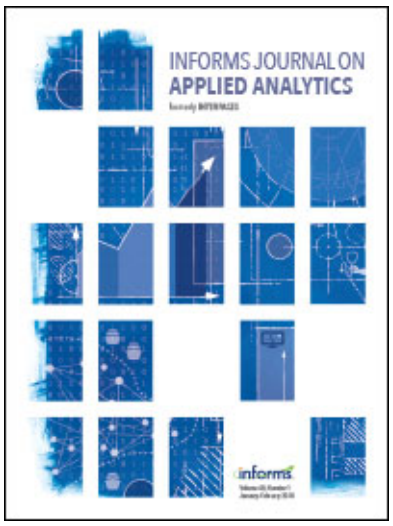


INFORMS Journal on Applied Analytics

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

Regionalize and Scale: Amazon's Fulfillment Network Design for Faster and Cheaper Delivery

Amitabh Sinha, Jeremy Agte, Russell Allgor, Cristiana L. Lara, Ashish Agiwal, Semih Atakan, Lourdes Campo, Tolga Cezik, Daniel Chen, Qi Chen, Jesse Fischer, Eitan Gor, Nader Kabbani, Ryan Kennedy, Kaushik Krishnan, Yuan Li, Shahbaaz Mubeen Mamadapur, Tanmay Mathur, Nick McCabe, David Mildebrath, Eric Powell, Andrea Qualizza, Denton Schroeder, Nityansh Seth, Xiaoyan Si, Kaushik Sinha, Darren Stegner, Jun Xiao, Ling Zhang, Shanshan Zhang, Jikai Zou

To cite this article:
Amitabh Sinha, Jeremy Agte, Russell Allgor, Cristiana L. Lara, Ashish Agiwal, Semih Atakan, Lourdes Campo, Tolga Cezik, Daniel Chen, Qi Chen, Jesse Fischer, Eitan Gor, Nader Kabbani, Ryan Kennedy, Kaushik Krishnan, Yuan Li, Shahbaaz Mubeen Mamadapur, Tanmay Mathur, Nick McCabe, David Mildebrath, Eric Powell, Andrea Qualizza, Denton Schroeder, Nityansh Seth, Xiaoyan Si, Kaushik Sinha, Darren Stegner, Jun Xiao, Ling Zhang, Shanshan Zhang, Jikai Zou (2026) Regionalize and Scale: Amazon's Fulfillment Network Design for Faster and Cheaper Delivery. INFORMS Journal on Applied Analytics 56(1):23-41. https://doi.org/10.1287/inte.2025.0295

INFORMS JOURNAL ON APPLIED ANALYTICS
Vol. 56, No. 1, January–February 2026, pp. 23–41
ISSN 2644-0865 (print), ISSN 2644-0873 (online)
https://pubsonline.informs.org/journal/ijaa

THE FRANZ EDELMAN AWARD
Achievement in Operations Research

Regionalize and Scale: Amazon's Fulfillment Network Design for Faster and Cheaper Delivery

Amitabh Sinha,[a,*] Jeremy Agte,[a] Russell Allgor,[b] Cristiana L. Lara,[a] Ashish Agiwal,[a] Semih Atakan,[a] Lourdes Campo,[c] Tolga Cezik,[a] Daniel Chen,[d] Qi Chen,[a] Jesse Fischer,[a] Eitan Gor,[a] Nader Kabbani,[e] Ryan Kennedy,[a] Kaushik Krishnan,[a] Yuan Li,[a] Shahbaaz Mubeen Mamadapur,[a] Tanmay Mathur,[f] Nick McCabe,[g] David Mildebrath,[a] Eric Powell,[h] Andrea Qualizza,[i] Denton Schroeder,[j] Nityansh Seth,[f] Xiaoyan Si,[a] Kaushik Sinha,[k] Darren Stegner,[a] Jun Xiao,[a] Ling Zhang,[a] Shanshan Zhang,[a] Jikai Zou[l]

[a] Amazon.com, Bellevue, Washington 98004; [b] Auger.com, Bellevue, Washington 98004; [c] Amazon.com, Arlington, Virginia 22202; [d] Amazon.com, Singapore 018916; [e] Hims & Hers, Bellevue, Washington 98004; [f] Blue Origin, Kent, Washington 98032; [g] Amazon.com, Tempe, Arizona 85281; [h] Cape, Washington, District of Columbia 20003; [i] Amazon.com, Washington, District of Columbia 20001; [j] Amazon.com, New York, New York 10001; [k] Amazon.com, Boston, Massachusetts 02210; [l] Coupang, Inc., Seattle, Washington 98101

*Corresponding author

Contact: amitabsi@amazon.com, https://orcid.org/0000-0002-6035-3671 (AS); agte@amazon.com (JA); allgor@alum.mit.edu (RA); larcrist@amazon.com, https://orcid.org/0000-0002-1990-3299 (CLL); aagiwal@amazon.com (AA); atakans@amazon.com, https://orcid.org/0000-0002-9699-0314 (SA); lourcamp@amazon.com (LC); cezikm@amazon.com (TC); chonglic@amazon.com, https://orcid.org/0000-0001-6538-5608 (DC); chnqc@amazon.com, https://orcid.org/0000-0002-2389-2238 (QC); jessef@amazon.com (JF); eitangor@amazon.com (EG); nader.kabbani@gmail.com (NK); rykenn@amazon.com (RK); krikaush@amazon.com, https://orcid.org/0009-0003-7125-5676 (KK); yyuanli@amazon.com (YL); shnb@amazon.com (SMM); tanmayamathur@gmail.com (TM); mccabn@amazon.com (NM); dmildebr@amazon.com (DM); ericpowell80@gmail.com (EP); qualizza@amazon.com (AQ); dents@amazon.com (DeS); nityansh5@gmail.com (NS); xiaoyasi@amazon.com (XS); sinhakau@amazon.com (KS); dstegner@amazon.com (DaS); juxia@amazon.com (JX); lnz@amazon.com (LZ); zshansha@amazon.com (SZ); jikai.zou@gmail.com (JZ)

Accepted: November 7, 2025

https://doi.org/10.1287/inte.2025.0295

Abstract. This paper presents Amazon's implementation of regionalization, a strategic transformation of its fulfillment network in the United States that partitioned the country into eight interconnected but largely self-sufficient regions to address growing complexity and inefficiencies in order fulfillment. By 2021, Amazon's unprecedented network growth had led to nonlinear increases in transportation system complexity and suboptimal equilibria that increased costs while reducing delivery speeds. The core principle behind regionalization involves matching demand with capacity through geographical partitioning, moving away from a flexible national network toward a more structured regional approach. The development leveraged extensive operations research methodologies over 1.5 years, encompassing region design, network optimization modeling, inventory-speed trade-off analysis, and significant software and operational changes. Following successful pilot and full network deployment by March 2023, regionalization helped deliver substantial improvements: a 15% reduction in distance between sites and customers, 12% fewer middle-mile touchpoints, increased in-region fulfillment from 62% to 76%, and the first reduction in cost-to-serve per unit since 2018, with over $0.45 per-unit savings in the United States alone while simultaneously improving delivery speeds with over nine billion items delivered the same or next day globally in 2024.

Keywords: Edelman Award • real-time resource allocation • retail operations • transportation network optimization

Introduction

By the end of 2021, Amazon in the United States was shipping over one billion items per year, through a network of hundreds of fulfillment centers (FCs), each carrying millions of unique items, spread throughout the continental United States. These items were delivered through a single national transportation network consisting of internal resources and external delivery partners, leveraging a variety of facilities, for example, FCs, sortation centers (SCs), and delivery stations (DSs), that were multimodal and densely interconnected. The nationally connected network offers flexibility in choosing the location from which to fulfill each order. This can be an advantage because doing so can reduce the risk of backlogs and disruptions in a single site and increase Amazon's ability to handle variations in demand.

However, Amazon's proposition of offering a vast selection and fast delivery, with low cost-to-serve, faced challenges as a result of its unprecedented network growth in the past few years. The scale of growth led to a nonlinear

increase in complexity of our transportation systems. The need for speed requires systems that make real-time decisions on the location from which to fulfill the order, based on the state of the network and inventory availability, which in turn resulted in high variability in the set of FCs serving any destination area (e.g., city). That high variability translated to increased costs and reduced speeds.

In 2021, Amazon senior management identified this problem and assembled a multidisciplinary team to develop a solution. That solution is regionalization, where we partitioned the United States into eight interconnected but largely self-sufficient regions. The development and launch of regionalization were the result of 1.5 years of sustained operations research (OR) studies examining each aspect of regionalization, leveraging a multitude of OR methodologies.

The core idea behind regionalization is to reduce the variability of order assignment, by carefully matching fulfillment capacity with demand in a way that ensures sufficient capacity to serve each demand point and restricting order assignment systems to fulfilling largely from the matched capacity. There are many ways to do this matching; at Amazon, we implemented this through a geographical partition of customer locations into eight regions combined with the allocation of fulfillment capacity to each region.

Although the idea is conceptually simple, we needed to carefully define several other components of the network structure. We then evaluated these through a suite of custom-built OR tools, creating models of several different scenarios. Our network modeling revealed not just the potential cost savings and speed improvements from regionalization, but also the impact of critical inputs, which allowed us to prioritize efforts in steering those inputs (such as inventory levels) to deliver the right outcomes.

The network modeling stage was followed by testing through simulation in higher-fidelity systems. We then obtained senior management approval to proceed with a pilot, which was accompanied by a careful implementation schedule and risk mitigation actions. We had to make several changes in both software and operations. In our software systems, we made changes in how we manage inventory, generate customer promises, and assign orders to FCs. We also had to change the configuration of our FCs, network connectivity, and schedules of our transportation and delivery resources. The implementation took two months with dozens of teams working in close coordination. In the remainder of this paper, we first describe Amazon's operations, and follow with the modeling, execution, and impact of regionalization.

Core Principles of Regionalization

As we mention above, the core idea behind regionalization is matching demand with capacity. To conceptualize this, we first briefly describe the critical elements of customer order delivery, from the perspective of both the physical and the software systems involved.

Overview of Amazon's Operations and Systems

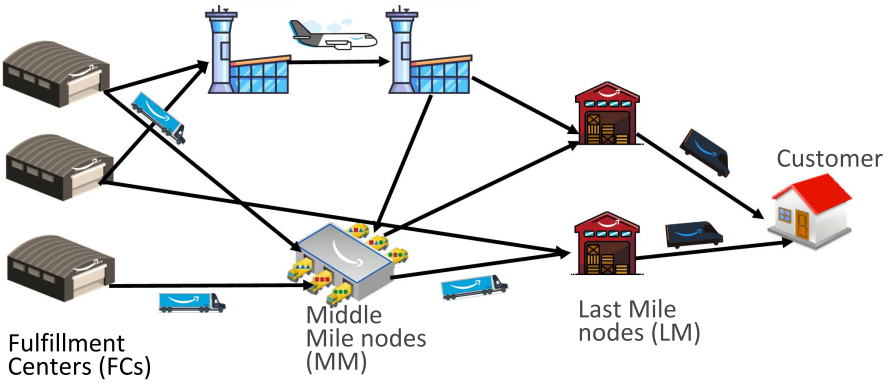
In the United States, Amazon has several FCs, which form the heart of our network. FCs store unsold inventory, which typically arrives from vendors and third-party sellers in trucks. When a customer places an order, it is assigned to a specific FC for fulfillment. Each FC processes several orders in parallel and carries a queue or backlog of orders waiting to be processed. When an order reaches the beginning of the queue and is ready to be processed, items in that order are picked from inventory bins, and routed via conveyance systems to a packing station. The items are then packed into a box ("box" is a simplification—Amazon also uses bags and ships some items in their original packaging), and a shipping label is applied. Packages are then loaded onto trucks for different destinations. The trucks carry these packages through a network of nodes, which are called the middle-mile (MM) nodes, until the packages arrive at a last-mile (LM) delivery node. At the LM node, packages are transferred to delivery vans, which make the final delivery to the customer.

For both the MM and LM, Amazon uses a combination of first-party and third-party networks. The first-party network consists of Amazon-operated SCs and delivery stations, with aircraft, trucks, and delivery vans operated by delivery partners. The third-party network consists of commercial carriers such as the United States Postal Service (USPS) and United Parcel Service (UPS); Amazon does not have visibility into these operations. In this paper, we will focus on our FCs and first-party MM and LM networks. For any item in an order, we will refer to the sequence of selected FC, MM, and LM nodes, and schedule of actions, as the fulfillment "path" of the item. Figure 1 shows a brief schematic.

At any instant in time, the network has a "state." This state can be described by several elements, such as available processing capacity at each node (FCs, MM, and LM), available transportation capacity on each transportation resource, and available inventory at each FC. In addition, many resources have backlogs and future reservations: for instance, an FC typically has a backlog of orders waiting to be processed, and a transportation resource (such as a truck scheduled for tomorrow) may have some capacity already allocated to future packages. For orders waiting to be processed at an FC, we will use the words "backlog" and "queue" interchangeably.

A customer's journey in making a purchase at Amazon is fairly complex, but again we focus on the critical elements necessary for this paper, which starts shortly before the moment of purchase. Consider a customer who is evaluating a single item and trying to decide whether to purchase it. At that time, we

Figure 1. Schematic of the Fulfillment Network Showing the Different Building Types from Fulfillment Centers to the End Customer



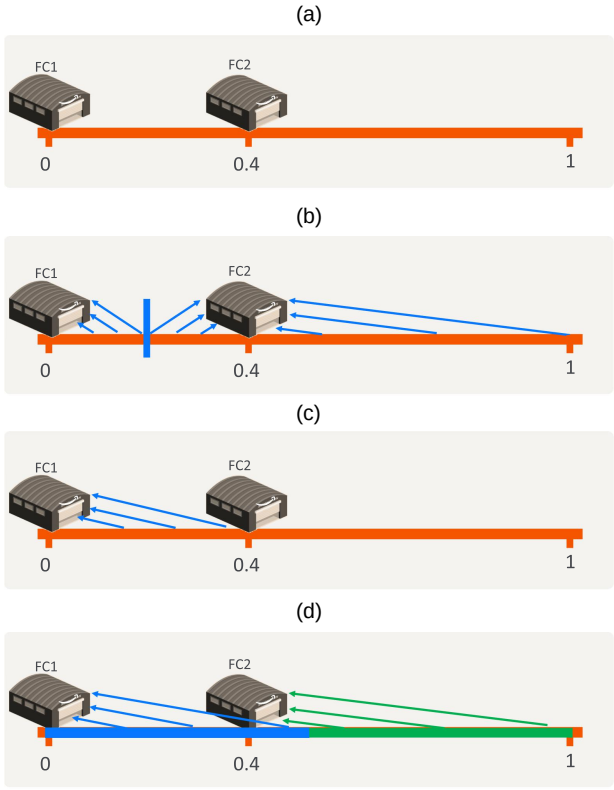
show the customer a "promise"—the deadline by which the item will be delivered, if the customer decides to buy then. Internally, the promise is given based on an analysis of the network state, and verification that at least one feasible delivery path exists. Once a customer makes a purchase decisions, all ordered items are placed in a virtual queue, and at a subsequent point in time (called "order assignment"), the items are assigned to specific FCs and paths. This assignment is described in more detail in the "Fulfillment Execution Software Before and After" section. The assigned FC and path at order assignment time may be different from the feasible path that was used to generate the customer promise, because the final FC assignment is an opportunity for Amazon to optimize the network performance. The assigned path may change more than once until the item is picked in the FC, after which the path stays locked until delivery (unless there are disruptions, which are not in the scope of this paper). For simplicity, in this paper, we focus on the order assignment system and ignore how customer promises are generated.

Matching Demand with Capacity

If we want to minimize time to delivery, order assignment must be done in near real time, to enable sufficient time for processing, transportation, and delivery within the same or next day. The key question is: how should we construct our transportation network and do order assignment to find a solution that is optimal for cost and speed?

We start with describing a simple, "greedy" policy, and demonstrate how it leads to undesirable outcomes. We first define a general cost function $cost_p$, which is a combination of the transit time and transportation cost along path p. At its simplest, as soon as a customer orders an item, the greedy policy evaluates all feasible paths, and selects the one that minimizes this cost function, breaking ties uniformly at random (i.e., each possible outcome has an equal chance of being chosen). The cost function may include various offsets, opportunity costs, or dual variables that are designed to steer order assignment toward specific outcomes, which we will examine later in the paper; however, we first illustrate the core principle through a stylized example.

Figure 2. Simple Example with Customers Distributed on the Unit Interval [0, 1] and Two Fulfillment Centers



Notes. (a) Setup with two FCs. (b) Time $t = 0$: FC1 gets 0.2 load, and FC2 gets 0.8. (c) Equilibrium. FC1 has 0 backlog. FC2 has 0.2 backlog. Customers in [0.4, 1] are indifferent between both FCs. Both FCs get 0.5 load in equilibrium. (d) Regionalization. Customers in [0, 0.5] are assigned to FC1, and customers in (0.5, 1] are assigned to FC2. Both FCs have 0.5 load and 0 backlog.

Consider the network in Figure 2. Figure 2(a) shows our setup. Customers are uniformly distributed in the continuum $[0,1]$, with total demand of one per unit time. There are two FCs, labeled FC1 and FC2, at locations 0 and 0.4, respectively, each with a capacity of 0.5 per unit time. The cost function is simply the total service time (i.e., waiting plus transit), with transit speed of one per unit time.

As Figure 2(b) shows, at time $t = 0$, FC1 is the cost minimizer for customers in [0, 0.2], and FC2 is the minimizer for [0.2, 1]. This means that the load on FC2 is 0.8 per unit time. However, its processing capacity is only 0.5 per unit time; therefore, orders start adding to its backlog. Consequently, FC2 progressively becomes less attractive to customers in [0.2, 0.4], until we reach an equilibrium.

Figure 2(c) shows the equilibrium, where FC2 has a backlog of 0.2, which takes it 0.4 units of time to process. For customers in [0, 0.4], FC1 is the cost minimizer, because it has no backlog. For customers in (0.4, 1], both FCs are equal in terms of cost. For example, for the customer at location 1, FC1 has cost 1 (all transit), whereas FC2 also has cost 1 (0.4 waiting plus 0.6 transit). As a result, customers in the interval (0.4, 1] are allocated in a 1:5 ratio between FC1 and FC2 in equilibrium. The average service time (equal to average cost for this example) is 0.5. At Amazon, at a scale of dozens of FCs and millions of items sold every day, we found ourselves in a similar situation with an unsatisfactory equilibrium speed and increasing transportation costs.

Now consider Figure 2(d), where we do a forced matching of demand to capacity. Customers in [0, 0.5] are matched to FC1, and customers in (0.5, 1] are matched to FC2. Both FCs have an equal load of 0.5 (which matches their capacity); therefore, no backlog is ever developed. The average service time is now 0.3, which is a 40% speed improvement relative to the greedy policy.

This balanced matching is at the core of regionalization. In reality, however, there are several additional factors. Customer demand in the United States is not a one-dimensional uniform continuum; instead, it is a two-dimensional surface with highly nonuniform demand density. Amazon has dozens of FCs, not only two. Customers order millions of different items, all of which cannot be stocked in each FC. There is significant variability in customer orders along dimensions such as time, what items customers are ordering, and geography. We operate a first-party transportation network as we show in Figure 1, with a cost function that is much more complex than the example illustrates.

One way to steer order assignments away from the greedy equilibrium is to add artificial cost offsets to FCs that incur too much load under a greedy policy, as we will discuss briefly later. Such cost offsets can be used in general to prevent overuse of constrained resources, including scarce inventory and expensive transportation assets, and can be computed using the dual solutions of the corresponding primal resource allocation problems. Alternatively, we could change the objective function of the order assignment problem to also minimize backlog, or incorporate other terms to drive toward favorable outcomes. However, as Amazon grew, we found that such solutions were unfavorable because they led to race conditions (where offsets of two or more constrained resources were continuously increasing), which negated their effectiveness. In addition, these cost-offset-based solutions lacked explainability, because cost offsets were computed in real time and therefore changed often. Regionalization, with its rigid structure, brought a level of explainability and auditability that not only led to the cost and speed efficiencies, but also allowed us to quickly build on it with additional improvements such as in inventory placement.

Regionalization at Amazon, at its simplest, is *an operational (hardware) and system (software) design and implementation of a roughly balanced partition that matches demand with fulfillment capacity*. To implement this, however, we had to address several technical challenges. This paper describes some of the most significant challenges and thus provides a succinct story of the development and implementation of regionalization.

Literature Review

E-commerce brings the opportunity to delay fulfillment decisions past the customer's purchase time and opens up many opportunities to exploit the order-to-delivery time window to reduce costs, as Xu (2005) eloquently documents. Within Amazon, we have gone through multiple evolutions of our fulfillment policies. Initially, it was largely batch optimization with reoptimization (Xu et al. 2009). In parallel, Amazon's operations have also evolved, from a largely third-party operation, toward our own first-party fulfillment network with our own transportation capabilities (Lara et al. 2023). With this evolution, our original fulfillment systems needed to evolve in parallel to explicitly take into account the fixed-cost nature of our transportation network. The early part of this work is documented in Amazon Science (2022), and regionalization is a refinement of that work (O'Neill 2023). More general descriptions of supply chains, such as at Amazon, include optimization and control (Kassmann and Allgor 2006) and picking within fulfillment centers (Allgor et al. 2023). In the academic literature as well, significant work in the area of e-commerce fulfillment has been done in the past two decades. Acimovic and Farias (2019) and Jasin et al. (2019) provide excellent broad surveys.

The fulfillment flexibility in e-commerce can prove advantageous under certain situations, as DeValve et al. (2023) show. However, the manufacturing

literature includes a long stream of research dating back to Jordan and Graves (1995) that describes how limiting the flexibility into well-defined structures can bring significant efficiencies. The regionalization model can be thought of as a small extension of the bipartite model of Jordan and Graves (1995).

Fulfillment itself is constrained by the delivery speed promise that we make to our customers. This promise aspect can be best modeled as allocating incoming jobs to servers, in a multiserver spatial setting where total delivery time to a customer depends both on the time at the server (i.e., waiting plus service) and the time to transit from the server to the customer. Recently, several papers have explored equilibria and optimization of such spatial multiserver queues, including Afèche et al. (2022), Besbes et al. (2022), and Carlsson et al. (2024). Our work contributes to this literature by showing a real-world implementation of these phenomena within a first-party transportation network.

Regionalization's Design and Predeployment Evaluation

The first step in creating our regionalization structure was to partition the customer demand, and match FC capacity with it.

Region Design and FC Mapping

The basis of a regionalized network is to strategically decompose the customer geography into a set of distinct regions and maximize fulfillment from FCs within the region. Although smaller region sizes potentially lead to expedited customer deliveries because of the proximity of FCs to customers, they also present a challenge in terms of inventory breadth. Conversely, larger regions encompass a greater number of FCs, potentially offering a more comprehensive inventory selection. However, this may come at the cost of increased delivery times. Therefore, the optimal region size represents a trade-off between these competing factors: sufficiently small to facilitate rapid deliveries, yet large enough to incorporate an adequate number of FCs that can maintain a diverse and comprehensive inventory to effectively satisfy customer demands.

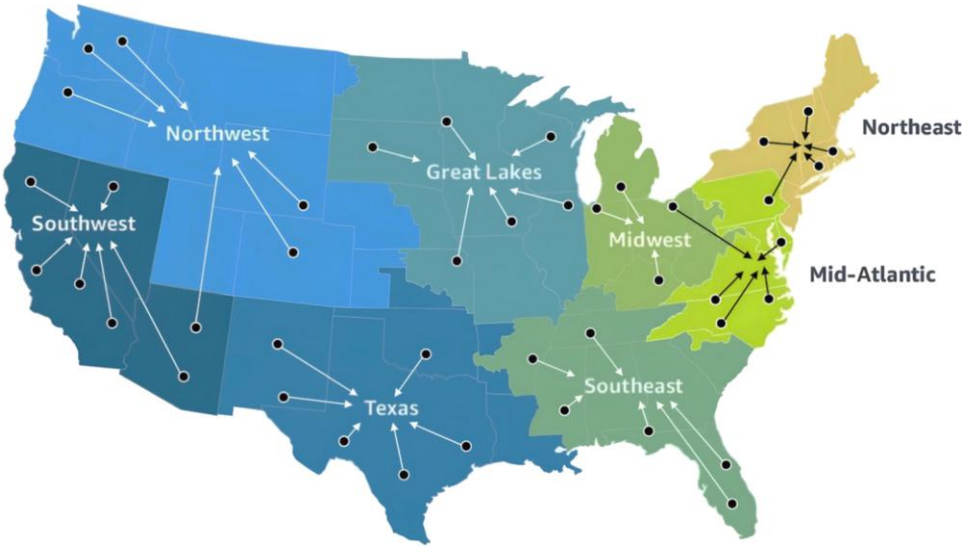
From an outbound fulfillment perspective, customers could be served faster if packages are routed through direct fulfillment lanes from FCs to DSs, or by cross-docking through intermediate SCs. These processes involve presorting packages destined for DSs at a limited number of sortation points within the FCs. Consequently, the quantity of DSs in each region should be constrained to ensure that FCs possess sufficient sortation points for presorted pathways to all in-region DSs. Smaller regions are preferable because they facilitate the establishment of dense, direct fulfillment paths with continuous flow, thereby contributing to accelerated delivery speeds. Of course, regions should still be large enough to offer economies of scale with respect to first-party transportation resources. Based on the expected network capacity in 2023, 8–10 regions offered a good balance between the factors discussed above. As a secondary objective, the regions' boundaries should ease management complexity by following natural topographical boundaries (e.g., time zone lines) and operational boundaries (e.g., Amazon Logistics (AMZL)/USPS jurisdictions) wherever possible.

To explore approaches for creating regions, we framed the problem as a clustering task, where the customer geography would be partitioned into k distinct clusters. We designated five-digit zip codes as the "points" to be grouped into regions. We employed two capacitated clustering algorithms (Mulvey and Beck 1984): one is based on the k-means algorithm and the second is based on integer programming, with additional constraints on maximum demand and size per region. The first algorithm incorporates the concept of regret, defined as the difference between the distances to the closest and second-closest region centroids, for each point. Points are assigned to clusters in descending order of regret, continuing until the constraints on maximum demand share per region and maximum distance per region are satisfied.

However, the initial results proved unsatisfactory for two primary reasons. First, the generated regions lacked proper shape and geographical compactness. Second, a single algorithm struggled to effectively handle geographies with diverse characteristics, such as varying population densities and geographical features. For instance, the model failed to consistently cluster the most densely populated region (Northeast) and the most sparsely populated region (Northwest) in an efficient manner. Consequently, we implemented substantial manual adjustments to the clustering results and evaluated multiple region designs. Figure 3 illustrates the eight-region design selected for further refinement, along with the corresponding FC assignments, which we describe next.

After defining regions, we developed FC Regionizer (FCR), a mixed-integer optimization model, to allocate FC resources. FCR prioritizes alignment-driven objectives while minimizing transportation costs and balancing regional demand with FC capacity (see Appendix A for model details). FCR generates multiple scenarios by allowing trade-offs between capacity deficits and transportation costs, acknowledging unmodeled constraints and stakeholder preferences. This approach creates a Pareto-surface of solutions with diverse visuals, facilitating consensus among business partners. The goal is to enable informed decision making by providing a range of options, addressing complex logistics challenges rather than pursuing

Figure 3. Map Showing the Approximate Structure of the Eight-Region Design



Notes. The dots represent FCs assigned to serve each region. Boundaries and FC locations are approximate and intended to be only directionally representative.

a single optimal solution. Figure 3 provides an illustration of one of the proposed solutions selected for further refinement.

Network Design and Optimization

Given the extensive scope of planned changes, it was essential to develop a comprehensive understanding of regionalization at various levels and its impact on key network performance metrics. To achieve this, we utilized an internal custom-built suite of network optimization tools, called Integrated Transportation and Topology Optimization (ITTO). ITTO integrates mathematical models, data pipelines, and other analytics tools to assess the impact of major network transformations and conduct in-depth scenario analyses. Using the ITTO suite, we evaluated multiple factors such as various region definitions, regional assignment shares, and potential structures for both regional and cross-regional (XR) networks.

ITTO is structured around three core components, each designed to model a major category of fulfillment decisions. The standard modeling process considers topology, customer demand profiles, and network design parameters as inputs. Executed in sequence, these products generate an optimized network design, complete with performance metrics such as speed and cost. We show a breakdown of the network design process in Figure 4.

The first component, "Shipment Generation," models the assignment of customer demand to FCs, generating shipments that we refer to as origin-destination (OD) flows. This product leverages machine learning and mathematical programming models to generate realistic outcomes, parametrized based on inputs that represent inventory states and other network characteristics.

The second component, "Network Connectivity," involves a static, path-based network optimization model, Middle Mile Topology Optimizer (MMTO), which captures the assignment of OD flows to fulfillment paths through our MM and LM nodes. Each potential path embodies a sequence of facilities through which shipments travel, along with the resources utilized at each step of fulfillment. The model optimizes the flow of shipments by determining the most efficient routes and resource allocations, producing structural metrics such as the share of packages that go in direct FC-to-DS arcs (without the additional SC touch).

The final component, "Network Timing," addresses a network optimization problem on a time-expanded graph, modeling intraday linehaul movements, processes, and hourly capacities. This generates hourly schedules of package processes and movements and enables a fine-grained evaluation of the designed network in terms of cost and speed metrics. An early version of this component is described in Lara et al. (2023).

Figure 4. The Network Design Suite (ITTO) Has Three Major Components

In the remainder of this section, we go into more detail about the second component, which optimizes network connectivity.

At a high level, the network design model MMTO takes MM topology and fulfillment strategy as inputs. Topology inputs specify the process capabilities and capacities of each facility, whereas the fulfillment strategy is represented through path decisions. The model, formulated as a mixed-integer program, minimizes transportation and handling costs under operational constraints. Its outputs provide insights into capacity utilization and fulfillment metrics, such as distance traveled, enabling planners to evaluate the impact of various topology and fulfillment strategies and identify the optimal approach.

Before regionalization, Amazon's fulfillment network was generally connected with cost and delivery speed in mind, resulting in an unstructured yet efficient network based on an assumed set of shipments. With regionalization, the presumed localization of shipments created several opportunities to enhance fulfillment productivity. One such opportunity was the specialization of SCs, which are one class of MM nodes, based on specific processes to improve productivity. Prior to regionalization, SCs managed a variety of tasks, including package sortation, container cross-docking, and the sortation and palletization of shipments destined to different last-mile nodes. Specializing SCs into subcategories was expected to improve productivity, leading to their split into two facility types: hubs and last-mile sort centers (LMSCs).

Hubs were primarily responsible for receiving XR fulfillment from FCs and distributing it to LMSCs. Hub processes were tailored toward achieving improved fill rate over long hauls, such as minimizing empty space in trailers. These facilities were further classified into three subcategories: origin hubs, destination hubs, and hybrid hubs, each serving a specific function in the distribution network using specialized processes. In contrast, LMSCs acted as regional points of consolidation and distribution, positioned before shipments reached last-mile facilities.

Next, the fulfillment strategies were defined based on these new facility types. For regional fulfillment, we designed the network to be as simple as possible, minimizing the number of nodes traversed to reduce fulfillment costs. Furthermore, a cost-effective, regional fulfillment structure would further increase the share of regional flows, because real-time execution systems would likely prioritize these options because of their lower costs. However, XR fulfillment was generally considered to involve longer-haul transportation, requiring the use of one or more hubs. As before, the more costly fulfillment structure would be less preferable by our execution systems, deprioritizing XR fulfillment.

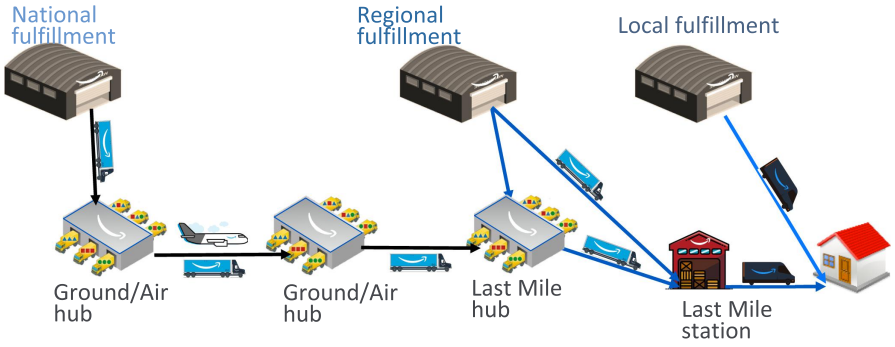
The final state of SC specialization and the fulfillment structure, as we show in Figure 5, were established through multiple rounds of experiments on network design performed using the ITTO toolset. The required capacities for origin, destination, and hybrid hubs were determined using our network design model, which optimizes one- or two-hub touch path assignments for XR flows (i.e., it determines which XR flows should be assigned to paths that go through a single hub and which to paths that go through two hubs). For instance, a higher share of cross-regional fulfillment would result in more packages passing through only a single hub, prompting operators to invest more in hybrid and destination hubs that can handle such flows.

Inventory vs. Speed Trade-Offs

Although we expected regionalization to result in delivery speed improvements, we also anticipated that a one-size-fit-all implementation, where all stock-keeping units (SKUs) were stocked in each region, would not be the most cost-effective. Stocking sufficient inventory of a given SKU to meet a desired regional fulfillment service level (each region stocks to a specified quantile of its regional demand) would likely increase speed (by ensuring fulfillment over shorter distances) but would require more inventory to be held (because of no demand pooling at the national level), compared with stocking to meet a national fulfillment service level (the network as a whole stocks to a specified quantile of national demand). Intuitively, higher-velocity SKUs were the best candidates for a regional fulfillment strategy because their relatively higher demand-to-inventory ratio led to relatively higher speed gains relative to increased inventory requirements, whereas slow-moving (also known as tail) SKUs were better suited for a national fulfillment strategy. However, another complicating factor was the desired service level according to which inventory was to be stocked for each fulfillment strategy. Therefore, we conducted a simulation, varying the fulfillment strategy (national versus regional, and the desired service level) and observing the resulting trade-offs between speed and total inventory to be held.

We segmented SKUs into 10 deciles of equal expected demand to create inventory velocity bands. For each decile, we then assigned a fulfillment strategy: regional, superregional, or national (in decreasing granularity), where inventory is stocked to meet a desired service level in each region, each superregion (the superregions comprise a partition of the regions), or across the nation, respectively. To simplify the decision space, we introduced a monotonicity condition: each faster-velocity band had to be at least as granular in its fulfillment strategy and meet a service level at

Figure 5. The Regionalized Transportation Network Structure Reduces Touchpoints for In-Region Fulfillment

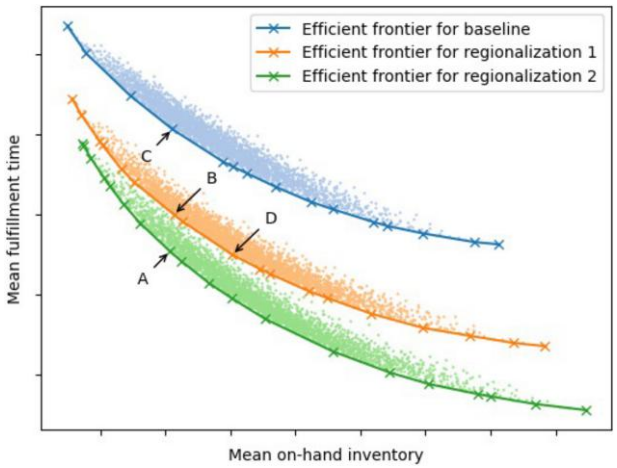


least as high as any slower-velocity band, where we assumed the same service level for all velocity bands using the same fulfillment strategy. Enumerating over this decision space, we calculated the interregional flow using simulation. Achievable speed was, thus, computed as a weighted average of the intraregional and interregional speed for each velocity band, where the weights were determined by the interregional flow for each band and the out-of-stock percentage.

For a fixed assignment of fulfillment strategies, we varied the desired service levels to produce an efficient frontier with different inventory-speed combinations. Iterating over different fulfillment strategy assignments allowed us to compare these efficient frontiers. We found that significant speed could be gained through improved management of inventory positioning (and service levels) enabled by regionalization—even for the same total on-hand inventory. We illustrate this in Figure 6, where each line represents the efficient frontier trade-off between mean fulfillment time and mean on-hand inventory as service level changes for three different regionalization strategies: baseline (national fulfillment for all velocity bands), regionalization strategy 1 (a mix of regional, superregional, and national fulfillment), and regionalization strategy 2 (a mix of regional and national fulfillment). Comparing point A with point C shows shorter fulfillment times (i.e., higher speed) for the same inventory requirements by using some regional fulfillment compared with the baseline. The speed gain results primarily from a focused allocation of inventory by region, as well as from a shift in the distribution of service level by velocity band. Some regionalization strategy assignments are better than others: comparing point A with points B and D shows that strategy 2 can result in attaining higher speed at lower inventory, whereas strategy 1 would require choosing between higher speed and lower inventory (or accepting smaller degrees of improvement).

One of our key conclusions from this simulation analysis is that at higher inventory levels, for the same total inventory, it is possible to achieve the same or higher average delivery speed by operating a regional fulfillment strategy with a lower service level rather than a national fulfillment strategy with a higher service level. Comparing the efficient frontiers of various strategies, we were able to select a suitable threshold of demand velocity, above which we would want to fulfill SKUs with a regional strategy and below which we would want to fulfill them nationally; a key result is that a superregional fulfillment strategy is not as suitable for our current network compared with a regional one, as partially illustrated in Figure 6 by the efficient frontier of strategy 1 (orange) being Pareto dominated by the efficient frontier of strategy 2 (green). Comparing other scenarios, we also validated that the monotonicity condition on service levels was

Figure 6. Simulations Run to Determine the Trade-Off Between Mean Fulfillment Time and Mean On-Hand Inventory



Notes. The figure shows the results from two sample regionalization strategies compared to the baseline of using only national fulfillment. The dots of each color represent the outcomes from different sets of service levels corresponding to a given strategy, with the efficient frontier shown for each strategy. In this example, either regionalization strategy is better than the baseline, but strategy 2 attains a better efficient frontier than strategy 1.

a necessary condition for optimality. This is because the optimal distribution of a specified amount of inventory across velocity bands occurs when the incremental inventory and incremental speed have the same ratio for each band. Among the optimal points along the efficient frontier, we selected a set of service levels that achieved the desired balance between the cost of incremental inventory and the benefit of increased speed and reduced outbound costs arising from reduced interregional flows and reduced network-level stockouts.

Predeployment Order Assignment Simulations

Although regionalization showed great promise through network scenario modeling, senior management wanted to be confident that we could materialize the targeted regional assignment distributions under various proposed region mappings. Excess cross-region assignments would overload hub capacity in a regionalized network topology (see the "Network Design and Optimization" subsection), with the most obvious recovery path being a cascading rollout of cross-region arcs until the network no longer has a regional structure. Early analyses predicted inventory-driven cross-region flows far in excess of design capacities. To address this concern, we conducted a network-level simulation.

The Origin-Destination Load Balancer (ODLB) permutes the shipment cost input to the Order Assignment Model (OAM) (see the "Order Assignment Basics Before Regionalization" subsection) and allows us to test how our execution order assignment systems would perform in various situations. For a given customer demand cluster (destination), it lowers or raises the apparent fulfillment cost from each eligible FC (origin) to incentivize or dissuade shipment assignment. For example, for a given customer order and a specific FC, by adding a positive offset to increase the fulfillment cost from that FC, we may make other FCs appear cheaper and therefore more likely to be selected for fulfilling the order.

We created multiple simulation contexts; in each context, we obtained a counterfactual OAM output sequence corresponding to context-specific apparent fulfillment costs. This allowed us to compare OAM assignment outcomes across contexts, both decision-by-decision and in aggregate. To emulate regionalization, we injected a penalty-valued cost for out-of-region OD pairs and left unmodified in-region costs.

Each simulation decision executes in real time after its production counterpart, using the current production network inventory and capacity states. The simulation does not update its own inventory or capacity state, among other simplifications. Despite these limitations, the simulation allowed us to evaluate the full effect of regionalization inputs on OAM assignments.

We simulated a month of customer order assignment decisions from OAM, affirming the ability to achieve in-region assignment percentages compliant with design capacities. Further studies on individual decisions allowed us to characterize and address non-cost barriers to regionalization.

Evaluation of Expected Speed Impact

In the fulfillment process, the delivery speed is dictated by the interaction between order assignment strategies and network structure. We studied the potential impact on delivery speed, lane density, and average distance traveled by a shipment using a simplified speed simulation tool called SPST to compare the performance of the fulfillment network under nonregionalized and regionalized fulfillment. The delivery speed can be thought of in the context of a spatial queueing system allocating incoming orders to multiple servers (our FCs), where the total delivery time to a customer depends on both the total service time of the server, including any wait/backlog time, and the transit time from the server to the end customer.

As part of this simulation, the delivery capability was represented through a reference OD speed that encapsulates network connectivity, capacity, fulfillment path structure, and timing constraints and, therefore, differs across different fulfillment network design paradigms. For example, in a regionalized transportation network, the reference OD speed for in-region fulfillment is significantly higher than for out-of-region fulfillment. Compared with the nonregionalized strategy, the proposed regionalization strategy results in increased reference OD speed for all in-region fulfillment, while marginally degrading the reference speed for all out-of-region fulfillment. By improving inventory placement, the share of out-of-region fulfillment decreases and leads to an improvement in overall delivery speed.

We demonstrate the comparison between regionalized and nonregionalized fulfillment strategies using three key performance indicators (KPIs): (1) delivery speed—focusing on the share of packages delivered within one day, (2) OD consolidation—defined as the destination demand share fulfilled from each FC on average (we consider the top 80% of demand to exclude inventory tail effects), and (3) the average OD distance traveled by a shipment. For the first two KPIs, a regionalized-over-nonregionalized ratio larger than one indicates improvement over the nonregionalized paradigm, whereas a value smaller than one is desirable for the third. Table 1 shows the observed trend for these three KPIs under different outbound fulfillment capacity surplus levels.

The simulation results in Table 1 show that the relative improvement from regionalized fulfillment

Table 1. Sample Output of KPI Ratios Representing Improvement in Delivery Speed, OD Concentration, and Average Distance Under Regionalized Fulfillment

	Capacity surplus	Speed ratio	OD consolidation ratio	Average distance ratio
1	30%	1.10	1.09	0.98
2	20%	1.17	1.18	0.95
3	15%	1.25	1.27	0.92
4	10%	1.32	1.35	0.90
5	5%	1.38	1.42	0.89

Notes. The capacity surplus percentage indicates the percentage by which FC total fulfillment capacity exceeds total demand. The remaining three columns are the ratios of the respective KPIs in regionalization vs. the nonregionalized network. For example, the first row shows that if we have 30% surplus FC capacity, regionalization yields 10% faster speed, 9% more OD consolidation (flow concentration), and 2% lower distance compared with operating the same network without regionalization. As capacity becomes more constrained, the benefits of regionalization increase.

becomes stronger as the capacity becomes more constrained. For example, the absolute value of speed goes down as the network capacity becomes constrained, but the relative improvement under regionalization increases when compared with the nonregionalized strategy. A key differentiator under these two fulfillment regimes is the average rate of backlog accumulation at FCs; in all cases, backlog formation, if any, is delayed later into the day under the regionalized strategy. Overall, we observed lower FC backlogs under the regionalized strategy. and this inherently led to improved delivery capability because the effective service time (processing time plus backlog time) is reduced. Simulations also highlighted the importance of maintaining the regional capacity-demand balance as we highlight in the "Matching Demand with Capacity" subsection; any regional capacity deficit leads to increased out-of-region fulfillment that propagates through the network and degrades the overall performance.

Evaluation of Expected Cost Impact

All of the studies described so far showed enormous potential for regionalization to reduce fulfillment cost. However, there was still high uncertainty about the financial implications of the initiative because we did not have historical data upon which to base our estimations. To address senior management's concerns, we used scenario modeling to provide a range of possible outcomes. We used the network design models in ITTO (see the "Network Design and Optimization" subsection) to estimate cost savings under varying in-region assignments (IRAs) and established the IRA percentage needed to break even.

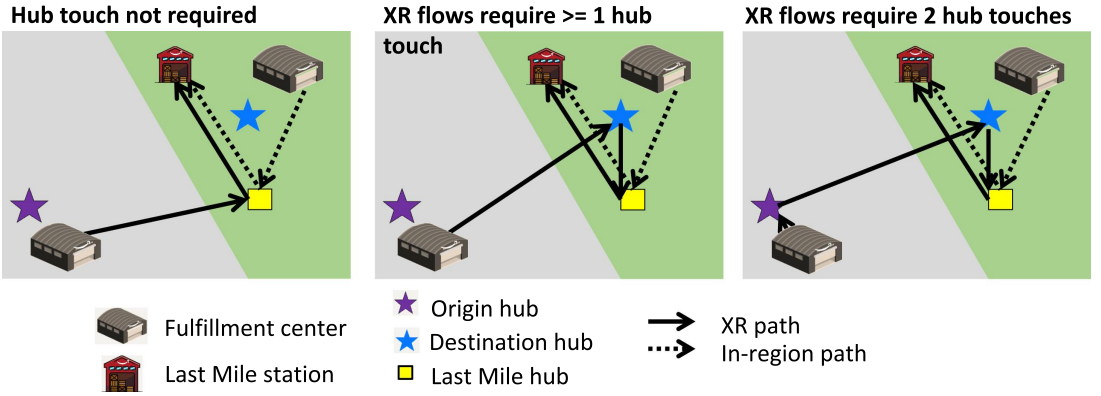
We modeled three network design alternatives, which differed in their treatment of XR flows: (1) all XR flows allowed to bypass the hub, (2) all XR flows required to go through at least one hub (a destination hub), and (3) all XR flows required to go through an origin and a destination hub. The XR connectivity has a direct impact in sortation requirements as well as detour miles (extra miles needed because of additional stops). Hence, it has significant impact on the potential savings. To better understand the connectivity options and decisions, we present an example in Figure 7, where we have two FCs close to the border between regions. Depending on the design alternative chosen, the XR flows get considerably costlier because of the additional touch and the detour.

We ran low-, medium-, and high-IRA scenarios for each XR design alternative. Figure 8 shows the expected savings based on the XR design and IRA assumptions. For low-IRA scenarios, a requirement to go through hub stops at least once completely offsets the savings that accrue from shorter distances and flow concentration. However, as the IRA increases and the XR flows become thinner, the additional hub touches improve consolidation and, in turn, truck utilization. We then combined the transportation cost analysis with additional considerations, such as software execution and robustness, and we developed our final network connectivity plan.

Fulfillment Execution Software Before and After

For each of the millions of customer orders Amazon processes daily, our order assignment system makes several discrete decisions to ensure demand is fulfilled within the promised delivery time window. These decisions include (1) which items are boxed together to form a shipment; (2) which FC will process each shipment; (3) whether we need to transfer items between FCs for shipment consolidation; (4) which transportation route (path) to use for each shipment, including whether we can synchronize deliveries to the customer destination; and, critically, (5) when each of these series of operations will take place. A single model, the OAM, is responsible for making the decisions on the above questions. OAM solves a variant of the set-partitioning problem for each order by making a choice from the discrete set of candidate shipments. A candidate shipment represents one way to fulfill a customer order in full or in part. Although

Figure 7. Example Illustrating the Network Design Alternatives Tested and Their Impact on Sortation Requirements and Detour Miles



Notes. The three figures show different options for connecting an FC in one region with a last-mile node in a different region. In the left panel, we allow the FC to connect directly to the other region's last-mile hub. In the middle panel, the FC connects to a destination hub, which in turn connects to the last-mile hub. In the right panel, we also add an origin hub connection right after the origin FC. In all three cases, the in-region path (dotted line) is the same: from in-region FC (top right), to last-mile hub, to last-mile station.

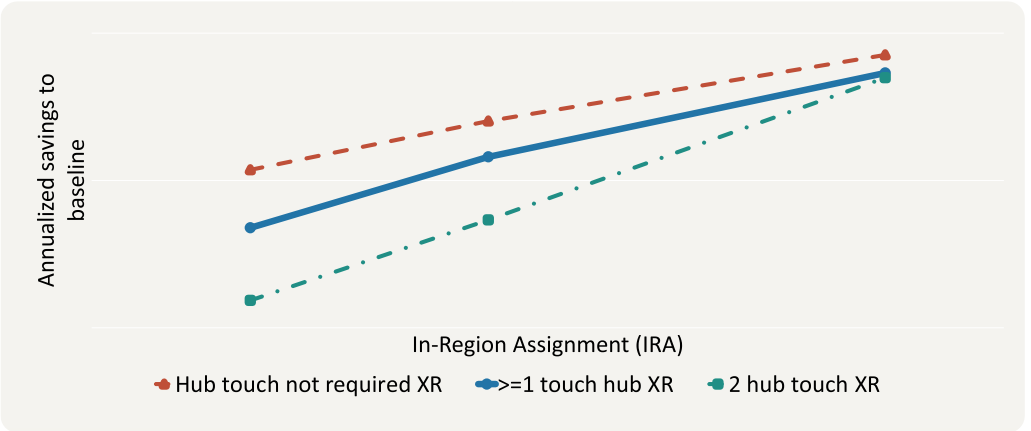
we can exhaustively enumerate all candidate shipments for small orders, we leverage local search heuristics to select a heterogeneous subset of candidate shipments for larger orders. This is necessary to keep the model size contained so we can solve it in less than one second for more than 90% of all new fulfillment orders, a strict requirement to ensure the timely reservation of inventory and transportation capacity that will meet each item's committed delivery time window. Naturally, order assignment decisions are subject to reevaluation at any point, including periodic reoptimization of all pending orders simultaneously to resolve inefficiencies of individual order assignment decisions. For brevity, we omit discussing these models, focusing instead on the main changes introduced to OAM to achieve regionalization.

Order Assignment Basics Before Regionalization

Amazon models the order assignment problem as a lexicographic goal-programming problem. Under FC inventory availability and transportation capacity constraints, the primary goals include the minimization of promised delivery time window violations and total fulfillment cost. In our example below, we consider only these two objectives for simplicity. Suppose the model is optimizing the assignment of an order for only two items, A and B, which have the candidate shipments as shown in Table 2.

1. The first problem answers the question: *what is the lowest delivery time window violation score I can achieve?* The answer is zero (i.e., we can fulfill both items A and B without violating promised delivery timeframes). For example, fulfilling both items from FC3 achieves

Figure 8. Cost Savings Relative to a Nonregionalized Baseline Network Increase as IRA Increases for all Three Connectivity Options



Notes. The three connectivity options shown are the same as in Figure 7. They describe three ways to connect an FC in one region with a customer in another region. The IRA on the x-axis is our primary metric for regionalization and measures the percentage of customer orders that are shipped from FCs in the same region as the customer.

Table 2. Simple Two-Item Example of Order Assignment Showing How the Model Makes Fulfillment Decisions

	FC1	FC2	FC3	FC4	FC5
Shipment item(s)	A	A	A and B	B	A and B
Objective 1: delivery timeframe violation	1	0	0	0	0
Objective 2: total fulfillment cost	$2.00	$1.00	$3.00	$1.00	$3.50

the minimum, as does fulfilling via two shipments from FC2 and FC4.

2. The second problem answers the question: *what are the least costly fulfillment assignments such that we do not make the prior objective value (i.e., delivery timeframe violation) any worse?* The answer is $2, which is achieved by selecting the candidate shipment from FC2 to fulfill A and the candidate shipment from FC4 to fulfill B.

The stylized model above is sufficient to optimize fulfillment assignments when fulfillment costs are tightly coupled to individual shipment construction. As the structure of Amazon's fulfillment costs evolved from shipment-focused rates to more resource-oriented expenses, such as the cost of procuring and utilizing a truck, the true operational costs were not adequately captured. As a result, we adapted the way OAM considered the shipment-level costs. We extended the model, adding more objectives to it, while allowing it to relax the existing shipment-level cost objectives. The idea behind this extension, dubbed cost relaxation (CR), was to allow a more expensive cost solution at the shipment level if the new assignments strictly improved other resource-oriented objectives such as the total number of shipments and/or the number of out-of-region shipments.

For illustration, suppose we allow the entire shipment cost to deviate by *up to* 50% from the optimal solution to reduce the total number of shipments. We can build upon the optimization model to answer the following question: can we fulfill the demand with a smaller number of shipments (i.e., only one) in a way that (1) we do not violate the delivery promise window and (2) the total cost is at most 50% worse than the optimal cost obtained previously (i.e., $2)? The solution we obtained earlier prescribed two shipments for a total cost of $1 + $1 = $2. The CR allowance permits a solution that has a cost of up to $3 (= $2 × (1 + 0.5)). With the allowance afforded by CR, the assignment fulfilling from FC3 at a cost of $3 is now identified as optimal.

In the context of regionalization, the structural changes to the fulfillment network naturally increased the shipment-level cost of XR shipments but did not fully capture the opportunity cost of the consumption of out-of-region inventory, warehouse resources, and transportation capacity. To maximize the benefits of a regional fulfillment strategy while balancing against shipment-level costs and consolidation opportunities, we leveraged an additional layer of cost relaxation to incentivize in-region shipments.

In Appendix B, we present the OAM formulation.

Order Assignment with Regionalization

Continuing with the example above, suppose that the selected FC3 is a warehouse in a region other than the customer's region, whereas FC5 is a warehouse in the same region as the customer. As Table 2 shows, the shipment cost for FC5 could be higher than the cost for FC3 although FC5 is an in-region warehouse. That can happen because the shipment costs consist of a variety of additional inputs. Some of these costs are "real" (e.g., third-party carrier rate cards), whereas others are "artificial" (e.g., load balancing costs or inventory or transportation resource opportunity costs) (Acimovic and Graves 2015). With regionalization, we built an additional layer of CR allowing us to relax up to, for example, 100% of the total cost for the initial solution, if doing so achieved a better in-region assignment while maintaining the same (or better) consolidation outcome. In the case above, CR would select FC5 at a cost of $3.5, within the 100% tolerance, which would allow costs to increase up to $4, that is, twice the cost of the original solution of $2. CR was easy to implement but was not guaranteed to be Pareto optimal; therefore, we later replaced it with a blended objective approach that better balanced trade-offs, although this improvement was not essential to regionalization's initial success.

Regionalization Software Enhancements and Simulation

The desire to regionalize Amazon's outbound fulfillment network necessitated the implementation of new software and systems that could assess the impact of modification to our operational capabilities in line with extensions to our fulfillment optimization models. The implementation process presented several critical challenges, including (1) the lack of a readily available software stack to evaluate regionalized network and configuration scenarios and order assignment models; (2) missing key features in the simulation software, which were needed to validate and operationalize a production pilot; and (3) a highly constrained timeline to develop new technology and validate hypotheses in time for the pilot launch.

In parallel, we invested in the integration of fulfillment optimization systems with what-if simulation

capabilities. This includes the ability to perform offline stateful simulations considering differing operational capabilities, and different parametrizations of the models. We used discrete event simulation making order assignment decisions on a predicted sample of orders while keeping track of transportation and labor capacities. Despite these complexities, the enhanced simulation software allowed us to simulate dozens of scenarios and helped us identify potential errors in the thousands of operational configurations that were set to change and achieve the timeline of the project. In particular, reorienting the network toward regionalization required substantial changes of new lane additions and lane deprecations to redefine the new operational configurations, which needed validation via simulation.

Leveraging these simulation capabilities, we built confidence in our approach and safely executed an aggressive plan to pilot these changes in a set of regions where the impact would be most pronounced according to our simulations, and eventually to launch network-wide.

Implementation

We divided the implementation of regionalization into two phases: (1) pilot launch in two regions in January 2023 and (2) network-wide launch planned three months later based on pilot results. The implementation plan included all parts of the outbound supply chain and required significant effort to ensure that the components of the supply chain (e.g., last mile and middle mile) were coordinated in the transition to the new network design.

Our pilot regions, Northeast (NE) and Mid-Atlantic (MA), comprised a large fraction of the U.S. network demand and relied on each other's fulfillment capacities; NE FCs fulfilled MA demand and vice versa. We clearly defined our success metrics up front, with the key output metrics being cost-to-serve and delivery speed. Within three weeks, cost to serve and speed in the pilot regions were noticeably improved. Therefore, we expedited the network-wide launch and moved the U.S. network to a regional construct a month prior to the initially planned timelines. Accordingly, the full implementation took place in March 2023. In the months that followed, we continued additional refinements to improve customer experiences and aligned inbound systems to support the regional outbound network. The regional network structure is fully deployed across the U.S. network as of this writing.

Implementation Challenges

We faced challenges at different stages of implementation. Initially, we had to provide evidence of the associated benefits to leaders of business units within our operations, such as linehaul trucking and sortation centers. This was especially difficult considering the potential immediate impact on our customer deliveries at a large scale. In addition, our systems are extremely complex and subject to variability from internal and external sources, and we could not model all these factors. Another key challenge was influencing mental model shifts: because most operational teams were used to the national fulfillment model, they initially found it difficult to see how restricting fulfillment flexibility through regionalization could actually lead to improved performance. For example, with regionalization, orders were steered to FCs allocated to serve the customer's region, rather than the closest FC. We had to help operations teams (such as those accountable for delivery speed and delivery costs) understand the drawbacks of myopic assignment policies, and how balancing capacity and demand resulted in a new equilibrium with faster speeds and lower costs.

To prove the concept, we initially used scientific modeling, which included analysis of several scenarios with varying inputs (such as varying the underlying inventory placement or varying different connectivity policies). We also dove deep into analyzing a large number of projected performance metrics, such as the projected performance of each of the eight regions on the truck fill rate and the total number of trucks required on each type of connection. Our stakeholders asked for several follow-ups, which we were able to deliver using scenario analysis leveraging our ITTO and simulation software for network modeling.

After stakeholders were sufficiently convinced, we were able to proceed with the pilot. As noted earlier, the results of the pilot quickly surpassed our projections, which led to increasing confidence and ultimately, nationwide launch.

Impact

Regionalization significantly simplified Amazon's network structure. Prior to regionalization, per Adam Baker, vice president of Global Transportation Services within Amazon, "it was like pushing on a giant spiderweb" (O'Neill 2023). As a result of regionalization, the number of lanes reduced significantly, facilitating decisions about when and how much to ship between regions. Items that were not available in the customer's region were still shipped to them through the cross-region network, but when items were available within the customer's region, we sharply increased the percentage that were shipped from within region.

To establish the impact of regionalization, we started by identifying KPIs that we expected to

improve, such as the distance the customer orders traveled from FCs to customers, the number of intermediate touchpoints, the percentage of orders fulfilled using FCs from within the customer's region, and the speed at which customer orders were delivered. We measured these KPIs at comparable timepoints in the years prior to and after regionalization to account for seasonality. We then translated this to costs using established methodologies that we consistently use for financial calculations. We make multiple changes to our network each year; therefore, we were not able to precisely allocate which action resulted in what fraction of the impact for some of our metrics.

In this way, however, we established that because in part of regionalization, the distance between our sites and customers decreased by 15%. We also found that touchpoints (transits through intermediate buildings) within our middle-mile network reduced by 12% in the first half of 2023 (Herrington 2023). The solution also helped us avoid driving nearly 16 million miles in 2023 (Amazon 2023). The percentage of customer orders being fulfilled entirely from FCs within each region increased from 62% to 76% in the first half of 2023 (O'Neill 2023). In the fourth quarter of 2023, we fulfilled 600 million additional units from in-region FCs compared with the same period in 2022 (Herrington 2024). In 2023, for the first time since 2018, regionalization helped Amazon reduce cost-to-serve per unit sold globally. In the United States alone, cost-to-serve was down by more than $0.45 per unit on a year-over-year basis (Jassy 2024). Regionalization also helped improve delivery speed relative to the previous several years, with more than seven billion items being delivered the same or next day globally in 2023 (Madan 2024). The success continued into the next year, with more than nine billion items delivered within the same or next day globally through 2024 (Amazon 2025). As that article states, "We've also shortened the distance our deliveries have to travel to reach customers by dividing our operations and transportation networks into smaller, easier-to-serve regions and stocking more of the products our customers want locally."

Regionalization also resulted in a simpler, more auditable operating structure, which enabled us to discover additional opportunities for optimization. Many of our systems were built for a different operating environment: for instance, whereas in 2005 we had many fewer FCs in the United States and relied exclusively on third parties for fulfillment, by 2023 we had hundreds of FCs and the majority of deliveries were through Amazon's own transportation network. As our network continues to evolve, we are building upon regionalization with greater emphasis on coordinating our inbound and outbound supply chain networks.

Conclusion

Regionalization was largely successful and is now part of the steady-state operational structure of our U.S. fulfillment network. The regionalized network structure is subject to continuous improvement, based on business growth, new FC launches, and changes in customer demand patterns. One of the largest efforts after the initial launch has been in restructuring our inbound network to ensure that the right inventory is routed to FCs within each region, thus enabling higher levels of in-region fulfillment (Jassy 2024). Our commitment to improving the customer experience, increasing delivery speed, and reducing costs requires us to continue exploring and developing further innovations.

Our work has significant potential for any industry operating in a spatial/geographic environment and facing real-time resource allocation decisions. This includes the broad domain of retail, where a variety of delivery modes exist (e.g., first-party delivery, third-party delivery, buying online, and in-store pickup). It also includes other service domains such as vehicle dispatch (e.g., ambulances and rideshare) and server allocation for cloud computing. The literature on how to design and operate a large-scale real-time retail network is fairly sparse, and many open research questions remain.

Acknowledgments

This paper describes the scientific research and modeling of the regionalized network, leveraging several tools. However, it builds on substantial prior research at Amazon, as well as mathematical modeling tools that have been built by many scientists and developers over the years. More importantly, translating the research into impact was the result of countless hours spent by operations managers, engineers, and associates on network configuration changes, software updates, building operations changes, troubleshooting, rapid problem solving, and creative thinking. The authors are immensely humbled by, and grateful to, the numerous people who enabled this success. The authors are also thankful for the leadership support they received, which included direction, encouragement, disambiguation, and unwavering trust in their mathematical modeling. Lastly, the authors are thankful to the Amazon internal review and support teams who helped them prepare this paper (including Legal, Public Relations, and Investor Relations), as well as their Edelman coaches for their feedback and support.

Appendix A. FCR Formulation

FCR is an optimization tool that we built to solve the problem of assigning which FCs serve each region. The decision of assigning FCs to regions is made twice a year for operating the network (once for the fourth quarter and once for the rest of the year), plus any time we want to evaluate future-year network scenarios (typically, every

few months). In this appendix, we describe the underlying mathematical model for FCR.

Let $\mathcal{R}$ be the set of all regions and $\mathcal{F}$ be the set of all FCs. We represent a region's fulfillment demand and storage demand over a given period by D_r and C_r, respectively, for each region $r \in \mathcal{R}$. For each $FC_i, i \in \mathcal{F}$, we use D_{ir} to represent the delivery cost (e.g., distance, in the simplest form) to serve one unit of demand in region r from FC_i. Let C_i^{ship} and $C_i^{storage}$ be the FC's shipment capacity (in units) and storage capacity (in cubic feet), respectively, over the same period as the demand. Let $R(i)$ denote the region where FC_i is physically located. Let $\beta \in (0,1]$ be a discount factor. We define the adjusted cost $\tilde{D}_{ir}$, showing the level of preference for in-region mapping, as $\tilde{D}_{ir} = D_{ir} \cdot \beta$ if $r = R(i), D_{ir}$ otherwise. The smaller β is, the smaller the adjusted cost for in-region mapping, relative to an out-of-region FC at the same distance. In this way, β allows us to approximate the true transportation cost, which is cheaper for in-region transportation legs because of fewer intermediate SC connections.

Let $\alpha_i \in [0,1]$ be the percentage of the capacity of FC_i (both shipment and storage) to be reserved for its physically located region, even if it is assigned to another region. Specifically:

- Storage capacity $\alpha_i C_i^{storage}$ and shipment capacity $\alpha_i C_i^{ship}$ will be reserved in region $R(i)$.
- The remaining $(1-\alpha_i)C_i^{storage}$ and $(1-\alpha_i)C_i^{ship}$ can be assigned to other regions.

When $\alpha_i = 0$ for all $i \in \mathcal{F}$, we call this the "hard mapping" scenario. The α_i parameter allows an FC to be multimapped to two regions (one always being its physical region).

Let $\mathcal{K}$ be the set of FC clusters in the vicinity of major metropolitan statistical areas (MSAs), and $\mathcal{F}(k)$ be the set of FCs in cluster $k \in \mathcal{K}$. Let $\widetilde{FR}$ be the set of FC-region pairs (i,r) that are not allowed. Although this set can be provided as a manual override, by default we define it based on a maximum distance threshold: $\widetilde{FR} = \{(i,r) \mid D_{ir} > \text{maxDistance}\}$, where maxDistance is a predefined parameter.

Let the decision variable x_{ir} be defined as $x_{ir} = 1$ *if* $(1-\alpha_i) \times 100\%$ of shipment and storage capacity of FC_i is assigned to region r and 0, otherwise.

The optimization problem is formulated as follows:

$$\textit{minimize } TransCost = \sum_{i \in \mathcal{F}, r \in \mathcal{R}} (1-\alpha_i) C_i^{ship} - D_{ir} x_{ir}$$

subject to

$$\sum_{r \in \mathcal{R}} x_{ir} = 1 \qquad \forall i \in \mathcal{F} \tag{A.1}$$

$$\sum_{i \in \mathcal{F}} (1-\alpha_i) C_i^{ship} x_{ir} + \sum_{i \in \mathcal{F} \mid R(i)=r} \alpha_i C_i^{ship} \geq D_r (1-\tau) \quad \forall r \in \mathcal{R} \tag{A.2}$$

$$\sum_{i \in \mathcal{F}} (1-\alpha_i) C_i^{storage} x_{ir} + \sum_{i \in \mathcal{F} \mid R(i)=r} \alpha_i C_i^{storage} \geq C_r (1-\omega) \quad \forall r \in \mathcal{R} \tag{A.3}$$

$$\sum_{i \in \mathcal{F}(k)} x_{i,R(i)} \geq 1 \quad \forall k \in \mathcal{K} : |\mathcal{F}(k)| > 1 \tag{A.4}$$

$$x_{ir} = 0 \quad \forall (i,r) \in \widetilde{FR} \tag{A.5}$$

$$x_{ir} \in \{0,1\} \quad \forall i \in \text{v}\mathcal{F}, r \in \mathcal{R}\,. \tag{A.6}$$

Constraint (A.1) ensures that each FC is assigned to exactly one region. Constraint (A.2) guarantees that sufficient shipping capacity is allocated to each region with a tolerance level $\tau \in [-1,1]$. Note that a negative tolerance implies a need for capacity surplus in a region. This constraint considers both the assignable capacity and the capacity reserved for the FC's physical region. Constraint (A.3), similar to (A.2), stipulates that for each region, the total storage capacity, including both assigned and reserved capacity, must exceed the specified storage target C_r with a tolerance level $\omega \in [-1,1]$. Different tolerances (for τ and ω) in different regions allow us to proactively plan for a network with asymmetries, where, for instance, we may choose to operate one region with higher levels of inventory than others to provide additional buffer for surplus inventory.

Constraint (A.4) mandates that at least one FC from each multi-FC cluster serves its local region, generally a major MSA, promoting local fulfillment speed. Finally, Constraint (A.5) prohibits FC-region assignments that either are manually excluded or exceed a predefined distance threshold, as defined by the set $\widetilde{FR}$ as the forbidden FC-region pairs.

This problem can be solved quickly because of its relatively small size as a mixed-integer programming (MIP) problem. Often, the mappings remain unchanged for certain ranges of τ and ω. In the analysis phase, we test several scenarios with different tolerance parameters to gain an understanding of different viable solutions and their structures. However, when we are in the final decision-making stage with leadership, we only cache and report scenarios that differ, which streamlines the discussion and frequently results in achieving alignment more quickly with fewer scenarios.

Appendix B. The Order Assignment Model

We next describe the mathematical formulation of the core OAM, excluding ancillary considerations that are beyond the scope of regionalization. The model is instantiated for each individual customer order to determine how it is fulfilled. For each customer order, we precompute a set of candidate shipments (i.e., ways to fulfill an order fully or partially, using a heuristic procedure). We employ a heuristic because the number of viable candidate shipments grows exponentially as the number of items in the order increases. Although we can exhaustively enumerate all candidate shipments for small orders, we leverage local search heuristics to select a heterogeneous subset of candidate shipments for larger orders. The OAM selects a subset of the candidate shipments to fulfill the demand while optimizing a cascade of linear objectives.

Set and Indices

- S: the set of candidate shipments
- O: the set of items in the customer order
- F: the set of FCs that have some inventory for any item in the customer order
- K: the set of objectives
- $s \in S$: the index for candidate shipment
- $i \in O$: the index for ordered items
- $k \in K$: the index for the objectives

Parameters

- f_s: FC associated with the shipment $s \in S$

- q_i: ordered quantity of item $i \in O$
- p_{si}: quantity fulfilled by candidate shipment s for item i
- I_{fi}: inventory available at FC f for item i
- v_s^k: objective score associated with shipment s for objective/goal k

Decision Variables

- x_s: binary variable indicating whether shipment s is used to fulfill the demand.

In its original form, OAM employs a goal-programming approach to solve for the cascade of objectives, for example, minimization of delivery time window violation, inventory readiness (e.g., preferring items on shelf, then in reserve), and minimization of total fulfillment costs, which are assumed to be additive across the chosen candidate shipments. The score v_s^k for objective k is computed when candidate shipment k is generated. For example, the objective score for delivery window violation is calculated as the number of days the shipment arrives later than the promised delivery date; the cost objective score is the sum of the fulfillment cost including, for example, FC operation cost and shipping cost. When OAM solves for the k-th objective, it minimizes the total objective score for this objective subject to demand satisfaction constraints, inventory availability constraints, and the constraints generated from the previous $k-1$ objective solves.

$$
\begin{aligned}
V^k := \min\ & \sum_s v_s^k x_s \\
\text{s.t.}\ & \sum_s p_{si} x_s = q_i, \quad \forall i \in O \\
& \qquad \text{(demand satisfaction constraint)} \\
& \sum_s 1_{f_s=f} p_{si} x_s \le I_{fi}, \quad \forall i \in O, f \in F \\
& \qquad \text{(inventory availability constraint)} \\
& \sum_s v_s^j x_s \le V^j, \quad \forall j = 1, 2, \ldots, k-1 \\
& \qquad \text{(constraints from previous } k-1 \text{ solves)} \\
& x_s \in \{0, 1\}, \quad \forall s \in S
\end{aligned}
$$

That is, V_k is the optimal objective function value when solving for the k-th objective, with previous $k-1$ objectives bounded by the corresponding optimal values when they were solved before the current k-th solve. The final fulfillment decision for a customer order is generated after OAM solves the entire cascade of $|K|$ problems and corresponds to the candidate shipments. The last objective in the above $|K|$ problems is to minimize the total fulfillment costs. As we mention above, the total fulfillment costs have a "real" and an "artificial" portion. The cost relaxation mechanism allows us to optimize additional objectives with a controlled relaxation of the artificial cost portion. For example, if we want to improve the in-region fulfillment quantity, we can define the "out-of-region" objective score (v_s^{oor}) for each shipment; that is, if the shipment is an "out-of-region" fulfillment, it has an objective score of one, otherwise zero. Then OAM solves the following problem to seek additional opportunities to improve in-region fulfillment.:

$$
\begin{aligned}
V^k := \min\ & \sum_s v_s^{\text{oor}} x_s \\
\text{s.t.}\ & \sum_s p_{si} x_s = q_i, \quad \forall i \in O \\
& \sum_s 1_{f_s=f} p_{si} x_s \le I_{fi}, \quad \forall i \in O, f \in F \\
& \sum_s v_s^j x_s \le V^k, \quad \forall k = 1, 2, \ldots, |K|-1 \\
& \sum_s v_s^{\text{real}} x_s \le V^{\text{real}} \\
& \sum_s v_s^{\text{artificial}} x_s \le \hat{V}^{\text{artificial}} \\
& x_s \in \{0, 1\}, \quad \forall s \in S.
\end{aligned}
$$

In the above formulation, V^{real} represents the total real cost calculated after the $|K|$-th solve for the total fulfillment cost, and $\hat{V}^{\text{artificial}}$ corresponds to the relaxation of the artificial cost portion derived from the same solve. This relaxation can be applied in different modes, such as a percentage-based adjustment or a fixed absolute amount adjustment.

We established the objective order and thresholds of CR through large-scale simulations that can assess the holistic fulfillment impact across a spectrum of configurations. In addition, we had to model various aspects, such as order consolidation, synchronization of deliveries, inventory depletion, load balancing across warehouses, and reoptimization of demand, as part of Amazon's order assignment systems in addition to the changes discussed to enable regionalization.

References

Acimovic J, Farias VF (2019) The fulfillment-optimization problem. Netessine S, ed. *Operations Research and Management Science in the Age of Analytics*, Tutorials in Operations Research (INFORMS, Catonsville, MD), 218–237.

Acimovic J, Graves SC (2015) Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing Service Oper. Management* 17(1):34–51.

Afèche P, Caldentey R, Gupta V (2022) On the optimal design of a bipartite matching queueing system. *Oper. Res.* 70(1):363–401.

Allgor R, Cezik T, Chen D (2023) Algorithm for robotic picking in Amazon fulfillment centers enables humans and robots to work together effectively. *INFORMS J. Appl. Analytics* 53(4):266–282.

Amazon (2023) Amazon sustainability report. Accessed November 1, 2024, https://sustainability.aboutamazon.com/2023-amazon-sustainability-report.pdf.

Amazon (2025) Amazon Prime members enjoyed fastest-ever delivery speeds in 2024, and US members saved on average over $500 on delivery fees. Accessed March 1, 2025, https://www.aboutamazon.com/news/retail/prime-members-us-savings-fastest-delivery-2024.

Amazon Science (2022) Maximizing the efficiency of Amazon's own delivery networks. Accessed November 1, 2024, https://www.amazon.science/blog/maximizing-the-efficiency-of-amazons-own-delivery-networks.

Besbes O, Castro F, Lobel I (2022) Spatial capacity planning. *Oper. Res.* 70(2):1271–1291.

Carlsson JG, Peng X, Ryzhov IO (2024) Demand equilibria in spatial service systems. *Manufacturing Service Oper. Management* 26(6):2305–2321.

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

Herrington D (2023) Amazon is delivering its largest selection of products to U.S. Prime members at the fastest speeds ever. Accessed November 1, 2024, https://www.aboutamazon.com/news/operations/doug-herrington-amazon-prime-delivery-speed.

Herrington D (2024) Amazon delivered to Prime members at the fastest speeds ever in 2023—And is working to get even faster in 2024. Accessed November 1, 2024, https://www.aboutamazon.com/news/operations/doug-herrington-amazon-prime-delivery-speed-2024-updates.

Jasin S, Sinha A, Uichanco J (2019) Omnichannel operations: Challenges, opportunities, and models. Gallino S, Moreno A, eds. *Operations in an Omnichannel World* (Springer, New York), 15–34.

Jassy A (2024) CEO Andy Jassy's 2023 Letter to Shareholders. Accessed November 1, 2024, https://www.aboutamazon.com/news/company-news/amazon-ceo-andy-jassy-2023-letter-to-shareholders.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Sci.* 41(4):577–594.

Kassmann D, Allgor R (2006) Supply chain design, management and optimization. *Comput. Aided Chemical Engrg.* 21:101–106.

Lara CL, Koenemann J, Nie Y, de Souza CC (2023) Scalable timing-aware network design via Lagrangian decomposition. *Eur. J. Oper. Res.* 309(1):152–169.

Madan U (2024) Amazon is delivering at its fastest speeds ever for Prime members in the U.S. and globally. Accessed November 1, 2024, https://www.aboutamazon.com/news/retail/amazon-prime-same-day-delivery-speed-2024.

Mulvey JM, Beck MP (1984) Solving capacitated clustering problems. *Eur. J. Oper. Res.* 18(3):339–348.

O'Neill S (2023) Sizing down to scale up: How Amazon reworked its fulfillment network to meet customer demand. Accessed November 1, 2024, https://www.amazon.science/news-and-features/how-amazon-reworked-its-fulfillment-network-to-meet-customer-demand.

Xu PJ (2005) Order fulfillment in online retailing: What goes where. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Xu PJ, Allgor R, Graves SC (2009) Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing Service Oper. Management* 11(2):340–355.

Amitabh Sinha is a director of research science at Amazon's Modeling and Optimization group, leading scientists developing algorithmic solutions for supply chain optimization. Previously, he was an associate professor at University of Michigan's Ross School of Business and co-led the Tauber Institute. He holds a PhD in algorithms, combinatorics and optimization from Carnegie Mellon and an MS in mathematics and computer applications from IIT Delhi.

Jeremy Agte is the director of Amazon's Modeling and Optimization organization, working on operation research problems from last-mile planning to inbound network design, and helped lead Amazon's 2023 outbound regionalization effort. Before joining Amazon in 2017, he served 20 years in the U.S. Air Force as an Experimental Flight Test Engineer. He specializes in complex systems design and optimization, holding an MS from George Washington University and PhD from MIT.

Russell Allgor (NAE member and INFORMS Fellow) is chief supply chain scientist at Auger.com, integrating optimization, simulation, and AI. Previously chief scientist at Amazon for 24+ years, he led mathematical modeling experts and is among the most influential scientists in e-commerce logistics. His team used data analysis, modeling, and optimization to improve Amazon operations across several domains. He holds a PhD in chemical engineering from MIT and BS from Princeton.

Cristiana L. Lara is a principal research scientist at Amazon in the Modeling and Optimization team. Since joining, she has led the development of models and algorithms for network design, line haul scheduling, and inventory placement. She now leads a team focused on delivering speed to customers through inventory placement and scheduling optimization. She holds a PhD in process systems engineering from Carnegie Mellon University and received the INFORMS Early Career Practitioner Award in 2024.

Ashish Agiwal is vice president of Amazon Fulfillment Technologies (AFT). Over the last 15 years at Amazon, he has led supply chain engineering, technology, and science across inventory sourcing, inventory planning, delivery promise and fulfillment optimization, network design, and flow planning. As part of AFT, he currently manages Amazon's Warehouse Management Systems that enable safe and optimal building flows by orchestrating all processes, labor, machines, and robots.

Semih Atakan is a senior applied scientist at Amazon with 7 years of experience in supply chain design. He built scalable science models to optimize large-scale network decisions and contributed to several major Amazon initiatives impacting fulfillment, profitability, and customer experience. He holds a PhD in industrial & systems engineering from the University of Southern California, specializing in mathematical modeling and stochastic optimization.

Lourdes Campo is a finance director at Amazon, responsible for North America Operations Financial Planning & Analysis (FP&A). Throughout the last 10 years, she has held various operations finance and strategy roles at Amazon. Prior to joining Amazon in 2016, she worked at PricewaterhouseCoopers as an auditor. She holds a bachelor of science and a master's degree in accounting from the University of Florida and is an active Florida certified public accountant.

Tolga Cezik is a senior principal research scientist in Amazon's Modeling and Optimization group, where he focuses on process design and algorithmic innovation for fulfillment centers, transportation networks, and inventory systems. His scheduling algorithm for Amazon's robotic fulfillment centers was an INFORMS Edelman Competition finalist. He holds a BS in industrial engineering from Middle East Technical University and PhD in operations research from Columbia University.

Daniel Chen is a senior research scientist in the Modeling and Optimization group at Amazon. His research has influenced strategic thinking in network design and contributed optimization tools for effective operations within fulfillment centers. Separately, he has also worked on projects in the energy and transportation sectors with the Agency of Science, Technology and Research in Singapore. He holds a PhD in operations research from MIT and a BA in mathematics from the University of Cambridge.

Qi Chen is a senior applied scientist at Amazon, deploying tactical planning solutions across the supply chain, from replenishment to fulfillment. His work spans analytics, modeling, and graphics processing unit–accelerated optimization. He brings a process systems perspective from chemical engineering, with a PhD from Carnegie Mellon University, an MSc from Imperial College London, and a BS from the University of Kansas.

Jesse Fischer is a principal engineer in the Fulfillment Optimization organization within Supply Chain Optimization Technologies at Amazon. Since joining Amazon in 2008, he has worked across Amazon's supply chain technology stack, including fulfillment optimization, execution systems, and promise systems. He holds a BS in computer engineering from the University of Washington.

Eitan Gor is a director and product leader for warehouse design and technology at Amazon's Transportation Tech & Services (TTS) organization. He led the vision for system-driven facility execution that enables granular shipment optimization. He led a Warehouse

Management System (WMS) team that created software enabling sorting, segmenting, and processing shipments at a precise capacity for Amazon's 2023 outbound regionalization network design. He holds a BA in economics and business from Bar-Ilan University, and an MBA from the Massachusetts Institute of Technology.

Nader Kabbani is an operations and optimization leader with extensive experience designing large-scale systems. At Amazon he served 18 years, leading regionalization as the single-threaded leader and driving major innovations across Logistics, Amazon Flex, Kindle, Kindle Direct Publishing, and Amazon Pharmacy. He later held leadership roles at Flexport, Symbotic, Hims & Hers; and earlier roles at McKinsey, Trilogy, CALEB, and American Airlines.

Ryan Kennedy is a technical product manager on the Modeling and Optimization team at Amazon where he designs, builds, and orchestrates segmented optimization systems to solve network design and supply chain problems. Since joining Amazon in 2017, he has led multi-disciplinary teams supporting Amazon's inbound and outbound networks. He holds BS and MS degrees in industrial and operations engineering from the University of Michigan.

Kaushik Krishnan is a senior data scientist at Amazon, developing large-scale optimization and machine learning solutions for connectivity and order-fulfillment in the outbound network. Prior to Amazon, he was an operations research scientist at Optym where he worked on optimization models for transportation. He holds an MS in industrial engineering from the University of Illinois at Urbana-Champaign and a BE in industrial engineering from the College of Engineering, Guindy.

Yuan Li is a senior research scientist in Amazon's Modeling and Optimization group, focusing on strategic network design and algorithm development for Amazon's outbound operations. Since 2017, he has led research on customer pickup topology, last-mile planning in demand-sparse regions, and integrated middle-mile modeling, while also contributing to major inbound and outbound regionalization initiatives. He holds a PhD in operations research from Georgia Tech.

Shahbaaz Mubeen Mamadapur is a principal product manager who works with scientists, engineers, and business leaders to operationalize and scale operation research models across Amazon's global supply chain. He specializes in network topology design, science-driven product strategy, and diagnosing complex network issues to resolve systemic problems. He has led key initiatives such as upstream package containerization and delivery-speed optimization. He holds a BS in computer science and an MBA.

Tanmay Mathur is vice president of supply chain at Blue Origin. He holds degrees from IIT Roorkee and Harvard University. His expertise spans network design, optimization algorithms, and predictive modeling. In past, he led Amazon Transportation planning and drove manufacturing optimization at Shell. He holds two patents and advances scalable, reliable logistics through data-driven methods. His work focuses on capacity management and network de-bottlenecking to improve operational efficiency.

Nick McCabe is director of supply chain and network design at Amazon with 20+ years of experience in supply chain, transportation, and logistics engineering. He led the 2023 outbound regionalization effort that restructured the fulfillment network into regional zones, bridging scientific network design with operational execution and ensuring planning systems evolved to support the new architecture at scale. He holds a BS in industrial engineering from Purdue University and an MBA from the University of Memphis.

David Mildebrath is a senior applied scientist in the Modeling and Optimization team at Amazon. Since joining Amazon in 2021, he has developed optimization models and other tools for network design and supply chain management. He holds a PhD in computational and applied mathematics from Rice University.

Eric Powell is the head of internal operations at Cape. Prior to that, he held various operations and supply chain roles at Amazon. He regularly moved between field execution and corporate planning roles, and developed a particular skillset in implementing new initiatives. In his role overseeing network design and long-term planning, he worked on the redesign of Amazon's outbound transportation network. Prior to Amazon, He served overseas and at sea with the U.S. Navy. He has a law degree and an economics degree from Harvard, and serves his community as a volunteer emergency medical technician.

Andrea Qualizza is a senior principal scientist in Supply Chain Optimization Technologies at Amazon. Since 2011, he has led optimization initiatives across promise, order assignment, network topology, capacity planning, and inventory placement. His work drives major impact on Amazon's global supply chain, including a framework that redefined outbound operations planning and execution, saving billions in costs. He holds a PhD in operations research from Carnegie Mellon, specializing in combinatorial optimization.

Denton Schroeder is a senior manager of supply chain management at Amazon Shipping with a decade of experience driving innovation in Amazon's outbound supply chain. His expertise includes operations, engineering, network design, customer experience, and external logistics. He led the planning and execution of Amazon's regional network and holds degrees in industrial engineering (New Jersey Institute of Technology), engineering management (University of Parma), and industrial organization (University of Extremadura).

Nityansh Seth is a director of analytics and science at Blue Origin, where he works on procurement planning and transportation for the supply chain. Previously, he was an analytics manager at Amazon, where he was responsible for building heuristics tools to design connectivity between regional designs and developing all associated analytics. Nityansh also developed forecasting tools for Amazon's middle mile national network volume. He holds a master's degree in analytics from Georgia Tech.

Xiaoyan Si is a senior research science manager at Amazon's Modeling and Optimization team, building models and tools for supply chain and fulfillment network design at scale. She has solved high-impact strategic network design problems from inventory inbound to order fulfillment. Before Amazon, she worked in the railroad industry for 10 years on optimization, data analytics, and computer vision. She received her PhD in operations research and industrial engineering from the University of Texas Austin.

Kaushik Sinha is a principal research scientist at Amazon, specializing in network structure design and optimization. Since joining Amazon in 2018, he has developed analytical models that drive fulfillment, profitability, and customer experience through network re-design initiatives. He holds a PhD in engineering systems from MIT, with expertise in systems engineering, design optimization, and technology valuation across multiple industries.

Darren Stegner is a senior principal engineer at Amazon with 20+ years of experience optimizing outbound fulfillment networks as part of Supply Chain Optimization Technologies. He architected and led the creation of Amazon's real-time delivery date calculation and discovery services that have set hundreds of trillions customer delivery expectations. He holds a BS in computer science from Seattle Pacific University and provides ongoing strategic and technical guidance on network regionalization.

Jun Xiao is a senior applied scientist in the Modeling and Optimization group at Amazon. Since joining in 2020, he has focused on strategic network design and capacity optimization across Amazon's end-to-end logistics network. His work spans inbound operations, middle-mile network connectivity, and last-mile delivery. He holds a PhD in transportation engineering from Arizona State University.

Ling Zhang is a senior manager of product management technical in supply chain optimization technologies at Amazon. Since 2017, she has led retail and supply chain optimization initiatives across Retail Category, Last Mile Delivery, and Fulfillment. She now leads strategic efforts in fulfillment execution, capacity

planning, and capability optimization, reducing costs, accelerating delivery, and improving customer experience. She holds an MBA from Brigham Young University.

Shanshan Zhang is a principal applied scientist in the Modeling and Optimization group at Amazon. Since joining in 2015, she has built large-scale science products and developed core models and analyses that inform strategic decisions across capacity planning, network design, and under-the-roof process optimization. She holds a PhD in operations research and information engineering from Cornell University, where she focused on structured non-smooth optimization.

Jikai Zou was formerly a principal applied scientist at Amazon's Supply Chain Optimization Technology organization for 4.5 years, driving network optimization, coordination between network planning and fulfillment operations, and improving Pareto efficiency in order assignment. He holds a PhD in operations research from Georgia Institute of Technology and is currently a senior staff scientist in global operations technology at Coupang.