Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Managing Multitier Inventory Networks with Expediting Under Normal and Disrupted Modes

Yanyang Zhao, John R. Birge, Levi DeValve, Robert R. Inman

To cite this article:
Yanyang Zhao, John R. Birge, Levi DeValve, Robert R. Inman (2026) Managing Multitier Inventory Networks with Expediting Under Normal and Disrupted Modes. Manufacturing & Service Operations Management

Published online in Articles in Advance 26 Mar 2026

. https://doi.org/10.1287/msom.2023.0249

https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Articles in Advance, pp. 1–19
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Managing Multitier Inventory Networks with Expediting Under Normal and Disrupted Modes

Yanyang Zhao,[a] John R. Birge,[a] Levi DeValve,[a,*] Robert R. Inman[b]

[a] Booth School of Business, University of Chicago, Chicago, Illinois 60637; [b] Sales, Service and Customer Intelligence, General Motors Company, Warren, Michigan 48093

*Corresponding author

Contact: alexzhao@chicagobooth.edu, https://orcid.org/0000-0003-2293-4907 (YZ); john.birge@chicagobooth.edu, https://orcid.org/0000-0002-7446-0953 (JRB); levi.devalve@chicagobooth.edu, https://orcid.org/0000-0002-4730-4541 (LD); robert.inman@gm.com (RRI)

Received: April 30, 2023
Revised: June 19, 2024; September 19, 2025; December 13, 2025
Accepted: January 22, 2026
Published Online in Articles in Advance: March 26, 2026

https://doi.org/10.1287/msom.2023.0249

Abstract. ***Problem definition*:** We collaborate with an industrial partner whose supply chain uses multiple tiers, locations, and shipping speeds to efficiently serve customers. In practice, our partner also faces the possibility of upstream disruptions, which limit inventory availability. We model these key features of our partner's network as a multiechelon distribution system (central warehouse and retailers) with expediting and disruptions. ***Methodology/results*:** We prove a novel stochastic program lower bound on optimal cost in this model and use this program to develop a heuristic base-stock policy. Our analysis demonstrates that there is a pronounced benefit from centralized inventory (i.e., holding inventory at the central warehouse) in distribution systems with expediting and disruptions as it can be used to both clear backlogs through expediting and hedge against future disruptions. Further, in the disrupted mode, we provide a simple criterion to determine when decentralization (i.e., holding inventory at the retailers) is preferred over complete centralization. Then, we validate our policies using data from our partner's nationwide distribution network in the United States. ***Managerial implications*:** We provide novel inventory policies for managing a distribution system with expediting and disruptions that are understandable and implementable in practice. Our analysis provides the insight that facilitating the right level of central warehouse inventory is a critical hedge for improving performance in these systems. Finally, our industrial partner's data suggest that our policies can provide significant cost savings in practice.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2023.0249.

Keywords: inventory theory and control • math programming • simulation • supply chain management

1. Introduction

To effectively make products available to customers, supply chains often operate in multiple tiers (or "echelons") as well as in multiple locations. This allows the supply chain to take advantage of both the operational efficiency of various tiers as well as the responsiveness of locations closer to customers. For example, Seven-Eleven Japan supports large clusters of traditional brick-and-mortar retail stores through a centralized distribution center (Chopra 2017), whereas Chinese e-commerce giant JD.com operates a similar network consisting of several front distribution centers supported through a larger regional distribution center (RDC) (DeValve et al. 2023); in both cases, a central facility efficiently supports large product volumes, whereas a set of dispersed facilities offers quick response times to customers. This ubiquitous supply chain setting is often called a *distribution system* in the operations management literature (e.g., Zipkin 2000, Gallego et al. 2007, Federgruen et al. 2018), and it is the focus of this paper.

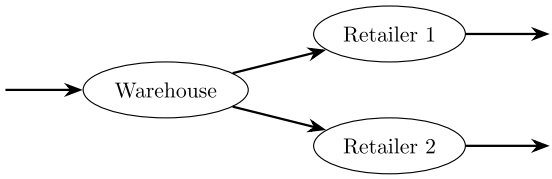
In particular, we analyze a distribution system operated by our industrial collaborator, a large U.S. automotive manufacturer, for distributing service (i.e., repair) parts used in their automobiles. Operating this supply chain presents several practical challenges relative to existing distribution system research, including expediting from the central warehouse, multiple demand classes, stochastic lead times, and stochastic nonstationary demand. In addition to these challenges, our

industrial partner also faces the risk of supply disruption to its operations from various major identifiable causes. In just the past few years, these have included natural disasters, labor disruptions, and pandemics (among others), which effectively limit the supply available to operate the system. Therefore, we consider two "modes" of operation: the *normal mode* and the *disrupted mode*, where the supplier is disrupted and unable to fulfill new orders. The automaker's goal is to design inventory management policies that effectively control the system in both normal and disrupted modes.

This is an inherently challenging problem as optimal inventory policies remain intractable to compute even in the classic distribution system model without the additional practical complications faced by our partner (Gallego et al. 2007). Thus, in this paper, we focus our main analysis on a parsimonious model incorporating the two most salient features of the automaker's context: expedited delivery and supply disruptions. In a distribution system, these features interact to create a unique opportunity to exploit centralized inventory; expediting from the warehouse offers an effective way to fulfill demand from a central location and thus, provides a hedge against the negative impacts of a disruption. The classic analysis of distribution systems can miss this opportunity in simple parameter regimes, such as when all locations have the same holding cost, and hold no inventory at the central warehouse.

To better illustrate these issues, consider an example of a simple distribution system with one warehouse and two retailers as depicted in Figure 1. Each retailer independently faces exogenous periodic Poisson demand with a rate of one, which can be backlogged at a cost of \$10 per unit per period, and holding inventory at either the warehouse or retailers costs \$1 per unit per period. The warehouse orders from an external supplier with a lead time of four periods and then, ships to the retailers with a lead time of two periods. We consider two additional features on top of this classic distribution system. First, inventory at the warehouse may be expedited (i.e., used to instantly fill demand and/or backlog at either retailer for a cost of \$15 per unit). Second, we model supplier disruptions with a Markovian state variable; in each undisrupted period, there is a 1% chance for a disruption to begin in the following period, and once a disruption begins, it will last for a Poisson distributed number of periods with a mean of 15. A disrupted supplier takes no new orders, but orders already in the pipeline are delivered as scheduled. The decision maker minimizes long-run-average expected holding, expediting, and backlog costs (see Section 2 for full details of our formal model).

Figure 1. A Simple Distribution System with One Warehouse and Two Retailers



We first illustrate what goes wrong if expediting and disruptions are ignored using a classic approach from the distribution system literature. In particular, a standard heuristic of Federgruen and Zipkin (1984c) and Gallego et al. (2007) sets a system-wide base-stock level of 17 units for this system and pushes all inventory to the retailers (see Online Appendix C.1.1 for a general description of this policy and Online Appendix A for detailed simulations for this example). A direct implementation of this heuristic that simply expedites to clear any backlogs before allocating inventory to retailers (as backlog cost over the retailer lead time, $2 \times \$10$, is higher than expediting, \$15) in a simulation gives average daily cost of \$27.15, with an average daily backlog per retailer of 0.65 units per day, accounting for 48% of the average daily cost, and it expedites an average of 0.19 units per retailer per day. Clearly, a system where backlog cost is 10 times higher than holding cost should not incur nearly half its cost in backlog. This inflated backlog cost is driven by ignoring the possibility of expediting and disruptions, and it can be addressed with centralized inventory.

To do this, we develop a novel stochastic programming approach for setting base-stock levels that calibrates warehouse and retailer inventory to the possibility of both expediting and disruptions. In particular, the stochastic program approximates the average cost of a policy across normal and disrupted periods, and it can be used to set base-stock levels, taking both into account (see Section 3.2.1 for policy details). In this example, the policy sets local base-stock levels of 4 units at the warehouse and 10 units at each of the retailers. Thus, our policy plans for more inventory than Gallego et al. (2007) to hedge against disruptions, and also, it centralizes a small portion of the inventory at the warehouse to use for expediting. This reduces the average backlog to 0.32 units per retailer per day with only 0.10 units of expediting per retailer per day while achieving an average cost of \$25.08 per day. Thus, in this example, planning for expediting and disruptions with centralized inventory reduced average cost by 7.6%, and it cut costly backlogs and expediting in half. Clearly, there is a benefit from exploring policies for exploiting these observations, which is the goal of this paper. We summarize our main contributions next.

1.1. Stochastic Program Lower Bound and Base-Stock Policy Design

In our parsimonious model of a distribution system with expediting and disruptions, we derive a novel

stochastic program lower bound on the cost of an optimal policy. Our analysis overcomes several technical hurdles, including the different timescales of fulfillment caused by expediting and the stochastic periods of supply unavailability caused by disruptions. But, more importantly, the stochastic program that we derive provides two important practical benefits. (i) It can be used in an intuitive way to decide on base-stock levels for an effective inventory policy, and (ii) it is flexible enough to adapt to the various constraints faced by our industrial partner in practice. We also develop a simpler heuristic that sets base-stock levels using recursive newsvendor solutions, which can also be adapted to a range of practical constraints. We demonstrate in simulation that these approaches offer significant cost improvements over existing policies. The primary managerial insight from this analysis is that holding inventory at the central warehouse has a pronounced benefit in distribution systems with expediting and disruptions, and our heuristics effectively take advantage of this opportunity.

1.2. Disrupted-Mode Policies

During a disruption, our industrial partner currently takes centralized control over all inventory at the warehouse and in the pipeline (but not pulled from the retailers), a strategy that we call "centralization." This raises the question of whether policy interventions are advantageous during disruptions. In the automaker's context, the main motivation for intervention is to centralize control of disrupted parts in order to prioritize customers for fulfillment. For example, when supply is limited across the network, a later arriving customer at one retailer may receive a part sooner than a customer who has been waiting longer at another retailer simply because the former retailer has not stocked out yet. To avoid this situation, the automaker takes centralized control of the inventory during a disruption to ensure that customers are prioritized appropriately. The clear trade-off, however, is that although centralizing inventory gives the automaker better control over backlogs, it also leads to increased expediting costs and centralized coordination efforts. Our industrial partner, therefore, recognizes that not all disruptions are created equal (e.g., short disruptions or those with ample starting inventory may not justify the increased cost and effort of an intervention). We address this question by developing a criterion for deciding when centralization is an effective strategy during a disruption.

To do so, we compare the two extremes[1] of complete centralization versus "decentralization" or keeping all inventory at the retailers. We derive a straightforward criterion for comparing the backlog and expediting costs to determine whether decentralization will achieve lower cost than centralization. Intuitively, the condition favors decentralization when expediting cost is high; however, it also shows that when either supply or demand is large, decentralization is preferred because inventory is less likely to get siloed locally, while backlogs exist elsewhere in the network. Thus, our result suggests that centralization is not always preferred during disruptions, and also, it provides a simple condition for deciding when to keep inventory decentralized, which is useful for our industrial partner. Moreover, in our disrupted-mode analysis, we derive a new concentration bound on the sum of Poisson random variables that requires a novel analysis of the incomplete gamma function, which may be of independent interest.

1.3. Validation via Data-Driven Simulation

Finally, we adapt our policies to accommodate the full gamut of practical constraints faced by our industrial partner (including nonstationary demand, stochastic lead times, etc.). In simulating these policies with our industrial partner's data, we find that the stochastic program and newsvendor base-stock policies provide more than 4% and 2% improvement over the status quo policy, respectively. We observe that the stochastic program policy does better when the disruption probability or demand variance is high because of a novel cost accounting scheme that we derive to more accurately capture the cost of backlogs during a disruption. The simpler newsvendor policy, on the other hand, provides satisfactory performance when average disruption lengths are small relative to system lead times. Intuitively, this is because newsvendor calculations naturally set large inventory buffers for systems with large lead times, and so, these systems are better situated to handle disruptions. Further, we find that our simple cost condition provides an effective tool for deciding when to centralize inventory during disruptions and maintain low overall costs. Thus, our simulations validate the adaptability of our approach to designing effective inventory policies, and they suggest that holding centralized warehouse inventory can provide significant cost savings in practical distribution systems with expediting and disruptions.

1.4. Literature Review

There is a large amount of literature on inventory management studying each of distribution systems, expediting, and disruptions, but (until now) no work considered these three interacting features together. To position our contributions relative to this literature, we next discuss all three topics.

Distribution systems have received the most attention, and we mention only the most relevant work here (see, e.g., Axsäter 2003, Federgruen et al. 2018 for reviews). The early work of Clark and Scarf (1960) identified that the main technical issue is allocating inventory among retailers, and it suggested relaxing this part

of the problem to derive practical heuristics. A series of works, including Eppen and Schrage (1981), Federgruen and Zipkin (1984a, b, c), and Gallego et al. (2007), formally developed these relaxations to demonstrate that the analysis becomes tractable if inventory positions can be reallocated between the retailers in any period. In the long-run average setting, this relaxed model can be solved as a multistage stochastic program, and through this lens, our stochastic programming method incorporates disruptions and expediting into this classic approach. Other techniques in the distribution system literature include Lagrangian relaxation (Kunnumkal and Topaloglu 2008, 2011) and recursive optimization using newsvendor-type calculations (Rong et al. 2017), the latter of which we also adapt to our setting with expediting and disruptions.

There are already a few studies that explicitly consider expediting in distribution systems, including the work by Moinzadeh and Aggarwal (1997) and Drent and Arts (2021), which both assume local control (retailers independently control their own orders) via augmented base-stock policies with thresholds for expediting and then, propose various methods to optimize the policy within this class. Although our model differs in some key details (e.g., centralized control, only retailers' shipments are expedited), our work builds on this literature by establishing a lower bound on the true optimal policy under central control and incorporating disruptions. Huggins and Olsen (2010) demonstrate that the difficulty of analyzing inventory models with expediting is comparable with that of lost-sales models. Other work studying expediting in various inventory systems includes Lawson and Porteus (2000) (the serial system of Clark and Scarf 1960 with an expediting option), Muharremoglu and Tsitsiklis (2008) (supermodular expediting costs), and Shen et al. (2022) (multitier model with a single location in each tier). In general, this literature demonstrates that computing the optimal solution for multitier inventory systems with expediting is nontrivial, even without the possibility of disruptions.

This brings us to the literature on inventory management with supply disruptions, which also has a rich history, with only a sampling of the most salient features provided here (we point the reader to Snyder et al. 2016 for a more detailed review). Our paper is most closely related to the periodic review inventory control setting of Song and Zipkin (1996), who introduce a Markovian state variable governing information about supply disruption conditions. This work and the subsequent literature consider many important problem features, including partial backorders (Arreola-Risa and DeCroix 1998), dual sourcing (Tomlin 2006), and discrete versus continuous supply uncertainty (Schmitt et al. 2010). Although this existing work treats inventory management for a single location, we are the first to incorporate such a Markovian disruption model into a distribution system.

2. Distribution Network Model

In this section, we introduce our base model of a distribution network with expediting and disruption. We will consider further details of our industrial partner's setting in Section 6.

2.1. Two-Tier Distribution Network

The distribution network consists of two tiers. The upper tier is a warehouse that places orders for the product with an exogenous supplier, whereas the lower tier consists of n retailers (indexed by i) that receive shipments from the warehouse and that fulfill exogenous demand from customers. Both the warehouse and the retailers may hold inventory of the product. Each retailer can use its inventory to fulfill demand from only those customers arriving at its location (i.e., no transshipment). The warehouse can use its inventory to either replenish the retailers' inventories or to expedite a shipment to directly fulfill a customer's demand. We will refer to this latter form of delivery as "expediting." Expedited shipments are differentiated from normal replenishment shipments in their cost and the time that they take to arrive. We will specify further details of the time dynamics when presenting the evolution equations below.

2.2. Demand at the Retailers

Demand for the product occurs only at the retailers (the warehouse receives no exogenous demand stream). Time is discrete and indexed by period t. Demand at retailer i in period t is denoted by the random variable $D_{i,t}$. In general, we assume that the random variables $D_{i,t}$ are independent and identical across time t but not independent nor identical across retailers i, and we let $\mathbf{D}_t$ represent the random vector of demand across retailers at time t. Unmet demand is backlogged at each retailer.

2.3. Supplier Orders and Disruptions

The firm may place replenishment orders with an exogenous supplier, and we let x_t denote the quantity of this order in period t, which arrives after a deterministic lead time of L periods. From time to time, the supplier may experience disruptions unrelated to the firm's operations that make new supply unavailable. We model this possibility using an exogenous (relative to the other state variables, random demands, and decisions made by the firm) state variable $r_t \in \mathbb{Z}_+$. If $r_t = 0$, the supplier is undisrupted and we say that the system is operating in "normal" mode in period t, whereas $r_t > 0$ indicates that the system is in a disrupted mode and that no new orders can be placed (i.e., $x_t = 0$).

The state variable r_t evolves as follows. It stays in normal mode ($r_t = 0$) for a geometrically distributed number of periods; then, when a disruption does occur, its length is drawn from a distribution with finite mean (and the decision maker sees the draw of this distribution). To model the evolution of this process, we use the nonnegative integer random variable $\tau \in \mathbb{Z}_+$, which is independent of all other processes in the system. The disruption state is updated as follows; starting in a period with $r_t = 0$, the system evolves in the next period to state $r_{t+1} = \tau_{t+1}$, where τ_{t+1} is an independent realization of τ. Thus, the probability of experiencing a disruption after an undisrupted period is $\alpha = \mathbb{P}(\tau > 0)$, whereas with probability $1 - \alpha = \mathbb{P}(\tau = 0)$, the system remains undisrupted in the next period. Then, for any period with $r_t > 0$, the state evolves in the next period to $r_{t+1} = r_t - 1$ with probability of one (i.e., once τ is drawn at the beginning of the disruption, the total length of the disruption is known to be τ periods, and we assume that the firm is able to observe this value when the disruption begins). This is reasonable in our industrial partner's setting, where suppliers often directly communicate an estimated downtime in the event of a disruption.

The stochastic process governing the state r_t is an irreducible and positive recurrent Markov chain on the nonnegative integers. Thus, there exists a stationary distribution, which is straightforward to compute using the typical balance equations, and it is characterized as follows:

$$\mathbb{P}(r_t = k) = \frac{\mathbb{P}(\tau \geq k)}{1 + \mathbb{E}[\tau]}, \quad \forall k \in \mathbb{Z}_+. \tag{1}$$

To simplify our analysis, we assume that the initial disruption state, r_1, is randomly drawn according to this steady-state distribution. We note that this assumption is without loss of generality as we consider long-run average costs, but it is made for convenience. A disruption only impacts new orders so that in any period with $r_t > 0$, the ordering policy must obey $x_t = 0$, but any orders placed before this period will still arrive at the end of their lead time L. Also, shipments from the warehouse to the retailers (regular or expedited) can still take place during a disruption as it only impacts the external supplier.

2.4. Evolution Equations

We now describe how the firm's inventory system evolves from one period to the next. Let I_t and $X_{i,t}$ denote the warehouse's and retailer i's inventory on hand at the end of period t, respectively. Similarly, let $B_{i,t}$ denote retailer i's backlog level at the end of period t (backlogs are only accumulated at the retailers). We track these state variables at the end of each period because this is when associated holding and backlogging costs will be charged (described below). We assume that the system starts empty, so $I_0 = X_{i,0} = B_{i,0} = 0$ for all i.

To complete the state description, we must specify notation for the firm's allocation and fulfillment decisions in each period. Let $z_{i,t}$ denote the warehouse's replenishment shipment to retailer i in period t, which arrives after a deterministic lead time of $l \leq L$ periods. Let $y_{i,t}$ denote the warehouse's expedited shipment to retailer i in period t, which is assumed to arrive in the same period and directly fulfill demand at retailer i Finally, let $w_{i,t}$ denote retailer i's fulfillment of demand from its local inventory in period t, which occurs without a lead time. To be concrete on the timing of these decisions, we specify the following sequence of events in period t.

1. Realize state r_t.
2. Receive supplier order x_{t-L} at warehouse.
3. Receive warehouse shipment $z_{i,t-l}$ at each retailer i.
4. Realize demand $D_{i,t}$ at each retailer i.
5. Fulfill $w_{i,t}$ of demand at each retailer i.
6. Expedite $y_{i,t}$ from warehouse to each retailer i, which clears demand immediately.
7. Send shipment $z_{i,t}$ from warehouse to retailer i.
8. Order x_t from supplier if $r_t = 0$.

With these variables specified, we are now ready to characterize the system's evolution equations. In each period t, the inventory and backlog variables evolve as follows:

$$I_t = I_{t-1} + x_{t-L} - \sum_i (y_{i,t} + z_{i,t}),$$
$$X_{i,t} = X_{i,t-1} + z_{i,t-l} - w_{i,t}, \quad \forall i \tag{2}$$

$$B_{i,t} = B_{i,t-1} + D_{i,t} - w_{i,t} - y_{i,t}, \quad \forall i. \tag{3}$$

To be feasible, the decisions x_t, $z_{i,t}$, $y_{i,t}$, and $w_{i,t}$ must ensure that the state variables remain nonnegative (i.e., $I_t \geq 0$ and $X_{i,t}, B_{i,t} \geq 0$ for i and t). Finally, we introduce one more state variable to account for inventory in transit from the warehouse to the retailers (as we need to account for holding costs on this inventory below). In particular, define $IT_t = \sum_{s=t-l+1}^{t} \sum_i z_{is}$.

2.5. Cost Parameters

The firm incurs unit holding costs h_0 and h_i for inventory held at the warehouse and retailer i, respectively, at the end of each period. Likewise, the firm incurs a unit backlog cost of b_i for backlogged demand at retailer i at the end of each period. We assume that the costs of normal shipments, $z_{i,t}$, and normal fulfillment, $w_{i,t}$, are normalized to zero, whereas the unit cost of expedited fulfillment is f_i. We assume that the expediting cost to each retailer is larger than the warehouse's holding cost (i.e., $f_i \geq h_0$ for all i). The cost incurred in period t is then $h_0(I_t + IT_t) + \sum_i (h_i X_{i,t} + b_i B_{i,t} + f_i y_{i,t})$, where we charge a holding cost for inventory in transit

based on the warehouse holding cost. In this setting, because inventory can be replenished steadily over time, it is natural for the firm to consider a long-run average cost objective. In particular, the firm seeks to minimize the following objective:

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[h_0(I_t + IT_t) + \sum_i (h_i X_{i,t} + b_i B_{i,t} + f_i y_{i,t})], \tag{4}$$

subject to (2) and (3), $I_t, X_{i,t}, B_{i,t}, x_t, w_{i,t}, y_{i,t}, z_{i,t} \geq 0, \ \forall i,t$, and $x_t = 0, \ \forall t$ s.t. $r_t > 0$.

3. Designing an Effective Base-Stock Policy

Minimizing (4) is well known to be a hard optimization problem, even in the case with no expediting or disruption (Federgruen and Zipkin 1984a, b, c; Zipkin 1984) because of the large state space and challenging allocation decisions. Therefore, the typical approach to analyzing distribution systems in the literature has been to relax the problem to a simpler system that can be solved and then, use the simpler system's solution to design a heuristic policy for the original problem. In this section, we consider a few different ways to apply this technique to our setting with expediting and disruptions. The general policy form that we use is an echelon base-stock policy, which places external orders to bring the system-wide inventory position up to an echelon base-stock level and then, distributes inventory to the retailers according to their local base-stock levels. Our primary approach uses a stochastic program to derive base-stock levels in Sections 3.1 and 3.2, but we also adapt two heuristics based on simple newsvendor calculations in Section 3.3.

Our stochastic programming approach to policy design proceeds in the following steps: (i) developing a novel stochastic program lower bound and (ii) translating its solution into an effective base-stock policy. Although our main focus is on base-stock levels, we also note that distribution systems present challenging allocation decisions, which we address in Section 3.2.1. In pursuing this approach, we are guided by the practical considerations of our industrial partner, which operates a large service parts network with thousands of customers and many thousands of parts as well as many complicating problem features (including expediting, disruptions, nonlocal fulfillment, etc. as discussed in Section 1). As such, our partner wishes to minimize changes to their current operations while still providing cost reduction benefits from policy improvements. We, therefore, aim to maintain our partner's current policy form, a base-stock policy, while mainly providing improvements on determining where the inventory is held and in what quantities. With this goal in mind, we set out to develop a self-contained stochastic program lower bound that can be used to set base-stock levels and is flexible enough to adapt to our industrial partner's setting.

The main challenges in developing a lower bound in our model stem from the expediting and disruptions features considered. But, we show below that the stochastic programming approach to deriving a lower bound can be adapted to handle both of these challenges. Expediting allows for multiple fulfillment timescales, which we accommodate with a novel cost accounting and decision timing in the stochastic program. Meanwhile, disruptions create stochastic intervals with no replenishment, which we accommodate with a dynamic cost aggregation over periods (that interacts in a nontrivial way with the expediting cost and decision timing). We note that our dynamic cost aggregation scheme extends that of Song and Zipkin (1996) (who considered a single-location, single-fulfillment mode problem) to handle multiechelon, allocation, and expediting features.

3.1. A Stochastic Program Lower Bound

We begin our analysis in this section by developing a novel stochastic program that provides a lower bound on the long-run average cost of an optimal policy for (4). Similar bounding techniques have been developed for other systems (Reiman and Wang 2015, van Jaarsveld and Scheller-Wolf 2015, DeValve et al. 2020), but the combination of multiple echelons, disruptions, and expediting in our setting makes the analysis especially challenging. Therefore, to develop intuition for our stochastic program, it will be helpful to first review the classic distribution system relaxations mentioned in the literature review of Section 1.4 (Federgruen and Zipkin 1984a, b, c; Gallego et al. 2007). These papers demonstrate that the dynamic program for a distribution system simplifies considerably if costless transhipment is allowed to reconfigure inventory positions between the retailers in any period. Specifically, with no expediting or disruptions, Federgruen and Zipkin (1984c) relax $z_{i,t} \geq 0$ (noting that $z_{i,t} < 0$ implies transshipment between retailers, allowing lower echelon local inventory to be treated in the aggregate) to obtain a variant of the classic series system of Clark and Scarf (1960). The solution of this model can be interpreted as a three-stage stochastic program; the first stage decides a system inventory position, the second stage realizes the lead time L demand and allocates inventory to the retailers, and the third stage realizes the final lead time l demand and performs fulfillment. Our stochastic program builds on this framework in a few ways.

The first adjustment that we make is to the timescales of the stages in order to appropriately accommodate the expediting feature. In the timescale of Federgruen and Zipkin (1984c), the second stage and third stage include L and l periods, respectively, so that all inventory

in transit is delivered by the end of the respective stage (to the warehouse in the second stage and to the retailers in the third stage). However, when expediting is considered, an additional complication arises; it is now possible for an order to be placed with the supplier in the midst of the second stage and arrive at the warehouse in time for expediting at the end of the third stage. Thus, a stochastic program lower bound that considers second- and third-stage horizons of L and l, respectively, must also allow for a second order from the supplier to take place during the second stage. Through extensive numeric testing, we found that such a formulation leads to a very weak bound (typically an order of magnitude too low) because of the fact that the second order can be adapted unrealistically to the realization of second-stage demand.

This challenge can be addressed with the following observation; reducing the combined timescale of the second and third stages to L eliminates the possibility of a second order arriving in time for expediting in the third stage. Then, because the third stage must be l periods to allow for allocation shipments to arrive at the retailers, the second stage should encompass $L-l$ periods. This idea clearly presents a trade-off because the the lower bound now only considers costs from L periods of demand rather than $L+l$, but we lose the second ordering opportunity, which leads to increased costs. Our numerical tests indicate that eliminating the second ordering opportunity gives a much tighter bound (and is reasonable in our industrial partner's setting as L, typically months, is much greater than l, typically days). To control for this, we introduce an adjustment term, $0 \le \hat{l} \le l$, to model how much of the last l periods of demand we include in the stochastic program; the second-stage demand includes $L+\hat{l}-l$ periods and the third-stage demand includes l periods so that in total, we consider $L+\hat{l}$ periods of demand.

Our second adjustment to the stochastic program is to aggregate costs over the length of a disruption in order to account for the fact that ordering cannot take place during these periods. As mentioned above, this extends the approach of Song and Zipkin (1996) to accommodate distribution systems and expediting, and it significantly complicates the cost accounting and decision timescales; we will explain this further while presenting the stochastic program, which we do next.

3.1.1. Demand. To formulate the stochastic program and properly account for all of the various decisions and costs, we need to define several demand random variables. We have made an effort to keep these as parsimonious as possible, but they do require introducing a bit of new notation. Let $\mathbf{D}(j)$ denote a sequence of independent and identically distributed (i.i.d.) random vectors indexed by $j \ge 1$, each with the same distribution as $\mathbf{D}_t$. Recall that τ is a random variable denoting the length of a disruption and $0 \le \hat{l} \le l$ is a lead-time adjustment. We then define four random demand vectors representing successive stages of demand over the $L+\hat{l}+\tau$ periods that we consider in the stochastic program. These vectors are defined in such a way that different combinations provide relevant demand information needed to calculate costs in the stochastic program. In particular, the first and second vectors, $\mathbf{D}_{\hat{l}}^1$ and $\mathbf{D}_{\hat{l}}^2$, respectively, combine to give the first $L+\hat{l}-l+\tau$ periods of demand, whereas the third and fourth vectors, $\mathbf{D}_{\hat{l}}^3$ and $\mathbf{D}_{\hat{l}}^4$, respectively, combine to give the last l periods of demand. These represent the two stages of demand that we consider for the decisions made in the stochastic program. Further, the first and third vectors, $\mathbf{D}_{\hat{l}}^1$ and $\mathbf{D}_{\hat{l}}^3$, respectively, combine to give the first $L+\hat{l}$ periods of demand, whereas the second and fourth vectors, $\mathbf{D}_{\hat{l}}^2$ and $\mathbf{D}_{\hat{l}}^4$, respectively, combine to give the last τ periods of demand, which is needed to calculate the holding and backlog costs during a disruption. We now formally define the vectors

$$\mathbf{D}_{\hat{l}}^1 = \sum_{j=1}^{L+\hat{l}-(l-\tau)^+} \mathbf{D}(j), \quad \mathbf{D}_{\hat{l}}^2 = \sum_{j=L+\hat{l}-(l-\tau)^++1}^{L+\hat{l}-l+\tau} \mathbf{D}(j),$$

$$\mathbf{D}_{\hat{l}}^3 = \sum_{j=L+\hat{l}-l+\tau+1}^{L+\hat{l}+(\tau-l)^+} \mathbf{D}(j), \quad \mathbf{D}_{\hat{l}}^4 = \sum_{j=L+\hat{l}+(\tau-l)^++1}^{L+\hat{l}+\tau} \mathbf{D}(j),$$

where an empty sum is assumed to be zero. As discussed above, $\mathbf{D}_{\hat{l}}^1 + \mathbf{D}_{\hat{l}}^2 = \sum_{j=1}^{L+\hat{l}-l+\tau} \mathbf{D}(j)$, $\mathbf{D}_{\hat{l}}^3 + \mathbf{D}_{\hat{l}}^4 = \sum_{j=L+\hat{l}-l+\tau+1}^{L+\hat{l}+\tau} \mathbf{D}(j)$. That is, $\mathbf{D}_{\hat{l}}^1 + \mathbf{D}_{\hat{l}}^2$ represents demand over the first $L+\hat{l}-l+\tau$ periods, which will serve as the second-stage demand in the stochastic program, and $\mathbf{D}_{\hat{l}}^3 + \mathbf{D}_{\hat{l}}^4$ represents demand over the next l periods, which we will use as the third-stage demand in the stochastic program. We note that if there is no disruption (i.e., $\tau = 0$), then the second-stage and third-stage demands correspond to $L+\hat{l}-l$ periods and l periods, respectively, exactly as described above. When there is a disruption (i.e., $\tau > 0$), our cost aggregation scheme discussed above requires that we also include an extra τ periods of demand. These periods are included in the second-stage demand, $\mathbf{D}_{\hat{l}}^1 + \mathbf{D}_{\hat{l}}^2$, so that the third-stage demand $\mathbf{D}_{\hat{l}}^3 + \mathbf{D}_{\hat{l}}^4$ only includes the final l periods over which the allocation shipments arrive at the retailers.

The main complication then arises from the cost aggregation scheme, which requires accounting for costs incurred over the final $\tau+1$ periods. In particular, with no disruption, we include costs incurred in period $L+\hat{l}$ (which are derived from the demand over the first $L+\hat{l}$ periods), but if there is a disruption, we also include costs incurred over the whole disruption period τ. Thus, our formulation requires tracking demand over the final l periods (for the allocation

decision) as well as over the final τ periods (for cost accounting). See Online Appendix B.1 for further explanation of these demand distributions. To ease notation, we use $\mathscr{D}_{\hat{l}}^2 = (\mathbf{D}_{\hat{l}}^1, \mathbf{D}_{\hat{l}}^2)$ to denote second-stage demand and $\mathscr{D}_{\hat{l}}^3 = (\mathbf{D}_{\hat{l}}^3, \mathbf{D}_{\hat{l}}^4)$ to denote third-stage demand.

3.1.2. Stochastic Program. With our demand random variables defined, we are now ready to formulate the stochastic program as follows:

$$\underline{C}_{\hat{l}} = \min_{I,\mathbf{X},\mathbf{B}\geq 0} \mathbb{E}_{\tau,\mathscr{D}_{\hat{l}}^2}[g_1(I,\mathbf{X},\mathbf{B},\tau,\mathscr{D}_{\hat{l}}^2)], \tag{5}$$

where

$$g_1(I,\mathbf{X},\mathbf{B},\tau,\mathscr{D}_{\hat{l}}^2) = \left\{ \begin{array}{ll} \min\limits_{\mathbf{z},\mathbf{y}^2\geq 0} & \mathbb{E}_{\mathscr{D}_{\hat{l}}^3}[g_2(I,\mathbf{X},\mathbf{B},\mathbf{z},\mathbf{y}^2,\tau,\mathscr{D}_{\hat{l}}^2,\mathscr{D}_{\hat{l}}^3)] \\ \text{s.t.} & \sum\limits_i z_i + y_i^2 \leq I, \\ & y_i^2 \leq B_i + D_{i,\hat{l}}^1 + D_{i,\hat{l}}^2, \ \forall i. \end{array} \right\},$$

$$g_2(I,\mathbf{X},\mathbf{B},\mathbf{z},\mathbf{y}^2,\tau,\mathscr{D}_{\hat{l}}^2,\mathscr{D}_{\hat{l}}^3) = \left\{ \begin{array}{ll} \min\limits_{\mathbf{w},\mathbf{y}^3\geq 0} & g_3(I,\mathbf{X},\mathbf{B},\mathbf{w},\mathbf{y}^2,\mathbf{y}^3,\mathbf{z},\tau,\mathscr{D}_{\hat{l}}^2,\mathscr{D}_{\hat{l}}^3) \\ \text{s.t.} & \sum\limits_i z_i + y_i^2 + y_i^3 \leq I, \\ & w_i \leq X_i + z_i, \ \forall i, \\ & w_i + y_i^2 + y_i^3 \leq B_i + D_{i,\hat{l}}^1 + D_{i,\hat{l}}^2 + D_{i,\hat{l}}^3 + D_{i,\hat{l}}^4, \ \forall i. \end{array} \right\}$$

$$\begin{aligned} g_3(I,\mathbf{X},\mathbf{B},\mathbf{w},\mathbf{y}^2,\mathbf{y}^3,\mathbf{z},\tau,\mathscr{D}_{\hat{l}}^2,\mathscr{D}_{\hat{l}}^3) &= h_0(\tau+1)\Big(I - \sum_i (z_i + y_i^2 + y_i^3)\Big)^+ \\ &+ \sum_i h_i(X_i + z_i - w_i)^+ + \sum_i \frac{f_i}{l} y_i^3 \\ &+ \sum_i b_i \Bigg((B_i + D_{i,\hat{l}}^1 + D_{i,\hat{l}}^3 - w_i - y_i^2 - y_i^3)^+ \\ &+ \left(\frac{\tau+1}{2}(D_{i,\hat{l}}^2 + D_{i,\hat{l}}^4) + \tau(B_i + D_{i,\hat{l}}^1 + D_{i,\hat{l}}^3 - w_i - y_i^2 - y_i^3) \right)^+ \Bigg). \end{aligned}$$

Working backward, the third-stage decision variables are w_i, which denotes fulfillment of retailer i demand from local inventory over all $L + \hat{l} + \tau$ periods, and y_i^3, which denotes expedited fulfillment of retailer i demand over the final l periods. The second-stage decision variables are z_i, which denotes the inventory allocation to retailer i, and y_i^2, which denotes expedited fulfillment of retailer i demand over the first $L + \hat{l} - l + \tau$ periods. The first-stage decision variables are I, which denotes the starting inventory position of the warehouse; X_i, which denotes the starting inventory position of retailer i; and B_i, which denotes the starting backlog of retailer i (such a variable capturing initial backlog is often necessary for lower bounds in network inventory systems) (see, e.g., Doğru et al. 2010, DeValve and Myles 2025). Essentially, the stochastic program can be thought of as optimizing over these initial state variables. Further explanations of the stochastic program objective and constraints are in Online Appendix B.1.

With the stochastic program defined, we now introduce one more notation to aid our analysis in subsequent sections. We would like to refer to the stochastic program (5) in terms of the stochastic demand input, $\mathscr{D}_{\hat{l}}^2$ and $\mathscr{D}_{\hat{l}}^3$, used as it will be helpful to consider using various demand inputs in different situations. To that end and with a slight abuse of notation, we let $\mathcal{SP}(\mathscr{D}_{\hat{l}}^2, \mathscr{D}_{\hat{l}}^3)$ denote the stochastic program (5) with input second-stage demand $\mathscr{D}_{\hat{l}}^2$ and third-stage demand $\mathscr{D}_{\hat{l}}^3$. We are now ready to present our main result of this section; the stochastic program with $\hat{l} = 0$ provides a lower bound on the long-run average cost of any policy and hence, also a lower bound on the optimal long-run average cost.

Theorem 1. *The long-run average cost in* (4) *of any feasible policy is larger than* $\frac{\underline{C}_0}{1+\mathbb{E}[\tau]}$.

The impact of Theorem 1 is twofold. First, it provides a formal derivation of a stochastic program approximating the cost of an optimal policy for this complex system. We would like to emphasize that the stochastic program is an approximation (rather than an exact characterization) of the system cost because it relaxes the problem's decision-making horizons from the period level to the lead-time level. This leads to a more tractable stochastic programming formulation (as opposed to the full dynamic program that suffers from the curse of dimensionality), which suggests inventory positions for the warehouse and retailers that can be used to design effective heuristic policies (as we demonstrate in the next section). Second, the stochastic program provides the first lower bound on optimal cost for this model in the literature, and thus, it also offers the opportunity to indirectly compare the cost of the heuristic policy with an optimal policy. This can help give confidence in a policy's performance for practical applications.

3.2. A Base-Stock Policy

In this section, we use the solution of the stochastic program to design a base-stock policy. Let I^*, $\mathbf{X}^*$, $\mathbf{B}^*$, and $\mathbf{z}^*$ denote an optimal solution to the stochastic program $\mathcal{SP}(\mathscr{D}_0^2, \mathscr{D}_0^3)$. Let

$$S_0 = I^* + \sum_i (X_i^* - B_i^*)^+, \tag{6}$$

$$S_i = (X_i^* - B_i^*)^+ + \mathbb{E}[z_i^*], \ \ \forall i \tag{7}$$

denote the stochastic program's system-wide starting inventory position and average inventory position for each retailer i, respectively. We will use these quantities as base-stock levels.

3.2.1. Ordering, Expediting, and Allocation. First, when the supplier is undisrupted, the warehouse orders from the external supplier to bring the system-wide or

echelon inventory position up to S_0. Specifically, if $r_t = 0$, the warehouse places an order of size

$$x_t = S_0 - \left(I_t + IT_t + \sum_{s=t-L+1}^{t-1} x_s + \sum_i (X_{i,t} - B_{i,t}) \right). \quad (8)$$

Next, we specify the fulfillment, expediting, and allocation policy. First, we perform as much local fulfillment, $w_{i,t}$, as possible (i.e., the minimum of the current backlog and local inventory). Then, if there is remaining backlog, we myopically expedite, $y_{i,t}$, from the warehouse to clear as much backlog as possible (up to the warehouse on-hand inventory), breaking ties across retailers in descending order of the backlog costs, b_i (in our numerical experiments, we found that myopically prioritizing expediting to clear backlogs performed best). Finally, after fulfillment and expediting, we allocate remaining warehouse inventory to each retailer i to bring its inventory position up to S_i if possible. Specifically, letting $IT^-_{i,t} = \sum_{s=t-l+1}^{t-1} z_{i,s}$ denote retailer i's inventory in transit just before the period t allocation decision, the warehouse sends to retailer i a shipment of size $z_{i,t} = S_i - (X_{i,t} + IT^-_{i,t} - B_{i,t})$ as long as such a shipment is feasible for all i given the inventory on hand at the warehouse (i.e., if $\sum_i z_{i,t} \le I_{t-1} + x_{t-L} - \sum_i y_{i,t}$ for $z_{i,t}$ specified above). If this is not feasible, then we solve a simple allocation problem to break ties. In particular, we follow Federgruen and Zipkin (1984c) and Gallego et al. (2007) in defining a myopic allocation problem that chooses the current retailer inventory positions to minimize the expected backlogging and holding costs a lead time l later, with an additional constraint that the inventory positions be less than our base-stock levels S_i. In particular, letting $\mathbf{D}^l = \sum_{j=1}^{l} \mathbf{D}(j)$ denote a random vector with the distribution of lead time l demand, we solve

$$\begin{aligned} \min_{\mathbf{z}_t \ge 0} \ & \mathbb{E}\left[\sum_i (h_i - h_0)(X_{i,t} + IT^-_{i,t} + z_{i,t} - B_{i,t} - D^l_i)^+ \right. \\ & \left. + b_i (D^l_i - X_{i,t} - IT^-_{i,t} - z_{i,t} + B_{i,t})^+ \right] \quad (9) \\ \text{s.t.} \ & \sum_i z_{i,t} \le I_{t-1} + x_{t-L} - \sum_i y_{i,t}, \\ & X_{i,t} + IT^-_{i,t} + z_{i,t} \le S_i, \ \forall i. \end{aligned}$$

This completes the description of our base-stock policy derived from the stochastic program (5). Next, we consider the same policy with base-stock levels set by an alternative stochastic program.

3.2.2. Alternative Base-Stock Levels. In our numerical experiments, we observe that the stochastic program $\mathcal{SP}(\mathscr{D}^2_0, \mathscr{D}^3_0)$ can sometimes underestimate the base-stock levels because of the fact that the stochastic program only considers $L + \tau$ periods of demand rather than the full $L + l + \tau$ periods of uncertain demand that a supplier order may have to cover. For this reason, we also consider setting base-stock levels using the alternative stochastic program $\mathcal{SP}(\mathscr{D}^2_l, \mathscr{D}^3_l)$, which includes demand over a full $L + l + \tau$ periods. Although this stochastic program no longer provides a lower bound on optimal cost, we do find it to be helpful in designing a heuristic base-stock policy. Intuitively, if the stochastic program (5) can underestimate the base-stock levels, then one would expect using the longer time horizon of $L + l + \tau$ periods (without allowing a second order) to overestimate them. Indeed, this is precisely our observation in simulations, and so, we conjecture that the best base-stock levels would be somewhere in between those suggested by the two stochastic programs, which we explore via simulation in Section 5.

3.3. Newsvendor-Based Policies

We also adapt two other hueristics that rely on simple newsvendor-type calculations to our setting motivated by the distribution system literature. First, we include the classic base-stock ordering and myopic allocation policy developed in the stream of literature, including Federgruen and Zipkin (1984a, b, c), Zipkin (2000), and Gallego et al. (2007), which we call the FZ heuristic based on the original authors. Next, we develop a heuristic based on newsvendor calculations inspired by the recursive optimization approach of Rong et al. (2017) for setting base-stock levels in distribution systems, which we call the NV heuristic for newsvendor. Because of space constraints, we describe these policies in detail in Online Appendix C.1.

4. Disruption Intervention Allocation Policies

By default, our stochastic program policies plan inventory levels to account for the disruptions and therefore, maintain consistent allocation and fulfillment policies during disruptions. However, in practice, many firms intervene with altered control policies during disruptions in an attempt to minimize backlogs experienced by customers. A simple policy currently used by our industrial partner is to centralize all newly arriving inventory at the warehouse when a disruption occurs and then, fulfill all orders directly to the customer with expedited shipping (once any remaining local inventory at the retailer is exhausted[2]). This centralized control allows the automaker to ensure appropriate prioritization of fulfillment decisions, and it can, therefore, reduce the backlog costs during disruptions. But, it can also increase fulfillment costs because each unit of demand is filled through high-cost expedited shipping rather than going through the retailers. Thus, in this section, we focus on this trade-off between backlog and fulfillment costs in order to answer a fundamental

question for operating in disrupted mode. When is it better to keep inventory decentralized at the retailers rather than centralized at the warehouse? Our industrial partner can currently switch between these modes but desires a more principled criterion for when to switch; this section identifies the switch point. Specifically, we derive a simple and robust rule of thumb for making this comparison, so we next develop a simplified version of our model that captures the key trade-off.

First, as our analysis in this section focuses on the period when supply is disrupted, we assume that the system has just entered a new disruption phase of length τ periods (which conditioned on the disruption, is deterministic). To concentrate on the trade-off between backlog and fulfillment costs, in this section, we assume that the n retailers have identical cost parameters $b_i = b$ and $f_i = f$ for all i, and we also assume that the retailers have i.i.d. Poisson demand distributions. Further, to facilitate our analysis of the expected backlog costs, we approximate the periodic arrivals described in the Section 2 model with a continuous Poisson process. In particular, demand at each retailer i is approximated with a Poisson process on the interval $[0,\tau]$ with rate λ/τ per period so that the cumulative demand at retailer i over the course of the disruption is a Poisson random variable with rate λ, and we let D_i denote this demand. Thus, the total expected demand across all retailers until the disruption ends is $n\lambda$, and we let $D^n = \sum_i D_i$ denote this system-wide demand. We also assume that there are $a \geq 1$ units of inventory available per retailer, so there are na units in the system (representing aggregate retailer inventory); for simplicity, we assume that they are all in the pipeline (i.e., none local at retailers), so they can all be centralized or be decentralized. Because a disruption is only meaningful for our industrial partner when it creates a supply shortage, we focus on the case when $\lambda \geq a$ (i.e., expected demand exceeds supply). We also assume that holding costs are identical across the retailers and warehouse and so, can be effectively ignored and that $n \geq 2$ (otherwise, decentralization is always optimal).

Our analysis critically relies on a novel concentration bound for tail probabilities of Poisson sums. The intuition is that when n grows, the Poisson variance of D^n grows more slowly than the mean, so the coefficient of variation decreases. Therefore, D^n is more concentrated around its mean than the single Poisson random variable D_i, and thus, when $\lambda \geq a+1$, there is a higher probability of D^n being larger than a value strictly below the mean.[3] The full proof of this result (along with others in this section) is in Online Appendix B.

Lemma 1. *For D_i, $1 \leq i \leq n$, i.i.d. Poisson random variables with rate $\lambda \geq a+1$ with $a \in \mathbb{Z}_+$, and $D^n = \sum_i D_i$, we have $\mathbb{P}(D^n \geq na) \geq \mathbb{P}(D_i \geq a)$.*

Theorem 2. *For $n \geq 2$ retailers, $a \geq 1$ inventory per retailer, and Poisson rate $\lambda \geq a$ per retailer, decentralization provides lower cost than centralization if $f \geq b\tau\left(\frac{0.621}{\lambda} + \frac{0.150}{a}\right)$.*

Theorem 2 verifies the intuition that decentralization is preferred when the expediting cost, f, is large relative to the backlog cost, b, and it also demonstrates the natural conclusion that shorter disruption times, τ, favor decentralization. It also suggests that decentralization becomes a better option as either demand or supply grows large, which intuitively follows because inventory is less likely to get trapped at one retailer as excess, while another retailer has backlog (see Online Appendix B.3 for more discussion). Primarily, Theorem 2 shows that decentralization can be an effective policy for some disruptions, and the cost condition provides an easy rule of thumb for making this decision, which we test in simulations in the next section.

5. Policy Benchmarking: Simulations

In this section, we assess the performance of the policies considered by comparing them against the benchmark heuristics FZ and NV from the distribution system literature (described in detail in Section 3.3). We perform these comparisons via numerical simulations on a synthetic data set in order to assess the best policies to test with our industrial partner's data in the next section.

5.1. Implementation of Our Base-Stock Policies

Following the discussion from Section 3.2.2 of what lead-time demand to include in our stochastic program, we test three alternatives for setting base-stock levels. The first is the policy described in Section 3.2.1 using base-stock levels derived from $\mathcal{SP}(\mathscr{D}_0^2, \mathscr{D}_0^3)$, which we denote by SP^L. The second is identical to the first except that the base-stock levels are calculated using an optimal solution of $\mathcal{SP}(\mathscr{D}_l^2, \mathscr{D}_l^3)$ in Equations (6) and (7), and we denote this policy as SP^{L+l}. The third is again identical to the first two except that it calculates base-stock levels using the average of the base-stock levels in SP^L and SP^{L+l}, and we denote this third policy by SP^{avg}. These policies can be seen as a first approximation to searching for the best base-stock levels in between those suggested by the two stochastic programs (similar to approaches suggested by Gallego et al. 2007, Reiman and Wang 2015). For brevity, we first test these three policies in Online Appendix C.2.2 and choose SP^{L+l} and SP^{avg} to focus on for the remainder of the synthetic simulations as they provide the most consistent performance across problem instances.

5.2. Disruption Allocation Policies

Building on our discussion in Section 4, we test three disruption allocation policies. The first policy we call "none" (N) as it simply leaves base-stock levels

unchanged during a disruption. The next policy is motivated by our industrial partner's practice of centralizing all inventory at the warehouse during a disruption, which we call "designate for intervention" (DFI), and it is achieved by setting all retailer base-stock levels to zero during the disruption. The next policy is a modified version of designate for intervention (which we call MDFI); this is motivated by the observation in Section 4 that it is often better to keep some inventory decentralized during a disruption. The high-level idea is to translate the condition for the stylized model in Theorem 2 to be applicable for the automaker's distribution network. To that end, consider the centralize/decentralize decision at the beginning of a disruption of length τ that has begun in period t. We approximate the condition of Theorem 2 using system averages for each parameter for each part; let $\tilde{f} = \sum_i f_i/n$ denote the average expedited fulfillment cost in the network, $\tilde{b} = \sum_i b_i/n$ denote the average backlog cost in the network, $\tilde{\lambda}_t = \sum_i \sum_{s=t+1}^{t+\tau} \lambda_{i,s}/n$ denote the average Poisson demand rate per retailer over the τ disruption periods, and $\tilde{a}_t = (I_t + IT_t + \sum_{s=t-L+1}^{t-1} x_s + \sum_i (X_{i,t} - B_{i,t}))/n$ denote the average system-wide inventory position per retailer. Then, following Theorem 2, our heuristic keeps inventory decentralized if

$$\tilde{f} \geq \tilde{b}\tau\left(\frac{0.621}{\tilde{\lambda}_t} + \frac{0.150}{\tilde{a}_t}\right). \tag{10}$$

In other words, if this condition holds, we simply maintain the existing retailer base-stock levels over the course of the disruption, whereas if this condition does not hold, we implement DFI by setting base-stock levels to zero during the disruption.

5.3. Our Simulation Setting

In conjunction with our industrial partner, we chose an initial set of 54 problem instances inspired by the range of parameters found in their data in order to perform an initial test of the policies. In this set, we consider $n \in \{4, 8, 16\}$ retailers and two different lead-time combinations of $(L, l) \in \{(10, 2), (20, 5)\}$. We assume that all retailers have independent Poisson demand, and we consider three different sets of rates. (i) All retailers have Poisson rate 1 per period. (ii) All retailers have Poisson rate 10 per period. (iii) Half of the retailers have Poisson rate 1 per period, and half of the retailers have Poisson rate 10 per period. For costs, all instances have $h_0 = h_i = 1$, $b_i = 10$, and $f_i = 15$ for all i. It remains to specify the distribution of the disruption parameter τ. To do so, it is helpful to break τ down into two components: one determining the binary probability of whether a disruption occurs (i.e., $\tau > 0$) and one for the length of the disruption conditional on a disruption occurring. To do so, define two independent random variables $\theta \sim \text{Bern}(\alpha)$ (i.e., θ has a Bernoulli distribution with $\mathbb{P}(\theta = 1) = \alpha$) and $\hat{\tau} \sim \text{Pois}(14) + 1$ (i.e., $\hat{\tau} \geq 1$ has a shifted Poisson distribution starting from one and a mean of 15). Then, let $\tau = \theta\hat{\tau}$ (i.e., with probability $1 - \alpha$, there is no disruption, $\tau = 0$, and with probability α, there is a disruption with a shifted Poisson distribution of mean 15, $\tau = \hat{\tau} > 0$). In our industrial partner's setting, disruptions are relatively rare, so we test disruption probabilities $\alpha \in \{0.005, 0.01, 0.02\}$. Thus, considering all combinations of three values of n, two combinations of lead times, three sets of demand rates, and three disruption parameters, we test 54 problem instances in all.

We note that in the problem instances considered, we have chosen $h_i = h_0$ (i.e., the warehouse and retailer holding costs are the same) as this reflects the accounting assumptions made by our industrial partner. In this cost regime, Zipkin (2000) notes that the FZ heuristic holds no inventory at the warehouse, which is indeed our industrial partner's current practice. This can be verified in our setting by inspection of the myopic allocation (9), which the FZ heuristic implements without the second set of constraints. Because $h_i = h_0$, the objective function has no positive coefficient on the values of $z_{i,t}$, and hence, they will always be set to a maximum cumulative value that makes the first constraint binding (i.e., no inventory will be left at the warehouse). On the other hand, because our allocation policy in the SP base-stock policies implements a limit on the retailers' inventory positions, it often leaves inventory at the warehouse, which is helpful to provide the option for future expediting. Similarly, the NV heuristic includes a specific newsvendor calculation accounting for the role of expediting, and also, it typically leaves inventory at the warehouse. Thus, at a high level, these simulations can be seen as identifying the value of holding some inventory at the warehouse (even when the classic analysis based on holding costs would suggest otherwise) in order to take advantage of expediting. See Online Appendix C.2.1 for details on the simulation implementation.

5.4. Simulation Results

5.4.1. Disruption Probability. Table 1 summarizes the performance of the four base-stock policies with the N disruption allocation policy. To provide a normalized and concise cost comparison between these policies, we denote $\pi^{\min}$ as the minimum cost base-stock policy among the four considered for each instance, and then, we compare the cost of each policy with $\pi^{\min}$. Table 1 illustrates the ratio of average daily costs between $\pi^{\min}$ and each of the policies[4] broken down by the different disruption probabilities, α, as well as the overall cost comparison on the final row. To report statistical significance of cost differences, in each row of Table 1, we compute a 95% confidence interval for the difference in average cost on instances in that row between each pair

Table 1. Average Cost Ratios of Policies with the "None" Disruption Allocation Heuristic Compared with $\pi^{\min}$

Disruption probability, α	SP^{L+l}-N	SP^{avg}-N	FZ-N	NV-N
0.005	1.013**	1.098	**1.005****	1.037*
0.01	1.078	1.021*	1.030*	**1.009****
0.02	1.033**	**1.006****	1.205	1.113*
Overall	**1.041****	1.042**	1.080	1.053*

of policies (i.e., each row has 6 = (4 choose 2) comparisons), and we report the number of confidence intervals that do not overlap zero (i.e., differences that are statistically significant at the 95% level) with the number of asterisks next to each number. Then, the policy (or policies) with the statistically significant lowest cost in the row in Table 1 is set in bold.

Table 1 gives a good sense of where each policy performs well. First, when the disruption probability is low, $\alpha = 0.005$, FZ-N has the statistically significant lowest cost. The fact that FZ-N performs well with low disruption probability is not surprising because it is a policy designed for distribution systems without disruptions, and indeed, its performance relative to $\pi^{\min}$ degrades significantly as the disruption probability increases. This illustrates our claim that disruptions need to be taken into account in inventory planning, but also, it highlights that if the disruption probability is low enough, then the classic approach of Gallego et al. (2007) still works well. Meanwhile, the SP^{avg}-N policy displays the opposite trend, having high cost when the disruption probability is low but improving significantly as the disruption probability increases so that it is the best policy alone when $\alpha = 0.02$.

To better understand this phenomenon, we consider Table 2, which lists the average system-wide base-stock level of each policy (i.e., the sum of base-stock levels for the warehouse and all retailers) as well as the average percentage of that total that is allocated to retailers (i.e., the sum of retailer base-stock levels divided by the total). Table 2 illustrates why each policy performs well at different disruption levels. When $\alpha = 0.005$, SP^{L+l}-N, FZ-N, and NV-N are each well calibrated to have an appropriate amount of system-wide inventory (a little over 1,000), mostly at the retail level, which is appropriate when there is little risk of disruption. We also see that SP^{avg}-N does not perform well here because although it maintains an appropriate *relative* level of decentralized inventory (98.8% at the retailers is in line with the other policies), its system-wide base-stock level is too low, which is a function of not considering the full $L + l$ periods of lead-time demand.

However, when the disruption probability increases to $\alpha = 0.02$, it is clear that the FZ-N and NV-N policies do not do a good enough job of calibrating either their system-wide base-stock levels or the percentage allocated to retailers (both of which remain mostly flat). Meanwhile, in this regime, the dominant SP^{avg}-N policy illustrates that both system-wide and centralized inventory levels need to be increased significantly to deal with the increased disruption risk. This naturally leads to the question of *why* the stochastic programming-based policies do a better job of calibrating centralized inventory to disruption risk than the newsvendor-based heuristics: FZ-N and NV-N. Although there are a number of issues at play in this setting, we focus on two intuitive explanations that may be helpful when deciding on useful policies: (i) demand variance and (ii) cost accounting.

5.4.1.1. Demand Variance. Our first observation is that the sporadic nature of disruptions leads to high variance demand distributions, which make it very difficult for newsvendor-type calculations to calibrate inventory for disruptions. For example, consider one of our simulation instances with four retailers, each with rate 1 Poisson demand, and lead times of $(L, l) = (10, 2)$. The NV heuristic considers demand over $L + l + \tau$ periods. Regardless of the value of τ, the demand over $L + l$ periods in this instance is a Poisson random variable with mean $48 = (10 + 2) \times 4$ and also, variance of 48 (because it is Poisson). Adding the additional demand over τ periods, however, significantly increases the variance. This is because this portion of demand has a low probability of materializing (0.5%, 1%, or 2%) but a large mean of $60 = 15 \times 4$ if it is realized. To illustrate this, we simulated 100,000 samples of the compound demand distribution for this system over $L + l + \tau$ periods and observe the variances to be 67, 86, and 123 for disruption probabilities of 0.5%, 1%, and 2%, respectively. This high variance impacts the newsvendor calculations carried out by the FZ and NV heuristics, which calculate percentiles of the demand distribution, the largest being $b/(b + h) = 10/11 \approx 91\%$ in this instance.[5] However, the empirical 91st percentile of the $L + l + \tau$ period system demand in this simulation is 58 for each of the disruption probabilities 0.5%, 1%, and 2%. In other words, simply finding percentiles of the distribution is not precise enough to distinguish between widely different variances. Intuitively, this is because the very highest demand realizations occur in roughly the top 0.5%–2% of scenarios when a disruption happens, but the 91st percentile of the distribution does not capture the impact of these scenarios. Therefore, as we can see in Table 2, the NV-N policy keeps roughly the same inventory as the disruption probability increases.

5.4.1.2. Cost Accounting. The previous observation obviously begs the question of why the SP-based policies do a better job of calibrating inventory to the higher variance distributions arising from higher disruption

Table 2. System Base-Stock Levels and Percentage Allocated to Retailers

	System base-stock level				Allocated to retailers, %			
Disruption probability, α	SP^{L+l}-N	SP^{avg}-N	FZ-N	NV-N	SP^{L+l}-N	SP^{avg}-N	FZ-N	NV-N
0.005	1,029	938	1,008	1,051	98.6	98.8	100.0	96.8
0.01	1,322	1,232	1,008	1,053	89.8	88.8	100.0	96.8
0.02	1,415	1,325	1,008	1,061	88.4	87.7	100.0	97.2
Overall	1,255	1,165	1,008	1,055	92.3	91.8	100.0	96.9

probabilities. The answer comes from our novel cost accounting in the stochastic program (5). In particular, our third-stage cost function g_3 estimates total backlogging costs for the entire duration of the τ period disruption using the coefficient $(\tau+1)/2$ (see details on the derivation of this coefficient in Online Appendix B). This has the effect of inflating the cost of backlogging during a disruption in our stochastic program, giving it an incentive to increase inventory levels to minimize this cost. This effect is observed in the rising inventory quantities of the SP policies in Table 2. Thus, we conclude that the newsvendor-based policies should be expected to do well in settings with low demand variance and low disruption probability, whereas the SP policies are able to better handle high variance and high disruption probability.

5.4.2. Lead Times. Given the previous section's observation that disruption probability and demand variance are key drivers of policy performance, in this section, we explore the nuance introduced by lead times. Table 3 reports policy performance split out by disruption probability and lead time with the same notation conventions as Table 1. Tables 1 and 3 tell a similar story, but Table 3 reveals the nuance that the advantage of the SP policies with high disruption probabilities is less pronounced with longer lead times; although an SP policy always performs best with lead times $(10,2)$, when lead times are $(20,5)$, an SP policy is best only with the most extreme 2% disruption probability. Intuitively, this is because the arguments of the previous section about $L+l+\tau$ period demand variance stemming from the τ disruption periods are muted when the $L+l$ periods make up a greater portion of demand. Thus, we refine our observations to say that the SP-based policies seem to perform well when the $L+l+\tau$ period demand draws a large portion of its variance from the τ disruption periods. Put in simpler terms, we expect the SP policies to do better when the disruption length is larger relative to the system lead times. This is intuitive as systems with shorter lead times typically have less inventory buffers and therefore, need to more carefully take disruptions into account.

5.4.3. Value of Centralized Inventory. Here, we illustrate more explicitly the value of centralized inventory. We modify the SP^{L+l} policy to solve the stochastic program with the added constraint that the central warehouse base-stock level is zero (allowing the retailer inventory levels to be set higher), with all other parts of the policy held the same, which we call SP^{L+l}_{nc}. We use "N" disruption allocation for both policies, but we drop this notation for brevity. Table 4 summarizes the results of this simulation comparing the two policies broken out by lead times and disruption probability. We see that the policy with centralized inventory reduces cost by 4.1% on average and consistently provides lower cost across the board, which is statistically significant in all but one row of Table 4. Further, the system base-stock levels show that the centralized inventory allows SP^{L+l} to be more efficient with less inventory as it is flexibly able to deploy this inventory across the network. We, therefore, conclude that centralized inventory is a key for successful inventory management in distribution systems with expediting and disruptions.

5.4.4. Demand Rates and Number of Retailers. Next, we consider the impact of demand rates and the

Table 3. Average Cost Ratios Compared with π^{min} Across Disruption and Lead-Time Parameters

Lead times (L,l)	Disruption probability, α	SP^{L+l}-N	SP^{avg}-N	FZ-N	NV-N
(10, 2)	0.005	**1.008*****	1.065	1.008**	1.029*
(10, 2)	0.01	1.032	**1.006****	1.053	1.008*
(10, 2)	0.02	**1.001*****	1.003**	1.332	1.198*
(20, 5)	0.005	1.017**	1.131	**1.002*****	1.045*
(20, 5)	0.01	1.124	1.036*	**1.007****	**1.010****
(20, 5)	0.02	1.066*	**1.009*****	1.078	1.028*

Table 4. Average Cost Ratio of SP^{L+l} with and Without Inventory Centralization and Average Base-Stock Levels

Lead times (L,l)	Disruption probability, α	$\frac{SP^{L+l}}{SP^{L+l}_{nc}}$	System BSL		BSL at retailers, %	
			SP^{L+l}	SP^{L+l}_{nc}	SP^{L+l}	SP^{L+l}_{nc}
(10, 2)	0.005	0.940*	685	750	99.1	100.0
(10, 2)	0.01	0.996	988	986	88.2	100.0
(10, 2)	0.02	0.983*	1,079	1,094	86.4	100.0
(20, 5)	0.005	0.895*	1,373	1,468	99.4	100.0
(20, 5)	0.01	0.972*	1,656	1,656	89.0	100.0
(20, 5)	0.02	0.970*	1,751	1,776	88.9	100.0
Overall		0.959*	1,255	1,288	91.3	100.0

Note. BSL, base-stock level.

number of retailers on policy performance as summarized in Table 5, which largely supports the observations made in previous sections. First, we see that the SP^{avg}-N policy does best when demand is low. Because we have assumed Poisson demand, the low demand setting has the highest standard deviation of demand relative to its mean (because Poisson mean and variance are equal). Thus, this observation is in line with our claim that the SP policies are expected to perform better under higher demand variance. Second, we observe that the average performance of all policies remains relatively stable across the number of retailers considered, suggesting that problem scale is not a main driver of policy differentiation in this context.

5.4.5. Disruption Allocation Policies. The policy comparisons to this point have focused on the "none" disruption allocation policy; here, we also consider DFI and MDFI. The performances of nine policy combinations (three base-stock policies SP^{avg}, FZ, and NV combined with three disruption allocation policies N, DFI, and MDFI) are summarized in Table 6.[6] Although DFI tends to increase costs regardless of the base-stock levels used, we note that our industrial partner's motivation for implementing this policy is more about prioritizing backlogs during disruptions than overall cost. In particular, Table 7 illustrates the impact of disruption allocation policies on backlogs during disruptions, a main concern of our industrial partner. Table 7 reports the average daily backlog and expediting quantities during normal and disrupted system states for the three disruption allocation policies (the table reports SP^{avg}, but other policies are similar). During a disruption, there are on average 130.6 customers per day waiting for service in these simulations, whereas the average demand is 51.3 customers per day, implying that about 2.5-days worth of customer demand is waiting for service on average during a disruption, which is clearly not a desirable situation for our industrial partner. This motivates the DFI policy, which cuts the average disruption backlog by 62% to 49.3 per day by increasing the expedited fulfillment during disruptions. Meanwhile, our modified DFI policy, MDFI, seeks to address the increased costs of the DFI policy by only implementing it when the condition in (10) is not met (which is not the case for most disruptions in the simulation, so the N and MDFI policies have very similar backlog levels). Furthermore, in terms of average cost, the MDFI policy significantly outperforms the DFI policy, bringing its cost much closer to the N policy. This provides an attractive option for our industrial partner: cost reduction through eliminating unnecessary interventions while maintaining the ability to control backlogs with centralization during more extreme disruptions.

Table 5. Average Cost Ratios Compared with π^{min} for Demand Rates and Number of Retailers

Parameter	SP^{L+l}-N	SP^{avg}-N	FZ-N	NV-N
Demand rates $(\lambda_{low}, \lambda_{high})$				
(1, 1)	1.039*	**1.006****	1.062	1.036**
(1, 10)	1.051*	**1.050****	1.076	1.048*
(10, 10)	**1.035****	1.069**	1.102	1.075*
No. of retailers				
4	1.039**	**1.030****	1.076	1.049*
8	**1.043****	**1.041****	1.079	1.052*
16	**1.042****	1.054**	1.084	1.059*

Our closing observation relating to these synthetic simulations is that our industrial partner's setting considered in Section 6 incorporates several additional problem features, including nonstationary demand, which requires solving multiple stochastic programs to set varying base-stock levels over time. Given the fact that SP^{L+l} and SP^{avg} perform similarly and that SP^{L+l} requires solving only one stochastic program, we choose to focus on this policy in Section 6.

6. Assessing Policies in a Practice-Based Context

In this section, we adapt our policies for use in the service parts distribution network of a large U.S. automobile manufacturer, which incorporates a number of

Table 6. Average Cost Ratios of Policy Disruption Allocation Combinations Compared with π^{min}

Base-stock policy	SP^{avg}			FZ			NV		
Disruption probability, α	None	DFI	MDFI	None	DFI	MDFI	None	DFI	MDFI
0.005	1.098	2.240	1.098	1.005	1.594	1.005	1.037	2.418	1.037
0.01	1.021	2.143	1.025	1.030	1.767	1.030	1.009	2.049	1.010
0.02	1.006	1.904	1.012	1.205	2.152	1.207	1.113	1.778	1.114
Overall	1.042	2.096	1.045	1.080	1.838	1.081	1.053	2.082	1.054

additional practical features relative to the model of Section 2, including a three-tier network, nonstationary demand, stochastic lead times with different distributions at the retailers, and transhipment between retailers. Further, much of this model was estimated directly from our industrial partner's data, which required significant effort, including decensoring demand, forecasting nonstationary demand, estimating dealer review periods and inventory policies, and estimating lead-time distributions (details are given in Online Appendix C.3).

6.1. Introducing Our Industrial Partner's Current Practice

We collected data on 23 of our industrial partner's service parts, with weights, labor costs, holding costs, shipping costs, and demand information representative of the broad spectrum of parts carried by our partner. Each part is supplied with some lead time by an external supplier in normal mode. In the following section, we describe the automaker's current practice for managing inventory of these parts.

6.1.1. Three-Tier Distribution Network. The automaker has a fulfillment network including a central warehouse and 15 regional distribution centers. In our partner's context, the RDCs operate in the retailer role of our base model in Section 2 because they receive shipments allocated from the central warehouse and fulfill demand from customers, and thus, we use the term RDC in place of retailer in this section. For each RDC, its customers are automobile dealers in the region, each of which in turn serves demand from their customers (i.e., end customers who need repairs on their automobile). There are a total of more than 4,400 dealers served by our industrial partner. Relative to our model in Sections 2–5, a new important feature here is that the dealers can hold their own inventory and thus, represent a third tier in the inventory network. However, because the dealers are independent businesses (i.e., the automaker does not control their inventory policy), our model in Section 2 still provides a good approximation because dealers are exogenous to our industrial partner's network and are served as customers of the RDCs. Moreover, we validate our approach by verifying that a base-stock policy is appropriate for a model with this new feature in Online Appendix B.4. Finally, in the automaker's network, some of the RDCs have assigned "partner" RDCs nearby, which act as a backup option for fulfillment (see Online Appendix C.3 for more details). We describe the partner RDCs with the fulfillment policy in Section 6.1.5.

Table 7. Average Daily Backlog and Expediting During Normal and Disrupted Periods for SP^{avg}

Metric	State	SP^{avg}-N	SP^{avg}-DFI	SP^{avg}-MDFI
Backlog	Normal	0.0	0.0	0.0
	Disrupted	130.6	49.3	130.0
Expediting	Normal	9.1	5.2	9.0
	Disrupted	0.3	10.8	0.3

6.1.2. Demand Distributions. The demand that an RDC sees each day (in the automaker's context, a period corresponds to one day) comes in two streams: normal (nonemergency) and customer (emergency) orders. This arises from the fact that dealers hold inventory. A normal order that a dealer places is a replenishment to get up to their base-stock level, whereas a customer order happens when the dealer is stocked out and has a customer waiting. Together with our industrial partner, we estimate daily end-customer demand (from those who arrive at dealers) as a heterogeneous Poisson process. Moreover, we estimate the dealers base-stock levels based on their frequency of normal orders versus customer orders. We provide more details on demand and base-stock levels estimation in Online Appendix C.3.

6.1.3. Lead Times. In practice, the automaker experiences uncertain lead times, so we adapt our model to incorporate this feature. The automaker models the supplier's lead time to the central warehouse as well as the warehouse's lead time to the RDCs as independent normal random variables (rounded to integers) whose means and variances are estimated from their data. We use these distributions to implement a standard model of stochastic sequential lead times in our simulations (see Online Appendix C.3 for a detailed description). Moreover, in accordance with our industrial partner's practice, we assume immediate fulfillment from local RDCs to dealers (i.e., dealers who place orders on day

t receive supply at the beginning of day $t+1$ to clear backlogs immediately; this reflects our partner's practice of daily milk-run deliveries from the RDC to dealers), and we assume a one-day lead time for the warehouse's expedited shipments and a two-day lead time for local partner RDC's fulfillment.

6.1.4. Disruptions. Our industrial partner faces infrequent supplier disruptions that limit the possibility of placing new orders for periods of time. Given that the automaker orders from the supplier on a weekly basis, the disruption lengths are modeled in units of weeks, and a typical disruption may range from one to several weeks long. Given the possibility of different disruption scenarios, our industrial partner finds it helpful to test several disruption model scenarios to get a sense of performance in each. Following the disruption model of Section 5.3, we model the random variable τ in two parts: $\theta \sim \text{Bern}(\alpha)$, which represents the chance of a disruption happening with probability α, and $\hat{\tau} \sim \text{Pois}(\beta)+1$, which is a shifted Poisson random variable representing the length of a disruption in weeks with a mean of $\beta+1$. Then, the disruption random variable in days is $\tau = 7\theta\hat{\tau}$. Together with our industrial partner, we choose to test three values of $\alpha \in \{0.005, 0.01, 0.02\}$ and two values for the mean disruption length $\beta+1 \in \{2,4\}$. This leads to six disruption model scenarios, each of which we test on 23 parts, resulting in a total of 138 simulation scenarios.

6.1.5. Industrial Partner's Current Policy. On each day, customers bring cars to dealers for repair. Dealers, based on their own base-stock policies and on-hand inventories, place normal and customer orders to their local RDCs. The RDCs prioritize customer orders and fulfill normal orders with any remaining inventory. In the event that customer orders exceed an RDC's inventory, a neighboring partner RDC would help fulfill these orders with its available inventory.[7] If the orders exceed the partner's inventory, then those orders are escalated and backlogged at the central warehouse (to be served via expediting when the next shipment arrives because the warehouse holds no inventory in our partner's current policy). For each service part, the warehouse orders weekly from its external supplier for that part using an independent base-stock policy for each RDC. In other words, the automaker maintains a separate base-stock level for each RDC and orders enough for each RDC to bring its inventory position up to its base-stock level. The RDC base-stock levels are set according to target service levels (ranging from 80% to 99.5%) that were set previously by the automaker for each part based on its specific characteristics (cost, customer value, etc.). We use these service levels as inputs to our simulation, which we combine with the forecasted lead-time demand to determine the automaker's incumbent (I) base-stock levels for all parts and RDCs. Then, when orders are received, in the automaker's current practice, the warehouse is generally merely a pass-through and does not hold inventory. So, once the warehouse receives supply, it first expedites shipments for customer order backlogs that were not fulfilled by the local RDCs previously, and then, the warehouse replenishes the RDCs following their independent base-stock levels. If there is insufficient supply at the warehouse after expediting to get each RDC up to their base-stock level, the program (9) is solved to allocate inventory to the RDCs (and we keep the allocation policy consistent across all base-stock policies considered). We denote the automaker's current policy for setting independent RDC base-stock levels as I (for "incumbent").

As discussed in Section 4, our industrial partner also adjusts its inventory allocation policy during disruptions. In particular, when a supplier of a particular part is disrupted, our industrial partner adopts what is called a "designate-for-intervention" policy; it prioritizes fulfilling customer orders via expedited shipping, and it holds newly arriving inventory at the central warehouse until the disruptive event terminates. In other words, the automaker currently centralizes inventory at the warehouse and does not allocate it to RDCs during a disruption, and it uses this centralized stock only to expedite customer orders. Our industrial partner designs DFI as a protection measure for its customers in case a disruption happens. Following our notation in Section 5, we denote the automaker's current policy (independent RDC base-stock policies plus DFI allocation policy during disruptions) as policy I-DFI.

6.1.6. Costs. To accurately capture the automaker's scale of costs, we include two additional costs in our simulations (full details are in Online Appendix C.3). Our partner incurs a differential cost for using the central warehouse as a pass-through (requiring one labor touchpoint for crossdocking) versus using the central warehouse as a storage point (requiring two labor touchpoints for storage and retrieval). We, therefore, incorporate a crossdocking cost for the automaker's incumbent policy and a higher processing cost for policies that store inventory at the warehouse. We also include a unit replenishment shipping cost for units shipped from the warehouse to the RDCs by truck.

6.2. Policy Summary

We test three base-stock policies: our industrial partner's incumbent policy I (which sets independent RDC base-stock levels) (see Section 6.1.5 for more details), our SP^{L+l} policy (which we denote simply as SP here for brevity) introduced in Section 3, and the NV heuristic introduced in Section 3.3. These policies are adapted from the those developed for the more stylized models of Sections 3 and 4 to accommodate the additional

problem features present in our industrial partner's setting; see Online Appendix C.3.1 for details on the policy implementation. For disruption allocation policies, in Online Appendix C.3.2, we compare the alternatives from Section 5 to determine our main policy combinations to focus on: I-DFI, I-MDFI, SP-N, and NV-N (the DFI and MDFI disruption allocation policies for SP and NV are dominated).

6.3. Simulating Policy Performance

In this section, we simulate the policies described above with our industrial partner's data in order to compare their performances (and we note that further simulation results are reported in Online Appendix C.3.2). As in the simulations of Section 5, in each scenario, all policies were simulated for 364 days after a burn-in period to allow for pipeline orders to be initiated. The initial disruption state was randomly drawn according to the stationary distribution of r_t characterized in (1). Costs for each policy were averaged across the 364 periods and then, across 100 separate simulations for each scenario. To calculate base-stock levels for the SP policy, the stochastic program is solved for each weekly ordering opportunity with a sample average approximation using 8,000 samples from the relevant demand distributions using Gurobi version 10.0.

Table 8 breaks down the average cost ratio of policies I-MDFI, SP-N, and NV-N compared with the automaker's incumbent policy I-DFI by the disruption probability parameter and the mean disruption length. Similar to Tables 1–7 in Section 5, we use asterisks to represent the number of statistically significant comparisons between a policy's average cost and other policies' average costs on the same row, and we bold the policies with the statistically significant lowest cost in each row. First, we acknowledge that although the I-MDFI policy consistently provides about a 0.2% improvement over I-DFI on average, these cost comparisons are not statistically significant. We highlight, however, that our industrial partner still finds MDFI to be an attractive policy for two reasons. First, as discussed in Section 4, the policy is easy to integrate with existing operations that already identify a switching point at the beginning of a disruption; MDFI simply filters this switching point decision through a simple calculation represented by (10). Second, as discussed in Section 5.4.5, our partner appreciates that the MDFI policy is robust to excessive backlogs during extreme disruptions by maintaining the option of centralizing inventory, even if it is rarely used. Therefore, even though more sophisticated policies may achieve lower average cost in simulations, our industrial partner finds the insights provided by the MDFI policy valuable for their operational context.

Moreover, we observe that the more advanced SP-N policy offers an even more significant improvement over the status quo, providing a 4.2% improvement over policy I-DFI on average and a statistically significant lower cost than all other policies tested. This illustrates a similar managerial insight to the simulations of Section 5; allowing the inventory policy to keep some centralized inventory at the warehouse can offer significant cost improvements in a distribution system with expediting and disruptions. Our industrial partner values this insight as it continues to improve its inventory policies and incorporate new inventory management capabilities across its network. Further, policy NV-N performs somwhere between policy SP-N and the automaker's incumbent policy, achieving more than 2% improvement over I-DFI on average, which is also statistically significant.

Next, we build on the observation from the synthetic simulations in Section 5 in Table 5 that the SP-N policy tends to perform well when demand is small because of the higher demand variance. Table 9 provides a breakdown among parts with different levels of average daily demand (the buckets were chosen to approximately equalize the number of parts in each bucket, and they are reported in the second column of the table) and focuses on the comparison between SP-N and NV-N as the two best-performing policies from Table 8 (an asterisk indicates a statistically significant lower average cost in that row for SP-N over NV-N). Table 9 reiterates a similar phenomenon; SP-N provides the most performance improvement when demand is small (corresponding to high variance relative to mean demand given our Poisson demand model). This is

Table 8. Average Cost Ratio Compared with I-DFI by Disruption Probability and Length

Disruption parameter	I-MDFI	SP-N	NV-N
Probability, α			
0.005	0.999	**0.959*****	0.980**
0.01	0.998	**0.957*****	0.977**
0.02	0.996	**0.957****	**0.970****
Length			
2	0.998	**0.954*****	0.977**
4	0.997	**0.961****	**0.974****
Overall	0.998	**0.958*****	0.976**

Table 9. Average Ratio of SP-N to NV-N by Average Daily System Demand

Average demand	No. of parts	SP-N/NV-N
0–4	7	0.977[a]
4–10	6	0.992
10–20	5	0.988
20+	5	0.991
Overall	23	0.987[a]

[a]Statistically significant lower average cost in this row for SP-N over NV-N.

useful for our industrial partner whose network mainly manages service parts with relatively low demand.

Finally, for average run times, base-stock levels for one part on one day take 3.3 seconds for NV-N and 32 seconds for SP-N. Although SP-N requires more time, our industrial partner appreciates SP-N's improved performance, especially noting SP-N's enhanced performance when the disruption probability is high. Taken together, our experiments demonstrate that modeling expediting and disruptions is critical for inventory planning in distribution systems, and our stochastic program is effective at centralizing inventory in these systems.

7. Conclusion and Discussion

In this paper, we collaborate with an industrial partner who operates a distribution system with expediting and disruptions, which are important practical features of the network but new to the academic literature. We develop a stochastic program that provides a lower bound on the cost of an optimal policy as well as guidance on how to set base-stock levels across the network. The main insight that we distill from this analysis is that holding inventory at the central warehouse can be advantageous, even in cases when the classic analysis from the distribution system literature would suggest to hold all inventory at the retailers. We close by highlighting the wider range of industries that operate distribution systems with expediting and disruptions, including electronics, retail, manufacturing, automotive, utility, military, and aerospace (Schmitt et al. 2017, Avci 2019). We, therefore, expect that our approach and insights will be applicable in a wide range of settings.

Endnotes

[1] This is because of the automaker's managerial strategy of using centralization for control over fulfillment priority; complete centralization allows full control over fulfillment priority.

[2] When a disruption occurs, there is a full L lead time's worth of supplier orders in the pipeline that still arrive as scheduled, and these (and only these) new arriving orders are centralized during the disruption. Further, retailers may have local inventory at the start of the disruption, which is exhausted before expediting from the warehouse.

[3] Although this is sufficient to prove our main result (Theorem 2), it is unclear whether the requirement $\lambda \geq a+1$ can be tightened to $\lambda \geq a+\epsilon_0$ for some $\epsilon_0 > 0$. However, a quick calculation confirms a nontrivial lower bound of $\epsilon_0 > 0.25$ as Lemma 1 fails to hold even for $a = 1$, $\lambda = 1.25$, and $n = 2$.

[4] The reason that there is not a one in each column is that the minimum cost policy can be different for each instance.

[5] NV is somewhat more complex than this, but its calculations effectively find percentiles that are at most 91%.

[6] We remove the asterisks for statistical significance to keep Table 6 from becoming cluttered and because the comparisons follow Table 1.

[7] We note that although this form of transshipment is included in our simulation, none of the policies considered directly take that fact into account when computing base-stock levels.

References

Arreola-Risa A, DeCroix GA (1998) Inventory management under random supply disruptions and partial backorders. *Naval Res. Logist.* 45(7):687–703.

Avci MG (2019) Lateral transshipment and expedited shipping in disruption recovery: A mean-CVaR approach. *Comput. Indust. Engrg.* 130(1):35–49.

Axsäter S (2003) Supply chain operations: Serial and distribution inventory systems. Graves SC, de Kok AG, eds. *Handbooks in Operations Research and Management Science*, vol. 11 (Elsevier, Amsterdam), 525–559.

Chopra S (2017) Seven-Eleven Japan Co. *Kellogg School Management Cases* 1(1):1–14.

Clark AJ, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. *Management Sci.* 6(4):475–490.

DeValve L, Myles J (2025) Approximation algorithms for dynamic inventory management on networks. *Management Sci.* 71(7): 5893–5909.

DeValve L, Pekeč S, Wei Y (2020) A primal-dual approach to analyzing ATO systems. *Management Sci.* 66(11):5389–5407.

DeValve L, Wei Y, Wu D, Yuan R (2023) Understanding the value of fulfillment flexibility in an online retailing environment. *Manufacturing Service Oper. Management* 25(2):391–408.

Doğru MK, Reiman MI, Wang Q (2010) A stochastic programming based inventory policy for assemble-to-order systems with application to the W model. *Oper. Res.* 58(4 Part 1):849–864.

Drent M, Arts J (2021) Expediting in two-echelon spare parts inventory systems. *Manufacturing Service Oper. Management* 23(6): 1431–1448.

Eppen G, Schrage L (1981) Centralized ordering policies in a multiwarehouse system with lead times and random demand. Schwarz LB, ed. *TIMS Studies in the Management Sciences, Vol. 16: Multi-Level Production/Inventory Control Systems: Theory and Practice* (North-Holland, Amsterdam), 51–67.

Federgruen A, Zipkin P (1984a) Allocation policies and cost approximations for multilocation inventory systems. *Naval Res. Logist. Quart.* 31(1):97–129.

Federgruen A, Zipkin P (1984b) Approximations of dynamic, multilocation production and inventory problems. *Management Sci.* 30(1):69–84.

Federgruen A, Zipkin P (1984c) Computational issues in an infinite-horizon, multiechelon inventory model. *Oper. Res.* 32(4):818–836.

Federgruen A, Guetta CD, Iyengar G (2018) Two-echelon distribution systems with random demands and storage constraints. *Naval Res. Logist.* 65(8):594–618.

Gallego G, Özer Ö, Zipkin P (2007) Bounds, heuristics, and approximations for distribution systems. *Oper. Res.* 55(3):503–517.

Huggins EL, Olsen TL (2010) Inventory control with generalized expediting. *Oper. Res.* 58(5):1414–1426.

Kunnumkal S, Topaloglu H (2008) A duality-based relaxation and decomposition approach for inventory distribution systems. *Naval Res. Logist.* 55(7):612–631.

Kunnumkal S, Topaloglu H (2011) Linear programming based decomposition methods for inventory distribution systems. *Eur. J. Oper. Res.* 211(2):282–297.

Lawson DG, Porteus EL (2000) Multistage inventory management with expediting. *Oper. Res.* 48(6):878–893.

Moinzadeh K, Aggarwal PK (1997) An information based multiechelon inventory system with emergency orders. *Oper. Res.* 45(5):694–701.

Muharremoglu A, Tsitsiklis JN (2008) A single-unit decomposition approach to multiechelon inventory systems. *Oper. Res.* 56(5): 1089–1103.

Reiman MI, Wang Q (2015) Asymptotically optimal inventory control for assemble-to-order systems with identical lead times. *Oper. Res.* 63(3):716–732.

Rong Y, Atan Z, Snyder LV (2017) Heuristics for base-stock levels in multi-echelon distribution networks. *Production Oper. Management* 26(9):1760–1777.

Schmitt AJ, Snyder LV, Shen Z-JM (2010) Inventory systems with stochastic demand and supply: Properties and approximations. *Eur. J. Oper. Res.* 206(2):313–328.

Schmitt TG, Kumar S, Stecke KE, Glover FW, Ehlen MA (2017) Mitigating disruptions in a multi-echelon supply chain using adaptive ordering. *Omega* 68(1):185–198.

Shen X, Yu Y, Song J-S (2022) Optimal policies for a multi-echelon inventory problem with service time target and expediting. *Manufacturing Service Oper. Management* 24(4):2310–2327.

Snyder LV, Atan Z, Peng P, Rong Y, Schmitt AJ, Sinsoysal B (2016) OR/MS models for supply chain disruptions: A review. *IIE Trans.* 48(2):89–109.

Song J-S, Zipkin PH (1996) Inventory control with information about supply conditions. *Management Sci.* 42(10):1409–1419.

Tomlin B (2006) On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Sci.* 52(5):639–657.

van Jaarsveld W, Scheller-Wolf A (2015) Optimization of industrial-scale assemble-to-order systems. *INFORMS J. Comput.* 27(3): 544–560.

Zipkin P (1984) On the imbalance of inventories in multi-echelon systems. *Math. Oper. Res.* 9(3):402–423.

Zipkin PH (2000) *Foundations of Inventory Management* (McGraw-Hill, Columbus, OH).