Mitigating Spillover in Online Retailing via Replenishment

JASON ACIMOVIC  
Smeal College of Business, The Pennsylvania State University,  
acimovic@psu.edu  

STEPHEN C. GRAVES  
Sloan School of Management, Massachusetts Institute of Technology,  
sgraves@mit.edu  

This Draft: February 17, 2015

Online purchases constitute about one tenth of US retail sales. The supply chains that support online retailing are fundamentally different from those that support traditional brick-and-mortar stores. Traditional solutions are not always appropriate to solve online retailing’s operations problems; thus, there is an opportunity to understand and improve these novel supply chains. One key characteristic of the inventory systems for online retailing is demand spillover, whereby a stockout at a fulfillment center (FC) results in demand spilling over to another FC. For this context we examine how to allocate inventory to the FCs under a periodic-review joint-replenishment policy, with an objective to minimize outbound shipping costs for a fixed amount of inventory. We first show how traditional decentralized allocation policies may perform suboptimally and induce dynamics (whiplash) that result in costly spillover. We find that this phenomenon increases with the prevalence of local stockouts and with the level of inventory imbalance. We then describe why inventory imbalance occurs in online retailing due to operational realities and provide evidence based on real data. Finally, we propose an implementable linear programming-based heuristic to replenish and allocate inventory accounting for possible spillover during the lead time. We test the heuristic by a simulation and show that it performs better than the status quo and is robust to operational realities.

Key words: Online retailing, inventory replenishment, centralized inventory control

1 Introduction

As it becomes easier to shop for items through the internet and mobile phone apps, online retailers will control a greater piece of the retail pie. In 2017, online retail sales are predicted to grow to $370 billion and constitute 10% of US retail sales (Forrester Research, Inc. 2013). Inventory management policies developed for other contexts such as brick-and-mortar retailing may not be applicable to an online retailing environment. Novel complexities exist in online retailing supply chains. Developing good inventory management and order fulfillment policies that account for these complexities is critical for online retailers (Agatz et al. 2008).

We describe the distinctive attributes of online retailing supply chains, and review relevant literature. We then examine the status quo replenishment policy at a large online retailer. When this status
The primary metrics we use to compare these policies are total outbound shipping cost and spillover cost. There is a demand spillover when the closest FC cannot serve the demand due to a local stockout, and the demand is served from a more distant FC, at a higher outbound shipping cost; the spillover cost is the additional shipping cost.

We show that our heuristic dampens the costly dynamics and can reduce spillover costs by between 5% and 8%, averaged over the scenarios we test (savings for individual scenarios may be significantly greater or smaller than this average). This translates to savings in total outbound shipping costs of between 0.4% and 1.2% averaged across these scenarios, depending on the magnitude of the operational realities the organization faces. We compare our heuristic to a partially clairvoyant replenishment policy; our heuristic achieves about a quarter of the potential improvement that this policy realizes. Outbound shipping costs can be significant for an online retailer. For instance, the online retailer Overstock.com spends about 5% of its net revenue on fulfillment (which includes outbound shipping), while Amazon.com spends about 8% of its net sales on outbound shipping costs (Amazon.com, Inc. 2012, 2013; Overstock.com, Inc. 2012, 2013). Thus, a 1% reduction in these delivery costs can have a significant impact on financial performance.

The work presented in this paper is some of the first to examine replenishment policies in online retailing. As such, we try to highlight how the online context differs from traditional retailing and what new challenges arise. In particular, a good portion of the paper is descriptive. We start with showing how the current replenishment policy at a large online retailer can go wrong, and how this can lead to both spillover and the phenomenon of whiplash. We then describe how inventory in the system has a tendency...
to become imbalanced, which leads to more frequent local stockouts and spillover. With this grounding we finally propose a simple replenishment heuristic and show how it can perform better than the status quo policy in this context. By describing the current system and some of the challenges, we hope to spur interest that will lead to further research, addressing the new opportunities.

2 Background on online retailing

Fundamental differences exist between the structure of brick-and-mortar and online retailing supply chains that necessitate rethinking how to manage online retail networks efficiently. Strategies that govern how inventory enters and exits the network must take these differences into account. Key characteristics that distinguish online retailing from brick-and-mortar retailing include:

- **No local or system-wide backorders** – As long as inventory exists somewhere in the system, the online retailer will fulfill a demand request. If a customer’s order cannot be fulfilled from his nearest FC due to a stockout, his order spills over to a more distant FC that does have the item in stock, and is fulfilled at a higher outbound shipping cost.

- **Delivery window heterogeneity** – Customers differ in their willingness to pay for fast delivery, creating multiple demand classes based on the requested delivery time. Furthermore, the shipping fees are actual-cost agnostic and depend only on the time window, not the actual cost to fulfill the order. Thus, the online retailer’s challenge is to satisfy the requested delivery time as cheaply as possible.

- **Significant nonlinear outbound shipping costs** – Outbound shipping costs can be as high as 8% of net sales. In addition, discontinuities and nonlinearities exist with respect to distance and time. This is due in part to the zone systems of many third party shippers and the fact that different transportation modes must be used on different facility-customer arcs.

The existence of these distinctive attributes in online retailing requires managers to make decisions that either do not exist in or differ from brick-and-mortar retailing. Three salient decisions that must be made are:

- **Which FCs should carry stock** - Online retailers must first decide how many and which FCs will hold inventory for a particular SKU. Not every building needs to carry inventory because of the flexibility to ship a customer’s order from any location. In general, slower (faster) moving SKUs are held in fewer (more) FCs.

- **From which FC to ship** – When a customer places a request, there may be several ways to fulfill it. It is not always optimal to ship a customer’s order from the nearest FC due to the heterogeneity in the delivery times and nature of the outbound shipping costs. For instance, see (Acimovic and Graves 2015; Mahar et al. 2009) for order fulfillment strategies in online retailing.

- **How much inventory to hold in each FC and how to plan replenishments** – An online retailer manages a network of FCs with a centralized inventory management system. As such, it will typically rely on a joint replenishment policy for each SKU, whereby it periodically places a system order from a vendor that then gets allocated to the FCs.
The focus of this paper is on the latter set of decisions, namely how to replenish the inventories into multiple FCs in light of the three distinguishing characteristics. While we concentrate primarily on pure online retailers in this paper, many of the results may be applicable to omnichannel retailers as well. For instance, an omnichannel retailer might have the flexibility to fulfill an online customer’s order from a set of FCs or from a brick-and-mortar store. Inventory policies that replenish inventory into both the FCs as well as the brick-and-mortar locations may benefit by taking into account this fulfillment flexibility and possible spillover effects.

3 Literature review

We divide the relevant literature into four streams: inventory policies in online retailing, inventory management when stockouts result in lost sales, transshipment of inventory among facilities at the same echelon, and replenishment under disruptions and sub-optimal conditions.

The first stream of literature relates to inventory policies in online retailing. (Acimovic and Graves 2015) examine the problem of deciding how to fulfill a customer’s order from a set of FCs assuming replenishment policies and inventory positions as given. This current paper builds on that work to develop better forward-looking replenishment policies. Agatz et al. (2008) provide a review related to inventory policies in online retailing. Much of the existing literature on the supply side (as opposed to the demand allocation and fulfillment side) focuses on drop shipping (Bailey and Rabinovich 2005; Khouja 2001; Netessine and Rudi 2006), inventory rationing (Ayanso et al. 2006; Cattani and Souza 2002), and returns handling (Mostard et al. 2005; Vlachos and Dekker 2003). We were not able to find any literature, however, that specifically prescribed replenishment policies tailored to an online retailing environment, as we aim to do in this paper.

A second stream of literature related to this paper is the lost sales literature. Looking at our online retail network as a whole, we assume that system-wide demand in excess of on-hand inventory is lost. Additionally, each FC, in effect, follows a lost sales model, but with a complicated demand distribution; its demand is a function of the inventory levels and demand distributions of the rest of the network and the spillover that might occur. Additionally, the demand distribution is a function of the fulfillment policy being utilized. If one could easily characterize the demand distribution function (which does not seem likely for general cases), then looking at one FC in isolation is analogous to a lost sales model: inventory is depleted until the FC has no more on-hand inventory. Future demand that would have been directed towards that FC will be either directed to another FC or lost. Inventory policies for lost sales models are, in general, intractable. Huh et al. (2009) and Zipkin (2008) provide reviews. A few exceptions include Reiman (2004)
who proves that for a sufficiently long lead time, constant order policies can outperform order-up-to policies in a continuous review lost sales environment and Goldberg et al. (2012) who prove that constant order policies are asymptotically optimal with respect to lead time in a periodic review context. Whatever hardness results apply to lost sales systems also apply to our spillover environment. While much of the existing literature focuses on developing sophisticated inventory replenishment policies that account for rather simple demand patterns, our work in this paper aims to better estimate the complicated demand pattern that each FC faces (complicated due to spillover patterns), and replenish inventory at the FCs accounting for those patterns.

A third stream of literature related to spillover involves transshipments (sometimes called emergency or lateral transshipments) in multi-location inventory problems. This class of problems assumes that several retail-type nodes exist. Each retail location serves a specific group of customers with its own random demand distribution. If one retailer stocks out, inventory can be reactively transshipped, at a cost, from a retailer with on-hand inventory, either at the time the demand is requested or at the end of the review period. The objective is to choose a replenishment and transshipment policy that minimizes the sum of holding, backorder (or lost sales), and transshipment costs. See Paterson et al. (2011) for a recent review. The literature in this stream that does concentrate on developing good periodic review replenishment policies (as opposed to focusing only on how to transship given a replenishment policy) is often focused on developing good stationary replenishment policies that account for transshipment within a period. For example, several papers focus on calculating static base-stock levels in a periodic review environment (Herer et al. 2006; Karmarkar 1981; Krishnan and Rao 1965; Robinson 1990). A common assumption in the transshipment literature is that the replenishment lead time is zero. Exceptions include Tagaras and Cohen (1992) (who develop a static base-stock policy) and Diks and de Kok (1996) (who develop static proportions of system inventory to allocate to each building). Unlike much of the prevailing transshipment literature, the replenishment policy we propose considers nonzero replenishment lead times and is dynamic; it takes into account the current inventory levels in the system and the possible spillover patterns that are likely to occur.

Inventory policies under disruptions and suboptimal realistic conditions have also been investigated. Some of this literature focuses on supplier disruptions risk and supply chain design (Craighead et al. 2007; Tang 2006; Tomlin 2006) while other portions focus on inaccuracies at the shelf or distribution center level (DeHoratius and Raman 2008; Kull et al. 2013). Specifically for the latter stream, DeHoratius et al. (2008) and Kök and Shang (2007) develop and analyze replenishment policies when inventory records are inaccurate. Our work is different from this latter stream in that we assume that
inventory records are accurate, but similar in the sense that operational realities (in their case inventory inaccuracy, in our case supply perturbations, inventory shifts, and demand re-routing) can have a significant impact on the organization, and replenishment policies should take these operational realities into account.

4 A local base-stock replenishment policy
In this section, we discuss the local base-stock policy utilized at our industrial partner. We then show on a small example how this policy may lead to costly dynamics if inventory becomes imbalanced. Finally, we find evidence of these dynamics by examining data from our industrial partner.

4.1 Description of the replenishment policy at our industrial partner
For many product lines, our industrial partner follows a periodic-review replenishment policy. At each review epoch, a system-wide order is placed, which is comprised of individual orders for each of the FCs that stock the SKU; each order arrives at the FC after a lead time. Specifically, our industrial partner employs a base-stock, or order-up-to, policy, both locally and system-wide. We describe here the policy under the assumption that the lead time to each FC is the same. The system-wide base-stock level is set equal to the expected demand over the lead time and review period plus safety stock; that is,

\[ B_{SYS} = (L + r)d_{SYS} + SS_{SYS} \]  (1)

where \( B_{SYS} \) is the system-wide base-stock level, \( SS_{SYS} \) is the system-wide safety stock, \( d_{SYS} \) is the expected daily demand, \( r \) is the review period, and \( L \) is the lead time to replenish each FC.

The system-wide safety stock is modeled as:

\[ SS_{SYS} = \Phi^{-1}(\alpha_{SYS})\sigma_{SYS} \sqrt{L + r} \]  (2)

where we define \( \sigma_{SYS} \) as the system standard deviation in daily demand, \( \alpha_{SYS} \) as the system-wide in-stock service level target (Type I service level, or cycle service level), and \( \Phi^{-1}(\cdot) \) as the inverse normal cumulative distribution function. For each FC that stocks the SKU, its local base stock level is

\[ B_i = \lambda_i B_{SYS} \]  (3)

where we associate a positive load factor \( \lambda_i \) to each FC \( i \), where \( \sum \lambda_i = 1 \). Nominally, the load factor \( \lambda_i \) represents the fraction of demand to be served by that FC; it is common to set it equal to the proportion of all customers for whom FC \( i \) is the nearest facility among the set of FCs stocking the SKU.

At each review epoch, the online retailer places a system-wide order to bring the inventory position (on-hand plus pipeline inventory) up to the base-stock level:
\[ y_{\text{SYS}} = B_{\text{SYS}} - x_{\text{SYS}} \]  

(4)

where \( x_{\text{SYS}} \) denotes the system inventory position and \( y_{\text{SYS}} \) is the system order quantity. The order amount for each individual FC will be:

\[ y_i = B_i - x_i = \lambda_i \cdot B_{\text{SYS}} - x_i \]  

(5)

Most likely, the quantity computed in (5) will be fractional and require some form of rounding in a way that ensures the sum of the individual FC orders equals the desired system-wide order amount. In practice, it may also be adjusted to account for case quantities, minimum order quantities, and total system-wide order limits, factors which we do not consider here.

Operating with an order-up-to level for each building in isolation may not be optimal, as this does not account for the interaction between FCs in serving customers. Nevertheless, a local base-stock policy is simple to deploy and relatively intuitive. As specified above, it is quite flexible: if customer dynamics change or warehouse topology changes, only the \( \lambda_i \)'s need to be adjusted. Furthermore, a local base-stock policy can be optimal in a decentralized multi-echelon system under some conditions (see for example Karmarkar (1981)). Additionally, in omnichannel environments, many systems are still decentralized due to the tendency for online channels and brick-and-mortar channels to operate under separate departments (Forrester Research, Inc. 2014).

### 4.2 Whiplash

When an FC stocks out, demand that would have been served by that FC spills over to another FC. This spillover can result in a chain reaction affecting all FCs: if one FC stocks out, extra demand is routed to a second FC, which now has a higher probability of stocking out, and so on. In addition to creating lateral interactions among FCs, spillover can also create temporal interactions when a local base-stock replenishment policy is used. We observe a spillover-induced phenomenon we call whiplash: if an FC serves a greater (smaller) proportion of demand in a review cycle than its target \( \lambda_i \), then in the next period, it is more likely to serve a smaller (greater) proportion of demand than its target. We first demonstrate this phenomenon on a stylized example to gain intuition; we will then report on empirical evidence of whiplash.

Imagine an online retailer with deterministic constant demand, and with two FCs that serve demand from two corresponding customer regions. The cost of each FC serving its own region is low, but there is a larger cost associated with one FC shipping a package to a region other than its own.
Each facility orders inventory according to a local base-stock policy. The load factors \( \lambda_i \) are set equal to the proportion of demand realized within each of the two regions. Let \( d = d_{SYS} \) denote the exact daily system demand as there is no variability. Additionally, there is no safety stock in the system because a 100% service level can be achieved without it. We assume that the lead time \( L \) is less than or equal to the review period \( r \). (Similar results hold when \( L > r \), but the equations become more tedious). We can express the base-stock levels and order amounts for each FC \( i \) as:

\[
B_i = d\lambda_i (r + L) \tag{6}
\]

\[
y_{it} = B_i - x_{it} = d\lambda_i (r + L) - x_{it} \tag{7}
\]

where we have added a time index for the inventory and order replenishment variables.

To get some insight into the system dynamics, we examine a numerical example with the review period \( r = 7 \) days, the lead time \( L = 3 \) days, the system demand \( d = 10 \) units per day, and the load factors \( \lambda_1 = 0.4, \lambda_2 = 0.6 \). Thus, the base stock levels are given from (6) by \( B_1 = 40, B_2 = 60 \).

In this example, days 1, 8, 15, and 22 are review days, with inventory arriving on days 4, 11, 18, and 25 (due to \( L = 3 \)). If the starting inventory positions on the initial review day 1 equals \( L \cdot d \cdot \lambda_i \) for \( i=1,2 \), no spillover will occur. This corresponds to FC 1 starting with 12 units, and FC 2 starting with 18 units. Figure 1 shows the resulting inventory levels through time of the two FCs. Both on-hand and on-order inventories are displayed with the latter stacked on the former; the sum equals the inventory position.
In the preceding example, each FC started with exactly the right amount of inventory to serve its own regional demand through the lead time: the inventory was balanced. However, it is possible that the system may have begun imbalanced. The thirty units of starting inventory might not be allocated optimally as 12 in FC 1 and 18 in FC 2 for a variety of reasons: one of the FCs might be temporarily at physical storage capacity and cannot accept replenishments, one of the FCs might have experienced a system outage and the other FC temporarily served demand from both regions, the vendor might have mixed up the FCs and delivered the wrong amount to each FC, or other operational realities. For any other starting inventory positions other than 12 in FC 1 and 18 in FC 2, spillover will occur. Suppose that FC 1 starts with 20 units, and FC 2 starts with 10 units. Figure 2 shows the resulting inventory levels in this example. On day 1, each facility orders up to its own base-stock level (40 and 60 respectively). Then, in the middle of the second day, FC 2 runs out of inventory. Because demand spills over rather than being backlogged, FC 1 fills this unsatisfied demand, draining its inventory faster than when it was serving only one region. When the inventory arrives on day 4, the inventory level in FC 1 is lower than what the local base-stock policy had accounted for (because of the lack of backlogging at the local level), leading to a low on-hand inventory position in FC 1 on day 4. Conversely, the inventory position on day 4 in FC 2 is high. This then leads to FC 1 running out of inventory in the second week, with the demand in its region being served by FC 2. This spillover oscillation occurs ad infinitum.
For this example, we find that this two-period oscillation can be avoided only if the starting inventory at the initial review period is $Ld\lambda_i$ for $i=1,2$. Otherwise, under the assumptions that there is no safety stock and each starting inventory is not greater than its base stock, then there will be oscillation and there will be spillover indefinitely. The magnitude of the spillover depends on how far away the starting inventory levels are from the ideal inventory levels.

More formally, let $t$ denote a review epoch with on-hand inventory $X_t$ at FC $i$. The assumption of no safety stock implies that there is sufficient inventory to cover demand over the lead time, but no more; thus, we assume that

$$x_{it} + x_{2it} = dL$$

Furthermore, suppose that each inventory is no greater than its base stock level, i.e., $x_{it} \leq B_i$. Then we show in Appendix A that the inventory follows a two-period cycle, with

$$x_{i,t+2kr} = x_{i,kr} \text{ for } k = 1,2,...$$

$$x_{i,t+(2k+1)r} = B_i - x_{it} - d\lambda_i(r - L)$$

$$= 2d\lambda_iL - x_{it} \text{ for } k = 0,1,2,...$$
We will have no oscillation iff $x_t = x_{t,(2k+1)}$; from (8) and (9) we see immediately that this occurs iff $x_t = d \lambda_t L$. Otherwise the system will oscillate with a two-period cycle. We also show in the appendix that the amount of spillover in each period, in terms of the number of units shipped from one FC to the other region, can be bounded by $\min(\lambda_1, \lambda_2) \cdot dL$.

From this example we observe that in the simplest system we can have spillover whenever there is any imbalance in the inventories. Furthermore, we find an oscillatory pattern in which the inventory at an FC at a review epoch cycles between two points, one high and the other low. This also produces an oscillation in the size of the replenishment orders: when the inventory is high (low), the size of the order is small (large). We term this behavior as a *whiplash*, in which the orders from an FC will cycle between too large and too small.

Obviously, for this particular setup, the best thing to do is to start with balanced inventory positions. However, what this example suggests is that if a retailer uses a local base-stock policy with little safety stock, then the system can be very sensitive to events that might imbalance the inventories across the system. Once the inventories become imbalanced, there will be increased amounts of spillover, resulting in increased outbound shipping costs, as well as a whiplash in the replenishment orders. Furthermore, in the absence of safety stock, the system is not self-correcting, as the oscillations continue forever without dampening. When safety stock is present, it acts as a dampener: the oscillations persist through time, but reduce in absolute value in each period. The general findings remain the same in systems with more than two FCs, with stochastic demand, with heterogeneous lead times, and with lead times longer than the review period (Acimovic 2012).

### 4.3 Empirical evidence of whiplash

We examine data for 2,604 SKUs from our industrial partner to see if there is any evidence of whiplash. The analysis is performed on SKUs from a single product line over a five month period, from the end of October 2011 until the beginning of April 2012. The SKUs in this product line are stocked in anywhere from three to a dozen FCs. The online retailer was able to order each SKU in single units as the purchases were not restricted to cases nor constrained by a minimum order quantity; thus, in the data set there are no “lumpy” order effects. At each review epoch, the replenishment system determined for each FC whether an order should be placed, and if so, for how many units. This was done according to the local base-stock policy outlined in section 4.1 above.
For each SKU and FC, we examine every pair of consecutive review periods and compare how the proportion of the total system order assigned to each FC changes from one period to the next in each pair. For a specific SKU, let $\rho_{it}$ represent the proportion of the total order placed at review period $t$ that was assigned to FC $i$. Recall that $\lambda_i$ is the load factor associated with FC $i$ for the specific SKU, i.e., the targeted portion of total demand served by FC $i$. In our analysis, we use the load factors utilized by our industrial partner in their local base-stock policy calculations. We then define the deviation $\Delta_{it} = \rho_{it} - \lambda_i$. This represents how far an FC’s actual order deviated from its target in time period $t$.

For instance, imagine that an SKU is stocked at two locations. The following table shows hypothetical sample data and the resulting $\Delta_{it}$’s and $\rho_{it}$’s.

<table>
<thead>
<tr>
<th>FC</th>
<th>Load factor $\lambda$</th>
<th>Replenishment amount to each FC</th>
<th>Proportion of total order $\rho_i$</th>
<th>Deviation $\Delta_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>July 1 (t-1)</td>
<td>July 15 (t)</td>
<td>$\rho_{i,t-1}$</td>
<td>$\rho_{it}$</td>
</tr>
<tr>
<td>Nashville</td>
<td>0.7</td>
<td>15</td>
<td>40</td>
<td>0.25</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>0.3</td>
<td>45</td>
<td>10</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Nominally we expect the deviations to center around zero, as the replenishment orders should reflect the demand fulfilled in the previous review period and the demand should be proportional to the load factors. As a null hypothesis one might expect the deviations to not be correlated over time, as they reflect the relative proportion of demand from the FC’s region; indeed, if there were correlation, one might surmise that it would be a positive correlation reflecting that the regional demand for an SKU was consistently higher or lower than the assumed load factor. An alternate hypothesis is that there is negative correlation, which is consistent with the phenomena of spillover and whiplash, as illustrated with the stylized example from the prior section.

From the 2,604 SKUs we generated a dataset of 46,040 observations, where each observation corresponds to a pair of successive replenishments of a specific SKU at a specific FC. The correlation between $\Delta_{i,t-1}$ and $\Delta_{it}$ for these 46,040 observations is -0.14. If an FC orders more (less) than its load factor in one period, it orders less (more) than its load factor in the subsequent period. We also perform a regression in order to check the robustness of this observed negative autocorrelation. The result is negative and significant (coefficient = -0.157, t-value = -26.2) when several fixed effects are included and errors are clustered at the SKU level. The details of this regression as well as other robustness checks can be found in Appendix B.
We infer that this negative autocorrelation is evidence of whiplash and demand spillover from the application of a local base-stock policy. We posit that this demand spillover results in higher outbound shipping costs. If an FC orders more units than its load factor, this suggests that this FC served more than its share of customers in the previous review period, namely it served customers outside of its preferred region. By definition, the system did so at a higher outbound shipping cost as compared to being able to serve those customers from a nearer FC. Similarly, when an FC orders less than its load factor, this suggests that another, more distant FC has served customers in its region, with a higher shipping cost.

5 Drivers of local stockouts and inventory imbalance
In the previous section (4), we showed on a stylized model how inventory imbalance and local stockouts result in spillover (compare Figure 1 and Figure 2). In this section, we examine what might drive system-wide inventory imbalance as well as frequent stockouts at individual FCs. If these phenomena are indeed common in actual online retailing systems, then replenishment policies that mitigate the associated negative effects (namely, high outbound shipping costs due to spillover) can provide value.

We define a local stockout as an individual FC running out of inventory and a system-wide stockout as every FC running out of inventory. When a local stockout occurs at an FC, a customer who would have normally had his order served from that facility must now have his order shipped from a more distant FC at a higher outbound shipping cost. Two characteristics of online retailing lead to local stockouts. The first is that the safety stock in the system is shared across the FCs; as a consequence the stockout rate at each FC is much greater than that for the system. We describe this effect in section 5.1 below. The second is the prevalence of operational realities that lead to network inventory becoming imbalanced. This imbalance, in turn, may result in local stockouts (see the example in Figure 2). We examine this phenomenon in section 5.2 below, by positing potential causes as well as observing imbalance on empirical data. To understand the extent of these operational realities and their effect on the system inventory balance, we build a simulation model of our industrial partner’s network, the details of which we describe in this section (5.2). We revisit this simulation model in section 7, and utilize it to compare the status quo replenishment policy with a heuristic replenishment policy we describe in section 6.

5.1 Pooled safety stock
In online retailing, inventory is centrally managed and controlled. As a result, there is pooling across the facilities and the system operates with a single safety stock that is spread over the FCs. While the system might have a low probability of stocking out, each individual FC might have a moderate probability of stocking out. As an illustration, assume that an online retailer replenishes according to a periodic review
base-stock policy, and operates \( n \) FCs with \( n \) corresponding regions, all identical and independent. Demand is assumed to be normal and backorders are allowed. The FC base-stock levels are set according to equation (3). The system-wide safety stock level is set according to (2), and distributed equally among the FCs such that each FC holds \( 1/n \) of the system safety stock. The probability that an FC stocks out during the replenishment cycle is

\[
\Pr\left(Z > \Phi^{-1}(\alpha) \sqrt{n}\right),
\]

where \( Z \) is the standard normal random variable. That is, the probability that an individual FC stocks out increases with \( n \): local stockouts may be frequent even if system-wide service levels are very high.

### 5.2 Operational realities

Oftentimes, the environment in which supply chains operate in practice is not as clean as it is in theory or even in simulations that attempt to capture realities of actual systems. Specifically, we are interested in the phenomena whereby operational realities in actual supply chains cause inventory to be out of balance: some facilities are overstocked, and others are understocked. These imbalances lead to additional stockouts, which lead to spillover and extra outbound shipping costs. These operational realities occur for both good and bad reasons. Some operational realities such as supply perturbations are due to errors on the part of the vendor. However, others such as demand re-routing can be the result of exercising flexibility in certain processes. For instance, the system might route extra customer orders from one FC, which is temporarily understaffed, to another FC with slack capacity. Exercising this flexibility allows the full utilization of the system-wide work force while at the same time meeting service commitments to customers. This flexibility is often not possible in traditional brick and mortar systems. However, this flexibility may come at a price by creating inventory imbalance.

In this paper, we do not attempt to analyze specific incidents nor do we determine what operational realities specifically occur at our industrial partner. Instead, we first describe the common causes, gleaned from conversations with our industrial partner. Second, we measure how balanced the inventory is among a set of actual SKUs. To do this, we utilize a balance metric that measures the geographical mismatch of supply and demand. Third, we compare how balanced the inventory was in the actual environment versus in a simulated environment. We observe that inventory is more out-of-balance in the actual system, suggesting that operational realities do occur at our industrial partner. Fourth, we generate synthetic inventory shifts in the simulated environment. These simulated inventory shifts are intended to replicate the level of imbalance observed in the actual system which experiences operational realities.
5.2.1 Common types of operational realities
Supply \textit{perturbations} occur when a supplier does not deliver what was ordered: examples include vendors shipping incorrect quantities to the network, vendors shipping the product to the wrong building or the wrong product to the right building, and very late shipments. \textit{Demand re-routing} occurs when a customer’s demand request is not fulfilled from her nearest facility. Common causes for demand re-routing include \textit{demand spillover} due to a stockout at the nearest FC, as well as:

- \textbf{Disruption at the FC} – An FC (or part of it) may go off line for a number of reasons, including planned down time for installing new equipment, a network outage, or adverse weather.
- \textbf{Labor capacity} – FCs can handle only so many customer orders in a day due to limitations on the labor force in the building. Surplus customer orders must be re-routed to a less burdened FC.
- \textbf{Multi-item orders} – An FC might be out of stock of – or does not carry – one or more of the items in a customer’s order and it is cheaper to send a single package from further away rather than break the order up.
- \textbf{Forward-looking fulfillment} – The nearest facility has the items in stock but the order is assigned elsewhere so as to protect the inventory at this facility, in light of possible future orders (see (Acimovic and Graves 2015)).
- \textbf{Physical capacity} – A building can fill up such that no additional inventory can be stored inside of it. Once this occurs, inventory that the online retailer would like to place in that FC must be placed elsewhere in the network, resulting in an inventory imbalance.

Re-routing a single demand from one building can affect the inventory position of every other building in the network, resulting in more frequent local stockouts and more demand spillover. It is unlikely that in a real system one can exactly characterize the impacts from supplier perturbations and demand re-routing or prevent these phenomena. Thus, good replenishment policies should be robust to these effects.

5.2.2 Measuring balance of inventory on actual SKUs
We measure the impact of operational realities by examining a sample of 2604 SKUs from our industrial partner. For each SKU and for each day on which a replenishment order was placed over a 5 month period, we have data for the on-hand inventory level and the inventory position just before the replenishment order was placed. We compute a balance metric that equals the normalized objective value to a transportation problem, representing the outbound shipping costs that match the actual inventory to expected demand by region and delivery-time request.

Specifically, the balance metric is the ratio of the objective values for two transportation linear programs (TLPs). The supply nodes for each TLP are the FCs. The demand nodes in both TLPs are the set of all geographical region and a customer’s time-window pairs (for instance, \{New York, 2-day\} or
{Oregon, 8-day} are potential demand nodes). The demand at each node in the TLP is the forecast of demand for that region-time pair. The cost from each supply node to each demand node represents the price of the cheapest transportation mode that will deliver the package in the requested time window. The total demand in the TLP is scaled such that it equals the sum of the total inventory. The TLPs differ based on the supply available at each supply node. For the first TLP (the numerator in the ratio) the supply at each FC equals the actual on-hand inventory. For the second TLP (the denominator in the balance metric’s ratio), the supply at each FC is the “ideal supply;” that is, we distribute the available on-hand inventory across the FCs so as to minimize the outbound shipping cost for the forecasted demand. The ideal value of the metric is one, in which case the actual inventory is in the right place and the right amount, and is termed to be balanced; a value greater than one is a measure of the inventory imbalance.

This balance metric was proposed and described in (Acimovic 2012) and in (Acimovic and Graves 2015). In those works, the authors showed that in an online retail environment, SKUs with a relatively higher (lower) value of the balance metric also incurred higher (lower) outbound shipping costs subsequently. Intuitively, if on-hand inventory for an SKU is in perfect balance, then the inventory is in the right geographical, relative to the expected demand, and the metric will correspond to the minimal outbound shipping cost for that stocking configuration (choice of FCs). If inventory is not in perfect balance then the measure can be interpreted as the percentage increase in the outbound shipping costs relative to the ideal for the stocking configuration, assuming no variability in demand.

For each SKU, for each date an order was placed, we calculate this balance metric. We then take an average of the balance metrics for each SKU across dates, weighted by the inventory levels on each date. For each SKU, we define its empirical average balance metric (E-ABM):

\[
E-ABM_s = \frac{\sum_t \theta_{st} \cdot I_{st}}{\sum_t I_{st}} = \frac{\sum_t \theta_{st}}{\sum_t I_{st}}
\]

where \( s \) denotes the SKU, \( t \) denotes the specific review epoch, \( \theta_{st} \) is the balance metric for SKU \( s \) on review epoch \( t \), \( I_{st} \) is the on-hand system inventory level of SKU \( s \) on review epoch \( t \). This provides a distribution of balance metric values with 2604 observations, one for each SKU.

5.2.3 Measuring balance of inventory on idealized SKUs

We just described a method whereby we measure how imbalanced SKUs in a real system actually are. But is this more or less imbalanced than we might expect if operational realities played no role? In order to answer this question, we create a simulation to mimic the inventory and fulfillment policies of our industrial
partner in all aspects except that no operational realities exist: a customer order is always fulfilled from the nearest FC with inventory and replenishments are always delivered to the correct FC. We postulate that if the actual system, as measured in section 5.2.2, is more imbalanced than the simulated system, then this additional imbalance is due to operational realities.

We simulate our industrial partner’s network for 100 review periods for each SKU and take a weighted average of the balance metric across the periods. In this simulation, the only source of stochasticity is the demand: how much and from where. For a given SKU, we use the empirical data to get the lead times to each FC, the review period, the mean demand, and the forecast error, which we use as a proxy for demand variance. We also record which FCs held inventory and the FC load factors set by the central planner. These data are utilized in an SKU-specific simulation that attempts to replicate the network of our industrial partner. In this paper, we utilize this simulation for two purposes: first, to estimate the level of inventory imbalance in the system by comparing the actual and idealized balance metrics (described in this section) and second, to compare the effectiveness of different replenishment policies in a realistic environment (discussed below in section 7). This simulation has the following characteristics:

- **FCs** – For each SKU, we utilize the actual FCs that held the item.

- **Demand nodes** - A demand node in our network consists of a cluster of Zip3’s (the geographical areas determined by the first three digits of a full zip code) and a requested shipping speed. For instance, ('205', 'Slow') would be the demand node for Washington, DC, for customers who are willing to wait a long time for their packages. We divide the United States into 100 geographical regions, and offer four delivery speeds to the customers, creating 400 demand nodes in total. The four delivery speeds are next day, second day, four day, and eight day. We group together geographical regions using k-means clustering, according to their costs to the FCs. We utilize the method described in (Acimovic 2012; Acimovic and Graves 2015)

- **Demand realization** - For the simulation tests, we assume a system daily demand mean ($\mu_{SYS}$) and standard deviation ($\sigma_{SYS}$) as input parameters which are derived from the empirical data for each SKU. We assume that each demand node $j$ generates on average a proportion of the system demand equal to $\eta_j$ (estimated from empirical data) where $\sum \eta_j = 1$. We model daily demand at each demand node as a negative binomial random variable with mean $\eta_j \mu_{SYS}$ and standard deviation $\sqrt{\eta_j \sigma_{SYS}}$. For each day, the simulation generates demand at each of the demand nodes. The simulation then randomizes the arrival order of these demands, and the demands arrive one by one to the system. We chose a negative binomial distribution for two reasons. First, the demand rates for some items are very low, necessitating a discrete distribution with non-negative support. Second, for some SKUs, the variance is greater than the mean; thus, we ruled out the Poisson distribution. The negative binomial distribution allows us to adjust both the variance and the mean of specific SKUs.
• **Shipping mode** - There are four possible shipping modes that the online retailer may utilize: air next day, air second day, premium ground, and United States Postal Service. Each ship mode has a different delivery time that may be based on distance. For instance, air second day can be utilized to ship anywhere in two days, whereas premium ground can deliver within one day for nearby locations and may take as long as a week to ship across the United States. These shipping time estimates are estimated from shipping company websites and data from our industrial partner.

• **Shipping costs** - We fit a linear function to each shipping mode based on distance, whose parameters (a fixed cost and a variable cost) are approximated from actual data from our industrial partner. From this cost data, the shipping time estimates by mode, and the fact that the retailer will use the cheapest feasible ship mode that will deliver a customer’s order on time, we can create a cost matrix from each FC to each demand node (location/delivery speed combination). Let $c_{ij}$ be the cost to serve demand $j$ from FC $i$.

• **Myopic fulfillment** - The simulation fulfills each demand from the nearest FC that has positive on-hand inventory. If no FC has positive on-hand inventory, the demand is not satisfied, and there is a lost sale. This is the policy utilized by our industrial partner at the time of this analysis.

• **Replenishment** – Here, we employ a local base-stock policy (that of our industrial partner). Review periods, FC-specific lead times, and load factors ($\lambda_i$’s) are derived from the empirical data. We assume a Type I service level for the system, and calculate local base-stock levels and safety stock as in equations (3) and (2) respectively.

Many of the above parameters are based on previous work we have performed with this same retailer. More detailed data on determining feasible ship modes and shipping costs can be found in (Acimovic 2012; Acimovic and Graves 2015).

We calculate a weighted average of the balance metrics for each SKU under an ideal simulated environment across the review epochs. We call this value for a specific SKU $S$-ABM, for *simulated average balance metric*, and it is calculated the same as $E$-ABM (see equation (10)).

### 5.2.4 Comparing actual balance to idealized balance

For each SKU, we then have the empirical average balance metric ($E$-ABM) and the simulated average balance metric ($S$-ABM). Demand stochasticity and demand spillover will lead to some imbalance, and would be observable in both systems. If operational realities (beyond spillover) are significant, then we might expect that $E$-ABM is bigger than $S$-ABM. We find that this is true on average: $E$-ABM is 9.9% higher proportionally than $S$-ABM. Figure 3 shows the distribution of the balance metric values for the empirical data and the simulated data, while Figure 4 shows the distribution of the differences (i.e., $\Delta_s = E$-ABM$_s - S$-ABM$_s$) in the balance metrics for the 2604 SKUs.
Figure 3: Distribution of balance metric values across 2604 SKUs for the actual observed on-hand inventory levels (above) and the simulated inventory levels (below). The scale has been removed for confidentiality reasons and 2% of the outliers have been removed.

Figure 4: Distribution of differences in balance metric across 2604 SKUs. The scale has been removed, although the tickmarks are on the same scale as in Figure 3.

This inventory imbalance experienced in real systems (and its causes) is difficult to model and account for explicitly. We believe the main driver behind the observed imbalance is the presence of operational realities that are not accounted for in typical forecasts. These operational realities may include demand re-routing, supply perturbations, or even unforeseen demand shocks. As a robustness check, we consider possible alternate explanations for the observed imbalance. We performed sensitivity on the input parameters to the simulation, correlation of the demand, and non-stationarity of the demand. Accounting
for these in the simulation did not explain the level of imbalance we observed in the empirical system (see Appendix C for more details). We infer from this that “stuff happens” leading to the inventory in the network being more imbalanced than we would otherwise expect. Some FCs have more inventory than we would expect, while others have less; and this increased imbalance should result in more frequent stockouts at each FC, leading to increased spillover and higher outbound shipping costs.

5.2.5 Simulating operational realities
In sections 6 and 7 below, we develop and test inventory policies that are robust to these operational realities. We do this by simulating the system of our industrial partner (described above in section 5.2.3) and applying both the status quo replenishment policy as well as an improved replenishment policy we describe in section 6. In order to determine how well each policy would work in a realistic environment subject to operational realities, we create synthetic inventory shifts that are meant to imbalance the inventories by approximately mimicking what might happen in real systems.

We do not have access to data regarding actual operational events that might have imbalanced the inventory. Therefore, we augment our simulation with a new class of random events that are intended to have a similar effect on the system as actual operational events. These events should adhere to three criteria: they should be simple to simulate, they should resemble the mechanics of a potential operational event, and they should be able to be tuned so that we can create a variety of magnitudes of “imbalance” to the system, including the magnitude of 9.9% described in section 5.2.4. We model the inventory shifts as follows. At the end of each day (after all demands are realized), we consider (in random order) each ordered pair of FCs \((i, j)\) from the set of FCs that stock the SKU. For each pair, we move all of the on-hand inventory at FC \(i\) to FC \(j\) with probability \(p\), where \(p\) is a user-chosen parameter.

We want to evaluate how well a local base-stock policy and an alternate policy perform in a realistic environment in the presence of these inventory shifts. Thus, we want to set \(p\) to a value that creates inventory shifts whose magnitudes are realistic. In sections 5.2.2 and 5.2.3, we examined a set of SKUs and were able to determine the inventory imbalance in the actual system (E-ABM) and in a cleaner simulated system without operational realities and inventory shifts (S-ABM). The balance metrics were higher (more imbalanced) in the actual system by 9.9%. In order to choose \(p\), we utilize that same set of SKUs as in sections 5.2.2 and 5.2.3 with the same simulation, except we simulate synthetic inventory shifts. When \(p\) is set to zero, the balance metric for the simulated system is at its lowest. As we increase \(p\), the simulated system becomes more imbalanced. When \(p=0.01\), then the simulated system with inventory shifts is about as out of balance as the actual system: that is, \(S-ABM(p=0.01) \approx E-ABM\).
There are many ways to simulate inventory shifts, and there is no guarantee that the method we chose exactly mirrors that of the actual system. However, this method adheres to the three criteria mentioned above. First, it is simple and requires the determination of only a single parameter. Second, we thought it mirrored what might happen in a real system. If an FC cannot ship out items due to labor constraints or an operational problem, then demand for that FC may be re-routed to another FC. When demand is routed from the first FC to the second FC, it is virtually as if inventory at this second FC is moved to the first FC (even though no physical inventory is actually moved between the FCs). Third, we were able to create imbalances on the order of what we observed (i.e., 9.9% more imbalanced with inventory shifts than a clean system without inventory shifts). Determining exactly how and why inventory shifts occur in a real system is an avenue for future research, with room for both empirical and analytical work.

6 Description of a heuristic replenishment policy
In this section we attempt to improve upon the current replenishment policy. For this development, a system-wide order amount is determined through a system-wide base-stock policy, as described in section 4.1. We then need to decide how to allocate the system-wide order to the FC’s to minimize the outbound shipping costs. One could formulate the optimal replenishment policy as a dynamic program (Acimovic 2012). However, this formulation is of limited practical value as solving it optimally is hard. It is at least as difficult as a single warehouse lost sales replenishment problem; even this simpler problem is difficult enough that plausible heuristics have been tested only on a limited range of systems (Zipkin 2008). Calculating optimal policies is even more difficult. Thus, we suggest a heuristic for the multiple FC replenishment problem, under the assumptions of equal lead times and holding costs for the FCs and equal lost sale costs for all customer classes.

6.1 Overview
A summary of the heuristic is this: at a review epoch \( t \), we estimate what the projected on-hand inventory levels will be in each of the FCs on the day the inventory will arrive (\( \chi_{it} \)), accounting for spillover. We also calculate for each FC the target inventory levels (\( \beta_i \)) that we would like to have on-hand at the FC after the receipt of the replenishment order. We then set the order to be the difference

\[
\gamma_{it} = \beta_i - \chi_{it}
\]

for each FC. We call this the projected base-stock policy (PB).

To implement the heuristic we need to specify how we will determine the projected on-hand inventory level (\( \chi_{it} \)) just prior to the replenishment (\( L \) days from a review epoch), and the inventory target
after the replenishment ($\beta_i$). The former is more difficult, whereas the latter is relatively straightforward to approximate. We will start with the latter.

We set the inventory target for FC $i$ as: 

$$\beta_i = r \lambda d_{sys} + SS_i$$

(12)

where $SS_i$ is the safety stock associated with FC $i$. (We define how this is calculated in equation (16) in section 7.2 below). Thus, we set the inventory target after a replenishment as the expected demand over the next review period, plus the safety stock allocation for the FC. We do not claim that this is optimal, but rather that it is a reasonable policy.

In the next section we describe how we determine $\chi_{it}$, the estimate of the on-hand inventory level at FC $i$ just prior to the replenishment, where the estimate is made at the review period $t$. Given this estimate, then the amount we order at review period $t$ is given by equation (11).

### 6.2 Estimating on-hand inventory

The determination of $\chi_{it}$, the estimated on-hand inventory positions $L$ days from $t$, is complicated as it depends on the effects of demand spillover. We suggest here one approximation based on solving a time-indexed transportation LP with a myopic fulfillment policy and fluid flows. The supply nodes are the FCs on a particular day with supply equal to their on-hand inventory plus on-order inventory scheduled to be received that day. The demand nodes are customer regions by day with the demand set to its daily expectation. We set the arc costs to assure that each demand node is served by its nearest facility if stock is available, then by its next nearest facility, and so on. We estimate $\chi_{it}$ as the inventory remaining in the FC at the end of the lead time, just prior to the arrival of a replenishment order. This, along with $\beta_i$ and equation (11), provides a suggested order amount for each FC. See Appendix D for more details.

### 6.3 Using sample paths to better estimate on-hand inventory

We also consider a more sophisticated version of the heuristic that takes into account the demand distribution; we call this the projected base-stock plus policy. We generate $M$ demand sample paths, where each sample path is comprised of a set of demand realizations for each region for each day over the next $L$ days. We then solve the time-indexed transportation LP (described in section 6.2) for this demand realization. For each sample path $m$, we can thus calculate the estimate of remaining on-hand inventory $\chi_{it}^m$. We average these values over all $M$ sample paths to obtain an estimate of remaining on-hand
inventory that takes into account demand stochasticity: $\chi_{it}^+$. We use this estimate to determine the order amount from (11).

### 6.4 Example of heuristic operating on a simple deterministic system

We show how this heuristic works on the stylized model from section 4.2 with two FCs and deterministic demand. No safety stock is held in the system. Assuming that FC 1 started with 20 units and FC 2 with 10, a local base-stock policy led to indefinite spillover (see Figure 2).

Employing the heuristic, on the other hand, will incur some spillover in the first period, but none thereafter. We can estimate the on-hand inventory levels in $L$ days, $\chi_{1t}$ and $\chi_{2t}$; each FC will have zero units on-hand just before the order arrives because the system will have zero units on-hand. Therefore, we calculate $\beta'$s, $\chi'$s and $y'$s as such:

\begin{align*}
\chi_{1t} &= \chi_{2t} = 0 \\
\beta_1 &= rd \lambda \beta + SS_1 = 28 \\
\beta_2 &= 42 \\
y_{1t}^{PB} &= \beta_1 - \chi_{1t} = 28 - 0 = 28 \\
y_{2t}^{PB} &= 42
\end{align*}

Note that in this specific example, the order amounts are independent of the state of the system. We see that once the inventory is brought into balance, no spillover will occur thereafter.
7 Simulation of replenishment policies

7.1 Simulation set-up

We now compare the performance of our heuristics to that of the local base-stock policy on examples involving stochastic demand with realistic system parameters, including positive safety stock, shipping costs related to time and distance, and inventory shifts.

We described the main aspects of the simulation above in section 5; here we specify how we adapted the simulation for these tests. In particular, we prescribe parameter values instead of deriving them from the empirical data. These include demand mean and variance, number of FCs, lead times, review periods, and service levels. Adjustments to the simulation are:

- **Which FCs hold inventory** – In the simulation tests, we vary the number of FCs at which we hold inventory (3, 6, and 12). If we stock an item in either 3 or 6 FCs, we must choose which of the 12 facilities should carry inventory. We solve a facility location problem to determine this.

- **Equal system order amounts** - The replenishment policies we consider might prescribe different system-wide order amounts for a given inventory position. However, in order to compare the effectiveness of each policy in distributing a system-wide order among the FCs, we require that each replenishment policy orders the same system-wide amount at each review day. Namely, the system-wide order for each replenishment policy must equal what the system-wide local base-stock policy would have ordered. We then scale up or down the individual replenishment orders for each FC according to the prescribed orders from the specific replenishment policy.
We utilize the following parameter values in our simulation, assuming a review period of one week ($r=7$):

- $n = \{3, 6, 12\}$
- $\mu_{SYS} = \{1, 4, 9\}$
- $\sigma_{SYS} / \mu_{SYS} = \{1.0, 1.5, 2.0, 2.5\}$
- $L = \{3, 10, 17, 24, 31\}$
- Type I Service Level $= \{0.90, 0.95\}$
- $p = \{0, 0.0015, 0.0025, 0.005, 0.01, 0.015\}$

We test every combination of the above parameters with the following exception: we exclude scenarios that would result in inventory being spread very thinly among the FCs, i.e., we require the weekly mean demand to be at least twice the number of buildings. Thus, we exclude all combinations in which $\mu_{SYS} = 1$ (weekly mean demand = 7) and the number of FCs is 6 or 12. The choice of $p$ sets the level of disruption in the simulation, with $p = 0.01$ reflecting our approximation for the actual operational realities. In all there are 1680 scenarios. In each scenario we simulate each replenishment policy for 250 review periods. We also tested lead times of $\{6, 13, 20, 27, 34\}$, which led to similar results. As a note, we utilize the above parameters as opposed to utilizing the parameters inferred from the industrial data described in sections 4.3 and 5 because the empirical data represents a specific product line whose characteristics do not represent the full range of the values that we actually want to test.

### 7.2 Comparing replenishment policies
We simulate five replenishment policies. For comparison purposes, each policy orders the same system amount each review period as dictated by equation (4) so that all simulations operate with exactly the same system inventory.

1. **Local base-stock policy (LB):** Each FC orders up to its own base-stock level as described in section 4 above. This is the status quo policy employed by our industrial partner.

2. **Constant order policy (CON):** This policy allocates the same proportion of inventory to each FC each period regardless of the actual inventory levels. The order amount for FC $i$ is $\lambda_{SYS} y_{SYS}$, rounded appropriately. The system order amount $y_{SYS}$ is determined by equation (4). This is the most naïve policy, as it utilizes only the system inventory position. We test this policy based on the observations of Reimann (2004) and Goldberg et al. (2012) that system-unaware policies work well under certain conditions.

3. **Projected base-stock policy (PB):** This is the heuristic defined in sections 6.1 and 6.2. We aggregate demand nodes into 25 geographical clusters when solving the LP for computational efficiency. This combined with the fact that there are four customer time-window requests results in 100 demand nodes for the LP.
4. **Projected base-stock policy plus (PB+):** This is the sample-path version of the heuristic defined in section 6.3 that approximately accounts for demand stochasticity. For the parameter $M$ (the number of sample paths utilized), we tested the performance of the heuristic against different values of $M$. We found good values of $M$ to be 100 when average daily demand was 1, and 50 when daily demand was 4 or 9.

5. **Instant re-allocation (IR):** This policy re-allocates the inventory among the FCs on the day it arrives into the system. The arriving inventory is allocated among the FCs to equalize their probabilities of stocking out before the next time inventory is scheduled to arrive (assuming no spillover and normal demand). This is an unrealistic clairvoyant policy, as it is equivalent to being able to see the demand perfectly between the day inventory is ordered and the day it arrives. However, the policy is not perfectly clairvoyant, as it is still unaware of the exact demand between the day inventory arrives for this replenishment and the day inventory arrives for the next replenishment. We believe it is a reasonable approximation for an upper bound on the best policy.

We calculate the safety stock levels similarly for the LB, PB, and PB+ policies. For the LB policy, we want to set the base-stock levels to minimize the sum of the spillover across all the FCs. But this problem seems very difficult. So as a proxy we assume that stockouts result in lost sales and we set the base-stock level to minimize the sum of expected lost sales over the FCs. For general differentiable demand distributions, setting the probabilities of stocking out to be equal across the FCs will minimize the lost sales. Thus, assuming normal demand, the safety stock for the LB policy is set as such:

$$SS_i = \Phi^{-1}(\alpha_{SYS}) \cdot \sigma_{SYS} \sqrt{L + r} \cdot \left( \frac{\sqrt{\lambda_i}}{\sum \sqrt{\lambda_i'}} \right)$$

(16)

Note that this is different from the policy utilized by the online retailing partner (compare equation (16) to equations (2) and (3)). The safety stock level for the PB and PB+ policies is set as in equation (16), except that $L$ is set to zero. The $\beta'$s for the PB and PB+ policies represent the target inventory levels for the FCs on the day inventory arrives; the next inventory arrival will occur one review period later. By maintaining similar, reasonable safety stock levels across the LB, PB, and PB+ policies, we are able to observe the quality of each policy itself.

### 7.3 Simulation results

Among most of the scenarios tested above, we find that the PB+ replenishment policy performs the best, although in many instances, the PB policy performs almost as well as the PB+ policy. The local base-stock policy performs worse relative to the other policies as the inventory shifts increase in intensity from Level 0 ($p=0$) to Level 5 ($p=0.015$), with Level 4 ($p=0.01$) being comparable to the actual system. Table 2 shows...
We note that the overall cost reduction increases with increased inventory shifts, whereas the spillover costs reduction does not vary as much. This is partly due to the fact that as inventory shifts are introduced, more spillover costs will occur. The heuristic saves more money in outbound shipping costs as inventory shifts are introduced, but the underlying costs are also increasing, especially the spillover costs. We can observe this overall cost increase in the first row of Table 2, which suggests that inventory imbalance and operational realities may be increasing outbound shipping costs as a whole on the order of 8%. The PB and PB+ policies achieve about 21% and 27% respectively of the improvement realized by the partially clairvoyant IR policy.
We also report how overall cost reduction (relative to a local base-stock policy) responds to the parameter settings, namely demand variance and lead time. Figure 6 shows the summary plots of “Overall costs reduction” under Level 0 (no disruption) and Level 4 (comparable to actual).

We make some general observations from this figure. The improvement is highest when the variability is the lowest. This makes sense if we assume that all policies will perform worse as variability increases, and that the possible improvement reduces along with the actual performance.

As lead time increases, so does the relative improvement of the heuristics and the constant proportion policy over the local base-stock policy. In fact when the lead time is very short (3 days), there is not much improvement from the alternate replenishment policies. However, as lead time increases, the alternate policies perform well because the PB and PB+ policies both account for not only the inventory position of each FC, but also the details of when each portion of the inventory which is on-order will arrive and what spillover and demand might occur in the meantime.
Not shown in the figure, we also observe that higher service levels lead to lower improvement gap between the heuristic and the local base-stock policy, although not dramatically. As the system service level increases, so does the safety stock held at each individual FC. Thus, each FC has a lower probability of stocking out, and less spillover occurs. We also observed that varying the number of buildings did not have a dramatic effect on the improvement gap. As the number of buildings increases, spillover occurs more frequently, but the cost of spillover is lower because buildings are nearer to each other on average. Demand rate did not have a practically significant effect on the performance of the heuristics.

The value of the heuristics increases as the level of inventory shifts increase, which we see in the differences between left and right columns of Figure 6 and also in Table 2. This suggests that the heuristic replenishment policies are more robust to operational realities than the status quo local base-stock policy.

The constant proportion policy performs well under some specific scenarios. In general, it leads to lower outbound shipping costs when: variance in demand is low, inventory shifts are most severe, lead times are long, and service levels are low. The observation that a naïve constant proportion policy performs better as lead times lengthen echoes the fact – proved by Goldberg et al. (2012) – that constant order policies are asymptotically optimal with respect to lead time in a lost sales environment. That is, as the lead time increases, no intelligent policy can beat the randomness associated with what happens to demand over the lead time. Thus, in systems that are expected to be severely chaotic for whatever reason, it may be best to ignore the system information all together and implement a naïve policy.

We compare here the extent to which whiplash exists among the four policies tested in these simulations. We show the results in Table 3 for the case when Level 4 inventory shifts (comparable to the actual system) are introduced. For each of the 280 simulated scenarios, we calculate the autocorrelation of the order deviation for each FC in the simulation. That is, for each FC we measure the correlation between the deviation in the current period and the deviation in the previous period, as defined in section 4.3. Over the 280 scenarios, we obtain 1800 observations.

<table>
<thead>
<tr>
<th>Policy</th>
<th>Breakdown of autocorrelation across 1800 observations</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Fraction &lt; 0</td>
<td>Fraction &lt; -0.1</td>
<td>Fraction &gt; 0.1</td>
</tr>
<tr>
<td>LB</td>
<td>0.09</td>
<td>0.76</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>CON</td>
<td>0.02</td>
<td>0.50</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>PB</td>
<td>0.03</td>
<td>0.41</td>
<td>0.09</td>
<td>0.23</td>
</tr>
<tr>
<td>PB+</td>
<td>0.00</td>
<td>0.53</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>IR</td>
<td>0.11</td>
<td>0.12</td>
<td>0.01</td>
<td>0.52</td>
</tr>
</tbody>
</table>

For this sample of observations, the LB policy has a negative autocorrelation for most of the cases and is more often negative than the other policies, suggesting that whiplash exists for this policy. This is consistent
with our finding in section 4.3 that whiplash exists on the empirical data from our industrial partner. For these simulation cases, we do not find evidence of whiplash from the heuristic PB and PB+ policies; rather the autocorrelation is slightly positive on average, with roughly an equal number of observations being positive as negative. The IR policy has positive autocorrelation. This seems to occur because it is not always possible to equalize stockout probabilities at each FC when inventory is instantly reallocated. Whenever this happens there will be at least one FC with “too much inventory,” perhaps because this FC’s inventory was not depleted over the previous review period. This FC will get a zero allocation and will have a lower stockout probability than the other FCs in the next review period. Consequently, at the next instant reallocation time, this FC is likely to get a smaller allocation again because it could still have “too much inventory.” Correspondingly the FCs that started with larger allocations will have a higher likelihood of getting a larger allocation at the next allocation epoch.

To summarize, both versions of the heuristic perform better than the local base-stock policy under most scenarios. The effect is especially pronounced when lead times are long, demand variability is low, and inventory shifts and operational realities are severe. These heuristics also lessen the impact of whiplash.

8 Conclusion and next steps
We investigated replenishment policies that better allocate inventory in an online retailing environment, guided in part through a collaboration with an industrial partner. After describing the distinctive aspects of these supply chains, we showed that the status quo decentralized replenishment policy is especially susceptible to inventory imbalance and local stockouts. We then showed why stockouts – and therefore spillover and excessive outbound shipping costs – might be common in online retailing. Specifically, our data analysis suggests that supply perturbations and demand rerouting may be regular occurrences and can have a significant impact on the balance of inventory. We proposed a heuristic, which is relatively simple and easy to implement, and showed that it results in lower outbound shipping costs and less whiplash. Additionally, a naïve policy that is system-unaware performs well when lead times are long and inventory shifts are frequent.

This research can be extended in several ways. In specifying the replenishment policy, we assumed a myopic fulfillment policy; one might determine how to recast the replenishment policy, allowing for a more sophisticated fulfillment policy, such as that described in (Acimovic and Graves 2015). Additionally, one would like to relax the assumption of deterministic and known lead times to allow for uncertain and heterogeneous lead times. One might also be able to calculate tighter bounds on the performance of the best possible policy: does our policy’s poor performance under certain scenarios suggest that the heuristic
can be improved upon or that every policy performs poorly. Also, our online retailing partner holds safety stock to guard against system-wide stockouts. An avenue for future research could be to determine the best amount (and allocation) of additional stock at each FC in order to guard against system-wide spillover: the tradeoff is lower outbound shipping cost versus higher holding cost. Finally, while characterizing operational realities and inventory shifts exactly probably remains impossible, there might be ways in which they can be more explicitly incorporated into replenishment heuristics.

One important takeaway from this research is that operational realities exist in real world supply chains, and can unbalance the inventory in a multi-location system. Some of these operational realities are due to variability and operational errors, like suppliers shipping to the wrong location. But in an online setting, other operational realities reflect other supply chain goals, like balancing the work across multiple, capacity-constrained facilities. In either case, however, organizations should understand the extent to which these operational realities affect the inventory balance in their supply chains, and then develop operational policies that mitigate and are robust to these imbalances.

References


Appendices

A Dynamic equations of two-FC deterministic system with spillover

Recall the following parameter and variable definitions:

- $i$ – Denotes system (SYS), or a fulfillment center (1,2)
- $B_i$ – Base-stock level
- $d$ – Daily system demand (deterministic and constant)
- $r$ – Review period
- $L \leq r$ – Lead time
- $x_{it}$ – On-hand inventory level in $i$ on day $t$
- $y_{it}$ – Order amount on review day $t$
- $\lambda_i$ – Load factor for fulfillment center $i$, also equal to proportional demand in region $i$ ($\lambda_1 + \lambda_2 = 1$)

as well as equations (6) and (7):

$$B_i = d \lambda_i (r + L)$$  \hspace{1cm} (6)
$$y_{it} = B_i - x_{it} = d \lambda_i (r + L) - x_{it}$$  \hspace{1cm} (7)

The dynamics of the system for FC 1 can then be defined as follows:

$$x_{1,t+L} = x_{1,t} - \left[ \min \left( x_{1,t}, d \lambda_1 L \right) + \left( d \lambda_2 L \min \left( x_{2,t}, d \lambda_2 L \right) \right) \right]$$  \hspace{1cm} (17)
$$= \left( x_{1,t} - d \lambda_1 L \right)^+ - \left( d \lambda_2 L - x_{2,t} \right)^+$$ \hspace{1cm} (18)
$$= 0$$ \hspace{1cm} (19)

where $(a)^+$ equals $\max(0,a)$ and $x_{it}^-$ denotes the on-hand quantity just before time $t$ (i.e., just before a replenishment arrives). Note the dynamics for FC 2 are the mirror image of those for FC 1.

In equation (17), the inventory level in FC 1 $L$ days after the review day (and just before the replenishment order arrives) is the inventory level on the review day minus the demand served by the FC over the lead time. The demand served consists of the demand realized in that region over $L$ days, up to the initial inventory, plus the demand realized in the other region that could not be fulfilled by its own FC (the other region’s spillover). Equation (18) is obtained through rearranging terms and algebra. For the last step in (19) we use the fact that there is no safety stock in the system, implying that

$$x_{1t} + x_{2t} = dL.$$ \hspace{1cm} (20)

From (7) and (19) we can now express the inventory after the receipt of the order replenishment as:

$$x_{1,t+L} = x_{1,t+L}^+ + y_{it} = d \lambda_i (r + L) - x_{1t}$$ \hspace{1cm} (21)
In equation (21), the order that was placed \( L \) days ago is added to the remaining inventory at the moment just before \( t+L \).

We write the on-hand inventory level in FC 1 one period into the future as:

\[
x_{1,t+r} = \left[ \left( x_{1,t+L} - d\lambda_1 (r - L) \right)^* - \left( d\lambda_2 (r - L) - x_{2,t+L} \right)^* \right]^* \tag{22}
\]

Equation (22) describes the inventory level on the next review day in terms of the inventory level when the last replenishment arrived, taking into account spillover during this time frame.

Having written the dynamics of the system in equations (17) through (22), we now show how whiplash occurs in such an environment. We write the inventory level in FC 1 on the next review day into the future as a function of the inventory level on this review day.

\[
x_{1,t+r} = \left[ \left( x_{1,t+L} - d\lambda_1 (r - L) \right)^* - \left( d\lambda_2 (r - L) - x_{2,t+L} \right)^* \right]^* \tag{23}
\]

From (23), we find one upper limit of the inventory in an FC on a review day:

\[
x_{1,t+r} = \left( 2d\lambda_1 L - x_{it} \right)^* - \left( x_{2it} - 2d\lambda_2 L \right)^* \tag{24}
\]

The above equation can be broken into three cases:

**Case 1:**

\[
(2d\lambda_1 L - x_{it}) \geq 0 \quad \text{and} \quad (2d\lambda_1 L - dL - x_{it}) \geq 0
\]

\[
\iff 0 \leq x_{it} \leq 2d\lambda_1 L - dL \tag{25}
\]

\[
x_{1,t+r} = dL
\]

**Case 2:**

\[
(2d\lambda_1 L - x_{it}) \geq 0 \quad \text{and} \quad (2d\lambda_1 L - dL - x_{it}) \leq 0
\]

\[
\iff 2d\lambda_1 L - dL \leq x_{it} \leq 2d\lambda_1 L \tag{26}
\]

\[
x_{1,t+r} = 2d\lambda_1 L - x_{it}
\]

**Case 3:**

\[
(2d\lambda_1 L - x_{it}) \leq 0 \quad \text{and} \quad (2d\lambda_1 L - dL - x_{it}) \leq 0
\]

\[
\iff 2d\lambda_1 L \leq x_{it} \tag{27}
\]

\[
x_{1,t+r} = 0
\]

These three cases outline how the inventory one period in the future depends on the inventory level on a review day in this period. We can now also place limits on the maximum and minimum inventory levels in an FC. From equation (23), we find one upper limit of the inventory in an FC on a review day:
\[ x_{i,t+r} = (2d\lambda_i L - x_{i,t})^+ - (x_{2,t} - 2d\lambda_2 L)^+ \]
\[ \leq (2d\lambda_i L - x_{i,t})^+ \]
\[ \leq (2d\lambda_i L)^+ \]
\[ = 2d\lambda_i L \quad \text{(28)} \]

We also know that because the system inventory is equal to \( dL \) on a review day, and because there are no backorders:

\[ 0 \leq x_{i,t} \leq dL \quad \text{(29)} \]

Lastly, we can calculate an additional lower bound on the inventory:

\[ x_{i,t+r} + x_{2,t+r} = dL(\lambda_i + \lambda_2) \]
\[ x_{i,t+r} + 2d\lambda_2 L \geq dL(\lambda_i + \lambda_2) \quad \text{(Because } x_{2,t+r} \leq 2d\lambda_2 L) \quad \text{(30)} \]
\[ x_{i,t+r} \geq dL(\lambda_i - \lambda_2) \]
\[ x_{i,t+r} \geq dL(2\lambda_i - 1) \quad \text{(Substituting } (1 - \lambda_i) \text{ for } \lambda_2) \]

Putting all the limits together from equations (20), (28), (29), and (30), and noting that limits on \( x_{i,t+r} \) also apply to \( x_{i,t} \) and \( x_{2,t} \):

\[ x_{i,t} \leq \min(2d\lambda_i L, \ dL) \quad \text{(31)} \]
\[ x_{i,t} \geq \max(0, \ dL(2\lambda_i - 1)) \]

To conclude, once the system is in steady state, the inventory level in a given FC will obey the limits in equations (31). The dependence of inventory for one review day on the previous review day is given by “Case 2” above:

\[ x_{i,t+r} = 2d\lambda_i L - x_{i,t} \quad \text{(32)} \]

It is also easy to see that the system adheres to a two period oscillation of spillover, that is, the system state in two inventory periods from now is the same as the state right now:

\[ x_{i,t+2r} = 2d\lambda_i L - x_{i,t+r} \]
\[ = 2d\lambda_i L - (2d\lambda_i L - x_i) \]
\[ = x_{i,t} \]
\[ = x_{i,t+(2k+1)r} \quad \text{for } k = 0, 1, 2, \ldots \]

From (32) and (33) it immediately follows that:

\[ x_{i,t+(2k+1)r} = x_{i,t} \quad \text{for } k = 1, 2, \ldots \]
\[ x_{i,t+(2k+1)r} = 2d\lambda_i L - x_{i,t} \quad \text{for } k = 0, 1, 2, \ldots \]

which correspond to equations (8) and (9) in section 4.
Let \( S_t \) define the amount of spillover in a review period served from FC 2 to region 1 (i.e., the unserved demand from FC 1 to region 1). Because the system always has enough inventory, it can be defined as such:

\[
S_t = (d\lambda_t L - x_t)\]

Plugging equation (34) into the bounds given by equations (31), we see that:

\[
0 \leq S \leq \min(\lambda_t, 1 - \lambda_t)dL
\]

From equations (32) and (34), we see that no spillover will occur if and only if \( x_t = d\lambda_t L \), that is, if there is exactly enough inventory in the FC to cover the demand over the lead time for that facility’s region.

**B Regression results for empirical whiplash analysis**

We perform a regression with \( \Delta \) as the dependent variable and \( \Delta_{t-1} \) as the independent variable of interest. Our intent is not to perform a detailed econometric analysis for the purpose of exactly defining the magnitude of the autocorrelation. Instead, we wish to perform due diligence to show that a sizable whiplash effect exists in practice, and to rule out other possible factors that might explain the negative autocorrelation. In order to control for other factors that might be affecting the deviation, we incorporate several sets of fixed effects and other possible explanatory variables into our model.

- FC fixed effects – Controlling for the FC \( i \) being replenished in each observation (dummy variable for each FC)
- Week fixed effects – Controlling for the week in which the order was made (dummy variable for each week)
- Product fixed effects – Controlling for each SKU (dummy variables for each SKU)
- Length of time between orders – How long (in days) elapsed between \( t-1 \) and \( t \) (continuous variable representing number of days)
- Number of items in first order – What was the total order size for a SKU at \( t-1 \) (continuous variable)
- Number of items in second order – What was the total order size for a SKU at \( t \) (continuous variable)

Table 4 shows the results of the linear model with and without controlling effects.
Table 4: The effect of a previous order’s deviation on the current order’s

<table>
<thead>
<tr>
<th>Coefficient for deviation in first order cycle $\Delta_{i,t-1}$</th>
<th>(1) No controlling effects</th>
<th>(2) All controlling effects (Date, SKU, Time between orders, Number of items in first and second order)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Coefficient}$</td>
<td>$-0.136$</td>
<td>$-0.157$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.019</td>
<td>0.057</td>
</tr>
<tr>
<td>$\text{Adj-R}^2$</td>
<td>0.025</td>
<td>-0.001</td>
</tr>
<tr>
<td>$\text{Observations}$</td>
<td>46040</td>
<td>46040</td>
</tr>
</tbody>
</table>

Observations exceed the number of SKU’s because for each SKU, there are multiple weeks and FC’s. Values in parentheses are standard errors with error clustering at the SKU level.

The dependent variable in both models is $\Delta_{it}$.

While linear regression might not be technically appropriate because $\Delta_{it}$ must lie between -1 and 1, we believe that a linear regression is a good approximation because 99% of the values of the dependent variable lie between -0.4 and 0.7.

We also consider computing the deviation relative to the average actual load factor, rather than the prescribed load factor $\lambda_i$. That is:

$$\Delta_{it} \equiv \rho_{it} - \frac{\sum_{t \in T_k} \rho_{it}}{|T_k|}$$

where $T_k$ is the set of order replenishment epoch times for SKU $k$. Under this new definition of $\Delta_{it}$, we also include only those SKU-FC pairs for which we observed at least 6 replenishment orders (i.e., for which $|T_k| \geq 6$). This results in 1132 SKUs and 29733 SKU-FC-DATE triplets. The correlation between $\Delta_{it}$ and $\Delta_{i,t-1}$ is -0.23. One reason the negative autocorrelation effect is stronger for this new definition of deviation is that we observe in the data SKU-FC effects. That is, some SKUs are not actually replenished to an FC in a quantity suggested by the current inventory levels and the load factors. This could be due to several operational factors, for instance, specific FCs are at physical capacity and a specific subset of SKUs are being diverted elsewhere.

Finally, we look at the probability of $\Delta_{it}$ being positive both unconditional and conditional on $\Delta_{i,t-1}$ being negative. The following table shows these results, based on 16395 observations:

<table>
<thead>
<tr>
<th>Fraction of time $\Delta_{it}$ is positive (negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over all 46040 observations</td>
</tr>
<tr>
<td>Conditional on $\Delta_{i,t-1}$ being negative (positive)</td>
</tr>
</tbody>
</table>

Note that some observations are equal to zero, and thus neither positive nor negative.

From this data, we infer that there is an inverse relationship between the deviation in one period and the deviation in the following period. That is, if an FC ordered more (less) than its load factor in one
review period, then it is more likely to order less (more) than its load factor the next period. This effect is robust to controlling for various effects, changing the definition of deviation, and performing a non-parametric analysis. This negative autocorrelation is consistent with the occurrence of whiplash, as predicted from the reliance on a local base-stock policy with demand spillover.

C Sensitivity analysis on the simulation

We examine four alternate hypotheses that might explain why the empirical inventory positions are more imbalanced than the simulated ones.

- The parameters of the simulation and balance metric are inaccurate. Some of the parameters we utilize are estimates based on aggregated data from our industrial partner (such as ship mode feasibility and costs). For other aspects of the simulation, we made specific decisions as modelers (such as utilizing a negative binomial distribution for demand). We cannot prove that our simulation is completely representative; we are not able to compare our results with actual system results. However, the parameters for this simulation were also utilized in a previous analysis that we performed for this same industrial partner in which we proposed operational changes. The improvement that the industrial partner experienced as a result of our analysis was about the same as we predicted (Acimovic and Graves 2015). To some extent, this validates the aspects of the simulation such as shipping costs and feasibility of shipping modes for different customers. The other aspects of the simulation were validated by the online retailer through extensive conversations, as well as analysis of the data. We also performed some robustness checks by varying the SKU input parameters to the simulation. We did the following checks: vary the mean demand (and change standard deviation to keep coefficient of variation the same) by factors of one half and two; vary the demand standard deviation (keeping the mean constant) by factors of one half and two; and vary the lead time by factors of one half and two on a set of the SKUs. We observe that the perceived imbalance still exists even when these parameters vary significantly. The S-ABM for each of these scenarios did not change from the baseline value by more than 0.9% relatively. (That is, if S-ABM’ represents the simulated average balance metric under a change of parameters and S-ABM represents the metric when we utilize the parameters as from the data, then the maximum of (S-ABM’- S-ABM)/S-ABM equals 0.009. Considering that (E-ABM - S-ABM)/S-ABM equals 0.099, we infer from this that our simulation is robust to input parameter inaccuracy, within reasonable ranges of the input parameters.

- The actual demand is correlated, which is amplifying the imbalance. For the simulation involving the idealized SKUs, we tested different types of positive correlation which might result in imbalances. We created three types of correlation using copulas.

  o First, we created positive correlation across demand regions for each day. We first drew a random sample from a standard multivariate normal distribution for each day. There are as many dimensions n as demand regions, and we set the correlation across dimensions to be 0.5. Then, for each dimension of each random draw, we apply the standard normal cumulative inverse distribution function. From one sample of an n-dimensional multivariate normal distribution, this creates n random variables uniformly distributed between zero and one that are correlated with each other. Note that the magnitude of the correlation is not preserved (it will not be 0.5 for the uniform random variables), but the direction and presence of correlation will be preserved.

Appendices: Page 6
Finally, on each of the $n$ uniformly distributed realizations, we apply the transform of the negative binomial distribution with the appropriate mean and variance for each region. In this way we create $n$ random draws from $n$ negative binomial distributions (one for each region) that are correlated with each other.

- We also test the impact of correlation throughout time in two different ways. The first time-correlation approach treats each demand region as independent from the other. For each of the $n$ demand regions, we then sample from a multivariate normal distribution of dimension $m$, where $m$ is the number of days to be simulated. The correlation matrix utilized for this approach has entries $(0.5^{(i,j)})$ where $i$ and $j$ are row and column indices of the matrix.

- The second time correlation approach samples correlated system demand over time. Then, we utilized a multinomial random variable to assign each day’s demand to the set of demand regions, with probabilities equal to the mean demand proportion. In this way, demand is positively correlated over both time and across regions.

Our observations from these three approaches suggest that the imbalance we observe is not caused by positively correlated demand. When demand was correlated across regions, the system was actually less imbalanced than when the demand was independent. Correlation through time did not have a practically significant effect on the balance.

- The actual demand is decreasing, which is amplifying the imbalance. When demand decreases, one or more FC might end up with excessive inventory. We tested the case of decreasing demand: at the start of the time horizon, the demand is twice the average demand rate for a specific SKU. Throughout the time horizon, we decrease the demand linearly (keeping the coefficient of variation constant), until at the end of the time horizon the demand is zero. Over the time horizons we tested (from 3 to 300 review period), the system was not more imbalanced than the baseline scenario by more than 1.1% relatively. Considering that the empirical data is 9.9% more imbalanced than the baseline scenario, demand that trends downward does not explain the observed imbalance in the actual data.

- Multi-item orders significantly affect the balance of inventory. If certain FCs are more likely to handle multi-item orders, then this might lead to imbalances. In the simulation, we restricted customer orders to be shipped only from a subset of the FCs, representing the case when a customer order may be shipped only from the set of FCs that house the other items in the order as well. We performed this experiment when the subset is selected randomly, and when the subset of FCs is more likely to be the same subset (representing the case where bigger FCs might have a higher likelihood of housing the other items in a random multi-item order). Less than a half of a percent of the imbalance was due to multi-item orders in our experiment.

We cannot explain away the observed imbalance through these sensitivity analyses. Thus, we infer that the observed imbalance on the empirical data is due to the presence of operational realities.

D Linear program used to estimate remaining on-hand inventory

Here we define the time-indexed LP which the heuristic employs to estimate on-hand inventory levels. Let $\tilde{c}_{ij}$ be the cost utilized in the transportation problem – not the actual cost – to serve a demand in a given region $j$ from FC $i$. Because fulfillment is myopic, each demand region has a prioritized list of spillover
FCs from which it would prefer to be fulfilled denoted as \( \Omega \). If FC \( i \) is in the \( k \)th slot in region \( j \)’s prioritized spillover list \( \Omega_j \), then we define \( \tilde{c}_{ij} \equiv \Pi(k) \). The cost function \( \Pi(k) \) - which is defined below after the formulation of the linear program – does not depend on the actual cost to ship an item from \( i \) to \( j \), but rather depends only on the FC’s place in line in a region’s desired set of FCs. For example, assume for region \( j \), \( \Omega_j = \{A, D, C, B\} \). This implies that demand in the region will first attempt to get inventory from \( A \), then from \( D \), then \( C \) and finally from \( B \). Thus, \( \tilde{c}_{iA} = \Pi(1) \) and \( \tilde{c}_{iB} = \Pi(4) \). Additional parameters and variables are defined as such:

- \( t \) – Time index (days). Takes on values from 0 to \( L \)
- \( w_{ijt} \) – Variable representing amount assigned from \( i \) to \( j \) in time \( t \)
- \( x_{it} \) – Variable representing the on-hand amount of inventory in \( i \) at the start of time \( t \)
- \( x_{i0} \) – The starting inventory in \( i \)
- \( \theta_{it} \) – Inventory that was previously ordered set to arrive in \( i \) on time \( t \). (Relevant if \( r < L \))
- \( \bar{d}_j \) – The expected demand in region \( j \) per day

The transportation problem is then formulated in this way:

\[
\begin{align*}
\min_{w, x} & \quad \sum_{i,j,t} \tilde{c}_{ij} w_{ijt} \\
\text{s.t.} & \quad \sum_j w_{ijt} \leq x_{it} \quad \forall i, t \\
& \quad \sum_i w_{ijt} = \bar{d}_j \quad \forall j, t \\
& \quad x_{it} = x_{i,t-1} + \theta_{it} - \sum_i w_{i,j,t-1} \quad \forall i, t > 0 \\
& \quad w_{ijt}, x_{it} \geq 0 \quad \forall i, j
\end{align*}
\]

For feasibility and computational reasons, the true formulation includes a dummy supply node to account for possible system stockouts and is time compressed over periods during which no inventory arrivals occur.

We defined the cost function \( \Pi(k) = \sum_{m=0}^{k} p^m \) for \( p < 0.5 \), which approximates the best myopic inventory allocation.