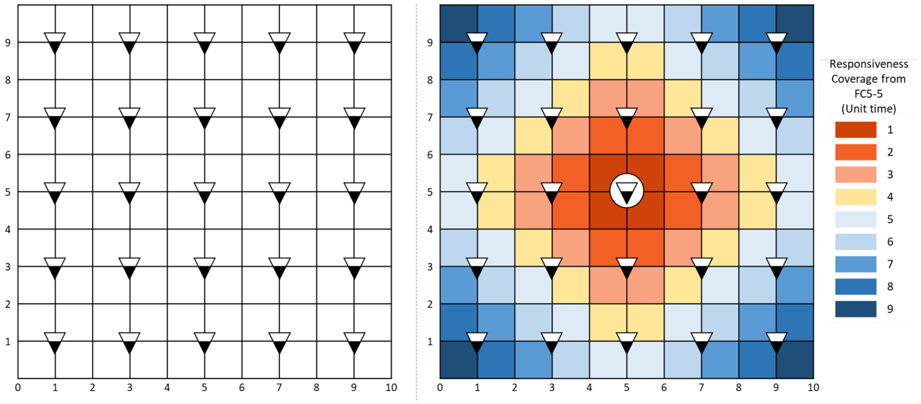


Figure 4.16: 10*10 Grid Network with 25 FCs (left) and coverage from FC5-5 (right)

To investigate how the solution performs when varying the responsiveness expectations, 16 responsiveness expectations scenarios are first set as shown in Figure 4.17. Scenarios 0 and 15 are extreme cases. Scenario 0 enables a complete pooling in the network, as all can be served from FC 5-5. Therefore, from an inventory point of view, the centralization is the most efficient with a minimum cost. In scenario 15, all demand segments can only be served by the nearest FC. Therefore, the pre-allocated solution on distributed network generates the optimal solution. The experiment is conducted under the default (high) margin and a low margin.

Under the default (high) margin, the total inventory using the W-solution is always lower than the pre-allocation solution, except in scenario 15 where the two solutions result in the same inventory as shown in Figure 4.18. The total inventory of the W-solution is

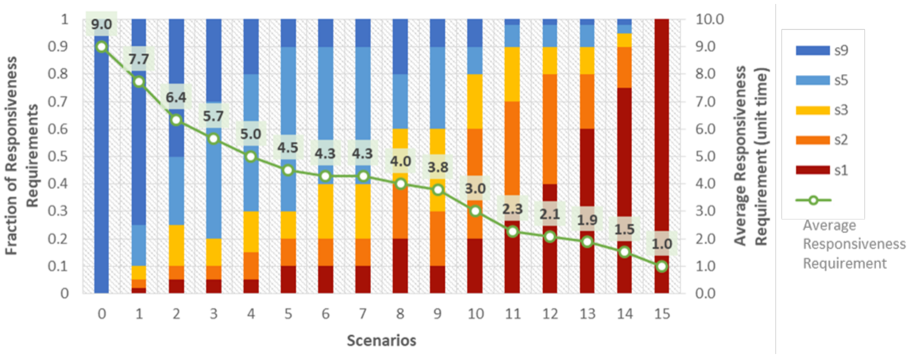


Figure 4.17: Responsiveness expectation scenarios: from slow to fast

the same as that of the centralized inventory under scenario 0. As expected from the total inventory level, the W-solution generates the optimal cost and profit in scenario 0 and 15 as the central and the pre-allocation solutions do in each scenario respectively. In scenarios 1-14, the W-solution has a superior performance than all the other solutions. Note that the performance of the centralized solution deteriorates sharply when a tighter responsiveness is required.

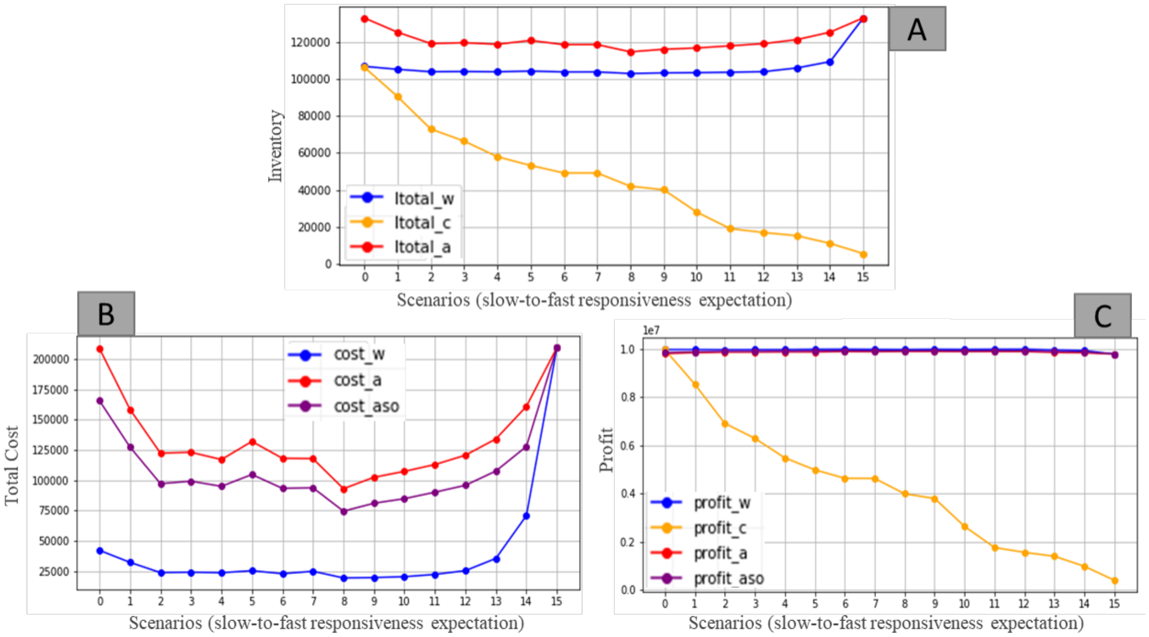


Figure 4.18: Total inventory (A), expected cost (B) and profit (C) by solution and by scenario under a high margin where $c_o = 5, c_u = 100$

Under the low margin, the total inventory using the W-solution is always higher than the other solutions with exceptions under scenario 0 and 15 as shown in Figure 4.19. Consistent to the earlier numerical results in the W-case and the 3-FC case, the W-solution generates more (less) inventory under high (low) margin than the pre-allocation solution. This is rather a counter-intuitive result as intuitively the more are pooled, the less inventories are needed due to variance reduction. We interpret the results as the role of pooling can be seen as a buffer to margin changes. In fact, the pooling has a 2-directional impact: (1) the inventory level in FC A can be reduced knowing that some demands can be covered by FC B and (2) the inventory level in FC A can be increased knowing that part of the inventory in A can be used to serve demands that are usually served by FC B. When the margin is high, the pre-allocation solution stocks up each FC to increase sales in demand segments allocated to it ignoring the potential to serve some of the demand from other FCs. On the other hand, when the margin is low, the pre-allocation solution tends to stock less to avoid the overage cost, which becomes relatively higher compared to the high margin case, ignoring the potential to use some of the excess inventories to serve the demand segments allocated to other FCs. However, the W-solution implies a pooling that is more robust to the margin change since the pooling works as a buffer to the change.

Although the plots of the inventory level were reversed, the relative cost and profit among the models remain the same to that of the case of the high margin as shown in plots B and C in Figure 4.19. With the low margin, the impact of lost demand due to the responsiveness requirement is relatively smaller, so the profit of the centralized inventory is rather comparable to the decentralized total inventory.

4.4.3 Case Study

The case study is further experimenting the impact of the decentralized fulfillment network, potentially using open FCs, and the W-solution under such environment using real data from our industry partner. The fulfillment network and historical sales of a selected product in USA are taken for the case study. The open FC network that consists of 186 FCs is set

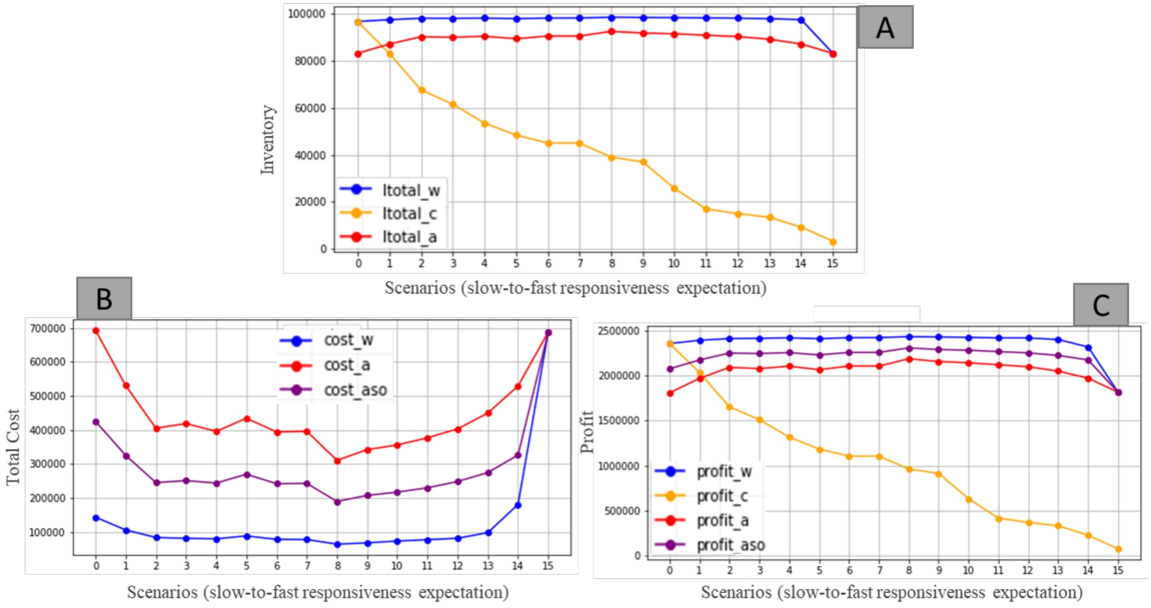


Figure 4.19: Total inventory (A), expected cost (B) and profit (C) by solution and by scenario under a low margin where $c_o = 100, c_u = 25$

using the public information on the FC locations of Amazon and Walmart. The resulting FC network used in the case study is shown in Figure 4.1.

The regions are estimated at the city level and the data includes 2021 cities. Each city is categorized into a metropolitan area or a non-metropolitan area based on its proximity to the 218 US metropolitan cities: a metropolitan area is when the estimated travel time is less than or equal to 2 hours and a non-metropolitan area is otherwise. Figure Figure 4.20 shows the cities by category on the map. In this case study, distance is calculated by multiplying circuity factor to great circle distance. Ballou, Rahardja, and Sakai [190] estimated average circuity factor as 1.2 and 1.21 with standard deviation of 0.16 and 0.17 in US East and West respectively. Noting that the circuity factor tend to increase as distance decreases and it tend to be larger within urban area ([191]; [192]), we used average plus one standard deviation of average US circuity factor, $\frac{1.2+1.21}{2} + \frac{0.16+0.17}{2} = 1.37$ to estimate travel distance from great circle distance. Travel time is estimated from travel distance assuming average travel speed is 40 mph.

Five responsive levels are considered: 4-hr, Same-day, Next-day, 2-day, and Next-week.

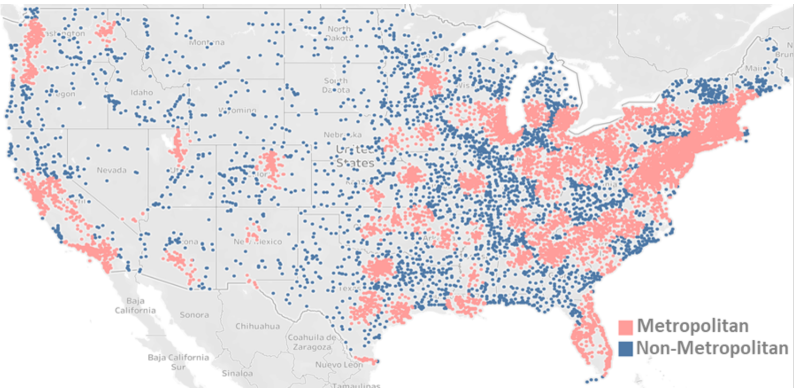


Figure 4.20: Extended metropolitan area and others on US map

The first 4 responsive levels require the travel time to be less than or equal to 2, 4, 10, and 24 hours respectively. Any location can be served from any FCs in the US by next-week. The travel time is shorter than the total delivery time required by the responsive level considering the pick-up time, the delay with traffic and the maximum driving time per day. With current network, the most responsive options are either 2-day or Next-week and 4-hr fulfillment is only feasible in few cities. However, with open FC network, majority of the cities can be fulfilled by next-day except few cities in inland states in Northwest. Moreover, the number of cities where 4-hr fulfillment becomes feasible increases significantly. Figure 4.21 shows the best coverage from the nearest FC in the current and open FC network respectively.

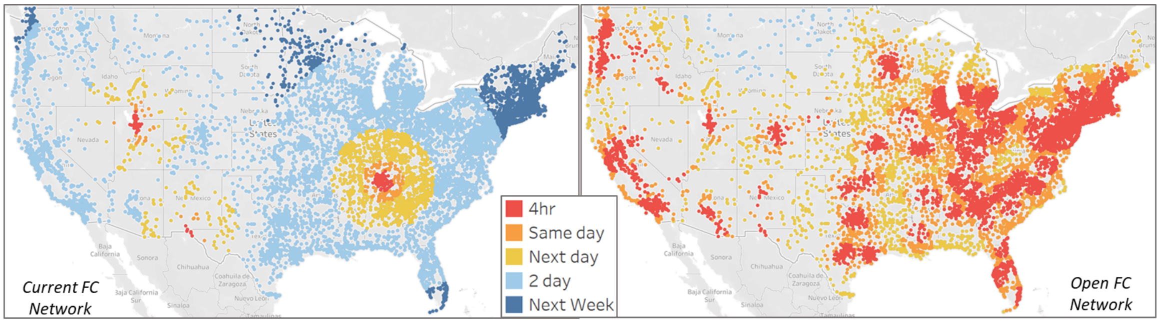


Figure 4.21: Contrasting the achievable time-sensitive coverage from current FC network (left, centralized) and Open FC network (right, decentralized)

For the experiment, 2 responsiveness requirement scenarios are considered as summarized in Table 4.6. For both scenarios, the requirements differ between the metropolitan and non-metropolitan areas. Customers in a metropolitan area expect a higher responsiveness level. In the moderate responsiveness scenario, the majority of customers expect next-day or 2-day delivery, whereas in the tight responsiveness scenario, more customers expect 4-hour or same day delivery. The inventory solutions using pre-allocation with spillover and the W-solution on the current and the open FC networks are compared for each scenario. For the ease of comparison, the results are compared as percentage differences to that of a base case. A centralized case where only next-week delivery is required is used as the base case, which is also an upper bound (best case). The overage and underage cost and Newsvendor ratio from the data are \$45.6, \$22.1, and 0.326 respectively.

Table 4.6: Fraction of demands in metropolitan and non-metropolitan areas by responsiveness requirements under 2 scenarios: Moderate and Tight responsiveness

Scenario	Moderate Responsiveness		Tight Responsiveness	
City Category	Metro	Non-Metro	Metro	Non-Metro
4-hr	0.1	0.01	0.3	0.05
Same-day	0.3	0.09	0.3	0.15
Next-day	0.2	0.2	0.2	0.2
2-day	0.2	0.3	0.15	0.3
Next-week	0.2	0.4	0.05	0.3

The Table 4.7 and Table 4.8 summarize the case study results. For both scenarios, the current FC network is insufficient to offer a satisfactory responsiveness and the limitation becomes more critical with tighter responsiveness requirements. As a result, it only captures $\sim 50\%$ and $\sim 35\%$ of the total demand under moderate and tight responsiveness requirements respectively and this is reflected in the profit directly. Also, due to the limited number of FCs, the pre-allocation and the W-solution have almost no difference under the current network. On the other hand, the open FC network has a significantly higher demand fill rate and profit in both scenarios. Moreover, in the open FC network, the W-solution is superior to pre-allocation in both scenarios. Even with the open FC network, pre-allocation loses $\sim 18\%$ and $\sim 23\%$ of demand in each scenario while W-solution fulfill more than $\sim 90\%$ of demand in both scenarios. The W-solution captures $\sim 5\%$ more profit than pre-allocation in both scenarios. With the moderate responsiveness requirement, it captures $\sim 95\%$ of the maximum profit (the base case profit). With the tight responsiveness requirement, it still captures $\sim 90\%$ of the maximum profit. The Figure 4.22 visualizes pre-alllcation solution and W-solution in the open FC network on the map under the tight responsiveness requirements. It can be seen that W-solution results in more uniform inventory levels at each FC than pre-allocation solution and this is because the demands that can be covered by each FC become similar with pooling.

Table 4.7: KPI changes with respect to the base case (upper bound) under moderate responsiveness requirements (PASO: Preallocation w/ spillover)

	Total Inventory		Demand Fill Rate		Cost		Profit	
	Current	Open	Current	Open	Current	Open	Current	Open
PASO	-48 %	-21%	-48%	-18%	67%	14%	-48%	-10%
W-solution	-51 %	-4%	-50%	-4%	67%	7%	-48%	-5%

Table 4.8: KPI changes with respect to the base case (upper bound) under tight responsiveness requirements (PASO: Preallocation w/ spillover)

	Total Inventory		Demand Fill Rate		Cost		Profit	
	Current	Open	Current	Open	Current	Open	Current	Open
PASO	-65 %	-25%	-65%	-23%	89%	21%	-65%	-16%
W-solution	-66 %	-8%	-66%	-9%	89%	15%	-65%	-11%

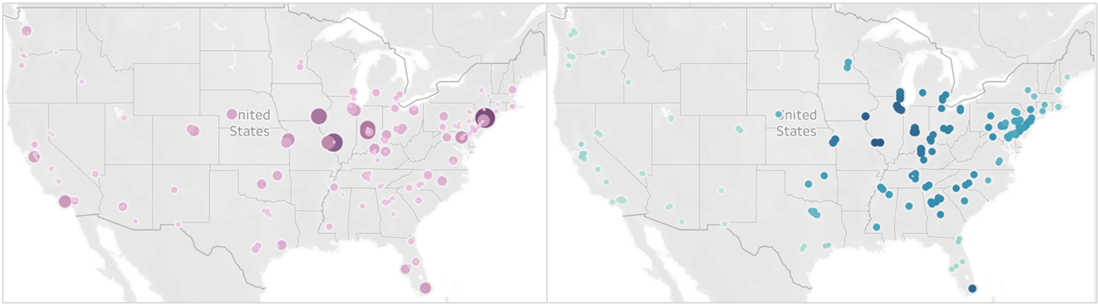


Figure 4.22: Inventory at each open FC from Preallocation (left) and w-solution (right) under tight responsiveness requirement scenario, sized by global relative inventory level and colored by relative level within each solution

4.5 Conclusion

We have presented an inventory problem where demands are partially pooled due to responsiveness requirements, as often observed in the current competitive e-commerce delivery. The problem is modelled as a variant of a Newsvendor model which determines the inventory requirements and allocation at the same time. Then we have derived an approximated solution, the W-solution, followed by a pragmatic heuristic to calculate the solution. The performance of the W-solution has been evaluated through numerical experiments and an empirical investigation based on the case of a drop-ship manufacturer in the US market. The results provide evidence on the importance of inventory models exploiting the pooling structure and the limit of the centralized inventory network under tight responsiveness requirements. Our investigation has also revealed the impact of the margin on the comparative inventory levels and profits.

The results lead to several insights on the inventory management in the market where customers expect a high responsiveness level. Firstly, the results show that the centralized network loses profit sharply as the customer expectation increases. The decentralized network seems to keep more inventory than a centralized network, but it covers more demand. However, it is clear that only a few companies can afford such a large decentralized fulfillment network under a traditional business model. The open FC networks and hyperconnected fulfillment system where the fulfillment centers are operated by third party service providers can open a new opportunity for companies to have access to a decentralized network without a huge capital investment. Analyzing the value of a decentralized network is the first step towards establishing a solid business model for such an open fulfillment network. Secondly, the relative superiority of the W-solution over the pre-allocation one addresses the importance of considering the partial pooling. An interesting observation is that the pre-allocation solution results in more inventory than the W-solution under a high margin whereas it generates less inventory under a low margin while the W-solution constantly costs less. A common intuition is that the optimal inventory level decreases as pooling increases. An interpretation of the rather counter-intuitive observation is that the solutions considering the pooling are more robust to the margin change as the pooling plays as a bumper that absorbs the impact of margin changes. On the other hand, the solution that does not consider pooling, such as the pre-allocation one, is susceptible to margin changes. It tends to overstock when the margin is high ignoring the possibility for the allocated demand to be partially served by other FCs and to understock when the margin is low, ignoring the possibility for its inventory to be used to serve demands allocated to other FCs.

We are to conclude the chapter by addressing the key limitations of the study which leaves the door open to avenues for a future research. The first avenue concerns the single-period nature of the Newsvendor model. Many products are continuously produced and replenished over time requiring a multi-period model. Because the Newsvendor model forms a base of many multi-period inventory models and the results of the study can be a solid building block of multi-period extension. The second avenue is the simplified cost structure. For example, a distance based shipping cost can be studied which can potentially decrease the degree of pooling on inventory as it increases the preference for shipping from a closer location. We believe that combining the inbound and/or outbound transportation costs is a promising further research topic. Again, the future research can be benefited by the results and insights of this chapter to tackle the complexity of the problem efficiently. The third avenue is that we take the fulfillment network as a given parameter and offers a solution for how to operate on the given network. However, it is often a critical question on which FC should be used, especially in the perspective of Physical Internet where a decentralized FC network is accessible via service providers. For example, when a company aims to decentralize its inventory by renting the space via Warehouse Anywhere, how many and which locations does it need? If an omnichannel retailer uses its stores as fulfillment locations as well, does it need to use all stores to get coverage or, if not, which subset of stores should be chosen? It is a problem that is gaining more and more practical and scholarly attentions and we believe that the modeling structure we presented in the chapter and the solution can be used to support the network selection problems. The last avenue for future research concerns the flat pricing of the services. Here we assume the profit does not depend on responsiveness. However, the faster delivery is often charged more and demand is often elastic to price. Although it is not within the scope of the chapter, the service design and pricing is a very interesting and pertinent future research extension.

CHAPTER 5
CONCLUSION

In summary, this thesis has investigated an alternative fulfillment solution, hyperconnected fulfillment, that enables open asset utilization and multi-player operations, facing the current logistics challenges under the prevalence of e-commerce and home delivery combined with growing needs for more responsive deliveries. Firstly, in Chapter 2, we provided academic foundations for the hyperconnected fulfillment system by designing the key system and decision architecture of the hyperconnected fulfillment. The numerical results clearly demonstrated the potential benefit of the hyperconnected fulfillment system by concurrently improving often-conflicting measures, such as induced costs, harmful atmospheric emissions, and customer service. In the following chapters, we turned our focus from a strategic to an operational decision support model. More specifically, we have studied operational models which optimally allocate inventories digitally and physically to maximize the benefit of the hyperconnected fulfillment. Chapter 3 presented the inventory allocation problem to multiple sales outlet under availability promising contract (APC) and retailers' promised availability threshold (PAT). We designed and evaluated three allocation policies and analyzed the impact of the PATs. We showed that the demand is partially pooled due to the customers' retailer substitution and preference, which also makes the per-retailer-demand endogenous. The experimental results revealed the importance of utilizing the pooling especially at a lean availability level. In chapter 4, we investigated the benefit of a decentralized fulfillment network and model a proper inventory ordering and deployment problem under tight responsiveness requirements. The model includes partially pooled risk and inventory structure induced by the responsiveness requirements. We showed the complexity of the model and the intractability of its solution. Then, we derived a pragmatic heuristic solution, named W-solution, and an efficient solution algorithm using the equal-fractile structure of W-solution, named W-heuristic. The numerical study demonstrated the benefit of decentralized fulfillment network and w-solution under responsiveness requirements. Together, it has enabled us to provide insights on the potential benefits of the hyperconnected fulfillment. We presented rigorous methodologies on how to design such a system and how to allocate and deploy inventories effectively under the system. From a practical view, the transition from asset-driven system to service-driven system can open a new business opportunity that a few companies have already started to exploit.

The works in this thesis have their own limitations and they open a rich potential for future research to be built upon them. We have addressed the limitations of each chapter at the end of the corresponding chapter. Here, we list the limitations which are the most relevant to the goal of the thesis, with respect to building and operating the hyperconnected fulfillment system. Firstly, additional study is needed to design the network of open FCs from the perspective of the service provider and how to select which FCs to use from the perspective of users who have the access to the full network. It includes the impact of the market size and/or the flow of products in the open FC network. Our work has taken the open FC network as given and assumed the given network is a subset of the open FC network selected for use. The operational models we presented can be used to support the network selection and expansion models. For example, the W-solution presented in chapter 4 can provide good insights on selecting a subset of open FCs by using it to calculate marginal gains of adding or removing each FC. Also the simulation framework presented in chapter 2 can be used to evaluate the performance of a selected sub-network. Secondly, pricing models are required to properly charge for fulfillment services that benefit both service providers and users. Such a pricing model is also closely related to the optimal subset of open FCs selected for use by each user. Such studies can also deliver critical insights for new businesses leading service-driven fulfillment. The models and insights from chapter 3 can support understanding the dynamics between players. Adding a pricing model is also an interesting extension of the inventory model presented in chapter 4. Thirdly, studies on facility design of such open FCs or hubs are necessary. The complexity, connectivity and dynamics of the open facilities with multi-player operations and on-demand utilization need to be analyzed carefully and suitable facility design needs to be accompanied to enable seamless operations. The cost of intra-facility operations will also affect the pricing models. Fourthly, from the perspective of inventory operations, expanding the scope to supplier and replenishment operations can enable more comprehensive design of inventory flow and bring additional insights. Also, expanding the scope to transportation, from a higher echelon to fulfillment network, between FCs, and from FCs to customers, is one of the most interesting future research avenue.

Despite its own limitations, the thesis contributes to the current literature in logistics and Physical Internet by being the first that explored the potential of the hyperconnected fulfillment system at systematic level and studied inventory operations under the setting revealing complex pooling structure in it under tight responsiveness requirements.

Appendices

APPENDIX A

CHAPTER 2

A.1 Experimental Results: Cost Details

Table A.1: Costs of each scenario ($)

	Labor Cost	Fuel Cost	Vehicle Cost	Inventory Holding Cost	Hub Usage Cost	Total Daily Cost	Total Cost Reduction (%)
Scenario1: DPFs	4,242	1,020	143	1,073	-	6477	
Scenario2: OPF	3,711	796	125	1,063	-	5696	-12%
Scenario3: OPFs	3,456	702	117	1,072	-	5346	-17%
Scenario4: OPFs & OUF	3,459	702	117	1,106	-	5384	-17%
Scenario5: OUH	3,531	679	90	1,068	866	6234	-4%
Scenario6: OUHs	3,001	616	104	1,068	421	5211	-20%
Scenario7: OPFs & OUH	3,559	686	91	1,069	873	6279	-3%
Scenario8: OPFs & OUHs	2,860	483	86	1,067	421	4917	-24%

Table A.2: Expected cost of scenario 1 and 8 by demand distribution ($)

	Demand Distribution	Labor Cost	Fuel Cost	Vehicle Cost	Inventory Holding Cost	Hub Usage Cost	Total Daily Cost	Total Cost Reduction (%)
Scenario1: DPFs	Center-Concentrated	4,242	1,020	143	1,073	-	6477	
	Uniform	4,610	1,175	155	1,068	-	7008	
	Suburb-Concentrated	4,670	1,214	157	1,070	-	7111	
Scenario8: OPFs & OUHs	Center-Concentrated	2,860	483	86	1,067	421	4917	-24%
	Uniform	3,013	537	89	1,066	410	5116	-27%
	Suburb-Concentrated	3,020	542	89	1,071	402	5124	-28%

APPENDIX B

CHAPTER 3

B.1 Linear modeling of lost demand

The demand lost by preferred retailer r due to inventory availability given original demand δ_r is $d_r^- = [\delta_r - [z_r - \theta_r]_+]_+$. To model it as a linear constraints, two indicator variables v_r and b_r are used. Note that there are the four possible cases as in table Table B.1.

Table B.1: Cases for linear d_r^-

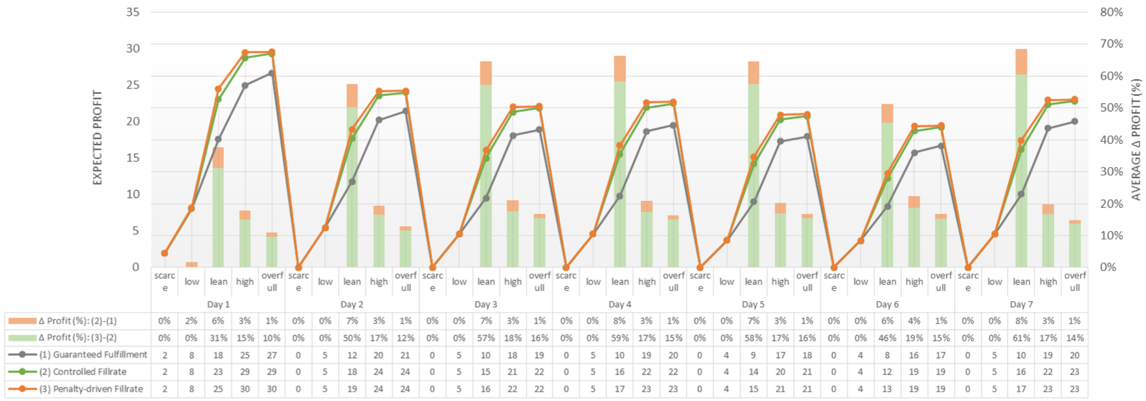
	$[z_r - \theta_r]_+$	v_r	$[\delta_r - [z_r - \theta_r]_+]_+$	b_r	d_r^-
Case 0	$z_r - \theta_r < 0$	$v_r = 1$	$\delta_r - [z_r - \theta_r]_+ \geq 0$	$b_r = 0$	$d_r^- = \delta_r$
Case 1	$z_r - \theta_r < 0$	$v_r = 1$	$\delta_r - [z_r - \theta_r]_+ < 0$	$b_r = 1$	$d_r^- = \delta_r$
Case 2	$z_r - \theta_r \geq 0$	$v_r = 0$	$\delta_r - [z_r - \theta_r]_+ \geq 0$	$b_r = 0$	$d_r^- = \delta_r - (z_r - \theta_r)$
Case 3	$z_r - \theta_r \geq 0$	$v_r = 0$	$\delta_r - [z_r - \theta_r]_+ < 0$	$b_r = 1$	$d_r^- = 0$

For simplicity, case 0 and 1 can be considered as one case, but they are separately described to avoid any confusion here. The 5 constraints in the model (Equation 3.5.16) - (Equation 3.5.20) ensure to achieve the non linear equality constraint for d_r^- under all cases. Note that $d_r^- \geq 0$.

Table B.2: Lost opportunity constraint by cases

Constraint	Case 0	Case 1	Case 2	Case 3
(Equation 3.5.16)	$d_r^- \leq \delta_r$	$d_r^- \leq \delta_r$	$d_r^- \leq M$	$d_r^- \leq M$
(Equation 3.5.17)	$d_r^- \geq \delta_r$	$d_r^- \geq \delta_r$	$d_r^- \geq -M$	$d_r^- \geq -M$
(Equation 3.5.18)	$d_r^- \leq M$	$d_r^- \leq 2M$	$d_r^- \leq \delta_r - (z_r - \theta_r)$	$d_r^- \leq M$
(Equation 3.5.19)	$d_r^- \geq -M$	$d_r^- \geq -2M$	$d_r^- \geq \delta_r - (z_r - \theta_r)$	$d_r^- \geq -M$
(Equation 3.5.20)	$d_r^- \leq 2M$	$d_r^- \leq M$	$d_r^- \leq M$	$d_r^- \leq 0$
Total	$d_r^- = \delta_r$	$d_r^- = \delta_r$	$d_r^- = \delta_r - (z_r - \theta_r)$	$d_r^- = 0$

B.2 Daily expected profit



	Day 1					Day 2					Day 3					Day 4					Day 5					Day 6					Day 7				
	scarce	low	lean	high	overfull	scarce	low	lean	high	overfull	scarce	low	lean	high	overfull	scarce	low	lean	high	overfull	scarce	low	lean	high	overfull	scarce	low	lean	high	overfull	scarce	low	lean	high	overfull
Δ Profit (%): (2)-(1)	0%	2%	6%	3%	1%	0%	0%	7%	3%	1%	0%	0%	7%	3%	1%	0%	0%	8%	3%	1%	0%	0%	7%	3%	1%	0%	0%	6%	4%	1%	0%	0%	8%	3%	1%
Δ Profit (%): (3)-(2)	0%	0%	31%	15%	10%	0%	0%	50%	17%	12%	0%	0%	57%	18%	16%	0%	0%	59%	17%	15%	0%	0%	58%	17%	16%	0%	0%	46%	19%	15%	0%	0%	61%	17%	14%
(1) Guaranteed Fulfillment	2	8	18	25	27	0	5	12	20	21	0	5	10	18	19	0	5	10	19	20	0	4	9	17	18	0	4	8	16	17	0	5	10	19	20
(2) Controlled Fillrate	2	8	23	29	29	0	5	18	24	24	0	5	15	21	22	0	5	16	22	22	0	4	14	20	21	0	4	12	19	19	0	5	16	22	23
(3) Penalty-driven Fillrate	2	8	25	30	30	0	5	19	24	24	0	5	16	22	22	0	5	17	23	23	0	4	15	21	21	0	4	13	19	19	0	5	17	23	23

Figure B.1: Daily expected profit by allocation policy under varying availability level over 7 days

APPENDIX C

CHAPTER 3

C.1 Newsvendor Solutions

C.1.1 Primal solutions for W-case

Table C.1: Primal solution by demand partition

Demand Partition	x_{11}	x_{1a}	x_{1b}	x_{22}	x_{2b}	x_{2a}
Ω_{11}	I_1	0	0	I_2	0	0
Ω_{12}	I_1	0	0	D_2	$I_2 - D_2$	0
Ω_{13}	I_1	0	0	D_2	D_b	0
Ω_{14}	I_1	0	0	D_2	D_b	$I_2 - D_2 - D_b$
Ω_{15}	I_1	0	0	D_2	D_b	D_a
Ω_{21}	D_1	$I_1 - D_1$	0	I_2	0	0
Ω_{22}	D_1	$I_1 - D_1$	0	D_2	$I_2 - D_2$	0
Ω_{23}	D_1	$I_1 - D_1$	0	D_2	D_b	$I_2 - D_2 - D_b$
Ω_{24}	D_1	$I_1 - D_1$	0	D_2	D_b	$D_a - (I_1 - D_1)$
Ω_{25}	D_1	$I_1 - D_1$	0	D_2	D_b	$D_a - (I_1 - D_1)$
Ω_{31}	D_1	D_a	$I_1 - D_1 - D_a$	I_2	0	0
Ω_{32}	D_1	D_a	$I_1 - D_1 - D_a$	D_2	$I_2 - D_2$	0
Ω_{33}	D_1	D_a	$D_b - (I_2 - D_2)$	D_2	$I_2 - D_2$	0
Ω_{34}	D_1	D_a	0	I_2	D_b	0
Ω_{35}	D_1	D_a	0	I_2	D_b	0
Ω_{41}	D_1	D_a	D_b	I_2	0	0
Ω_{42}	D_1	D_a	D_b	I_2	0	0
Ω_{43}	D_1	D_a	$D_b - (I_2 - D_2)$	D_2	$I_2 - D_2$	0
Ω_{44}	D_1	D_a	0	D_2	D_b	0
Ω_{45}	D_1	D_a	0	D_2	D_b	0

C.1.2 Dual solutions for W-case with shipping cost

Here, it is assumed that for unit sales shipped from FC1 to D_b or from FC2 to D_a, results in extra shipping cost s.

Table C.2: Dual solution by demand partition with extra shipping cost s

Demand Partition	λ_1	λ_2	μ_1	μ_2	μ_a	μ_b
Ω_{11}	p	p	0	0	0	0
Ω_{12}	p	p	0	0	0	0
Ω_{13}	p	p-s	0	s	0	s
Ω_{14}	p	0	0	p	p-s	p
Ω_{15}	p	0	0	p	p-s	p
Ω_{21}	p	p	0	0	0	0
Ω_{22}	p	p	0	0	0	0
Ω_{23}	p	p-s	0	s	0	s
Ω_{24}	s	0	p-s	p	p-s	p
Ω_{25}	s	0	p-s	p	p-s	p
Ω_{31}	p-s	p	s	0	s	0
Ω_{32}	p-s	p	s	0	s	0
Ω_{33}	0	s	p	p-s	p	p-s
Ω_{34}	0	0	p	p	p	p
Ω_{35}	0	0	p	p	p	p
Ω_{41}	0	p	p	0	p	p-s
Ω_{42}	0	p	p	0	p	p-s
Ω_{43}	0	s	p	p-s	p	p-s
Ω_{44}	0	0	p	p	p	p
Ω_{45}	0	0	p	p	p	p

C.2 Optimality condition for I_1 for case 3

$$\frac{p-c}{p} = P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] +$$

$$P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] + P\begin{bmatrix} D_1 \leq I_1 < D_1 + D_{1,2} + D_{1,2,3} \\ \&I_1 + [[I_2 - D_2]_+ - [D_{2,3} - [I_3 - D_3]_+]_+]_+ \geq D_1 + D_{1,2} \\ \&I_1 + [[[I_2 - D_2]_+ - [D_{2,3} - [I_3 - D_3]_+]_+]_+ - D_{1,2}]_+ \\ \quad +[I_3 - D_3 - D_{2,3}]_+ \geq D_1 + D_{1,2,3} \end{bmatrix} \tag{C.1.1}$$

$$\approx P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] + P\begin{bmatrix} D_1 \leq I_1 < D_1 + D_{1,2} + D_{1,2,3} \\ \&I_1 + I_2 + I_3 \geq D_1 + D_2 + D_3 + D_{1,2} + D_{2,3} \\ \&I_1 + I_2 + I_3 \geq D_1 + D_2 + D_3 + D_{1,2} + D_{2,3} + d_{1,2,3} \end{bmatrix} \tag{C.1.2}$$

$$\approx P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] + P[D_1 \leq I_1 < D_1 + D_{1,2} + D_{1,2,3}\&I_1 + I_2 + I_3 \geq D_{all}] \tag{C.1.3}$$

$$\approx P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] + P[D_1 \leq I_1 < D_1 + D_{1,2} + D_{1,2,3}]p[I_1 + I_2 + I_3 \geq D_{all}] \tag{C.1.4}$$

$$\approx P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1] + (1 - P[D_1 + D_{1,2} + D_{1,2,3} \leq I_1])p[I_1 + I_2 + I_3 \geq D_{all}] \quad \text{(C.1.5)}$$

C.3 Extra experimental results

C.3.1 W-solutions with respect to demand distributions

Here we show the impact of the having lower bound constraint. Finding W-case solution with or without lower bound constraint is illustrated in Figure C.1. For all three examples, probability functions of D_1, D_2 ,$D_{1assign}$, $D_{2assign}$, D_{1all}, D_{2all} and D_{all} is shown in the top graph with W-case solutions $I1_w, I2_w$ and $I1_{wnlb}, I2_{wnlb}$ are marked with vertical line. In example 1, lower bound didn't make any difference because $I1_w, I2_w$ is already larger than the bound. However, in example 2 and 3, values of $I1_w$ and $I2_w$ is increased due to lower bound constraint respectively. It can be seen that the impact of forcing lower bound differs depends on the shape of demand distributions. The bottom graphs shows how cumulative probabilities $F_1(I_1), F_2(I_2)$ and $F_{all}(I_1 + I_2)$ changes to $\eta = P[D_1 + D_a + D_b < I_1] = P[D_2 + D_a + D_b < I_2]$ which was defined in previous section, as well as the error, or the value of function g, which was also defined previously.

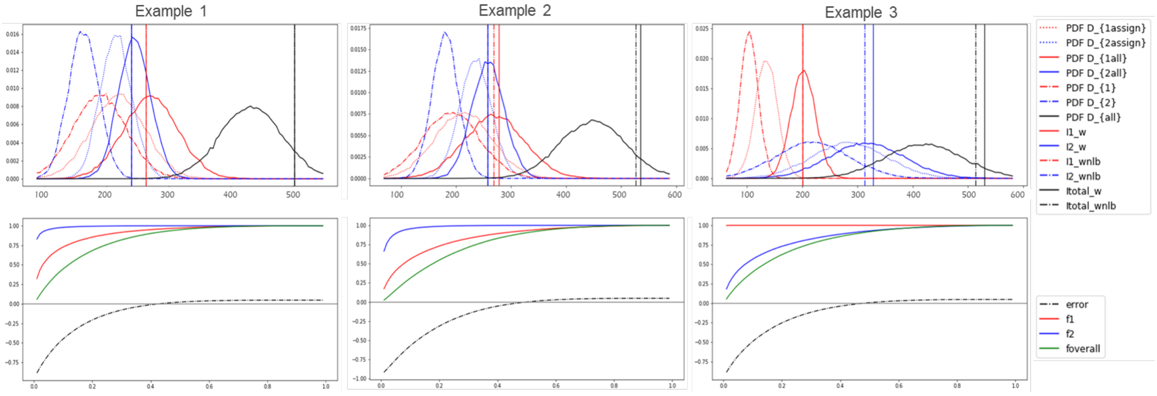


Figure C.1: Examples of w-solutions under W-case by varying demand distributions

REFERENCES

[1] Statista, *Retail e-commerce sales worldwide from 2014 to 2024*, https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/, Accessed: 2021-04-04.

[2] ——, *E-commerce share of total global retail sales from 2015 to 2023*, https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/, Accessed: 2021-04-04.

[3] ——, *Number of digital buyers worldwide from 2014 to 2021*, https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/, Accessed: 2021-04-04.

[4] ——, *Retail e-commerce revenue in the united states from 2017 to 2024*, https://www.statista.com/statistics/272391/us-retail-e-commerce-sales-forecast/, Accessed: 2021-04-04.

[5] A. Bhatti, H. Akram, H. M. Basit, A. U. Khan, S. M. Raza, and M. B. Naqvi, "E-commerce trends during covid-19 pandemic," *International Journal of Future Generation Communication and Networking*, vol. 13, no. 2, pp. 1449–1452, 2020.

[6] J. Wertz, *3 emerging e-commerce growth trends to leverage in 2020*, https://www.forbes.com/sites/jiawertz/2020/08/01/3-emerging-e-commerce-growth-trends-to-leverage-in-2020/?sh=3132f3d36fee, Accessed: 2021-04-04.

[7] DigitalCommerce360, *Us ecommerce grows 44.0% in 2020*, digitalcommerce360.com/article/us-ecommerce-sales/, Accessed: 2021-04-04.

[8] Invesp, *The importance of same day delivery – statistics and trends*, https://www.invespcro.com/blog/same-day-delivery/, Accessed: 2021-03-10, 2015-2017.

[9] N. A. Agatz, M. Fleischmann, and J. A. Van Nunen, "E-fulfillment and multi-channel distribution–a review," *European journal of operational research*, vol. 187, no. 2, pp. 339–356, 2008.

[10] B. Montreuil, "The physical internet: A conceptual journey," in *Keynote Presentation at 2nd International Physical Internet Conference, Paris, France*, Paris, France, 2015.

[11] B. Montreuil, R. D. Meller, and E. Ballot, "Physical internet foundations," *IFAC Proceedings Volumes*, vol. 45, no. 6, pp. 26–30, 2012.

[12] B. Hambleton and K. Mannix Scherer, *Customer 1st Supply Network: Logistics the Way It Should Be*. ES3 LLC, 2016.

[13] H. Sohrabi, B. Montreuil, and W. Klibi, "On comparing dedicated and hyperconnected distribution systems: An optimization-based approach," in *International Conference on Information Systems, Logistics and Supply Chain (ILS2016). Bordeaux, France*, 2016.

[14] G. D. Eppen, "Note—effects of centralization on expected costs in a multi-location newsboy problem," *Management science*, vol. 25, no. 5, pp. 498–501, 1979.

[15] UNHABITAT. (2016). "World cities report 2016."

[16] M. Savelsbergh and T. van Woensel, "50th anniversary invited article—city logistics: Challenges and opportunities," *Transportation Science*, vol. 50, no. 2, pp. 579–590, 2016.

[17] T. Bektaş, T. G. Crainic, and T. van Woensel, "From managing urban freight to smart city logistics networks," in *Network Design and Optimization for Smart Cities*, World Scientific, 2017, pp. 143–188.

[18] E. Taniguchi, R. G. Thompson, and T. Yamada, "Recent trends and innovations in modelling city logistics," *Procedia-Social and Behavioral Sciences*, vol. 125, pp. 4–14, 2014.

[19] K. K. Boyer, A. M. Prudhomme, and W. Chung, "The last mile challenge: Evaluating the effects of customer density and delivery window patterns," *Journal of business logistics*, vol. 30, no. 1, pp. 185–201, 2009.

[20] T. G. Crainic, "City logistics," in *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, INFORMS, 2008, pp. 181–212.

[21] B. Montreuil, "Toward a physical internet: Meeting the global logistics sustainability grand challenge," *Logistics Research*, vol. 3, no. 2-3, pp. 71–87, 2011.

[22] B. Montreuil, R. D. Meller, and E. Ballot, "Physical internet foundations," in *Service orientation in holonic and multi agent manufacturing and robotics*, Springer, 2013, pp. 151–166.

[23] T. G. Crainic and B. Montreuil, "Physical internet enabled hyperconnected city logistics," *Transportation Research Procedia*, vol. 12, pp. 383–398, 2016.

[24] M. Goyal, J. Cook, N. Kim, B. Montreuil, and C. Lafrance, "Hyperconnected city logistics for furniture and large appliance industry: Simulation-based exploratory

investigation," in *3rd International Physical Internet Conference, Atlanta, GA, USA*, 2016.

[25] N. Kim and B. Montreuil, "Assessing the impact of inventory deployment and sharing policies on hyperconnected last-mile furniture logistics," in *3rd International Physical Internet Conference, Atlanta, GA, USA*, 2016.

[26] N. Kim, N. Kholgade, and B. Montreuil, "Urban large-item logistics with hyperconnected fulfillment and transportation," in *5th International Physical Internet Conference, Groningen, Netherlands*, 2018.

[27] BESTUFS. (2007). "Bestufs good practice guide on urban freight."

[28] L. Faure, B. Montreuil, P. Burlat, and G. Marqués, "Ex ante sustainability improvement assessment of city logistics solutions: Learning from a simple interlinked pooling case," in *Proceedings of 1st International Physical Internet Conference, Québec, Canada*, 2014.

[29] J. Allen, M. Browne, A. Woodburn, and J. Leonardi, "The role of urban consolidation centres in sustainable freight transport," *Transport Reviews*, vol. 32, no. 4, pp. 473–490, 2012.

[30] J. R. van Duin, H. Quak, and J. Muñuzuri, "New challenges for urban consolidation centres: A case study in the hague," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 3, pp. 6177–6188, 2010.

[31] M. Browne, M. Sweet, A. Woodburn, and J. Allen, "Urban freight consolidation centres final report," *Transport Studies Group, University of Westminster*, vol. 10, 2005.

[32] W. van Heeswijk, M. R. Mes, and J. M. Schutten, "The delivery dispatching problem with time windows for urban consolidation centers," *Transportation science*, 2017.

[33] H. Quak and L. Tavasszy, "Customized solutions for sustainable city logistics: The viability of urban freight consolidation centres," in *Transitions towards sustainable mobility*, Springer, 2011, pp. 213–233.

[34] M. Janjevic and A. Ndiaye, "Investigating the financial viability of urban consolidation centre projects," *Research in transportation business & management*, vol. 24, pp. 101–113, 2017.

[35] W. van Heeswijk, R. Larsen, and A. Larsen, "An urban consolidation center in the city of copenhagen: A simulation study," *International journal of sustainable transportation*, vol. 13, no. 9, pp. 675–691, 2019.

[36] M. Björklund, M. Abrahamsson, and H. Johansson, "Critical factors for viable business models for urban consolidation centres," *Research in Transportation Economics*, vol. 64, pp. 36–47, 2017.

[37] L. Faure, P. Burlat, and G. Marqués, "Evaluate the viability of urban consolidation centre with regards to urban morphology," *Transportation Research Procedia*, vol. 12, pp. 348–356, 2016.

[38] E. Pérez-Bernabeu, A. A. Juan, J. Faulin, and B. B. Barrios, "Horizontal cooperation in road transportation: A case illustrating savings in distances and greenhouse gas emissions," *International Transactions in Operational Research*, vol. 22, no. 3, pp. 585–606, 2015.

[39] X. Wang and H. Kopfer, "Collaborative transportation planning of less-then-truckload freight," *OR spectrum*, vol. 36, no. 2, pp. 357–380, 2014.

[40] M. A. Krajewska, H. Kopfer, G. Laporte, S. Ropke, and G. Zaccour, "Horizontal cooperation among freight carriers: Request allocation and profit sharing," *Journal of the Operational Research Society*, vol. 59, no. 11, pp. 1483–1491, 2008.

[41] R. T. Wood, "Measuring urban freight in the tri-stae region," in *The urban movement of goods: Proceedings of the third Technology Assessment Review. Paris, OECD*, 1970, pp. 61–82.

[42] C. Cleophas, C. Cottrill, J. F. Ehmke, and K. Tierney, "Collaborative urban transportation: Recent advances in theory and practice," *European Journal of Operational Research*, vol. 273, no. 3, pp. 801–816, 2019.

[43] T. G. Crainic, N. Ricciardi, and G. Storchi, "Models for evaluating and planning city logistics systems," *Transportation science*, vol. 43, no. 4, pp. 432–454, 2009.

[44] T. G. Crainic, P. K. Nguyen, and M. Toulouse, "Synchronized multi-trip multi-traffic pickup & delivery in city logistics," *Transportation Research Procedia*, vol. 12, pp. 26–39, 2016.

[45] T. G. Crainic, F. Errico, W. Rei, and N. Ricciardi, "Modeling demand uncertainty in two-tier city logistics tactical planning," *Transportation Science*, vol. 50, no. 2, pp. 559–578, 2015.

[46] R. Cuda, G. Guastaroba, and M. G. Speranza, "A survey on two-echelon routing problems," *Computers & Operations Research*, vol. 55, pp. 185–199, 2015.

[47] M. Janjevic, M. Winkenbach, and D. Merchán, "Integrating collection-and-delivery points in the strategic design of urban last-mile e-commerce distribution networks,"

Transportation Research Part E: Logistics and Transportation Review, vol. 131, pp. 37–67, 2019.

[48] K. Toh, P. Nagel, and R. Oakden, “A business and ict architecture for a logistics city,” *International Journal of Production Economics*, vol. 122, no. 1, pp. 216–228, 2009.

[49] M. T. Melo, S. Nickel, and F. Saldanha-Da-Gama, “Facility location and supply chain management–a review,” *European journal of operational research*, vol. 196, no. 2, pp. 401–412, 2009.

[50] G. Nagy and S. Salhi, “Location-routing: Issues, models and methods,” *European journal of operational research*, vol. 177, no. 2, pp. 649–672, 2007.

[51] W. Klibi, A. Martel, and A. Guitouni, “The impact of operations anticipations on the quality of stochastic location-allocation models,” *Omega*, vol. 62, pp. 19–33, 2016.

[52] T. G. Crainic, N. Ricciardi, and G. Storchi, “Advanced freight transportation systems for congested urban areas,” *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 2, pp. 119–137, 2004.

[53] B. Montreuil, *Physical internet manifesto: Globally transforming the way physical objects are handled, moved, stored, realized, supplied and used, versions 1.1 to 1.11*, 2009-2012.

[54] S. Pan, E. Ballot, G. Q. Huang, and B. Montreuil, “Physical internet and interconnected logistics services: Research and applications,” *International Journal of Production Research*, vol. 55, 2017.

[55] A. Domínguez, J. Holguín-Veras, Á. Ibeas, and L. dellOlio, “Receivers’ response to new urban freight policies,” *Procedia-Social and Behavioral Sciences*, vol. 54, pp. 886–896, 2012.

[56] Lindawati, J. van Schagen, M. Goh, and R. de Souza, “Collaboration in urban logistics: Motivations and barriers,” *International Journal of Urban Sciences*, vol. 18, no. 2, pp. 278–290, 2014.

[57] I. Ben Mohamed, W. Klibi, O. Labarthe, J.-C. Deschamps, and M. Z. Babai, “Modelling and solution approaches for the interconnected city logistics,” *International Journal of Production Research*, vol. 55, no. 9, pp. 2664–2684, 2017.

[58] J. A. van Nunen, P. Huijbregts, and P. Rietveld, *Transitions Towards Sustainable Mobility: New Solutions and Approaches for Sustainable Transport Systems*. Springer Science & Business Media, 2011.

[59] E. Taniguchi, R. G. Thompson, and T. Yamada, “Emerging techniques for enhancing the practical application of city logistics models,” *Procedia-Social and Behavioral Sciences*, vol. 39, pp. 3–18, 2012.

[60] N. Firdausiyah, E. Taniguchi, and A. Qureshi, “Modeling city logistics using adaptive dynamic programming based multi-agent simulation,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 125, pp. 74–96, 2019.

[61] C. Osorio and M. Bierlaire, “A simulation-based optimization framework for urban transportation problems,” *Operations Research*, vol. 61, no. 6, pp. 1333–1345, 2013.

[62] J. R. van Duin, R. Kortmann, and S. van den Boogaard, “City logistics through the canals? a simulation study on freight waterborne transport in the inner-city of amsterdam,” *International Journal of Urban Sciences*, vol. 18, no. 2, pp. 186–200, 2014.

[63] A. Verbraeck, *Discrete modellen. Deel. 2. Simulatie theorie: SPM 2321*. TU Delft, 2010.

[64] W. van Heeswijk, J. Albertus, M. Mes, J. Schutten, and W. Zijm, “Evaluating urban logistics schemes using agent-based simulation,” *Transportation science*, 2020.

[65] O. Wangapisit, E. Taniguchi, J. S. Teo, and A. G. Qureshi, “Multi-agent systems modelling for evaluating joint delivery systems,” *Procedia-Social and Behavioral Sciences*, vol. 125, pp. 472–483, 2014.

[66] J. R. van Duin, A. van Kolck, N. Anand, E. Taniguchi, *et al.*, “Towards an agent-based modelling approach for the evaluation of dynamic usage of urban distribution centres,” *Procedia-Social and Behavioral Sciences*, vol. 39, pp. 333–348, 2012.

[67] M. J. Roorda, R. Cavalcante, S. McCabe, and H. Kwan, “A conceptual framework for agent-based modelling of logistics services,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 46, no. 1, pp. 18–31, 2010.

[68] R. Sarraj, E. Ballot, S. Pan, D. Hakimi, and B. Montreuil, “Interconnected logistic networks and protocols: Simulation-based efficiency assessment,” *International Journal of Production Research*, vol. 52, no. 11, pp. 3185–3208, 2014.

[69] S. Pan, M. Nigrelli, E. Ballot, R. Sarraj, and Y. Yang, “Perspectives of inventory control models in the physical internet: A simulation study,” *Computers & Industrial Engineering*, vol. 84, pp. 122–132, 2015.

[70] J. Holmgren, P. Davidsson, J. A. Persson, and L. Ramstedt, “Tapas: A multi-agent-based model for simulation of transport chains,” *Simulation Modelling Practice and Theory*, vol. 23, pp. 1–18, 2012.

[71] M. M. Solomon, “Algorithms for the vehicle routing and scheduling problems with time window constraints,” *Operations research*, vol. 35, no. 2, pp. 254–265, 1987.

[72] A. M. Campbell and M. Savelsbergh, “Efficient insertion heuristics for vehicle routing and scheduling problems,” *Transportation science*, vol. 38, no. 3, pp. 369–378, 2004.

[73] B. Liu, X. Guo, Y. Yu, and Q. Zhou, “Minimizing the total completion time of an urban delivery problem with uncertain assembly time,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 132, pp. 163–182, 2019.

[74] D. Weigel and B. Cao, “Applying gis and or techniques to solve sears technician-dispatching and home delivery problems,” *Interfaces*, vol. 29, no. 1, pp. 112–130, 1999.

[75] G. Clarke and J. W. Wright, “Scheduling of vehicles from a central depot to a number of delivery points,” *Operations research*, vol. 12, no. 4, pp. 568–581, 1964.

[76] V. N. Coelho, A. Grasas, H. Ramalhinho, I. M. Coelho, M. J. Souza, and R. C. Cruz, “An ils-based algorithm to solve a large-scale real heterogeneous fleet vrp with multi-trips and docking constraints,” *European Journal of Operational Research*, vol. 250, no. 2, pp. 367–376, 2016.

[77] C. Wang, D. Mu, F. Zhao, and J. W. Sutherland, “A parallel simulated annealing method for the vehicle routing problem with simultaneous pickup–delivery and time windows,” *Computers & Industrial Engineering*, vol. 83, pp. 111–122, 2015.

[78] S. Baldi, I. Michailidis, V. Ntampasi, E. Kosmatopoulos, I. Papamichail, and M. Papageorgiou, “A simulation-based traffic signal control for congested urban traffic networks,” *Transportation Science*, 2017.

[79] L. Chong and C. Osorio, “A simulation-based optimization algorithm for dynamic large-scale urban transportation problems,” *Transportation Science*, vol. 52, no. 3, pp. 637–656, 2017.

[80] C. Osorio and K. Nanduri, “Energy-efficient urban traffic management: A microscopic simulation-based approach,” *Transportation Science*, vol. 49, no. 3, pp. 637–651, 2015.

[81] E. Taniguchi and R. E. van Der Heijden, “An evaluation methodology for city logistics,” *Transport Reviews*, vol. 20, no. 1, pp. 65–90, 2000.

[82] B. Montreuil, “Omnichannel business-to-consumer logistics and supply chains: Towards hyperconnected networks and facilities,” in *14th IMHRC Proceedings (Karlsruhe, Germany)*, 2016.

[83] B. Montreuil, O. Labarthe, and C. Cloutier, “Modeling client profiles for order promising and delivery,” *Simulation Modelling Practice and Theory*, vol. 35, pp. 1–25, 2013.

[84] Y. Xing, Y. Xu, M. Shi, and Y. Lian, “The impact of pm2. 5 on the human respiratory system,” *Journal of thoracic disease*, vol. 8, no. 1, E69, 2016.

[85] S. Netessine and N. Rudi, *Supply chain structures on the internet and the role of marketing-operations interaction. d. simchi levi, sd wu, m. shen, eds. handbook of supply chain analysis in the e-business era*, 2004.

[86] Z. Zacharia, “Supply chain collaboration in transformative vertical industries: Implications of omnichannel and dropshipping,” *Di Central*, vol. url: https://go.dicentral.com/downloads/lehigh-study (access date: 2020-05-01), May 2019.

[87] Y.-K. Chen, F.-R. Chiu, W.-H. Lin, and Y.-C. Huang, “An integrated model for online product placement and inventory control problem in a drop-shipping optional environment,” *Computers & Industrial Engineering*, vol. 117, pp. 71–80, 2018.

[88] T. Randall, S. Netessine, and N. Rudi, “An empirical examination of the decision to invest in fulfillment capabilities: A study of internet retailers,” *Management Science*, vol. 52, no. 4, pp. 567–580, 2006.

[89] T. Cheong, M. Goh, and S. H. Song, “Effect of inventory information discrepancy in a drop-shipping supply chain,” *Decision Sciences*, vol. 46, no. 1, pp. 193–213, 2015.

[90] A. Ayanso, M. Diaby, and S. K. Nair, “Inventory rationing via drop-shipping in internet retailing: A sensitivity analysis,” *European Journal of Operational Research*, vol. 171, no. 1, pp. 135–152, 2006.

[91] S. Netessine and N. Rudi, “Supply chain choice on the internet,” *Management Science*, vol. 52, no. 6, pp. 844–864, 2006.

[92] C.-M. Chen and D. J. Thomas, “Inventory allocation in the presence of service-level agreements,” *Production and operations management*, vol. 27, no. 3, pp. 553–577, 2018.

[93] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[94] W. B. Powell, “A unified framework for stochastic optimization,” *European Journal of Operational Research*, vol. 275, no. 3, pp. 795–821, 2019.

[95] J. P. Bailey and E. Rabinovich, “Internet book retailing and supply chain management: An analytical study of inventory location speculation and postponement,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 41, no. 3, pp. 159–177, 2005.

[96] T. F. Rodrigues, M. M. Dantas, C. A. Cavalcante, *et al.*, “A dynamic inventory rationing policy for business-to-consumer e-tail stores in a supply disruption context,” *Computers & Industrial Engineering*, vol. 142, p. 106 379, 2020.

[97] C. A. Cavalcante *et al.*, “Order planning policies for business-to-consumer e-tail stores,” *Computers & Industrial Engineering*, vol. 136, pp. 106–116, 2019.

[98] S. Ma and Z. Jemai, “Inventory rationing for the news-vendor problem with a drop-shipping option,” *Applied Mathematical Modelling*, vol. 71, pp. 438–451, 2019.

[99] S. Ma, Z. Jemai, E. Sahin, and Y. Dallery, “The news-vendor problem with drop-shipping and resalable returns,” *International Journal of Production Research*, vol. 55, no. 22, pp. 6547–6571, 2017.

[100] M. Khouja and A. C. Stylianou, “A (q, r) inventory model with a drop-shipping option for e-business,” *Omega*, vol. 37, no. 4, pp. 896–908, 2009.

[101] S. T. Peinkofer, T. L. Esper, R. J. Smith, and B. D. Williams, “Assessing the impact of drop-shipping fulfilment operations on the upstream supply chain,” *International Journal of Production Research*, vol. 57, no. 11, pp. 3598–3621, 2019.

[102] Z. Y. Dennis, T. Cheong, and D. Sun, “Impact of supply chain power and drop-shipping on a manufacturer’s optimal distribution channel strategy,” *European Journal of Operational Research*, vol. 259, no. 2, pp. 554–563, 2017.

[103] G. Gallego, Ö. Özer, and P. Zipkin, “Bounds, heuristics, and approximations for distribution systems,” *Operations Research*, vol. 55, no. 3, pp. 503–517, 2007.

[104] A. Alptekinoğlu, A. Banerjee, A. Paul, and N. Jain, “Inventory pooling to deliver differentiated service,” *Manufacturing & Service Operations Management*, vol. 15, no. 1, pp. 33–44, 2013.

[105] G. P. Cachon and M. A. Lariviere, “Capacity choice and allocation: Strategic behavior and supply chain performance,” *Management science*, vol. 45, no. 8, pp. 1091–1108, 1999.

[106] X. Gan, S. P. Sethi, and J. Zhou, "Commitment-penalty contracts in drop-shipping supply chains with asymmetric demand information," *European Journal of Operational Research*, vol. 204, no. 3, pp. 449–462, 2010.

[107] G. P. Cachon and M. Fisher, "Supply chain inventory management and the value of shared information," *Management science*, vol. 46, no. 8, pp. 1032–1048, 2000.

[108] F. d. Véricourt, F. Karaesmen, and Y. Dallery, "Assessing the benefits of different stock-allocation policies for a make-to-stock production system," *Manufacturing & Service Operations Management*, vol. 3, no. 2, pp. 105–121, 2001.

[109] H. Wang, X. Liang, S. Sethi, and H. Yan, "Inventory commitment and prioritized backlogging clearance with alternative delivery lead times," *Production and Operations Management*, vol. 23, no. 7, pp. 1227–1242, 2014.

[110] E. J. McGavin, L. B. Schwarz, and J. E. Ward, "Two-interval inventory-allocation policies in a one-warehouse n-identical-retailer distribution system," *Management Science*, vol. 39, no. 9, pp. 1092–1107, 1993.

[111] W. Jang, D. Kim, and K. Park, "Inventory allocation and shipping when demand temporarily exceeds production capacity," *European Journal of Operational Research*, vol. 227, no. 3, pp. 464–470, 2013.

[112] Q. Geng and S. Mallik, "Inventory competition and allocation in a multi-channel distribution system," *European Journal of Operational Research*, vol. 182, no. 2, pp. 704–729, 2007.

[113] K. Kloos and R. Pibernik, "Allocation planning under service-level contracts," *European Journal of Operational Research*, vol. 280, no. 1, pp. 203–218, 2020.

[114] J. Chen, Y. Chen, M. Parlar, and Y. Xiao, "Optimal inventory and admission policies for drop-shipping retailers serving in-store and online customers," *IIE Transactions*, vol. 43, no. 5, pp. 332–347, 2011.

[115] P. Guo, F. Liu, and Y. Wang, "Pre-positioning and deployment of reserved inventories in a supply network: Structural properties," *Production and Operations Management*, 2019.

[116] Z. Hao, L. He, Z. Hu, and J. Jiang, "Robust vehicle pre-allocation with uncertain covariates," *Production and Operations Management*, 2019.

[117] W. Xie, Z. Jiang, Y. Zhao, and J. Hong, "Capacity planning and allocation with multi-channel distribution," *International Journal of Production Economics*, vol. 147, pp. 108–116, 2014.

[118] T. Santoso, S. Ahmed, M. Goetschalckx, and A. Shapiro, “A stochastic programming approach for supply chain network design under uncertainty,” *European Journal of Operational Research*, vol. 167, no. 1, pp. 96–115, 2005.

[119] I. Ben Mohamed, W. Klibi, and F. Vanderbeck, “Designing a two-echelon distribution network under demand uncertainty,” *European Journal of Operational Research*, vol. 280, no. 1, pp. 102–123, 2020.

[120] W. Klibi, F. Lasalle, A. Martel, and S. Ichoua, “The stochastic multiperiod location transportation problem,” *Transportation Science*, vol. 44, no. 2, pp. 221–237, 2010.

[121] A. N. Arslan, W. Klibi, and B. Montreuil, “Distribution network deployment for omnichannel retailing,” *European Journal of Operational Research*, 2020.

[122] N. Hibiki, “Multi-period stochastic optimization models for dynamic asset allocation,” *Journal of banking & finance*, vol. 30, no. 2, pp. 365–390, 2006.

[123] W. Klibi, A. Martel, and A. Guitouni, “The impact of operations anticipations on the quality of stochastic location-allocation models,” *Omega*, vol. 62, pp. 19–33, 2016.

[124] S. Ahmed, A. Shapiro, and E. Shapiro, “The sample average approximation method for stochastic programs with integer recourse,” *Submitted for publication*, pp. 1–24, 2002.

[125] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, “The sample average approximation method for stochastic discrete optimization,” *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

[126] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, “Scenario reduction in stochastic programming,” *Mathematical programming*, vol. 95, no. 3, pp. 493–511, 2003.

[127] M. Rothstein, “An airline overbooking model,” *Transportation Science*, vol. 5, no. 2, pp. 180–192, 1971.

[128] ——, “Or forum—or and the airline overbooking problem,” *Operations Research*, vol. 33, no. 2, pp. 237–248, 1985.

[129] I. Karaesmen and G. Van Ryzin, “Overbooking with substitutable inventory classes,” *Operations Research*, vol. 52, no. 1, pp. 83–104, 2004.

[130] K. Muthuraman and M. Lawley, “A stochastic overbooking model for outpatient clinical scheduling with no-shows,” *IIE Transactions*, vol. 40, no. 9, pp. 820–837, 2008.

[131] S. Rao, S. E. Griffis, and T. J. Goldsby, “Failure to deliver? linking online order fulfillment glitches with future purchase behavior,” *Journal of Operations Management*, vol. 29, no. 7-8, pp. 692–703, 2011.

[132] G. A. Chua, W. S. Lim, and W. M. Yeo, “Market structure and the value of overselling under stochastic demands,” *European Journal of Operational Research*, vol. 252, no. 3, pp. 900–909, 2016.

[133] M. O. Ball, C.-Y. Chen, and Z.-Y. Zhao, “Available to promise,” in *Handbook of quantitative supply chain analysis*, Springer, 2004, pp. 447–483.

[134] S. Derhami, B. Montreuil, and G. Bau, “Assessing product availability in omnichannel retail networks in the presence of on-demand inventory transshipment and product substitution,” *Omega*, p. 102 315, 2020.

[135] A. A. Syntetos, Z. Babai, J. E. Boylan, S. Kolassa, and K. Nikolopoulos, “Supply chain forecasting: Theory, practice, their gap and the future,” *European Journal of Operational Research*, vol. 252, no. 1, pp. 1–26, 2016.

[136] K. J. Ferreira, B. H. A. Lee, and D. Simchi-Levi, “Analytics for an online retailer: Demand forecasting and price optimization,” *Manufacturing & Service Operations Management*, vol. 18, no. 1, pp. 69–88, 2016.

[137] A. A. Tsay and N. Agrawal, “Channel conflict and coordination in the e-commerce age,” *Production and operations management*, vol. 13, no. 1, pp. 93–110, 2004.

[138] G. Bressolles and G. Lang, “Kpis for performance measurement of e-fulfillment systems in multi-channel retailing,” *International Journal of Retail & Distribution Management*, 2019.

[139] V. Goel and I. E. Grossmann, “A class of stochastic programs with decision dependent uncertainty,” *Mathematical programming*, vol. 108, no. 2-3, pp. 355–394, 2006.

[140] D. Honhon, V. Gaur, and S. Seshadri, “Assortment planning and inventory decisions under stockout-based substitution,” *Operations research*, vol. 58, no. 5, pp. 1364–1379, 2010.

[141] B. Montreuil, J. Bouchard, A. Morneau, and E. Brotherton, “Prévision de vente et aide à la décision de réapprovisionnement de produits à cycle rapide,” 2015.

[142] J. G. De Gooijer and R. J. Hyndman, “25 years of time series forecasting,” *International journal of forecasting*, vol. 22, no. 3, pp. 443–473, 2006.

[143] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber," in *International Conference on Machine Learning*, vol. 34, 2017, pp. 1–5.

[144] S. Chopra, "The evolution of omni-channel retailing and its impact on supply chains," *Transportation research procedia*, vol. 30, pp. 4–13, 2018.

[145] A. Snoeck and M. Winkenbach, "A discrete simulation-based optimization algorithm for the design of highly responsive last-mile distribution networks," 2020.

[146] J. Acimovic and S. C. Graves, "Mitigating spillover in online retailing via replenishment," *Manufacturing & Service Operations Management*, vol. 19, no. 3, pp. 419–436, 2017.

[147] E. Ballot, B. Montreuil, and Z. G. Zacharia, *Physical internet: First results and next challenges*, 2021.

[148] N. Kim, B. Montreuil, W. Klibi, and N. Kholgade, "Hyperconnected urban fulfillment and delivery," *Transportation Research Part E: Logistics and Transportation Review*, vol. 145, p. 102 104, 2021.

[149] S. Benjaafar, W. L. Cooper, and J.-S. Kim, "On the benefits of pooling in production-inventory systems," *Management Science*, vol. 51, no. 4, pp. 548–565, 2005.

[150] O. Berman, D. Krass, and M. Mahdi Tajbakhsh, "On the benefits of risk pooling in inventory management," *Production and operations management*, vol. 20, no. 1, pp. 57–71, 2011.

[151] A. J. Schmitt, S. A. Sun, L. V. Snyder, and Z.-J. M. Shen, "Centralization versus decentralization: Risk pooling, risk diversification, and supply chain disruptions," *Omega*, vol. 52, pp. 201–212, 2015.

[152] A. Govindarajan, A. Sinha, and J. Uichanco, "Joint inventory and fulfillment decisions for omnichannel retail networks," *Naval Research Logistics (NRL)*, 2018.

[153] C. J. Corbett and K. Rajaram, "A generalization of the inventory pooling effect to nonnormal dependent demand," *Manufacturing & Service Operations Management*, vol. 8, no. 4, pp. 351–358, 2006.

[154] W. J. Tallon, "The impact of inventory centralization on aggregate safety stock: The variable supply lead time case," *Journal of Business Logistics*, vol. 14, no. 1, p. 185, 1993.

[155] N. Erkip, W. H. Hausman, and S. Nahmias, "Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands," *Management Science*, vol. 36, no. 3, pp. 381–392, 1990.

[156] A. Federgruen and P. Zipkin, "Approximations of dynamic, multilocation production and inventory problems," *Management Science*, vol. 30, no. 1, pp. 69–84, 1984.

[157] C. Yang, Z. Hu, and S. X. Zhou, "Multilocation newsvendor problem: Centralization and inventory pooling," *Management Science*, vol. 67, no. 1, pp. 185–200, 2021.

[158] L. Dong and N. Rudi, "Who benefits from transshipment? exogenous vs. endogenous wholesale prices," *Management Science*, vol. 50, no. 5, pp. 645–657, 2004.

[159] N. Rudi, S. Kapur, and D. F. Pyke, "A two-location inventory model with transshipment and local decision making," *Management science*, vol. 47, no. 12, pp. 1668–1680, 2001.

[160] G. Tagaras and M. A. Cohen, "Pooling in two-location inventory systems with non-negligible replenishment lead times," *Management science*, vol. 38, no. 8, pp. 1067–1083, 1992.

[161] A. Alptekinoğlu and C. S. Tang, "A model for analyzing multi-channel distribution systems," *European Journal of Operational Research*, vol. 163, no. 3, pp. 802–824, 2005.

[162] L. Silbermayr and Y. Gerchak, "Partial pooling by independent firms with allocation according to contribution to pool," *International Journal of Production Economics*, vol. 218, pp. 375–385, 2019.

[163] A. Kranenburg and G.-J. Van Houtum, "A new partial pooling structure for spare parts networks," *European Journal of Operational Research*, vol. 199, no. 3, pp. 908–921, 2009.

[164] P. Guo, F. Liu, and Y. Wang, "Pre-positioning and deployment of reserved inventories in a supply network: Structural properties," *Production and Operations Management*, vol. 29, no. 4, pp. 893–906, 2020.

[165] A. M. Stroh, A. L. Erera, and A. Toriello, "Tactical design of same-day delivery systems," *Georgia Institute of Technology Working Paper*, 2019.

[166] M. A. Klapp, A. L. Erera, and A. Toriello, "The dynamic dispatch waves problem for same-day delivery," *European Journal of Operational Research*, vol. 271, no. 2, pp. 519–534, 2018.

[167] L. Wei, R. Kapuscinski, and S. Jasin, "Shipping consolidation across two warehouses with delivery deadline and expedited options for e-commerce and omni-channel retailers," *Manufacturing & Service Operations Management*, 2020.

[168] H.-W. Chen, D. Gupta, H. Gurnani, and G. Janakiraman, "A stochastic inventory model with fast-ship commitments," *Production and Operations Management*, vol. 25, no. 4, pp. 684–700, 2016.

[169] H.-W. Chen, D. Gupta, and H. Gurnani, "Balancing inventory and stockout risk in retail supply chains using fast-ship," *Production and Operations Management*, vol. 25, no. 12, pp. 2103–2115, 2016.

[170] L. Ozsen, C. R. Coullard, and M. S. Daskin, "Capacitated warehouse location model with risk pooling," *Naval Research Logistics (NRL)*, vol. 55, no. 4, pp. 295–312, 2008.

[171] Z.-J. M. Shen, C. Coullard, and M. S. Daskin, "A joint location-inventory model," *Transportation science*, vol. 37, no. 1, pp. 40–55, 2003.

[172] Y. Zhong, Z. Zheng, M. C. Chou, and C.-P. Teo, "Resource pooling and allocation policies to deliver differentiated service," *Management Science*, vol. 64, no. 4, pp. 1555–1573, 2017.

[173] L. DeValve, Y. Wei, D. Wu, and R. Yuan, "Understanding the value of fulfillment flexibility in an online retailing environment," *Available at SSRN 3265838*, 2018.

[174] J. Acimovic and S. C. Graves, "Making better fulfillment decisions on the fly in an online retail environment," *Manufacturing & Service Operations Management*, vol. 17, no. 1, pp. 34–51, 2015.

[175] W. C. Jordan and S. C. Graves, "Principles on the benefits of manufacturing process flexibility," *Management Science*, vol. 41, no. 4, pp. 577–594, 1995.

[176] S. C. Graves and B. T. Tomlin, "Process flexibility in supply chains," *Management Science*, vol. 49, no. 7, pp. 907–919, 2003.

[177] S. Gurumurthi and S. Benjaafar, "Modeling and analysis of flexible queueing systems," *Naval Research Logistics (NRL)*, vol. 51, no. 5, pp. 755–782, 2004.

[178] M. Khouja, "The single-period (news-vendor) problem: Literature review and suggestions for future research," *omega*, vol. 27, no. 5, pp. 537–553, 1999.

[179] A. Govindarajan, A. Sinha, and J. Uichanco, "Distribution-free inventory risk pooling in a multilocation newsvendor," *Management Science*, 2020.

[180] J. A. V. Mieghem and N. Rudi, “Newsvendor networks: Inventory management and capacity investment with discretionary activities,” *Manufacturing & Service Operations Management*, vol. 4, no. 4, pp. 313–335, 2002.

[181] J. M. Harrison and J. A. Van Mieghem, “Multi-resource investment strategies: Operational hedging under demand uncertainty,” *European Journal of Operational Research*, vol. 113, no. 1, pp. 17–29, 1999.

[182] J. A. Van Mieghem, “Investment strategies for flexible resources,” *Management Science*, vol. 44, no. 8, pp. 1071–1078, 1998.

[183] Ö. Özer and H. Xiong, “Stock positioning and performance estimation for distribution systems with service constraints,” *IIE Transactions*, vol. 40, no. 12, pp. 1141–1157, 2008.

[184] H. Sohrabi, B. Montreuil, and W. Klibi, “Collaborative and hyperconnected distribution systems: A comparative optimization-based assessment,” in *Proceedings of Proceedings of the 2016 Industrial and Systems Engineering Research Conference, Anaheim, CA*, 2016.

[185] H. Sohrabi, B. Montreuil, and W. Klibi, “An optimization-based investigation of exploiting physical internet-enabled interconnected distribution system,” in *2nd International Physical Internet Conference, July 6–July 8*, Mines ParisTech Paris, 2015.

[186] Y. Yang, S. Pan, and E. Ballot, “A model to take advantage of physical internet for vendor inventory management,” *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 1990–1995, 2015.

[187] B. Y. Ekren, A. Akpunar, and G. Mullaoglu, “Inventory control models towards physical internet: Lateral transshipment policy determination by simulation,” in *5th International Physical Internet Conference*, 2018.

[188] Y. Yang, S. Pan, and E. Ballot, “Mitigating supply chain disruptions through interconnected logistics services in the physical internet,” *International Journal of Production Research*, vol. 55, no. 14, pp. 3970–3983, 2017.

[189] S.-f. Ji, X.-s. Peng, and R.-j. Luo, “An integrated model for the production-inventory-distribution problem in the physical internet,” *International Journal of Production Research*, vol. 57, no. 4, pp. 1000–1017, 2019.

[190] R. H. Ballou, H. Rahardja, and N. Sakai, “Selected country circuity factors for road travel distance estimation,” *Transportation Research Part A: Policy and Practice*, vol. 36, no. 9, pp. 843–848, 2002.

[191] M. A. Qureshi, H.-L. Hwang, and S.-M. Chin, “Comparison of distance estimates for commodity flow survey: Great circle distances versus network-based distances,” *Transportation research record*, vol. 1804, no. 1, pp. 212–216, 2002.

[192] F. P. Boscoe, K. A. Henry, and M. S. Zdeb, “A nationwide comparison of driving distance versus straight-line distance to hospitals,” *The Professional Geographer*, vol. 64, no. 2, pp. 188–196, 2012.

VITA

Nayeon Kim was born in Seoul, South Korea. She earned B.S. degree in Industrial Engineering from Seoul National University in 2014, receiving National scholarship for 4 years. She studied at University of Oklahoma in 2011 through an exchange program. After graduation, she came to Georgia Tech to pursue Master's degree in Industrial Engineering, and continued to pursue Doctorate degree. Her main research interests are supply chain, Physical Internet, inventory management and fulfillment strategy.