Regionalize and Scale

Network design for faster and cheaper delivery

amazon

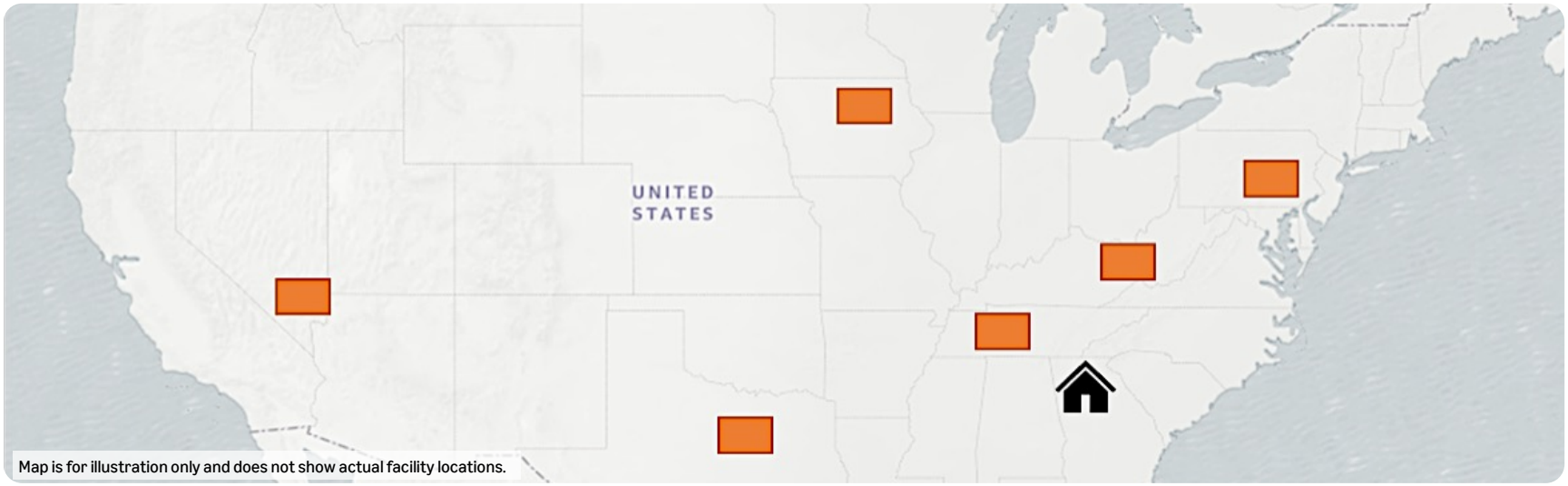
amazon

The Team

Sinha, Amitabh	Director, Research Science
Agte, Jeremy	Director, Research Science
Allgor, Russell	Formerly VP and Chief Scientist
Lara, Cristiana	Principal Research Scientist
Agiwal, Ashish	VP Technology
Atakan, Semih	Sr. Applied Scientist
Campo, Lourdes	Finance Director
Cezik, Tolga	Sr. Principal Scientist
Chen, Daniel	Sr. Research Scientist
Chen, Qi	Sr. Applied Scientist
Fischer, Jesse	Principal SDE
Gor, Eitan	Director, PMT
Kabbani, Nader	Formerly VP WW Operations
Kennedy, Ryan	Sr. Product Manager-Tech
Krishnan, Kaushik	Sr. Data Scientist
Li, Yuan	Sr. Research Scientist
Mathur, Tanmay	Formerly Director, SCM
McCabe, Nick	Director, Supply Chain Mgmt
Mildebrath, David	Sr. Applied Scientist
Mubeen, Shahbaaz	Principal PMT
Powell, Eric	Formerly Director, SCM
Qualizza, Andrea	Sr. Principal Scientist
Seth, Nityansh	Formerly Sr. Manager, PMT
Schroeder, Denton	Sr. Manager, SCM
Si, Xiaoyan	Principal Research Scientist
Sinha, Kaushik	Principal Research Scientist
Stegner, Darren	Sr. Principal Engineer
Xiao, Jun	Sr. Applied Scientist
Zhang, Ling	Sr. Manager, PMT
Zhang, Shanshan	Principal Applied Scientist
Zou, Jikai	Principal Applied Scientist

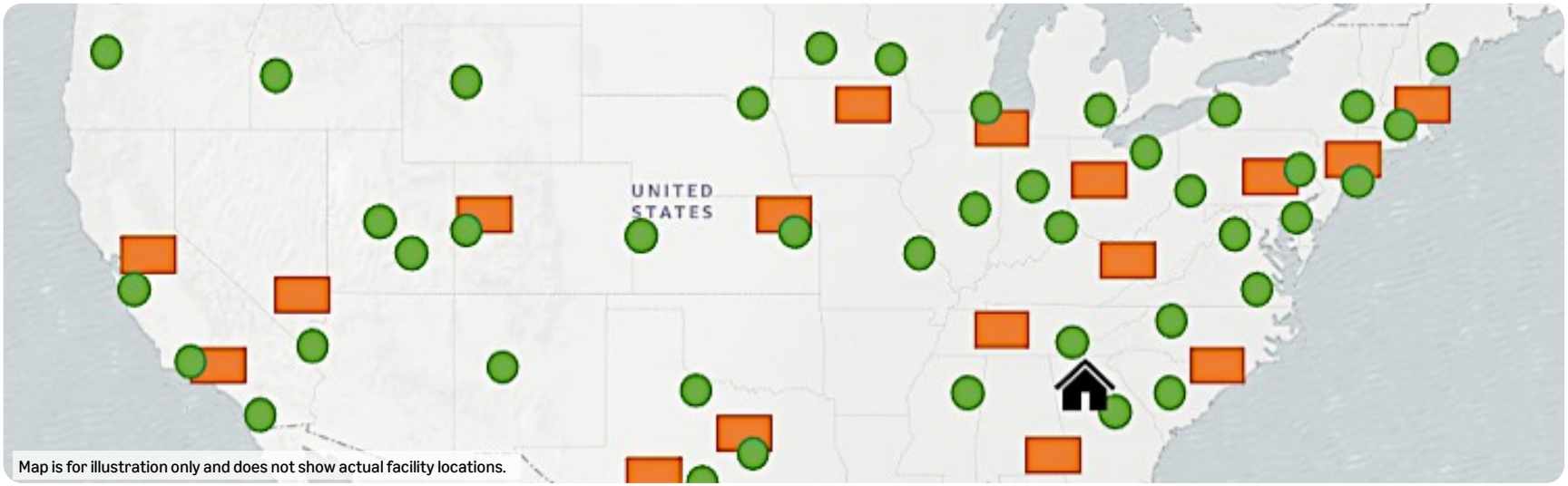
amazon

The problem



Amazon 2005

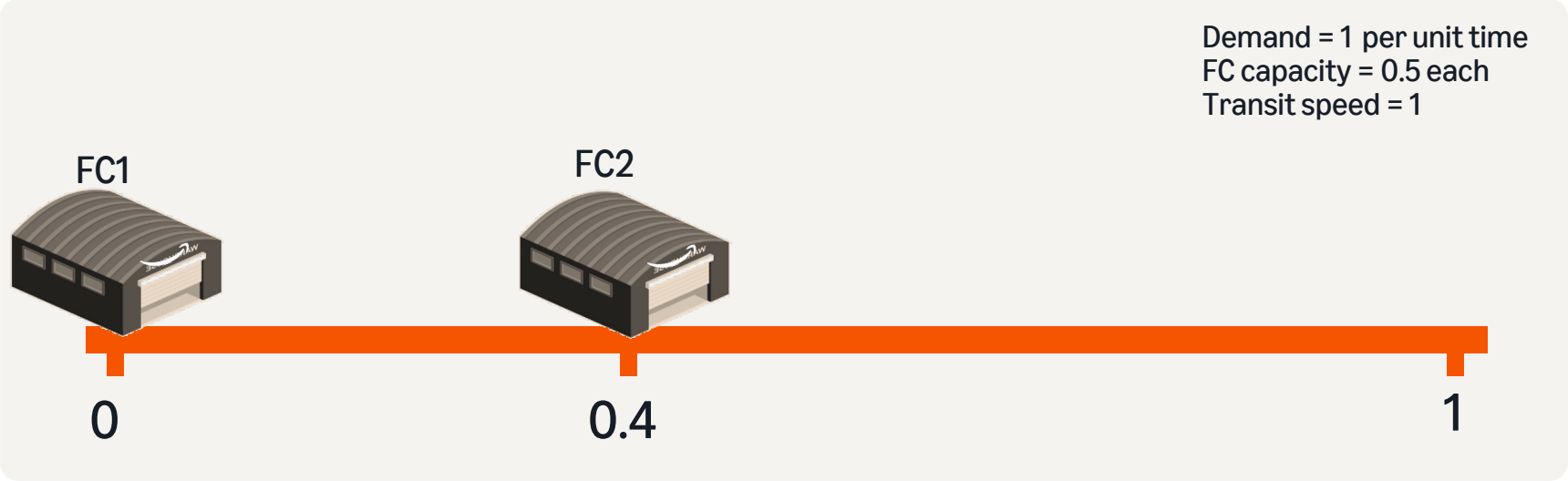
Approximately **6 FCs**, pure ecommerce operation, delivery only by 3P logistics companies.
Point-to-point makes sense.



Amazon 2021

Full-spectrum retail, **100+ FCs** with delivery using internal transportation network and 3Ps. Complexity leads to **diseconomies of scale**.

Core principles

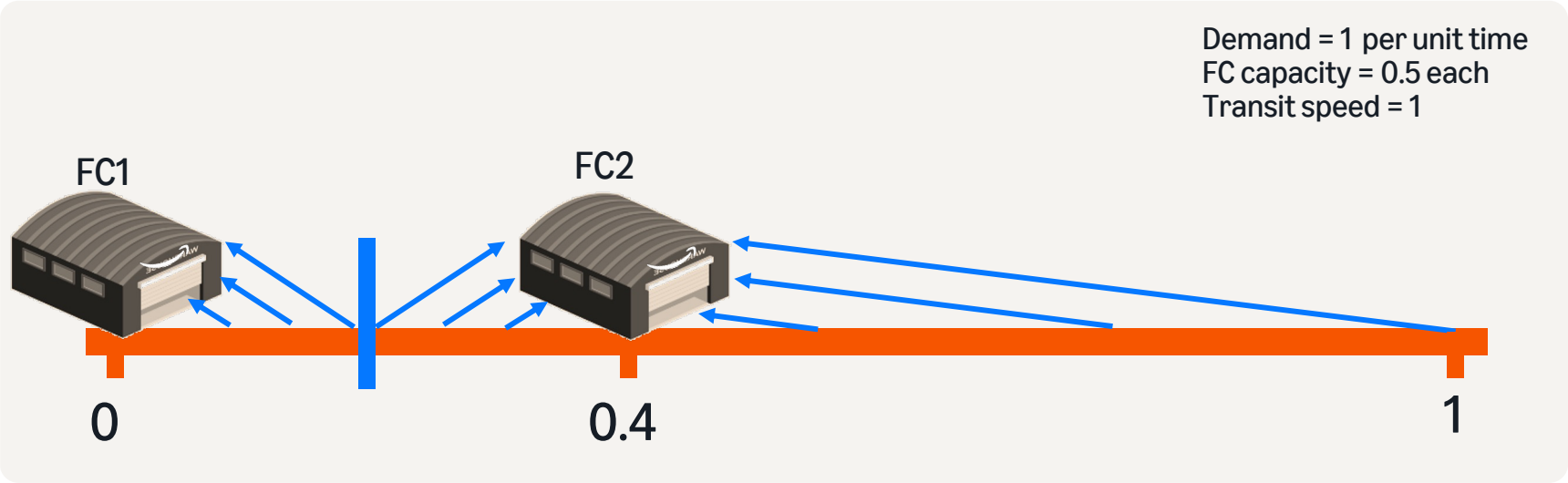


Matching capacity w/ demand

Order assignment policy: assign each order to FC which minimizes time to delivery (transit plus waiting)

amazon

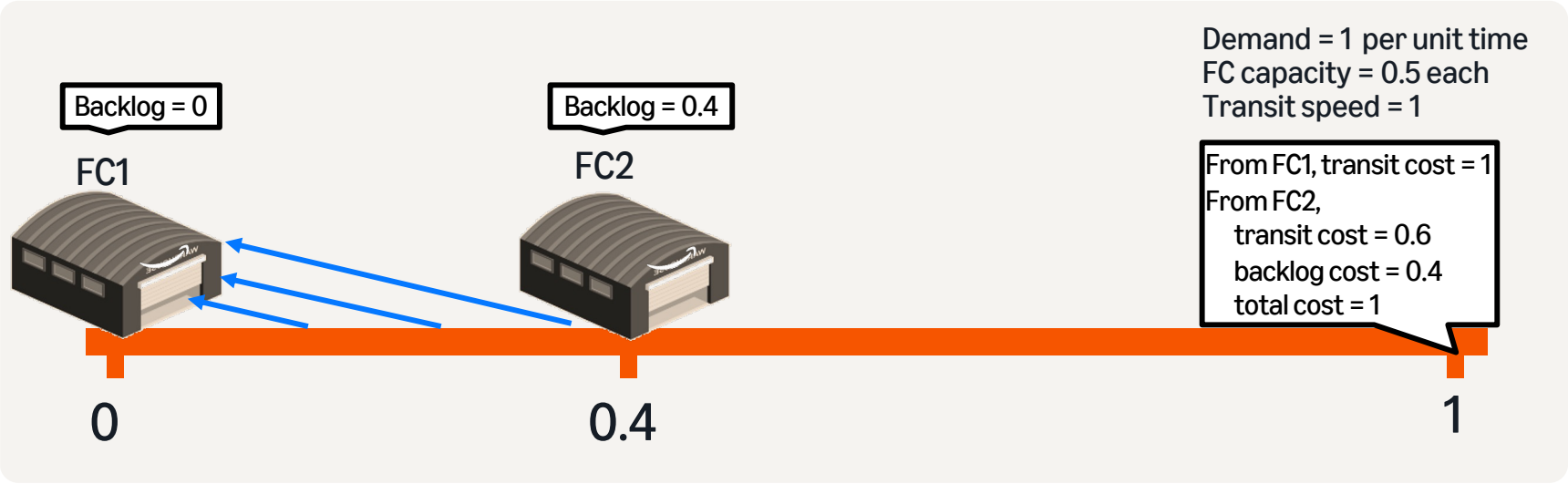
t = 0



Matching capacity w/ demand

FC1 load: 0.2
FC2 load: 0.8. Backlog builds up.

Equilibrium

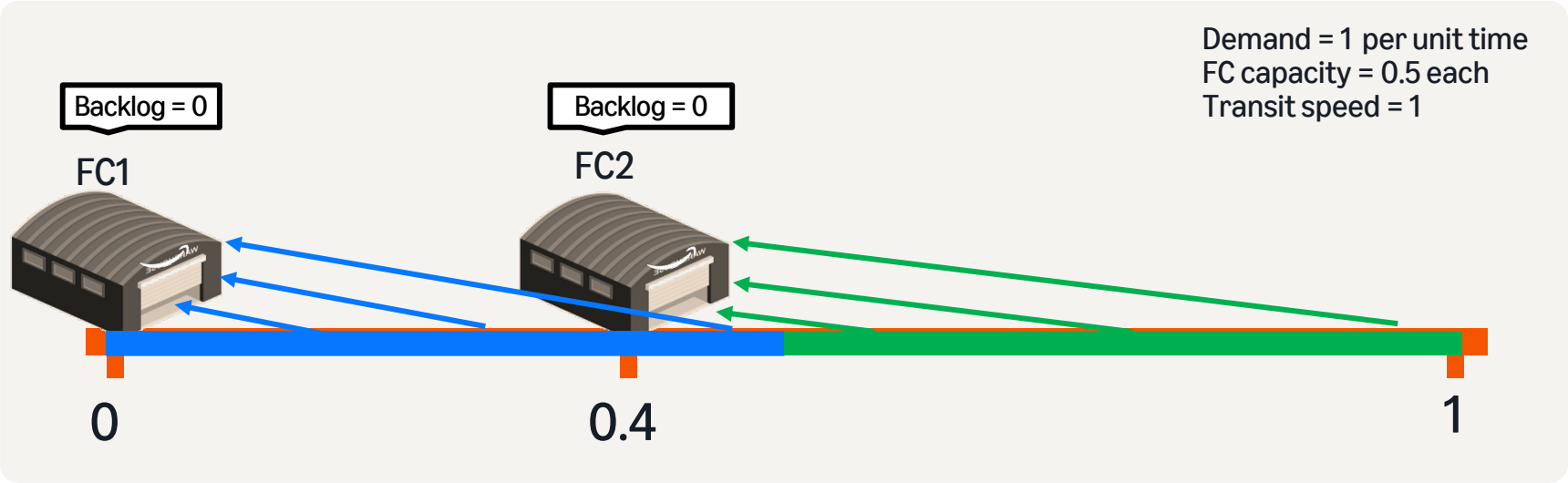


Matching capacity w/ demand

Customers in (0.4, 1] split 1:5 to FC1 vs FC2.

Average service time is 0.5 (topology lower bound is 0.22).

Regionalization

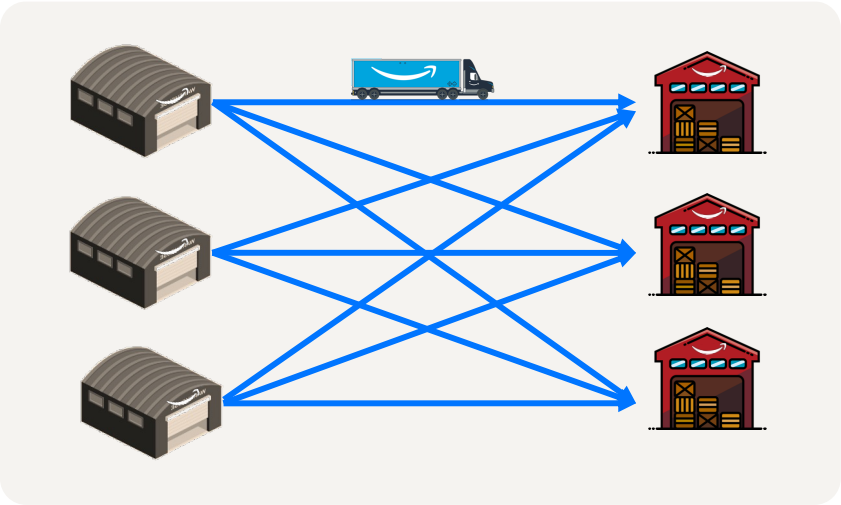


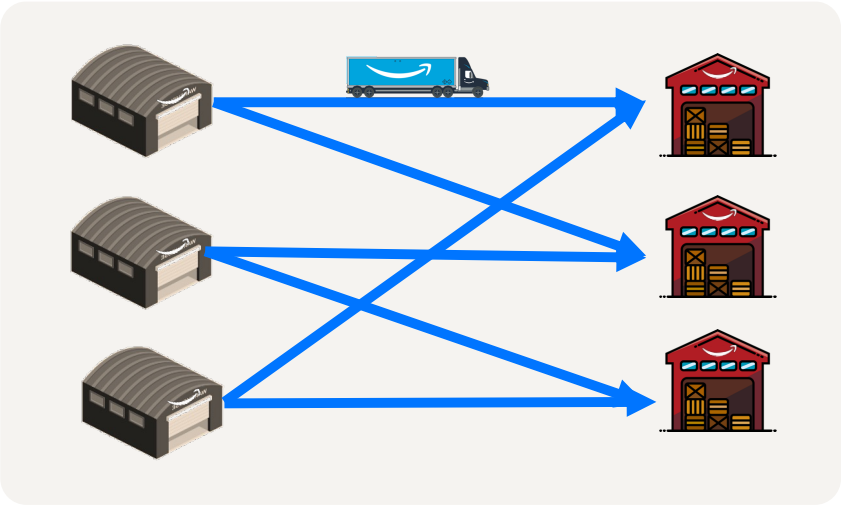
Matching capacity w/ demand

Average service time is 0.3 (compared to 0.22 topology lower bound, and 0.5 greedy).

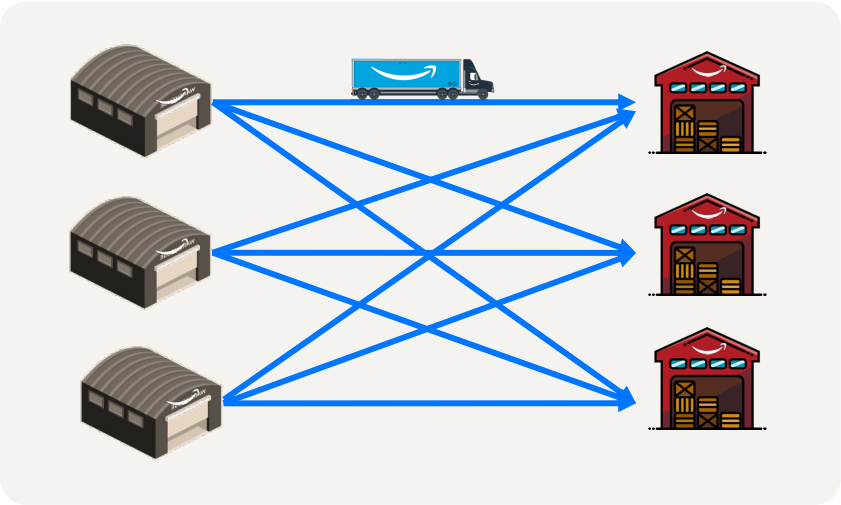
Several additional operational benefits.

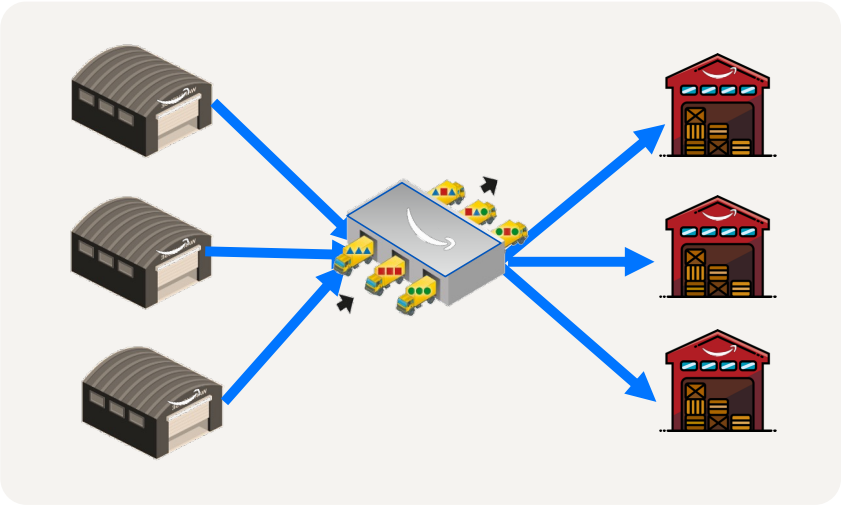
Better asset utilization through **flow concentration**



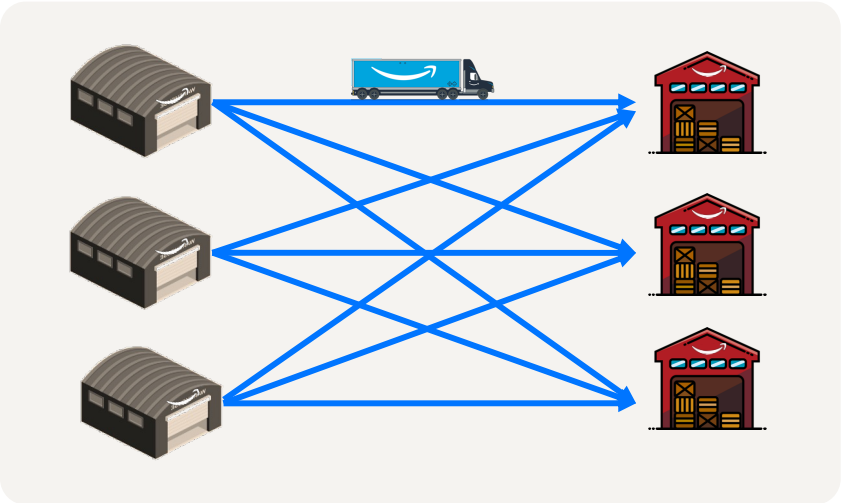


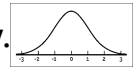
Better asset utilization through **consolidation**

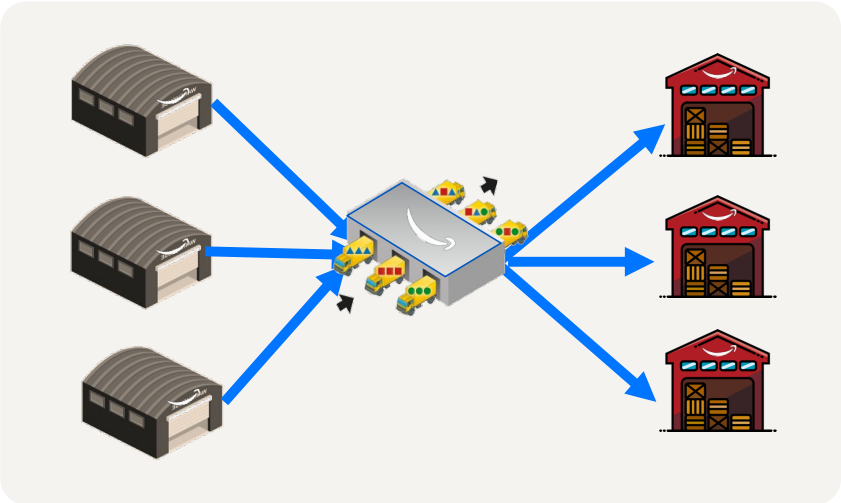




Better asset utilization through **variability pooling**



OD flow day-to-day variability. 



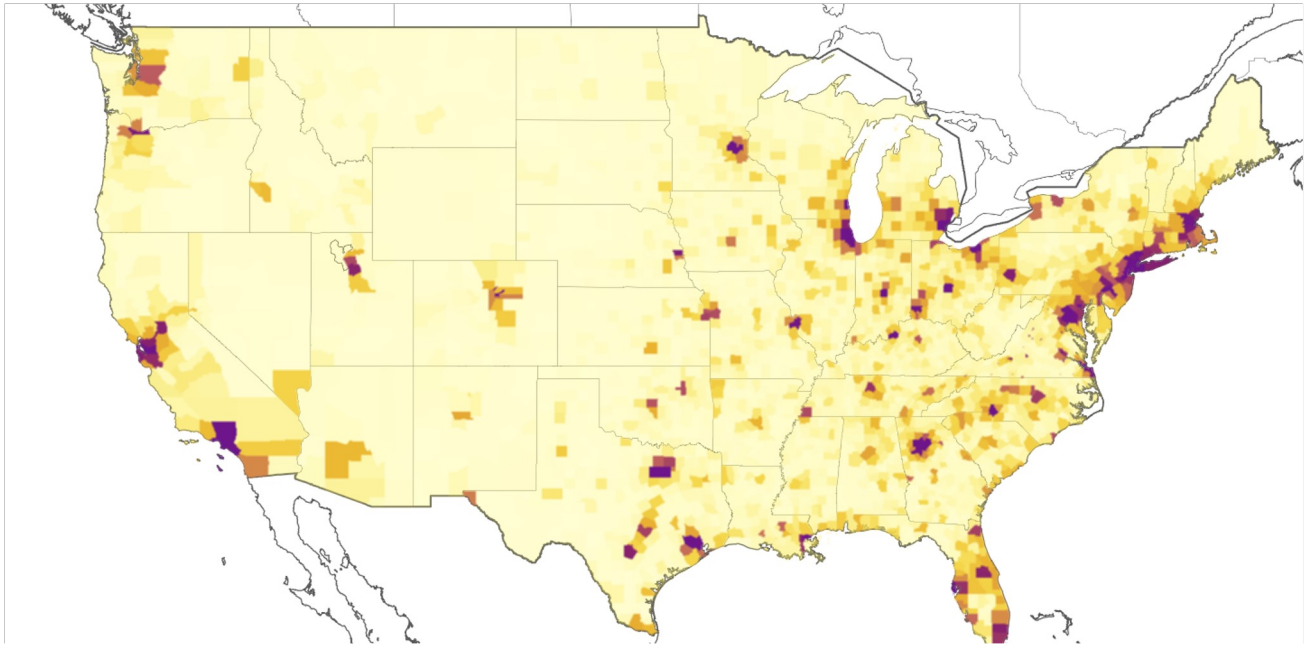
Better variability pooling. Less *ad hoc* and cancellations.

amazon

Region Definition

amazon

Partition demand



Population density in US 2020 Census for illustration only.

Partition demand

Fewer larger regions

(+) More FCs lead to inventory pooling
(-) Longer FC to customer distances
(-) Longer delivery times

vs.

Many smaller regions

(-) Fewer FCs limit inventory breadth
(+) Shorter FC to customer distances
(+) Expedited delivery times

Population density in US 2020 Census for illustration only.

amazon

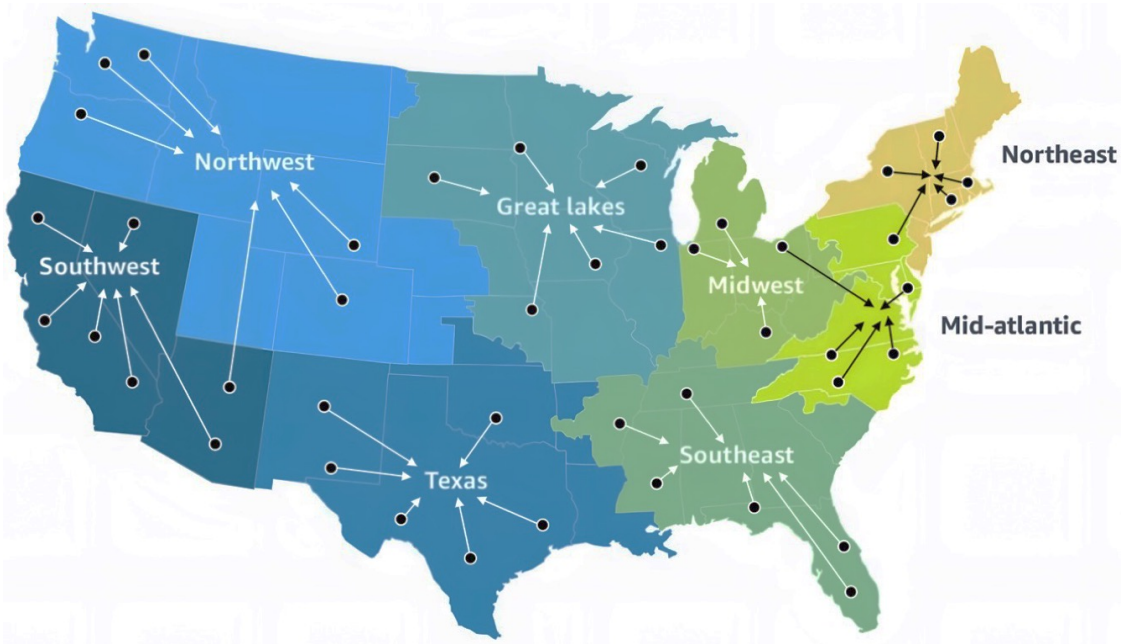
Demand

Capacity

Match Capacity with Demand

FC Mapping

Maps are for illustration only and do not show actual facility locations.



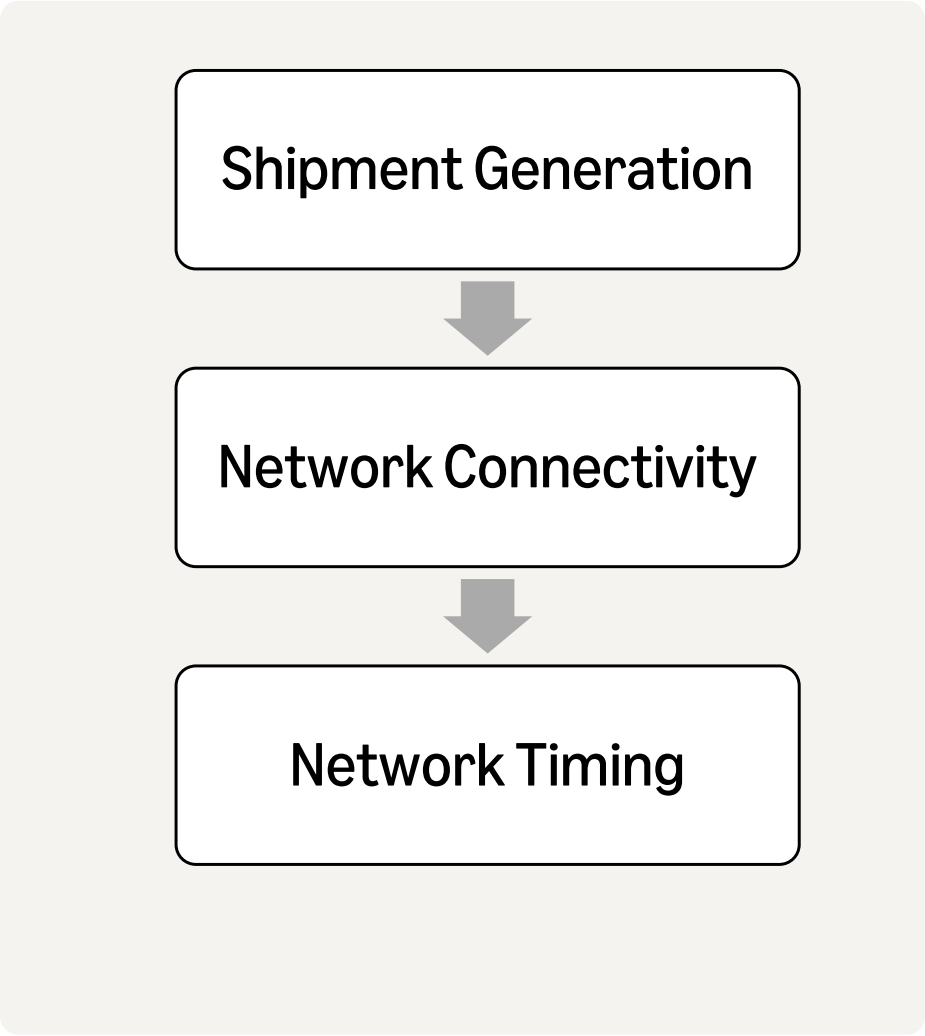
8 Region Design

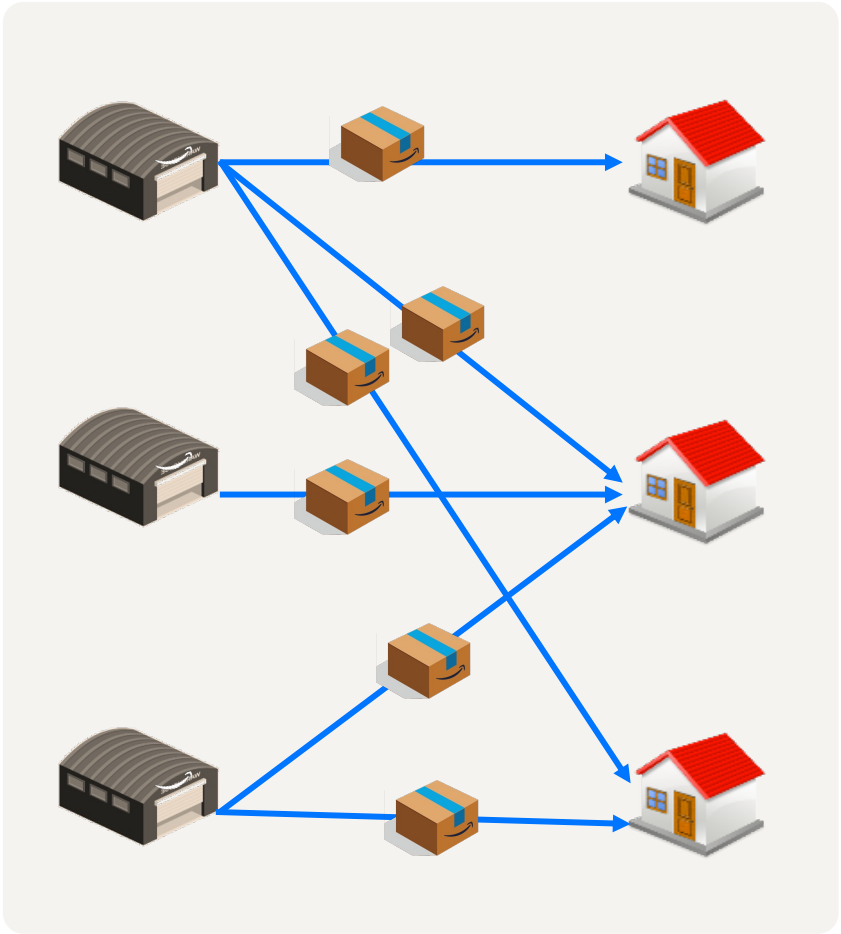
Maps are for illustration only and do not show actual facility locations.

amazon

Evaluation

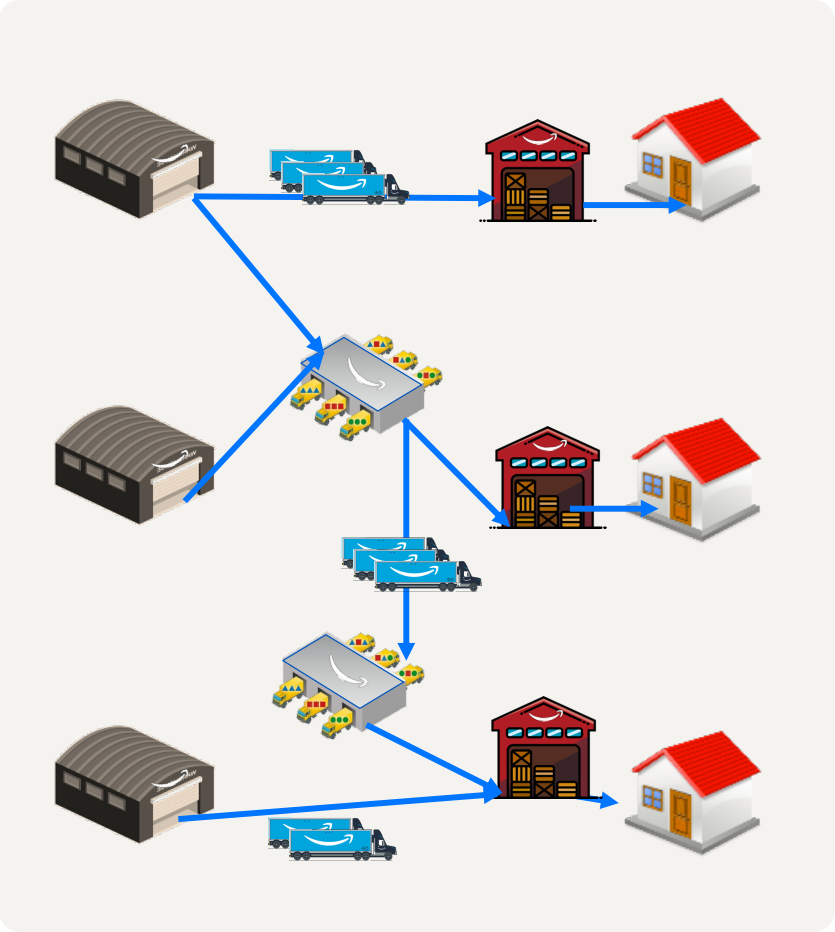
Network Design Suite





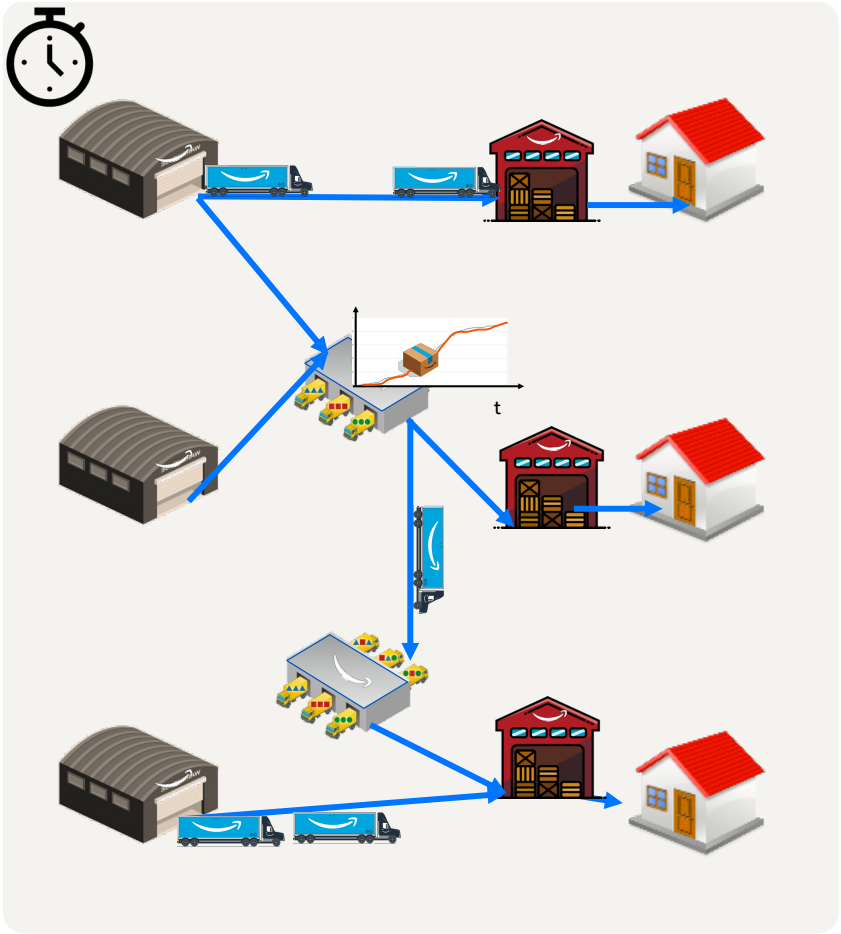
Shipment Generation

- Combination of Machine Learning and Mathematical Programming.
- Generates realistic origin-destination flow of shipments.



Network Connectivity

- Static, path-based network optimization model.
- Optimizes flow of shipments by determining most efficient routes and resource allocation.

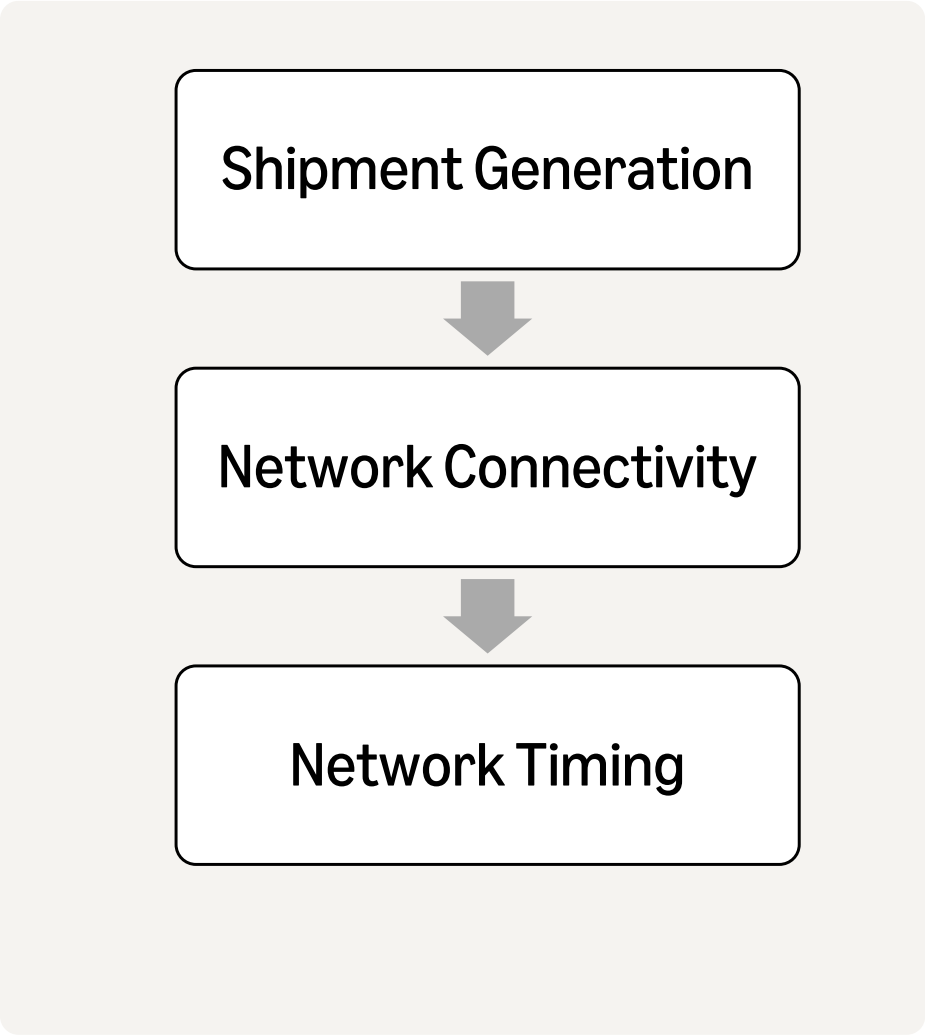


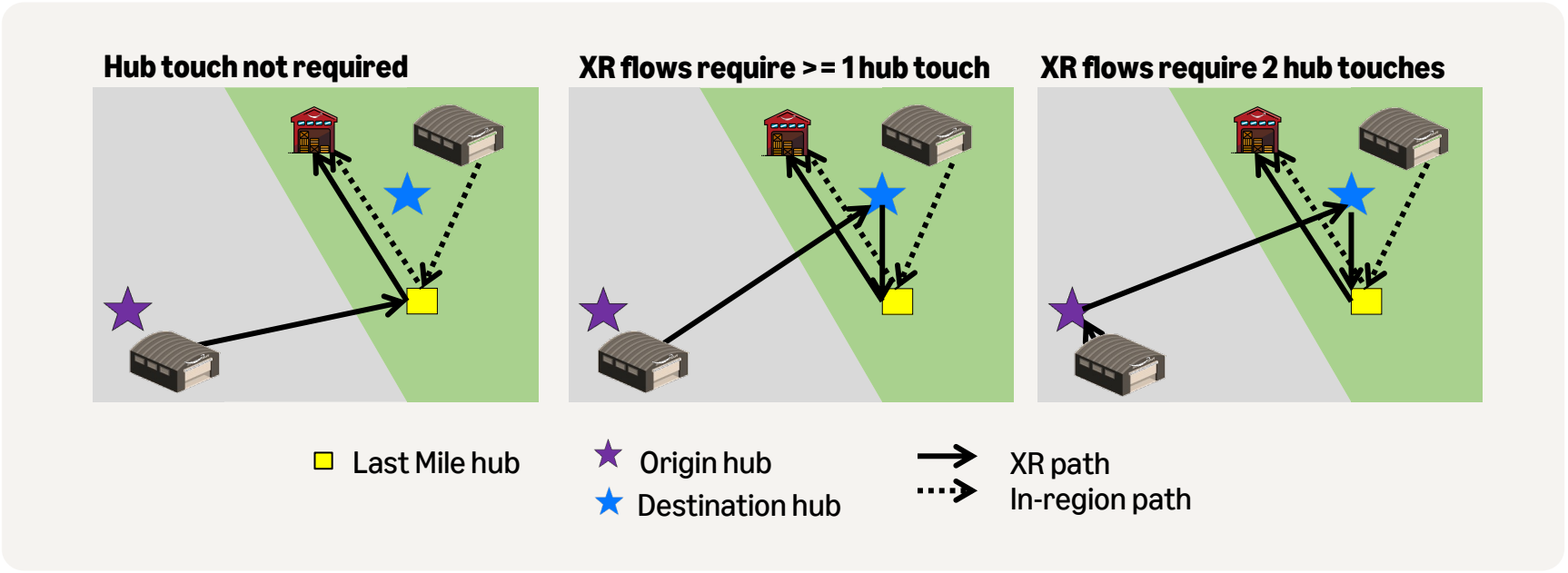
Network Timing

- Network optimization on a time-expanded graph.
- Schedules how shipments consume network resources.

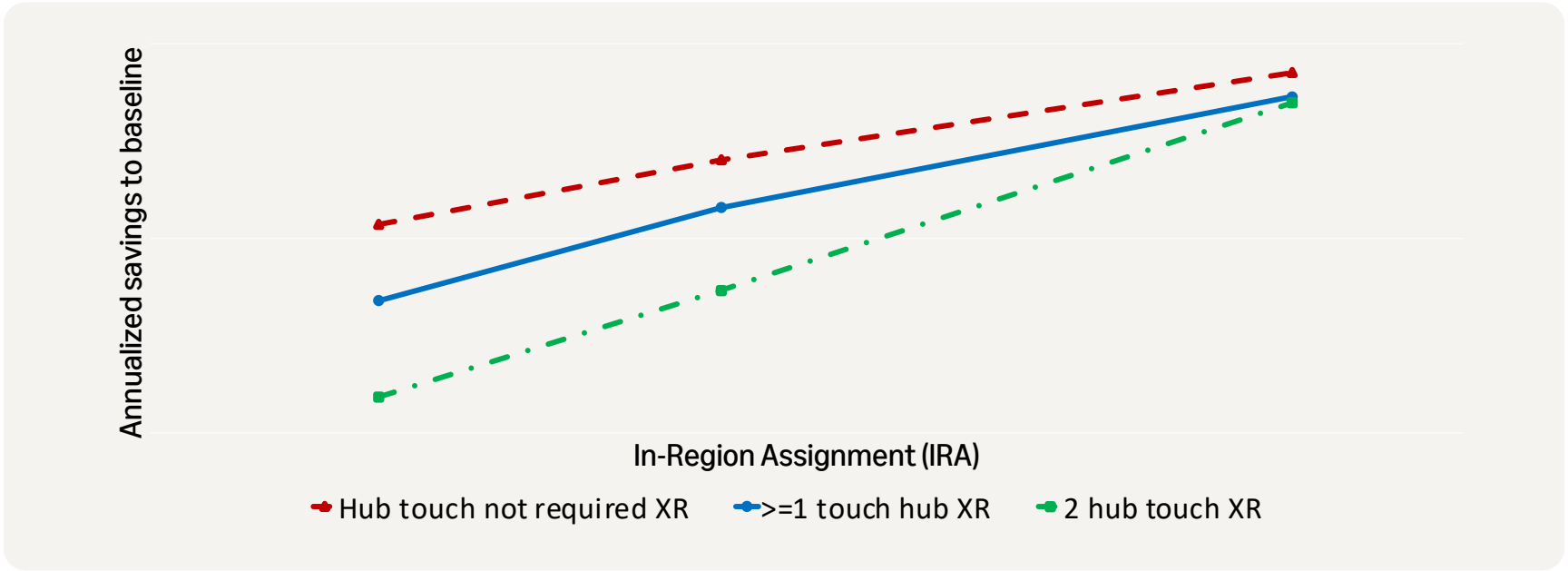
Lara, C., J. Koenemann, Y. Nie, and C. de Souza. 2023. "*Scalable Timing-Aware Network Design via Lagrangian Decomposition.*" European Journal of Operational Research.

Network Design Suite



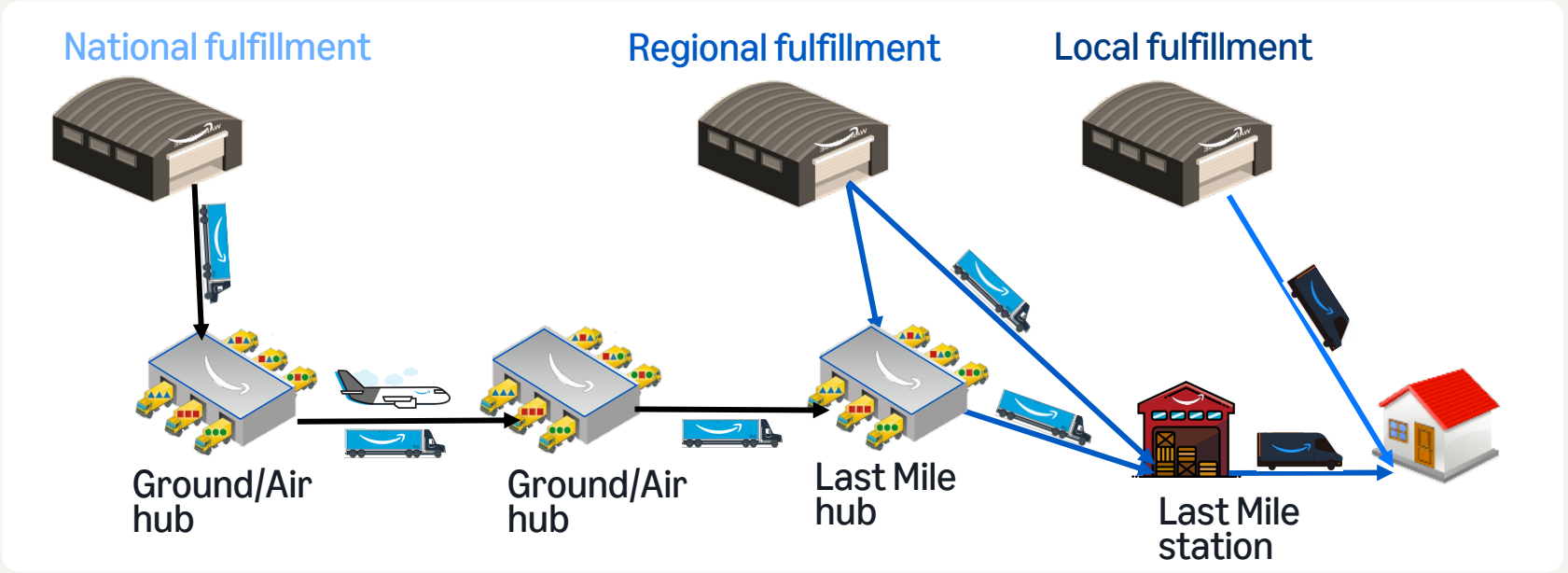


Cross-region Design Evaluation



Cross-region Design Evaluation

Regionalized Network Structure

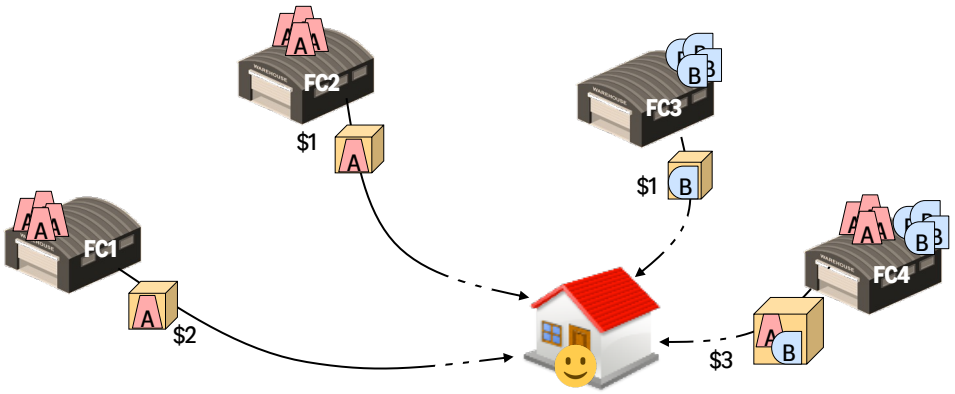


amazon

Software Implementation

Order Assignment Model

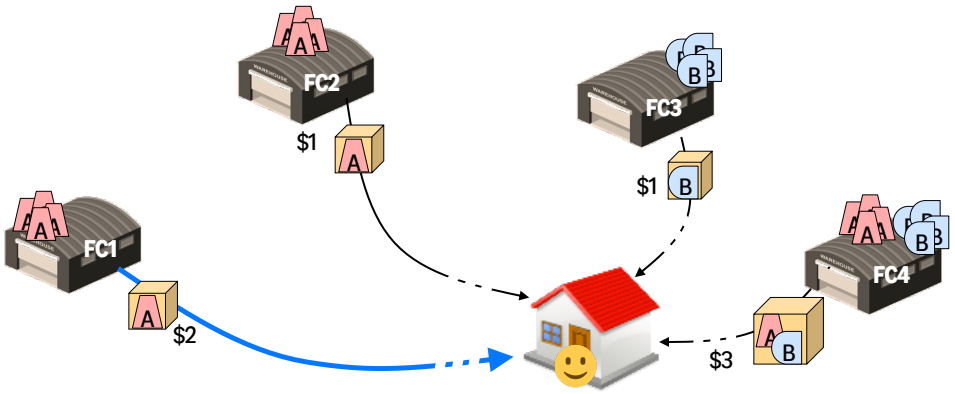
Pre-regionalization



- Order: 1 A and 1 B

Order Assignment Model

Pre-regionalization

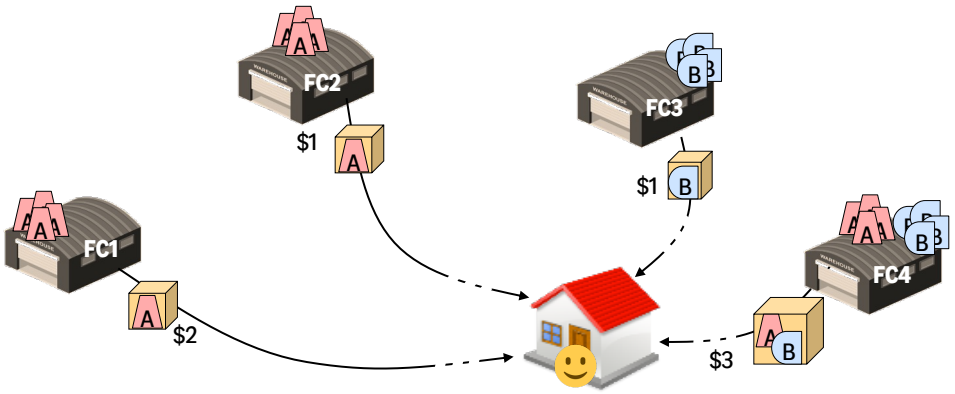


Candidate shipments	Promise met	Cost
1) A from FC1	✗	$2
2) A from FC2	✓	$1
3) B from FC3	✓	$1
4) A B from FC4	✓	$3

- Order: 1 A and 1 B
- Generate candidate shipments.
  For each we evaluate:
  violation of delivery window promised
  cost per shipment

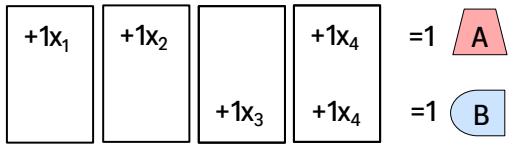
Order Assignment Model

Pre-regionalization



Candidate shipments	Promise met	Cost
1) A from FC1	✗	$2
2) A from FC2	✓	$1
3) B from FC3	✓	$1
4) A B from FC4	✓	$3

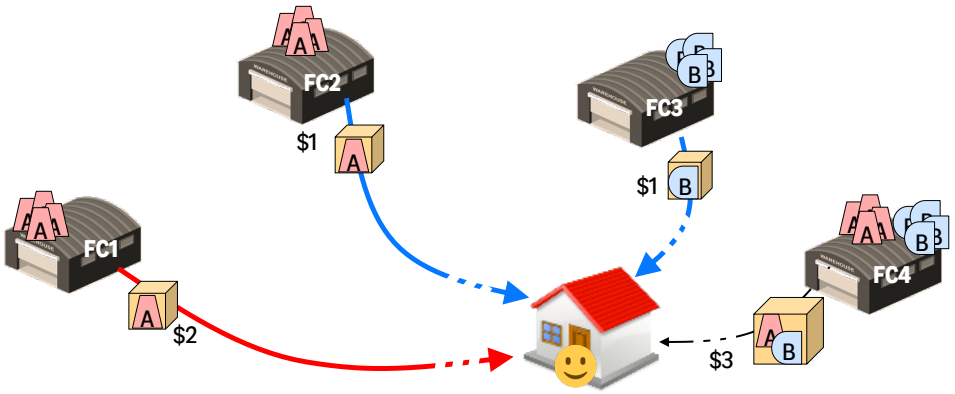
- Order: 1 A and 1 B
- Generate candidate shipments.
  For each we evaluate:
  violation of delivery window promised
  cost per shipment
- Instantiate the Order Assignment Model



x binary

Order Assignment Model

Pre-regionalization



Candidate shipments	Promise met	Cost
1) A from FC1	✗	$2
2) A from FC2	✓	$1
3) B from FC3	✓	$1
4) A B from FC4	✓	$3

- Order: 1 A and 1 B
- Generate candidate shipments.
  For each we evaluate:
  violation of delivery window promised
  cost per shipment
- Instantiate the Order Assignment Model
- Goal programming solves:
  Minimize promise violation

To exclude — A feasible solution

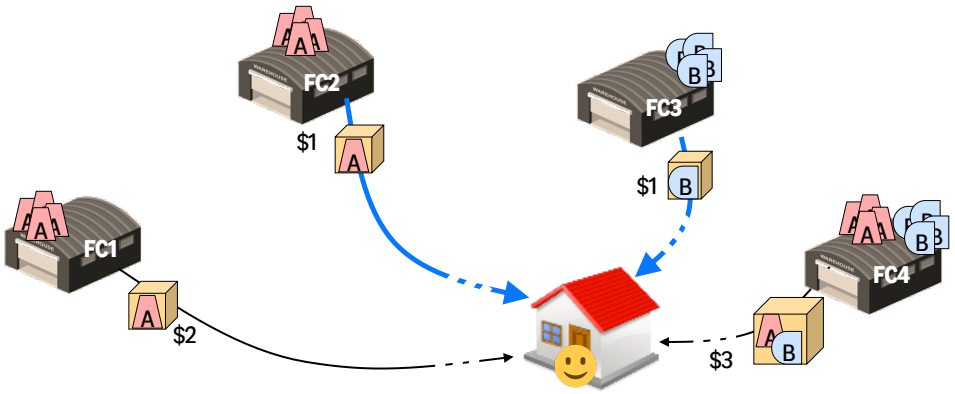
Promise viol. : $+1x_1$ $+0x_2$ $+0x_3$ $+0x_4$

$+1x_1$	$+1x_2$		$+1x_4$	=1 A
		$+1x_3$	$+1x_4$	=1 B

x binary

Order Assignment Model

Pre-regionalization



Candidate shipments	Promise met	Cost
1) A from FC1	✗	$2
2) A from FC2	✓	$1
3) B from FC3	✓	$1
4) A B from FC4	✓	$3

amazon

- Order: 1 A and 1 B
- Generate candidate shipments.
  For each we evaluate:
  violation of delivery window promised
  cost per shipment
- Instantiate the Order Assignment Model
- Goal programming solves:
  Minimize promise violation
  Minimize cost subject while keeping promise

Cheapest solution

Promise viol.:	$+1x_1$	$+0x_2$	$+0x_3$	$+0x_4$
Cost:	$+\$2x_1$	$+\$1x_2$	$+\$1x_3$	$+\$3x_4$

$+1x_1$	$+1x_2$		$+1x_4$	=1	A
		$+1x_3$	$+1x_4$	=1	B
$+1x_1$	$+0x_2$	$+0x_3$	$+0x_4$	≤0	

x binary

amazon

Shipment costs

Real costs

E.g. third-party carrier rate cards

Accurate and stable

+

Artificial costs

E.g. load balancing costs, inventory or transportation resource opportunity costs (Acimovic & Graves, 2015)

Volatile and often inaccurate.

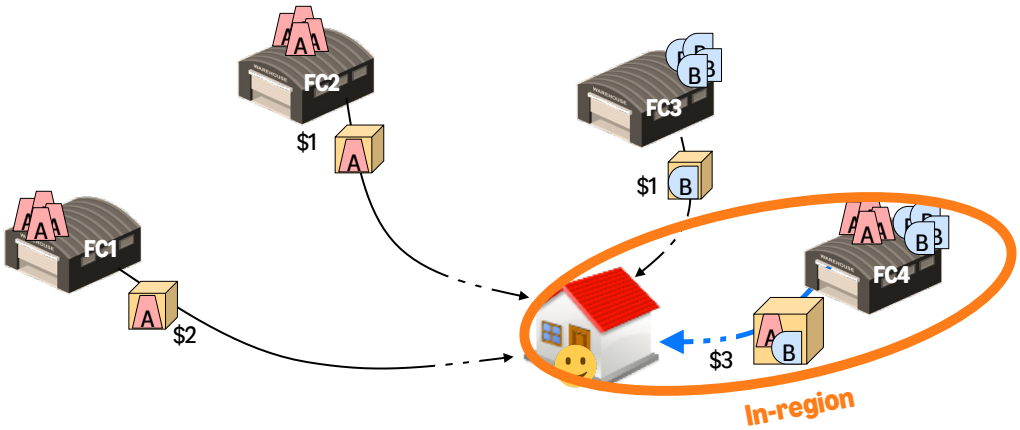
For years we solved Order Assignment Model assuming shipment costs were a good proxy for the strategic outcomes we sought.

However, with regionalization, in many situations artificial costs were driving assignments out-of-region.

Xu, P., R. Allgor, and S. Graves. 2009. "Benefits of Reevaluating Real-Time Order Fulfillment Decisions." *Manufacturing & Service Operations Management*.

Order Assignment Model

Post-regionalization



Candidate shipments	Promise met	Cost
1) A from FC1	✗	$2
2) A from FC2	✓	$1
3) B from FC3	✓	$1
4) A B from FC4	✓	$3

- Order: 1 A and 1 B
- Generate candidate shipments.
  For each we evaluate:
  violation of delivery window promised
  cost per shipment
- Instantiate the Order Assignment Model
- Goal programming solves:
  Minimize promise violation
  Minimize cost subject while keeping promise
  Minimize out-of-region while relaxing costs

…regionalized solution

Promise viol. :	$+1x_1$	$+0x_2$	$+0x_3$	$+0x_4$	
Cost :	$+\$2x_1$	$+\$1x_2$	$+\$1x_3$	$+\$3x_4$	
Out of region :	$+1x_1$	$+1x_2$	$+1x_3$	$+0x_4$	
	$+1x_1$	$+1x_2$		$+1x_4$	=1 A
			$+1x_3$	$+1x_4$	=1 B
	$+1x_1$	$+0x_2$	$+0x_3$	$+0x_4$	≤0
	$+\$2x_1$	$+\$1x_2$	$+\$1x_3$	$+\$3x_4$	≤$2(1+50%) = $3

x binary

In the real world...

In the example we made a number of simplifications.

In practice, order assignment models need to account for the following complexities:

- option to transship inventory across warehouses;
- delivery consolidation;
- synchronization at the customer destination;
- periodic re-optimization of decision up until work processing begins;
- tuning via discrete event simulations the cost relaxation parameters understanding the impact on systems that the artificial costs control.

amazon

Operational Implementation

Implementation Challenges

- Complex systems subject to variability.
- Not all factors could be modeled.
- Mental model shift.
- Convince leaders and operational teams through scientific modeling, then pilot.

amazon

Sep 2022

Start Planning

Began laying out the specifics of the regional structure and operational requirements to implement.

Dec 2022

Plans Finalized

Simulations completed. Detailed network design pilot details compiled and aligned across organization.

Jan 2023

Pilot launch

NorthEast and **MidAtlantic** regions.
Software changes to order assignment.
Network connectivity changes.

Mar 2023

Networkwide launch

Remaining 6 regions.
Software changes to order assignment.
Network connectivity changes.

amazon

Impact

In-region fulfillment

62% → 76%

The percentage of customer orders being fulfilled entirely from FCs within each region increased from 62% to 76% in the first half of 2023.

+600M

items **fulfilled from in-region** Fulfillment Centers YoY in the 4th quarter of 2023.

Distance to customers

-15%

Middle mile touchpoints

-12%

Due in part to Regionalization, the distance between our sites and customers decreased by 15%, with 12% fewer touchpoints within our middle mile network, in the first half of 2023.

-$0.45/unit

In 2023, for the first time since 2018, Regionalization helped Amazon reduce cost to serve on a per unit basis globally. In the U.S. alone, cost to serve was down by more than $0.45 per unit Year-on-Year.

-16M miles

Regionalization helped us avoid driving nearly 16 million miles in 2023.

Fastest speed

More than **7 billion** items delivered to Prime members the same or next day globally in 2023.

More than **9 billion** items delivered within the same or next day globally through 2024.

What is next?

For Amazon

- Continuous optimization.
- Re-structure inbound network to ensure right inventory is routed to FCs within region.

Future research directions

- How to design and operate a large-scale real-time retail network?
- Integration of inventory placement and network design to optimize for cost and speed.

Transferability

- Retail with a variety of delivery modes.
- Vehicle dispatch (e.g., ambulances and rideshare).
- Server allocation for cloud computing.
- Transportation network design, including role of hubs and multi-modal transport.

Acknowledgements

Huge thank you to many people for support.

- Collaboration among multiple teams, with lots of fluidity.
- Business planners and operators, for executing changes.
- Amazon leadership, for support, disambiguation, and driving priorities.
- PR, Legal, IR, Communications teams for dissemination help.
- Edelman coaches, for feedback and competition prep.
- INFORMS, for this wonderful conference and opportunity!

amazon

Thank you!