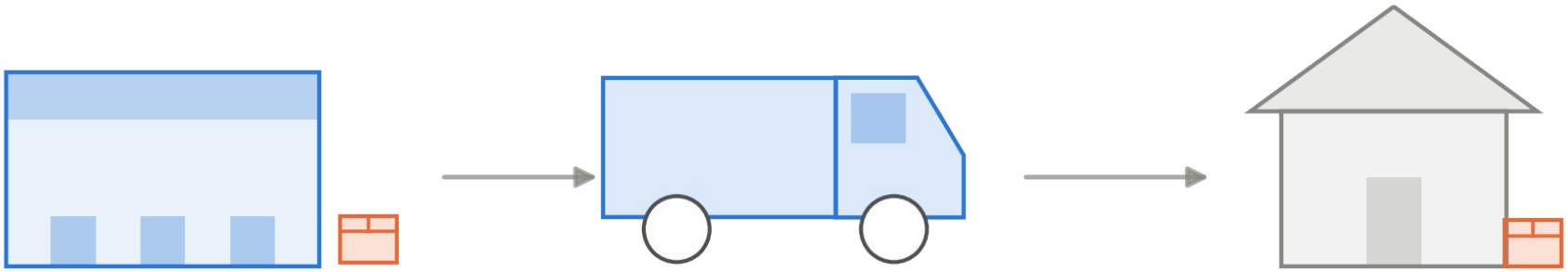
End-to-end Fulfillment Network Design

Lea Kapelevich and Shanshan Zhang

INFORMS Annual Meeting 2026, San Francisco

amazon

Outbound

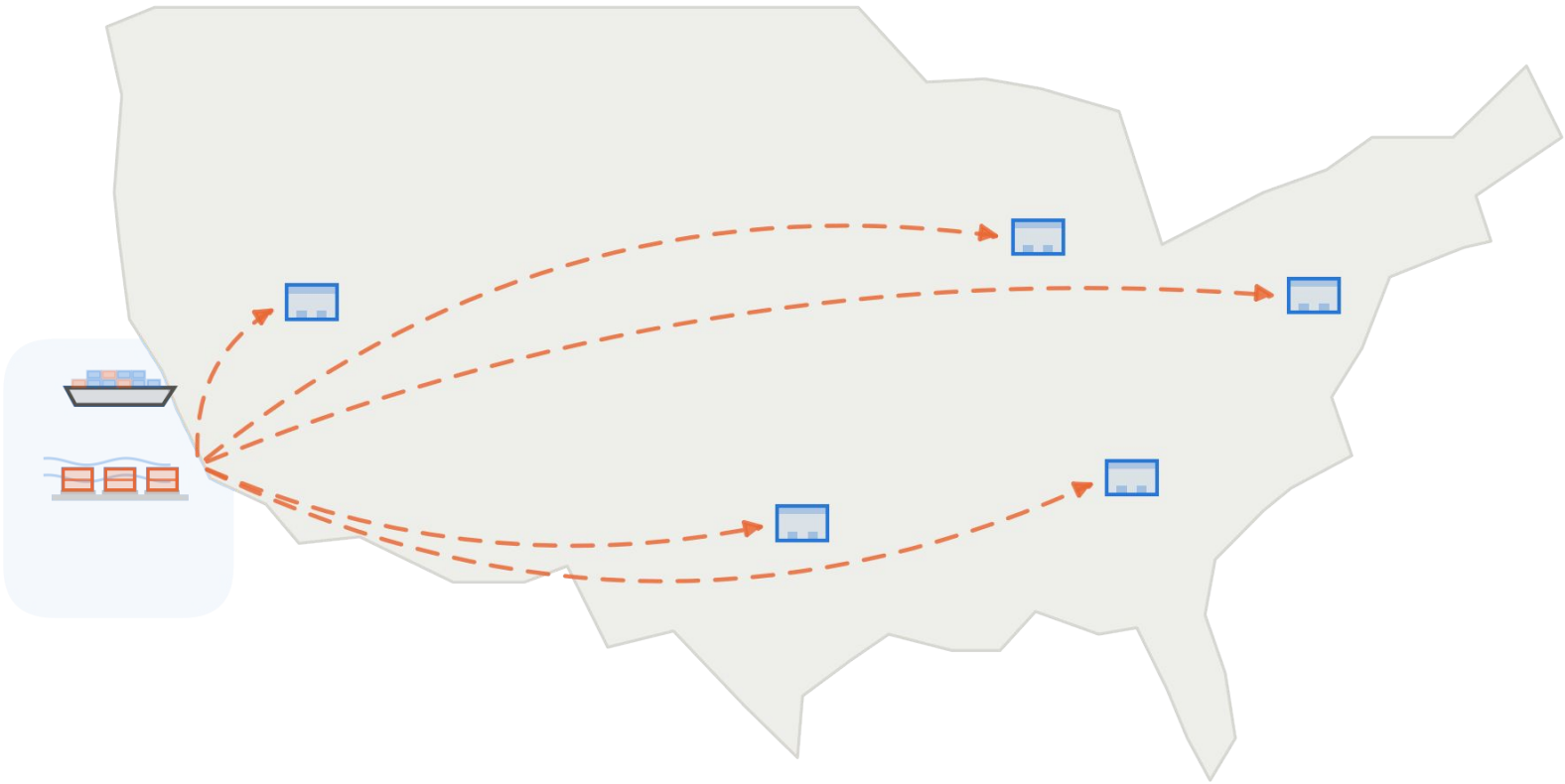


picked from the warehouse

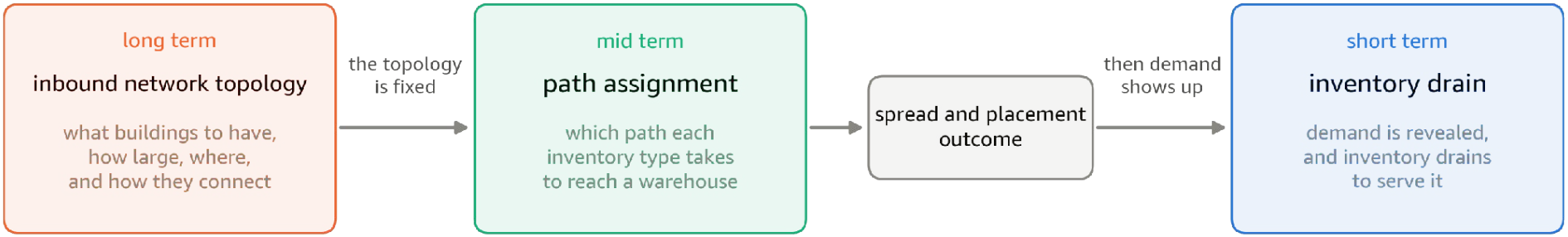
on the road

at your door

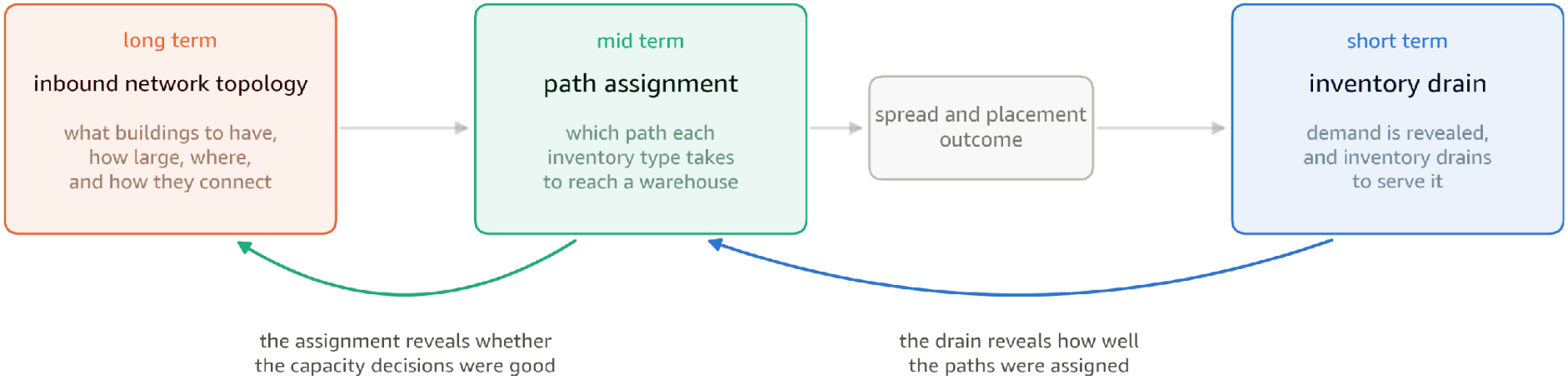
Inbound



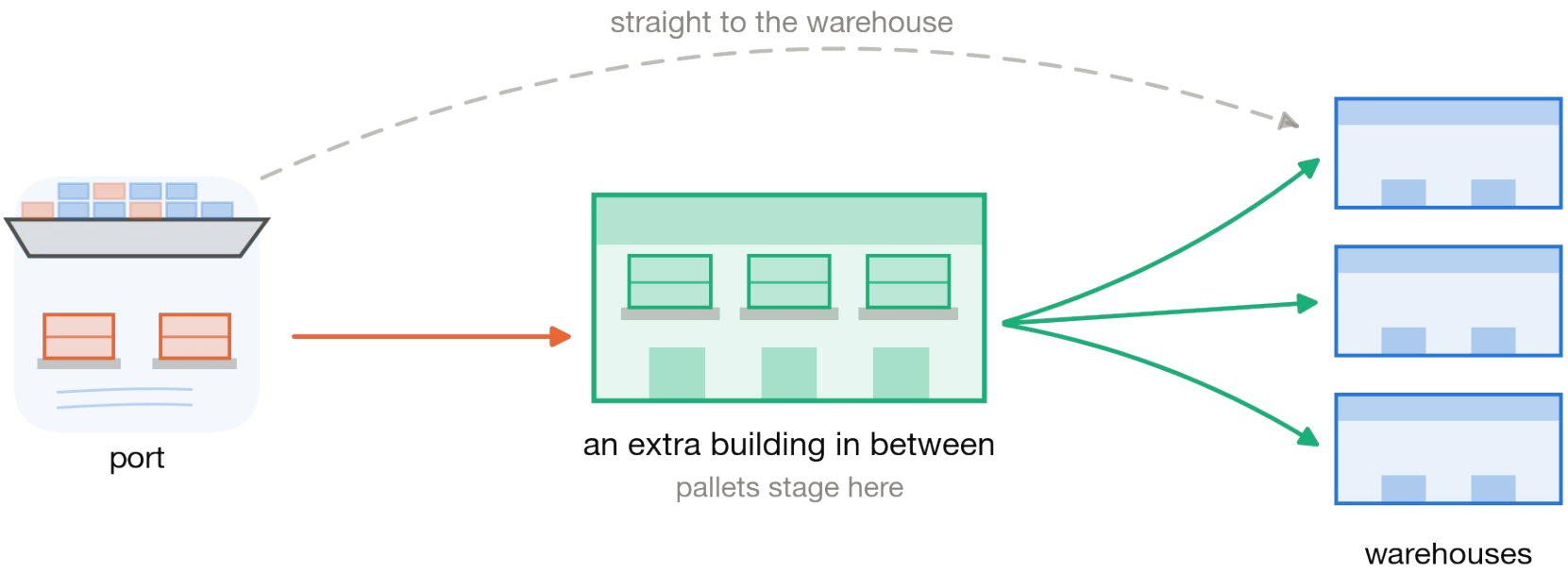
The big picture



The big picture

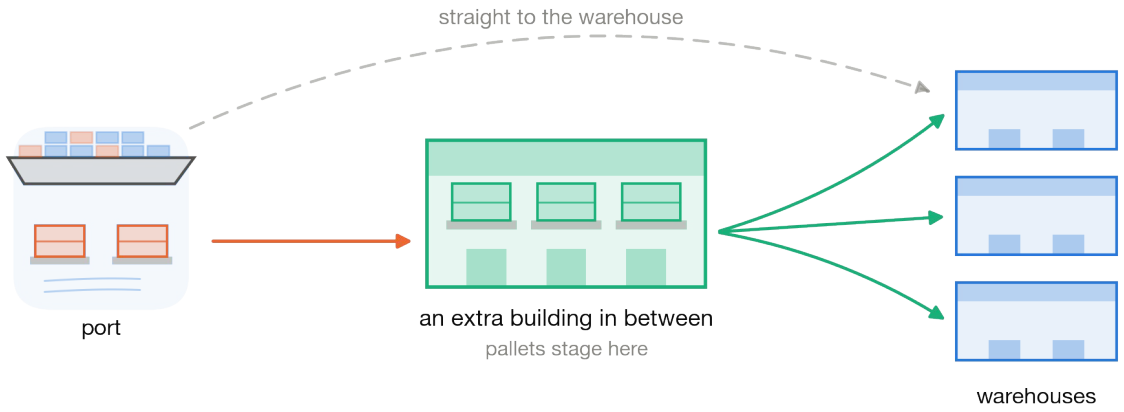


Today's talk


straight to the warehouse
port
an extra building in between
pallets stage here
warehouses

How to evaluate distribution center entitlement?

- A distribution center holds inventory before warehouses
- This talk: imagine heavy bulky items that are candidates
- My DCs are going to be
  - Slow
  - Handling pallets of goods at a time
  - Few in number
  - Well connected to all warehouses
  - With handling that is risky/expensive

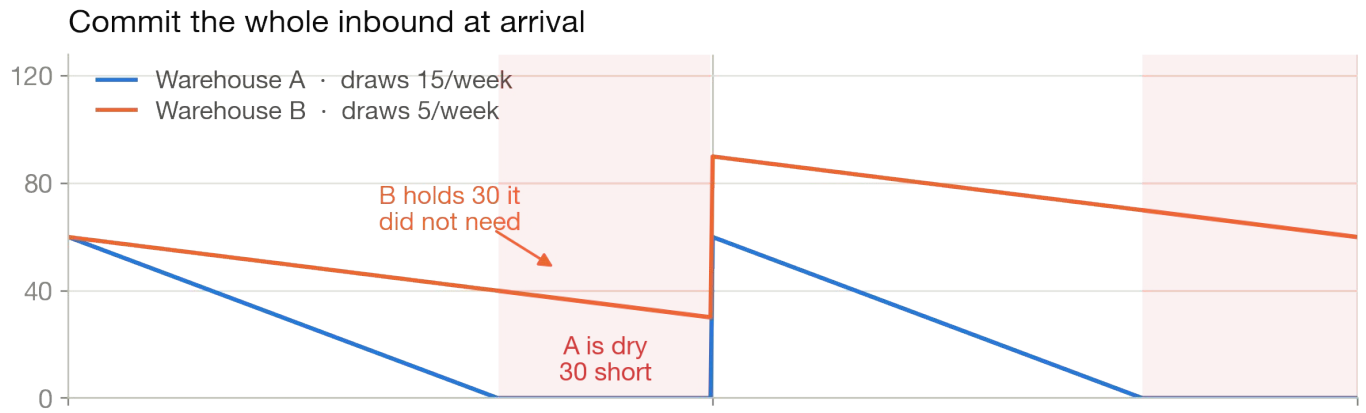


Adding an echelon

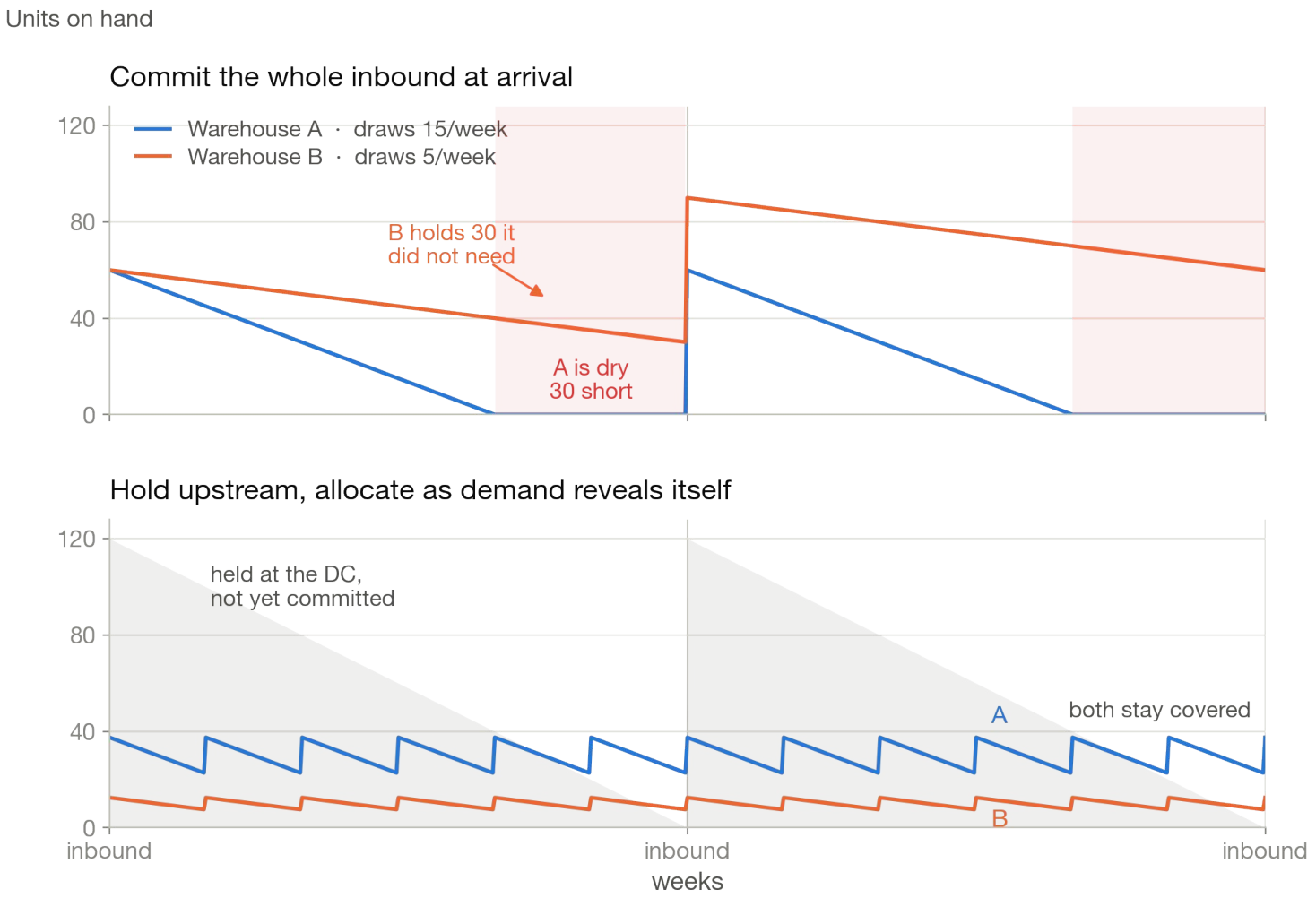
- Benefits…
  - Delayed allocation– less precommitment of volume
  - More stable lead times
  - Less cube stored in warehouses
  - More selection fits in warehouses
- Recall we want to justify the economics of expensive/risky extra handling
- For this talk focus on delayed allocation: what items can we stage to unlock the most goodness with unlimited capacity?

Delayed allocation

Units on hand



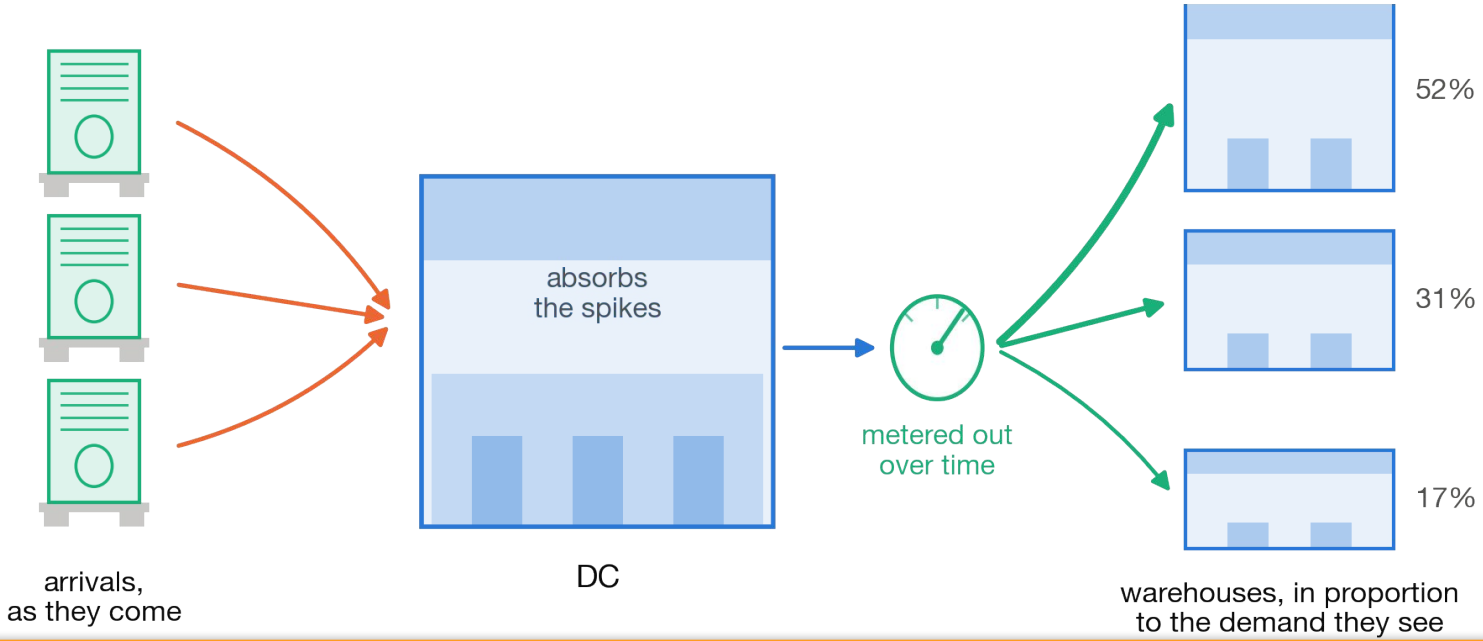
Delayed allocation



Illustrative. Same 120 units inbound every 6 weeks in both panels. A draws 15/week, B draws 5/week.

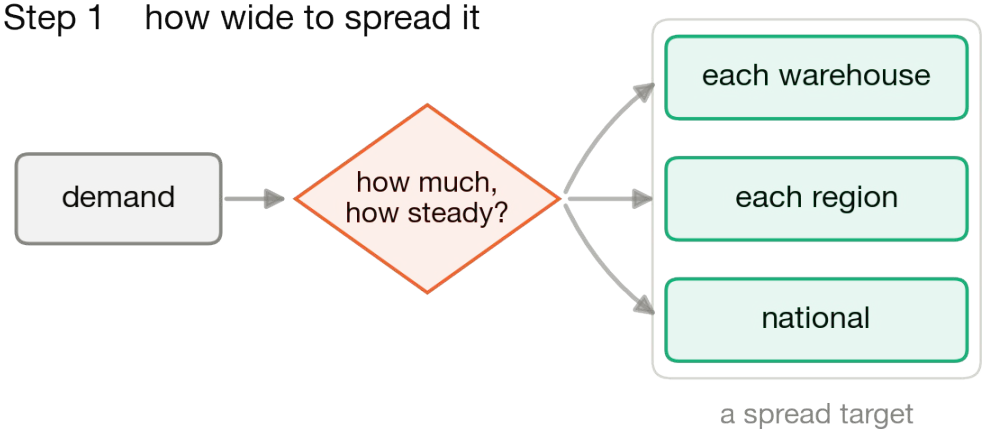
Segmenting inventory

- Key is we can identify products that are…
  - Inbounded infrequently
  - Popular in demand
  - Seasonal
- We absorb supply spikes and replenish warehouses with more accurate demand signal and lower transit variability

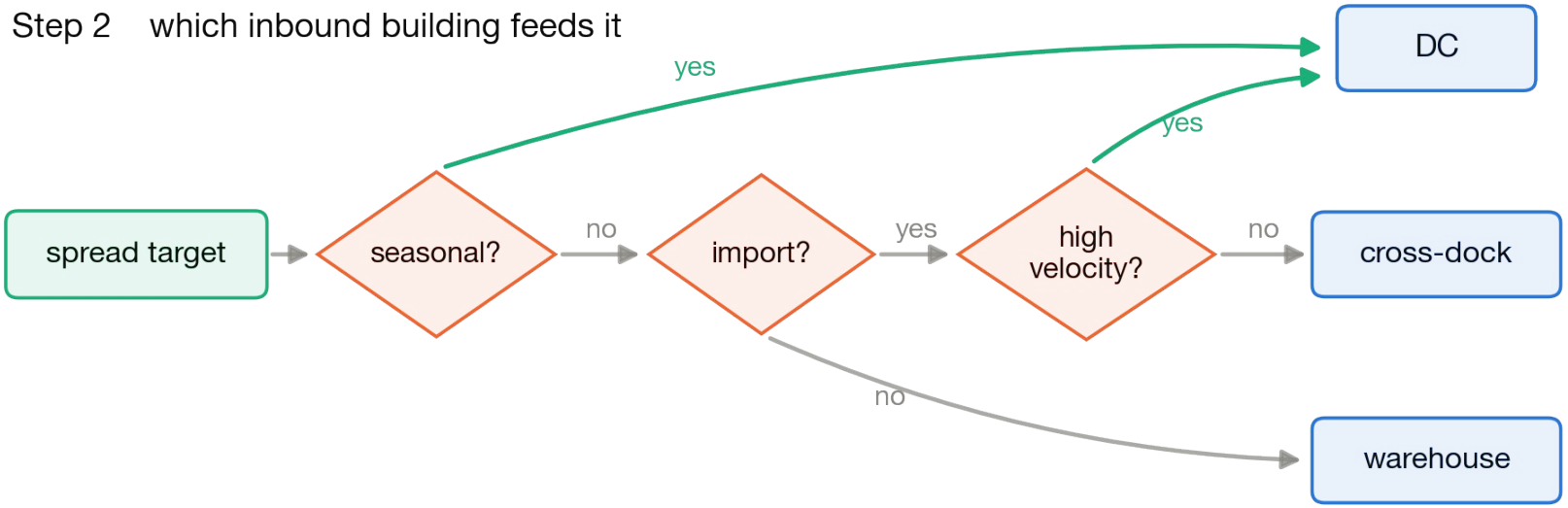


Segmenting inventory

Step 1 how wide to spread it



Step 2 which inbound building feeds it

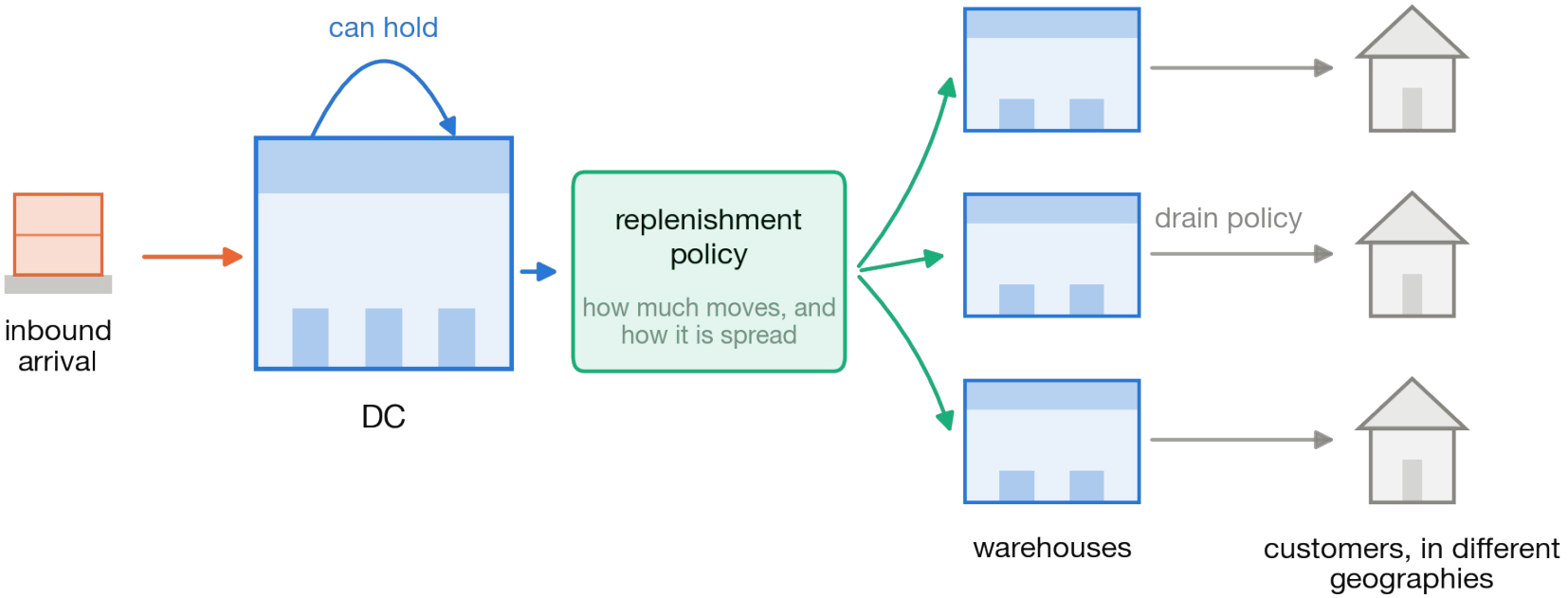


Quantifying expected benefits

- Given the inbound paths assigned in the previous slide, we want to quantify and validate the expected benefits
- Drain of inventory is complicated to model in closed form
- Network effects like discrete pallets of packages are also hard to model exactly
- We turn to simulation to validate our intuition that:
  - Highly ordered SKUs with more predictable demand unlock more benefit from frequent replenishment
  - SKUs that are inbounded infrequently benefit by being metered out to warehouses more regularly
  - Seasonal SKUs need not take unnecessary space outside of their season

Simulation

- Kudos to the developers on our team who created a performant simulator that took minutes to evaluate hundreds of thousands of candidate SKUs



How did we set this up?

$$I = \mu\frac{R}{2} + z\sigma\sqrt{R+L}$$

← total inventory requirement

cycle stock

demand between looks

safety stock

how wrong the forecast gets over review + lead time

$$\text{WOC} = \frac{I}{\mu} = \frac{R}{2} + z\frac{\sigma\sqrt{R+L}}{\mu}$$

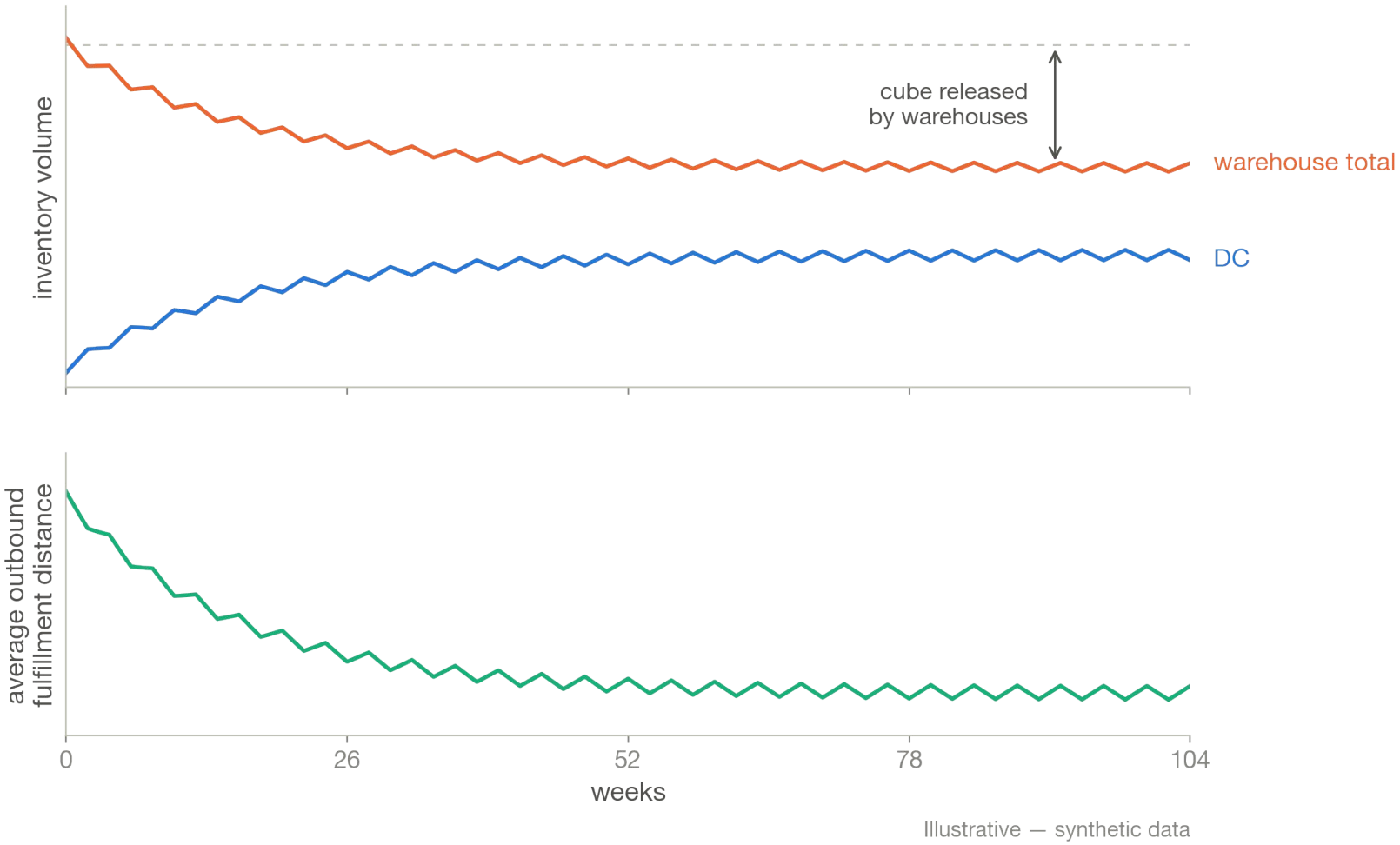
on hand ÷ average demand = weeks of cover

R	review period
L	lead time
μ, σ	demand mean and spread
z	service level: cover demand in X% of review periods

How did we set this up?

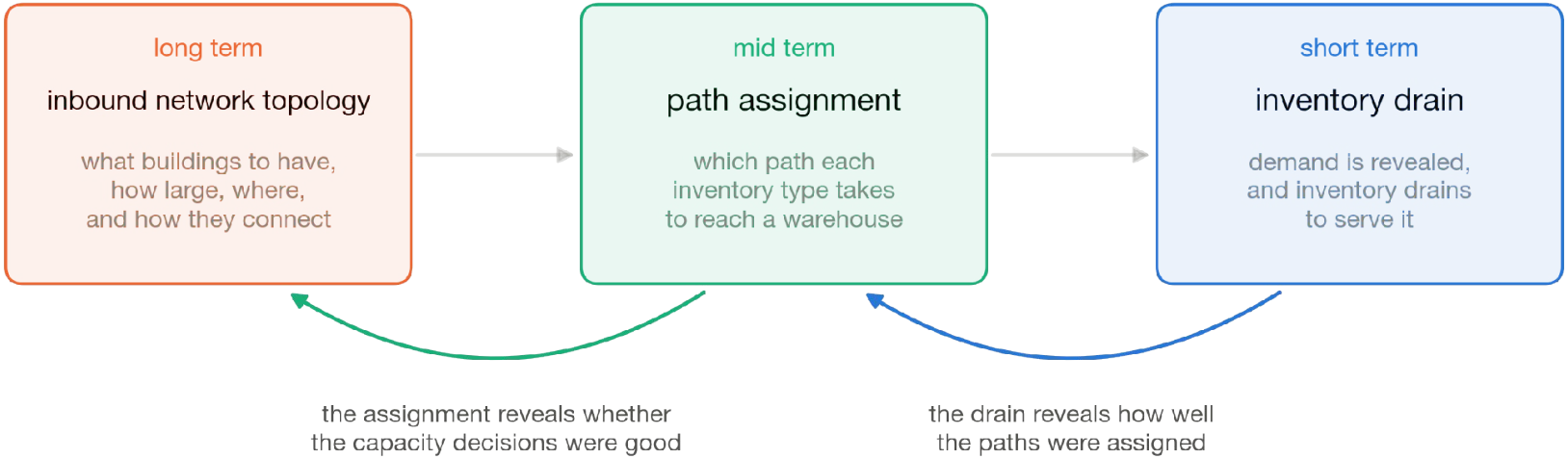
- Warehouses have a target inventory based on rolling demand and the WOC we calculated
- Replenish and distribute to warehouses proportionately to their target deficit
- Drain greedily by distance, look for improvement
- Many approximations, e.g. average pallet sizes, but simulation captures what closed form analysis can’t

Simulation behavior

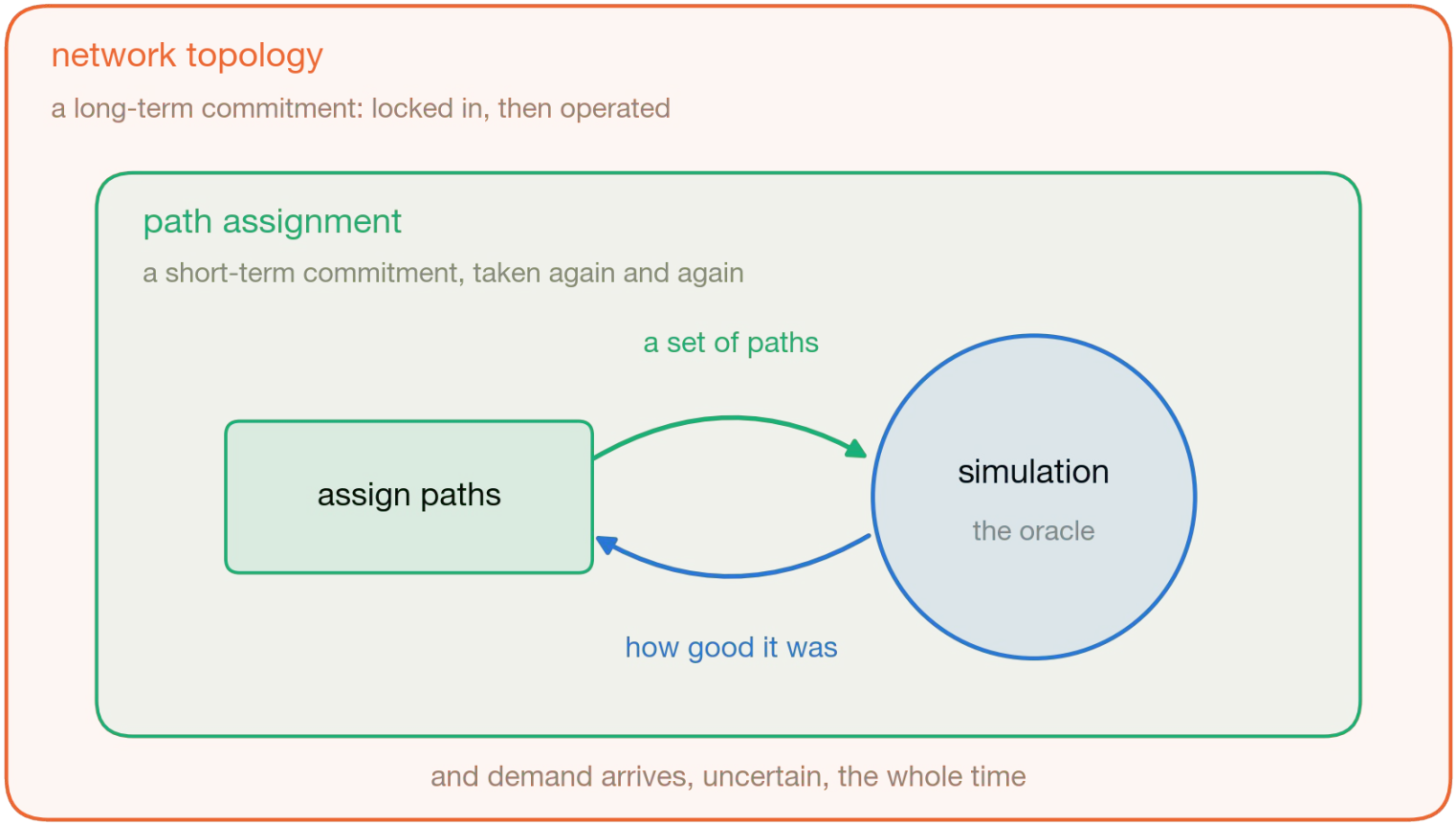


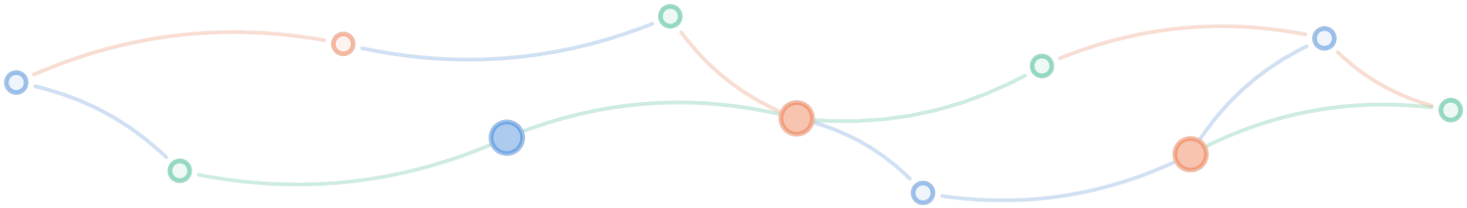
Connecting back

- We evaluated entitlement by a mapping of SKUs to paths, which can be formalized as an assignment
- Complicated network dynamics (e.g. pallet operations) call for simulation to validate our anticipated benefit
- The big picture is a combination of long and medium term commitments and assignments with complicated rewards



Connecting back





Thank you

Questions?

