Amazon: optimising the journey of a package

Examples of optimization and machine learning applications

Zsolt Csizmadia
Amazon, Modeling and Optimization
zsoltcs@amazon.co.uk

Some slides are from other Amazonians' Informs presentations
All photos are from www.aboutamazon.com

Agenda

- A glimpse of Amazon fulfilment
- Examples working backwards from the customer
  - Reinforced learning example in the last mile
  - Optimization example in the middle mile
  - Optimization example in the first mile

The journey of a package at a glimpse

What makes ordering special for you at amazon.com?
For me… the delivery estimate.

Give fast delivery estimate, then fulfil to delivery estimate
Raise the bar of customer's expectations
Customer expects even faster delivery

Customer looks at an item
Amazon gives a delivery estimate
Customer orders
Amazon decides where to fulfil from – decide what units to pack together if ordering multiple items
Amazon decides the route to take – Amazon 1P fulfilment / 3P fulfilment, what route, when

Fulfillment at Amazon

S$24.20 (S$0.61 / Count)

✓prime

FREE delivery **Tomorrow, 23 August**.
Order within 7 hrs 47 mins. Details

Add to Cart

Buy Now

Customer places order

Initial fulfillment plan

12:00 FC1 → DS1
16:00 AMZL last-mile

Picking

Packing

Final fulfillment plan

16:00 FC1 → DS1
05:00 AMZL last-mile

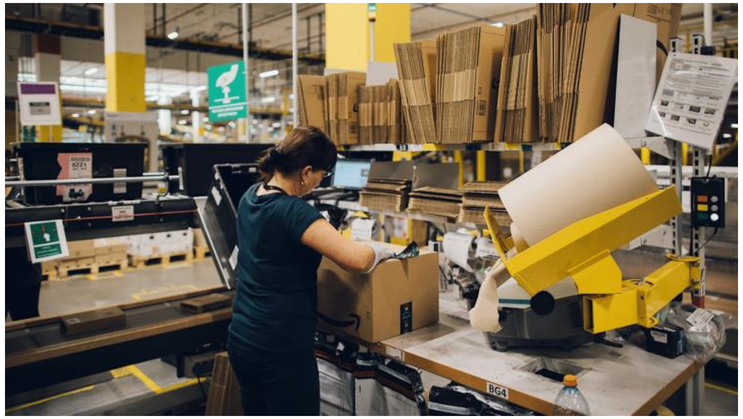
Delivered

Daniel Chen, Informs 2023

Inbound and stow

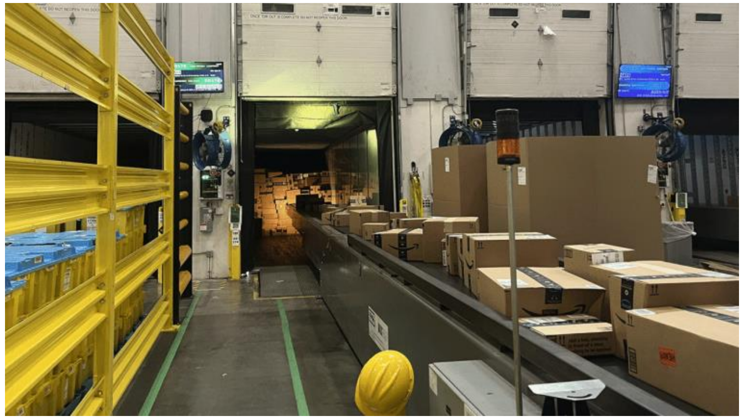
Older (orange) drives

Outbound, pick, package, SLAM – Ship Label Apply Manifest



www.aboutamazon.com

Docks and loading to linehaul trucks



ROBIN – ROBotic INduction

Sub-same day station

AR facility

Newer drives

www.aboutamazon.com

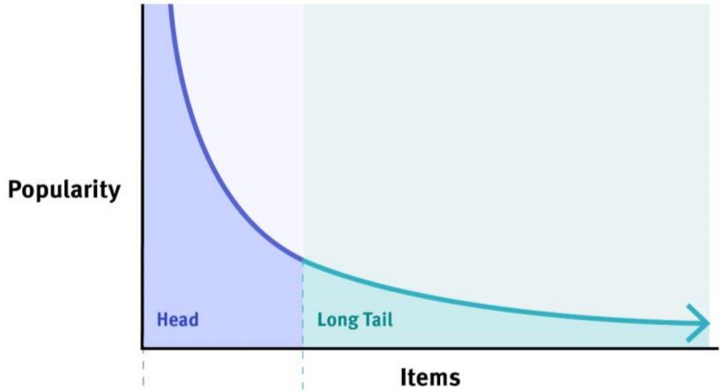
Selection is in a different order of magnitude

~10^5 unique products

vs.

~10^7 unique products

… and demand is very skewed



Also known as *the Long Tail.*[1]

[1]Anderson, C. *The Long Tail: Why the Future of Business is Selling Less of More*, 2006.

Cristiana Lara, Informs 2023

Vendors / suppliers

1000s

First mile:
inbound network,
distribute supply to
fulfilment centres.

Fulfilment centres.
Holds inventory.

100s of buildings
10s millions of unique items in inventory

10s millions of packages
1000s trucks
100s flights

Middle mile:
outbound network,
Maximise speed,
minimize cost

Last mile: delivery to customer

Delivery stations → Customer

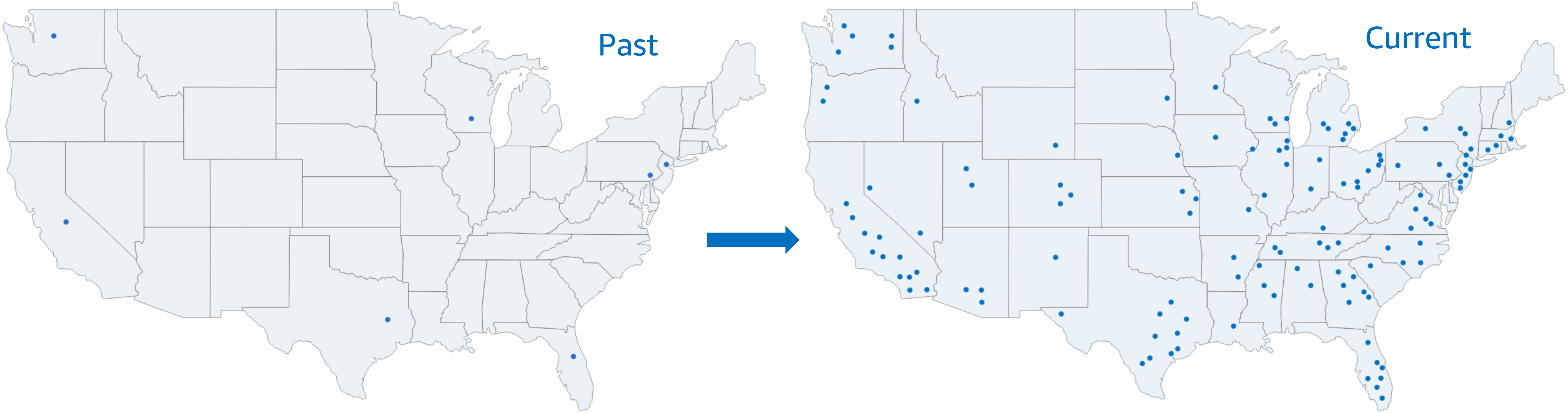
10, 000s vehicles

10s millions of orders

www.aboutamazon.com

The Everything Store Problem

- Start with few Fulfillment Centers (FCs), with thick lanes
- Fast-forward to 2024, 200+ FCs that can connect to any customer
- Network complexity spirals - hard to operate and get good signals for inventory placement



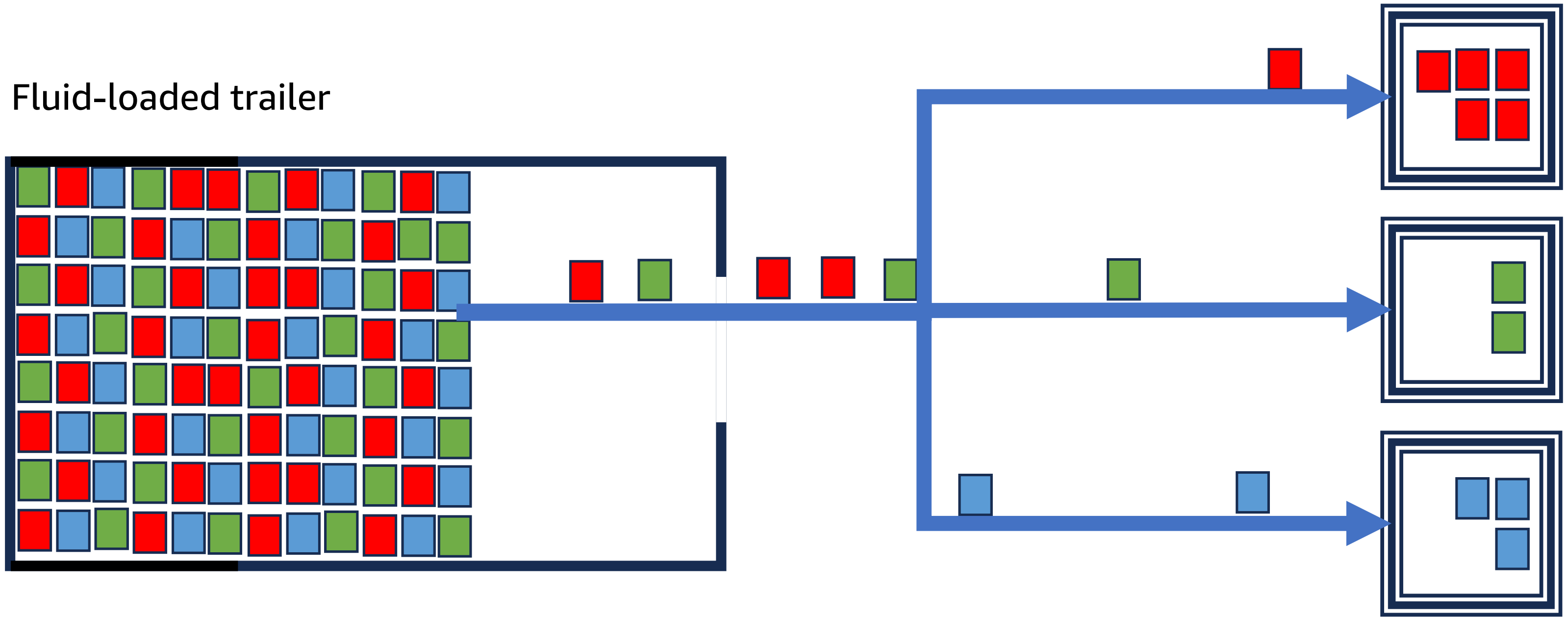
Not actual FC locations. Plot is only for illustration purpose.

- On the bright side, we now have many copies of FCs with fungible inventories
- A perfect opportunity to divide and conquer: **Regionalization**

Yuan Li, Informs 2023

Sortation centers (Inbound Cross Dock and Middle Mile Sort Centers)

Each Sort is a package level sortation process that sorts each package to its assigned destination



Shanshan Zhang, Informs 2023

Last mile

Last mile: delivery to customer

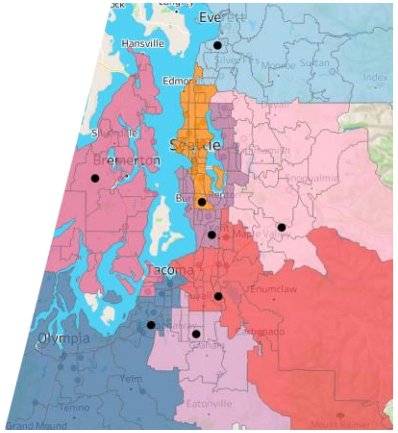
Delivery stations → Customer

10, 000s vehicles

10s millions of orders

Tactical:
- Vehicle routing problem
- Labour / driver demand

Strategic:
- Number and placement of delivery stations
- Throughput capacity
- Jurisdictions of delivery stations
  - how many customers
  - expected demand / workload
  - resources



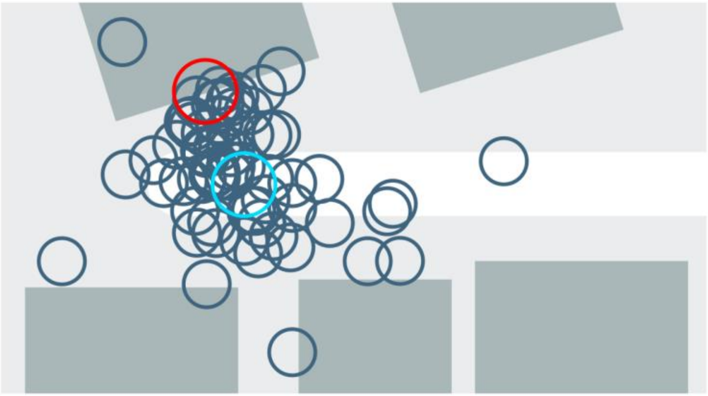
Louis Faugere, Informs 2023

Last mile example machine learning problem

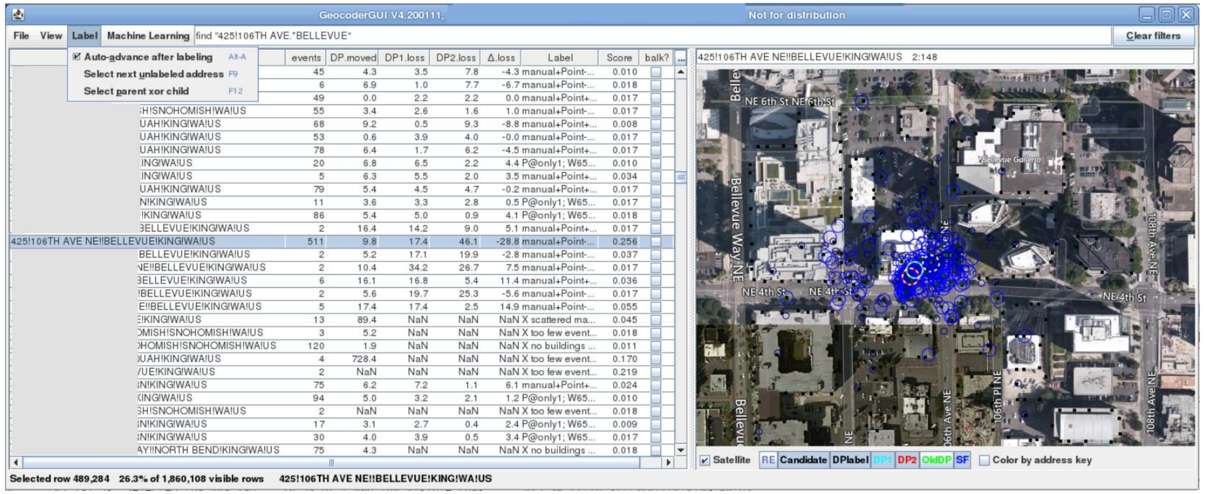
Getting Your Package to the Right Place: Supervised Machine Learning for Geolocation

Amazon Last Mile delivers millions of packages daily

The geolocation of each address is needed for partitioning and optimizing routes among vehicles



Learning to rank or machine-learned ranking is the application of machine learning, typically supervised, semi-supervised or reinforcement learning, in the construction of ranking models for information retrieval systems.
(Wikipedia)



George Forman, www.amazon.science, 2021

Middle mile

Tactical:
- where to fulfil from, including what to package together
- route selection, including 1p / 3p
- number of trucks (short / middle / long term)
- capacity management
- minimize fulfilment cost
- maximize speed

Strategic:
- where to place facilities
- capacities
- connectivity

Since early 2024: +Florida

100s of buildings
10s millions of unique items in inventory

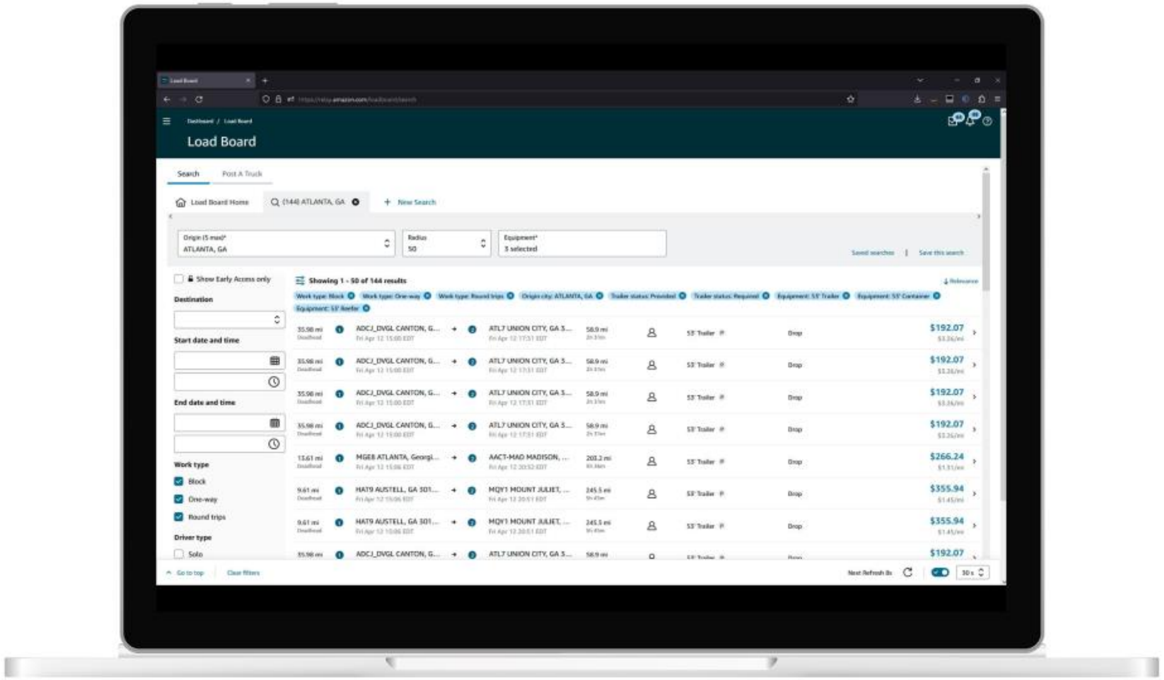
Fulfilment centres.
Holds inventory.

10s millions of packages
1000s trucks
100s flights

Middle mile:
outbound network,
Maximise speed,
minimize cost

Delivery stations

Middle mile example optimization and forecasting problem: The relay load board and the spot market



What does Relay offer?

For Carriers

A technology toolset with everything you need to haul loads for Amazon.

Short-Term Contracts

Secure full work-weeks for your drivers, and grow your fleet by locking in revenue for provided trucks with single or multi-week contracts several weeks in advance.

Auctions

Bid on contracts with full transparency into the timing remaining in the auction and the current lowest bid.

Loadboard

Enjoy exclusive access to Amazon's spot work and book loads with all-in pricing at the click of a button.

Post A truck

Reduce empty miles and idle time by sharing when and where your trucks are available. Relay will automatically book loads matching your criteria so you can focus your time on growing the business.

Rewards

Access exclusive discounts on fuels, tires, maintenance and more.

Join the Amazon Freight Partner program for long-term, consistent work.

Learn more

Pictures from relay.amazon.com

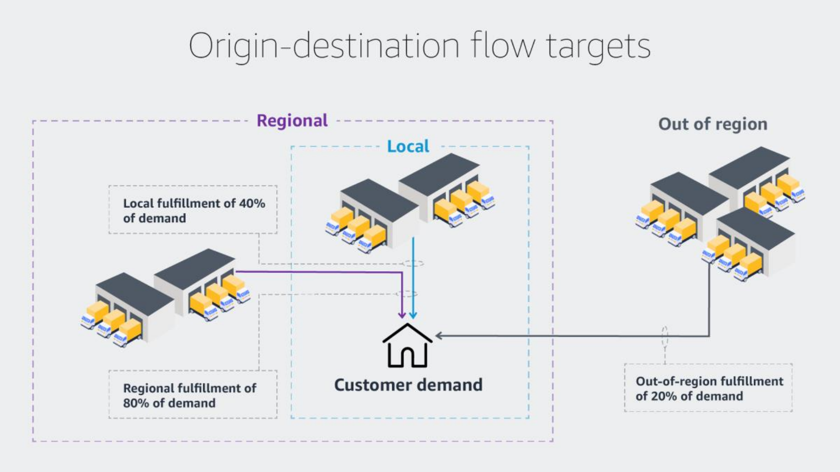
Middle mile example optimization problem: Origin and route selection with capacity management

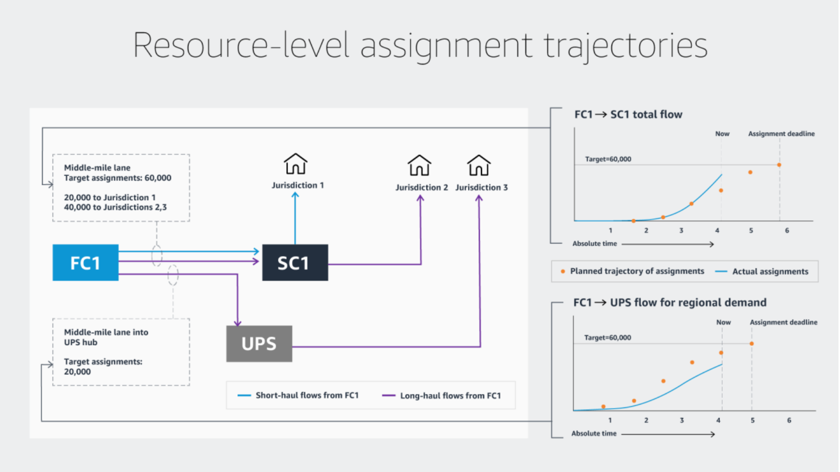
- A simplified example:

	Path A	Path B
Shipment 1 Costs	1	2
Shipment 2 Costs	1	10
Capacity	1	infinity

- Greedy online solution: shipment 1 -> Path A, shipment 2 -> path B
- Optimal solution: shipment 1 -> Path B, shipment 2-> Path A.

Special case: the cost is the same, but we can deliver faster in aggregate if we preserve capacity.





Ozlem Bilginer, Informs 2022

Lagrangian decomposition

Minimize $$\sum_{s\in S}\sum_{p\in P_s} c_{s,p}x_{s,p} + \sum_{r\in R} c_r o_r$$

s.t.

(upper-bound) $$\sum_{s\in S}\sum_{p\in P_s\cap P_r} \beta_{s,r}\, x_{s,p} \le U_{rt} + o_r \qquad \forall r \in R^+$$

(lower-bound) $$\sum_{s\in S}\sum_{p\in P_s\cap P_r} \beta_{s,r}\, x_{s,p} \ge L_{rt} - o_r \qquad \forall r \in R^-$$

(demand) $$\sum_{p\in P_s} x_{s,p} = 1 \qquad \forall s \in S$$

$$0 \le x_{s,p} \le 1, o_r \ge 0$$

Minimize $$\sum_{s\in S}\sum_{p\in P_s}\left(c_{s,p} + \sum_{r\in R_p\cap R^+}\alpha_r\beta_{s,r} - \sum_{r\in R_p\cap R^-}\alpha_r\beta_{s,r}\right)x_{s,p} + \sum_{r\in R}(c_r-\alpha_r)\, o_r - \sum_{r\in R^+}\alpha_r\, U_{rt} + \sum_{r\in R^-}\alpha_r L_{rt}$$

s.t.

(demand) $$\sum_{p\in P_s} x_{s,p} = 1 \qquad \forall s \in S$$

$$0 \le x_{s,p} \le 1,\ o_r \ge 0$$

Statistic	**Problem 1**	**Problem 2**	**Problem 3**
Number of Resource Constraints	~ 1.6 million	~ 1.6 million	~ 1.6 million
Number of Shipments	~ 3.5 million	~ 3.6 million	~4.3 million
XPRESS Time (seconds)	8,794	15,034	25,748
LD Optimality Gap at 1800 seconds	4.6%	3.4%	4.7%

Ozlem Bilginer, Informs 2022

Middle mile - topology design

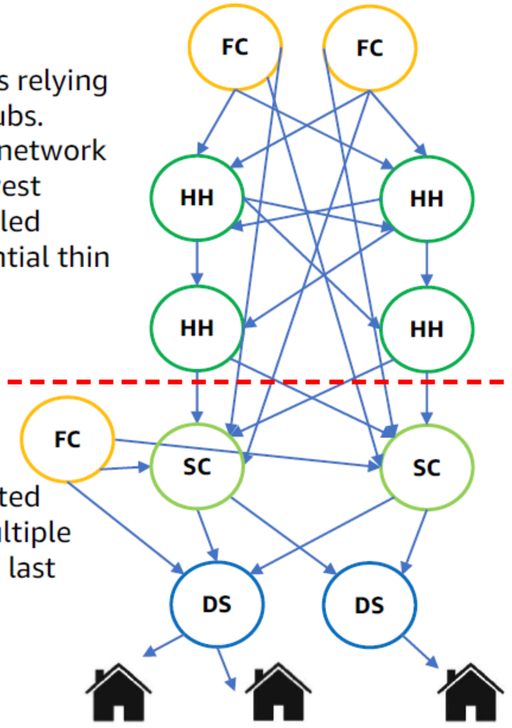
Demand variability resiliency: the case for consolidation

- Comparison of two network design principles.

[BAU] Business as Usual – shortest path

National
Multi-SC paths relying on national hubs. Complex hub network favorizing lowest distance traveled through potential thin lanes.

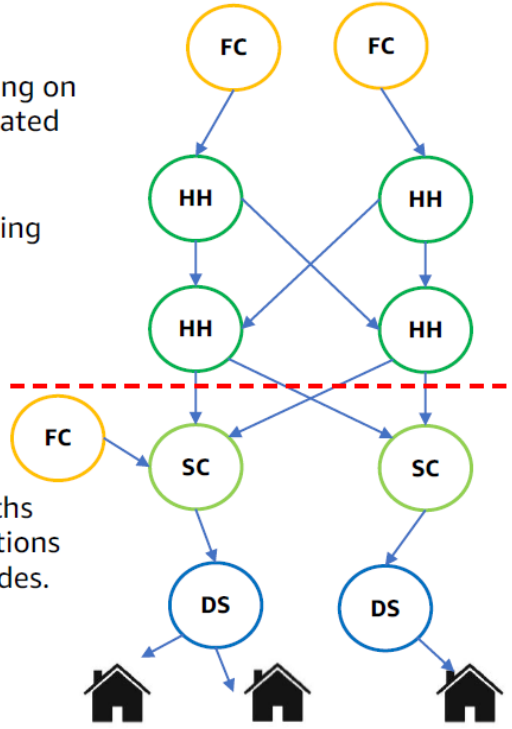
Regional
Mix of direct and consolidated paths with multiple injections into last mile nodes.



[Consolidation] Simplified connectivity - SC

National
3-SC paths relying on hubs with dedicated operations.
Simplified hub network favorizing thick lanes

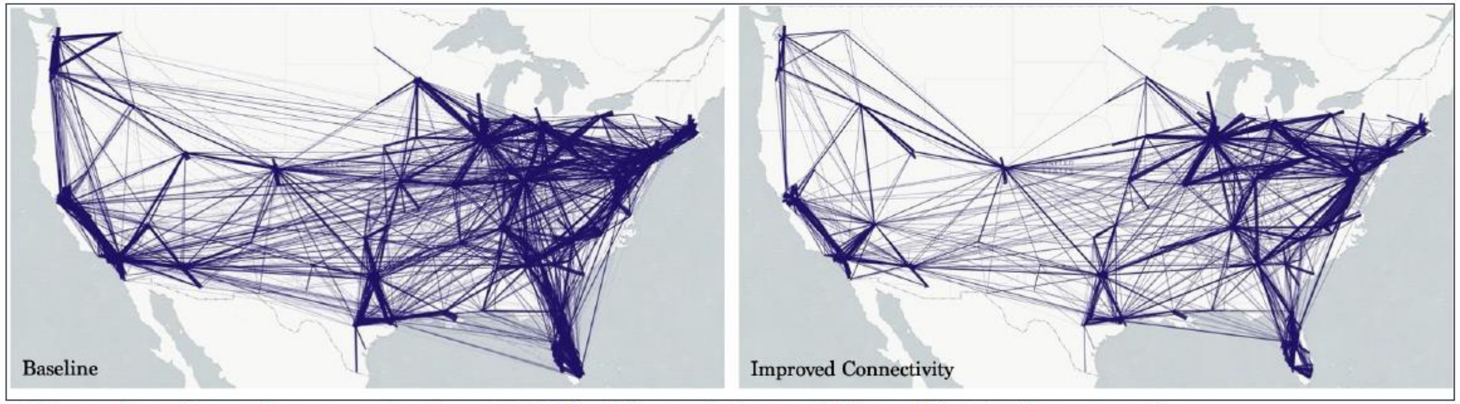
Regional
High frequency consolidated paths with single injections into last mile nodes.



Julie Poullet, Informs 2022

Middle mile – topology design

Consolidated network: less lanes with higher volume



Map showing the arcs in the *BAU* (left) and *Consolidated* (right) networks.

- More consolidated design leads to high Hub network utilization, both in terms of number of arcs and in volume.
- Overall reduction in number of arcs, which are also shorter, translating into a reduction in 'last' trucks and long distance trucks.
- Increase in number of touches.

Arc type	Arc count	Pkg count
FC->DS	-71.6%	-41.7%
FC->SC	-37.3%	7.8%
SC->DS	-37.5%	13.6%
SC->SC	25.4%	168.4%
Total	**-37.0%**	**16.9%**

Difference between *Consolidated* and *BAU* networks.

Arc count weighed package count

Julie Poullet, Informs 2022

Vendors / suppliers

1000s

First mile:
inbound network, distribute supply to fulfilment centres.

Fulfilment centres. Holds inventory.

Tactical
- Inventory replenishment
- Inventory placement / distribution

Strategic:
- Where to build warehouses
- Inbound facilities
- Inbound network connectivity

First mile example optimization problem: Topology design, vendor and FC connectivity, inventory distribution

Logistics Network Design Problem

- Multicommodity Capacitated fixed-charge network design problem (mcFND)

Output of an inventory placement model

- Commodities (o_i, d_i) with n_i packages and volume cb_i to be routed via path P_i along arcs a with u_a units of capacity. Variable cost c_p per path and fixed charge cost f_a per unit of arc capacity.

$$\min \sum_{a \in A} f_a n_a + \sum_{i \in \mathcal{O}, p \in \mathcal{P}_i} c_p f_p \quad \text{(IP)}$$

$$\text{s.t.} \sum_{p \in \mathcal{P}_i} f_p = n_i \quad (i \in \mathcal{O})$$

$$\sum_{p \in \mathcal{P}_a} \text{cb}_p f_p \le u_a n_a \quad (a \in A)$$

$$f \ge 0, n \in \mathbb{Z}_+^A$$

Path descriptions include under the roof constraints like sortation

- Assign commodities to feasible path so as to minimize cost

- NP-hard problem

Julie Poullet, Informs 2022

Thank you