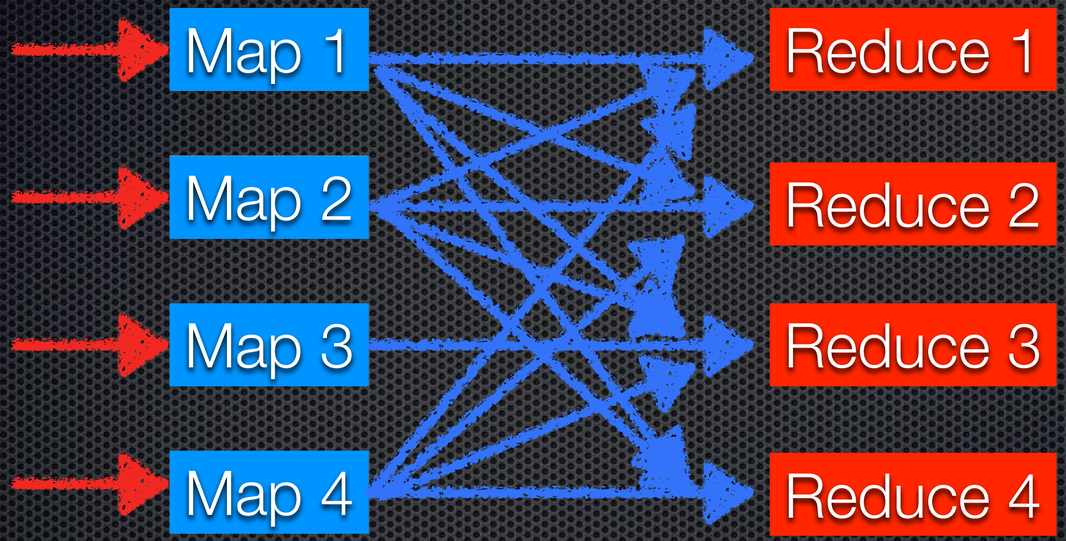
Siphon: Expediting Inter-Datacenter Coflows in Wide-Area Data Analytics

Shuhao Liu, ***Li Chen***, Baochun Li
University of Toronto
July 12, 2018

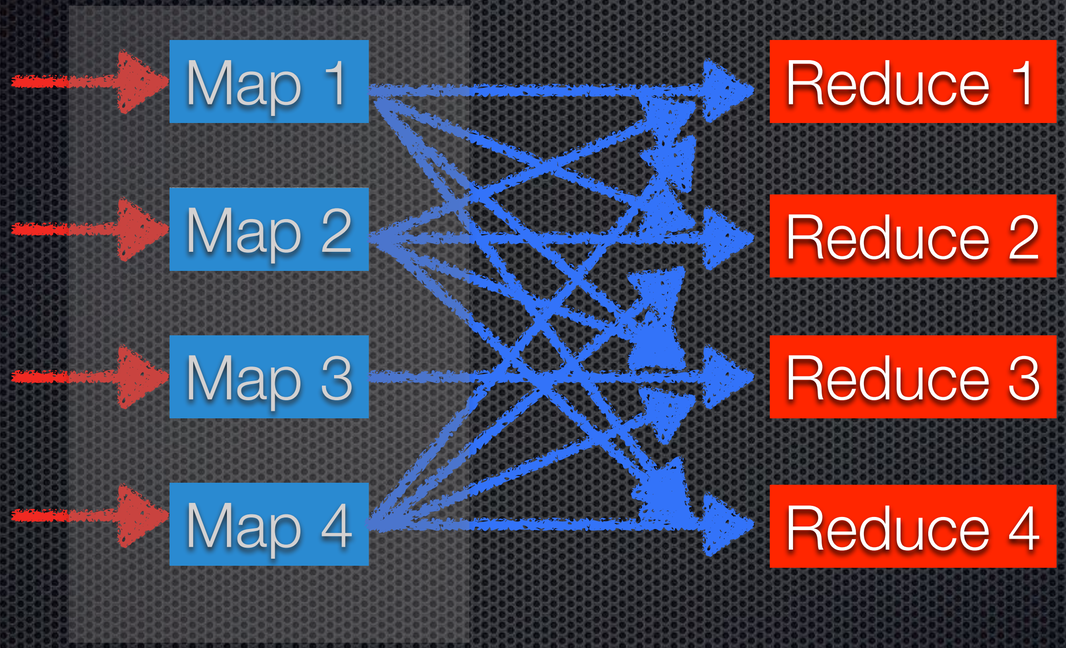
What is a ***Coflow***?

One stage in a data analytic job



What is a ***Coflow***?

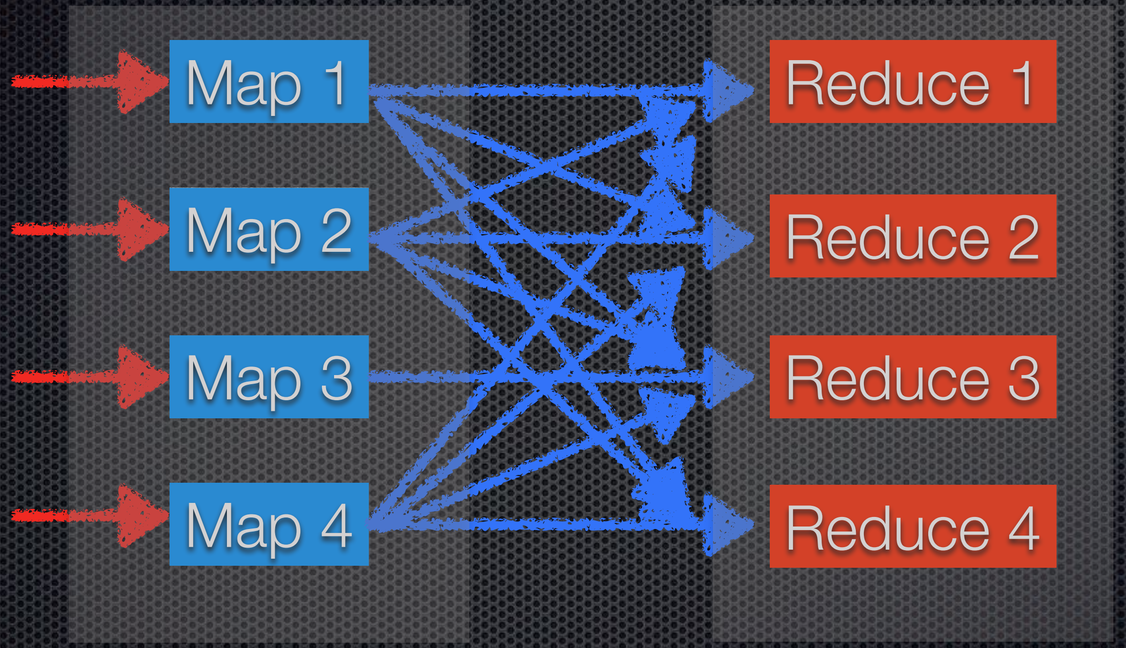
One stage in a data analytic job



Map tasks

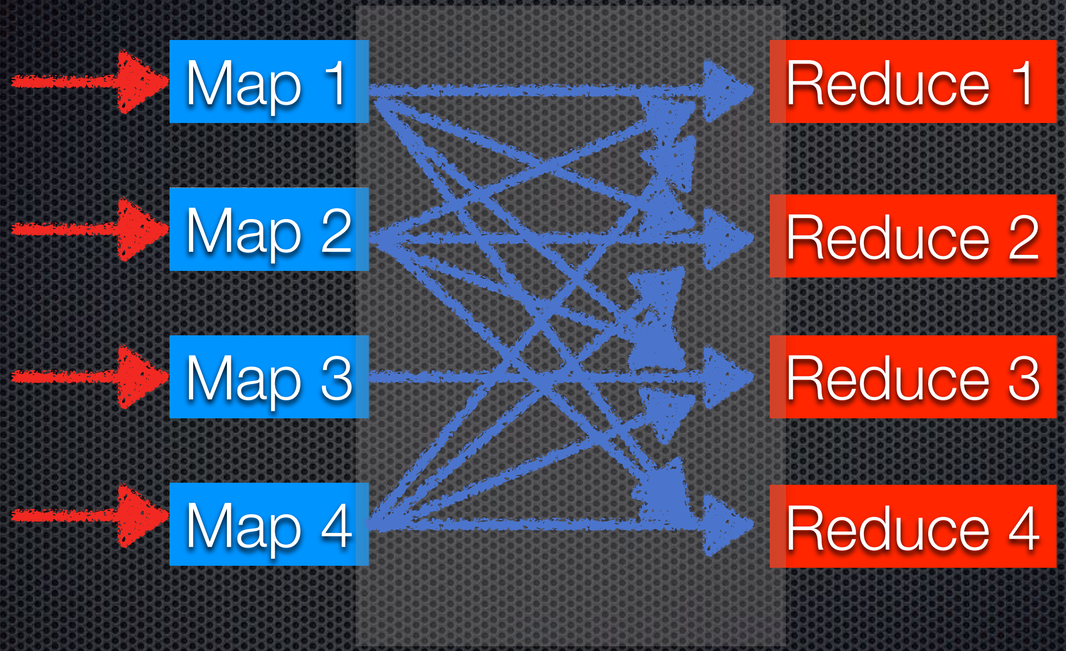
What is a ***Coflow***?

One stage in a data analytic job



What is a ***Coflow***?

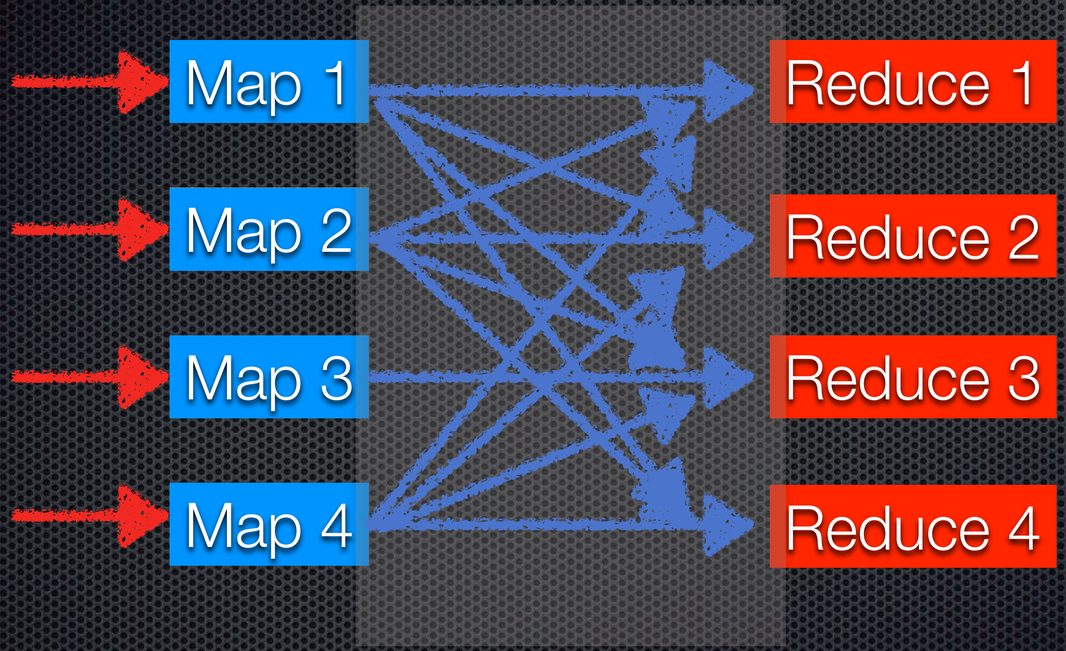
One stage in a data analytic job



all-to-all shuffle

What is a ***Coflow***?

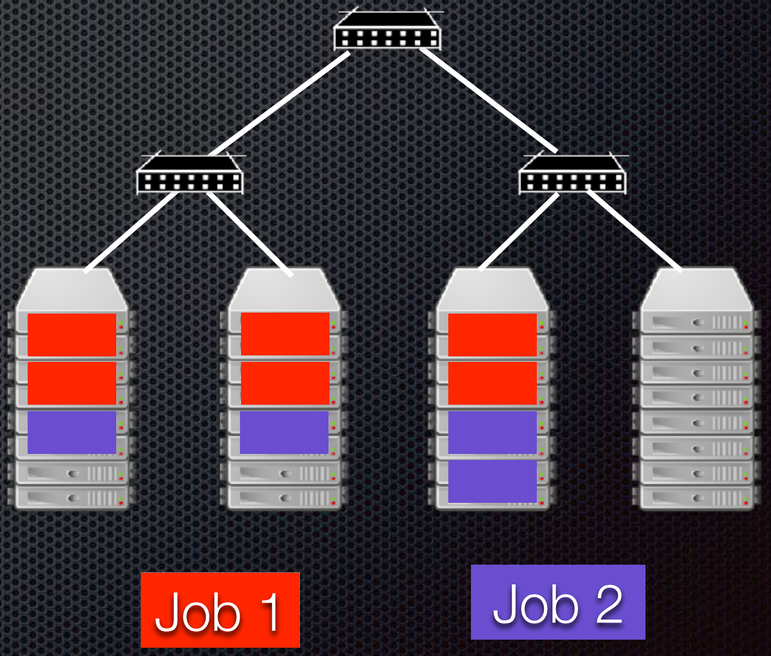
One stage in a data analytic job



Coflow: considered done only when all flows finish

Coflow Scheduling

- Objective: minimizing average coflow completion time
- Network model: datacenter networking
  - Big switch abstraction
    - network core is congestion-free

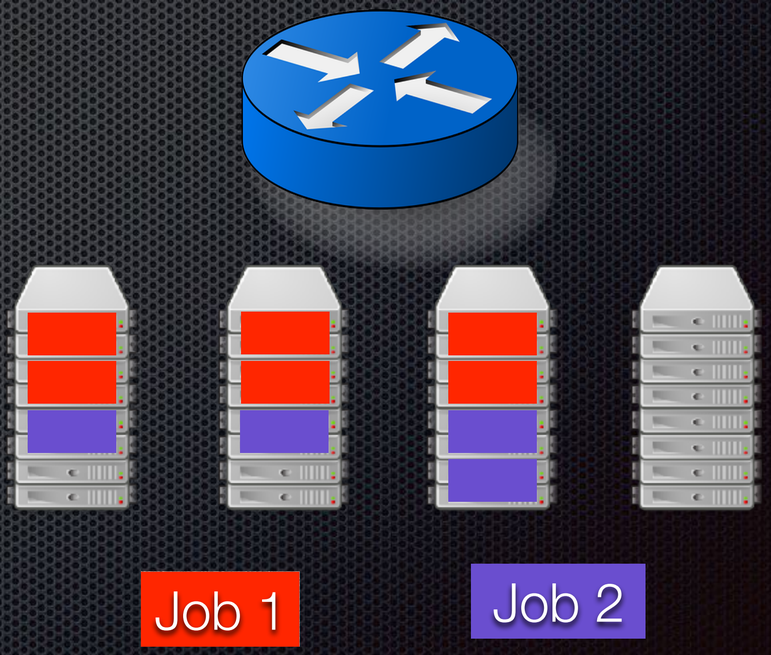


Coflow Scheduling

- Objective: minimizing average coflow completion time
- Network model: datacenter networking
  - Big switch abstraction
    - network core is congestion-free

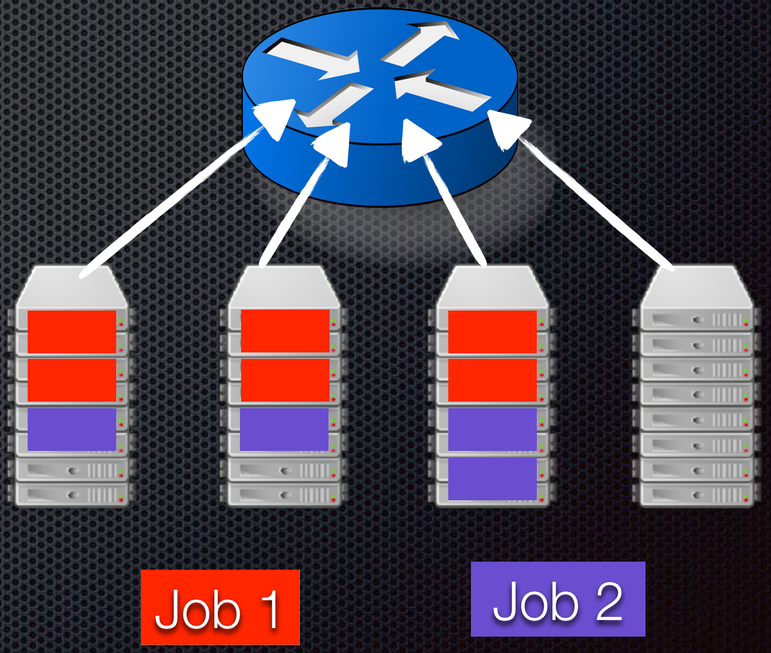
Coflow Scheduling

- Objective: minimizing average coflow completion time
- Network model: datacenter networking
  - Big switch abstraction
    - network core is congestion-free



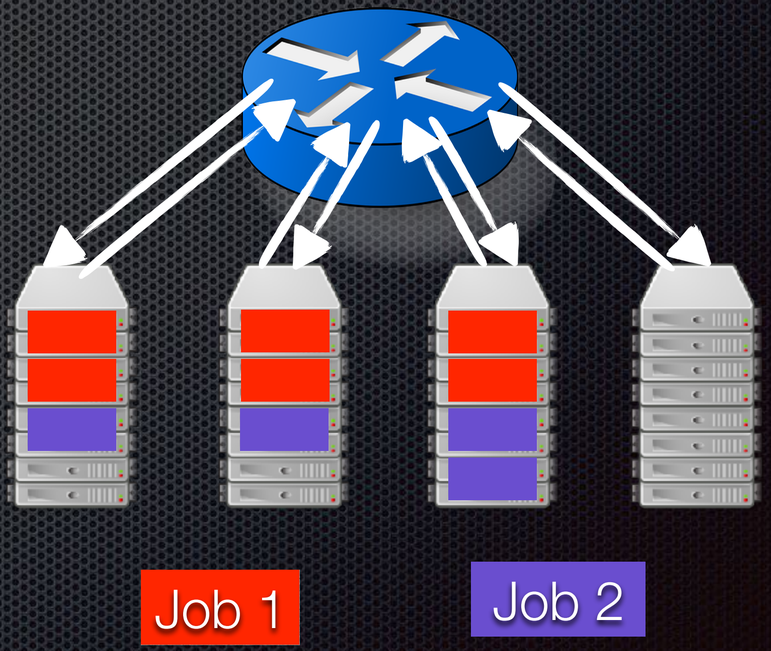
Coflow Scheduling

- Objective: minimizing average coflow completion time
- Network model: datacenter networking
  - Big switch abstraction
    - network core is congestion-free



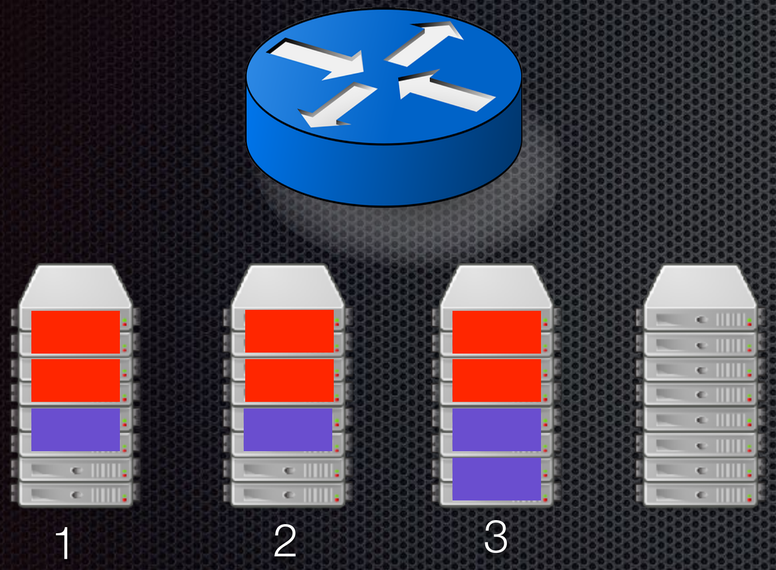
Coflow Scheduling

- Objective: minimizing average coflow completion time
- Network model: datacenter networking
  - Big switch abstraction
    - network core is congestion-free



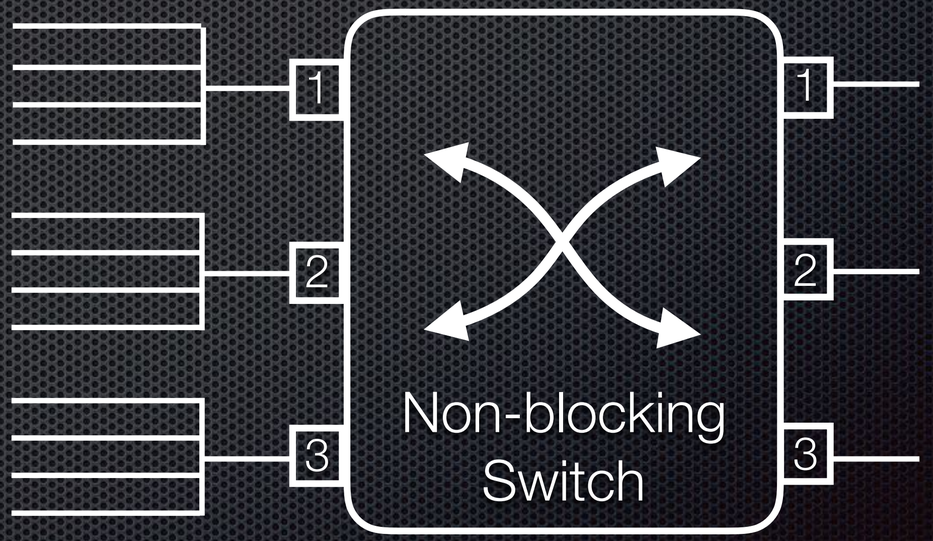
Coflow Scheduling

- Objective: minimizing average coflow completion time



Job 1

Job 2

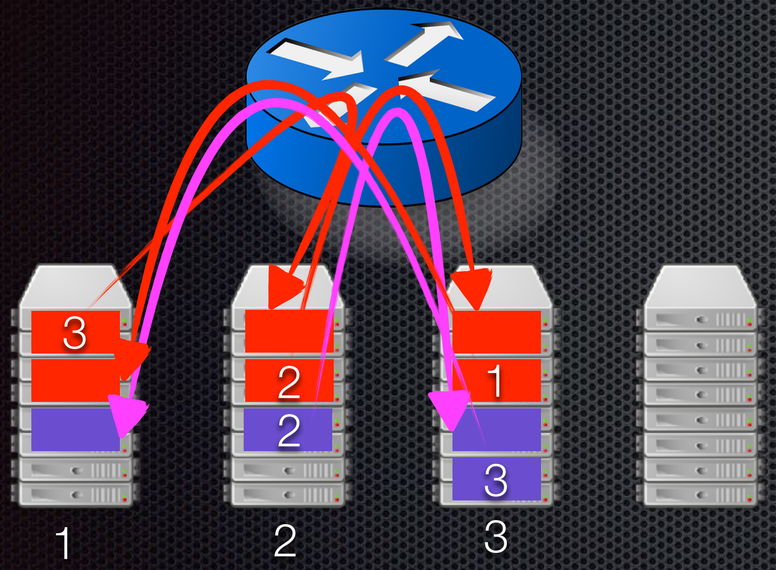


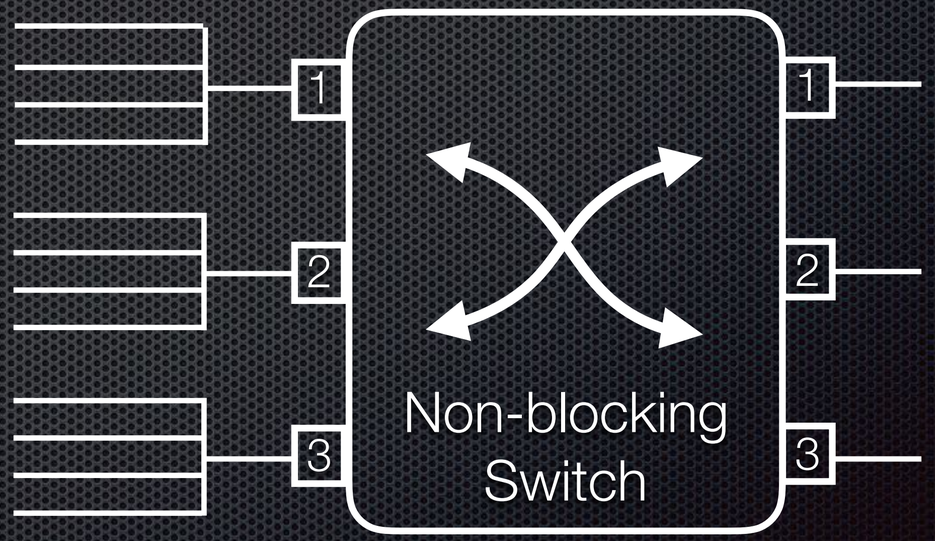
Coflow 1

Coflow 2

Coflow Scheduling

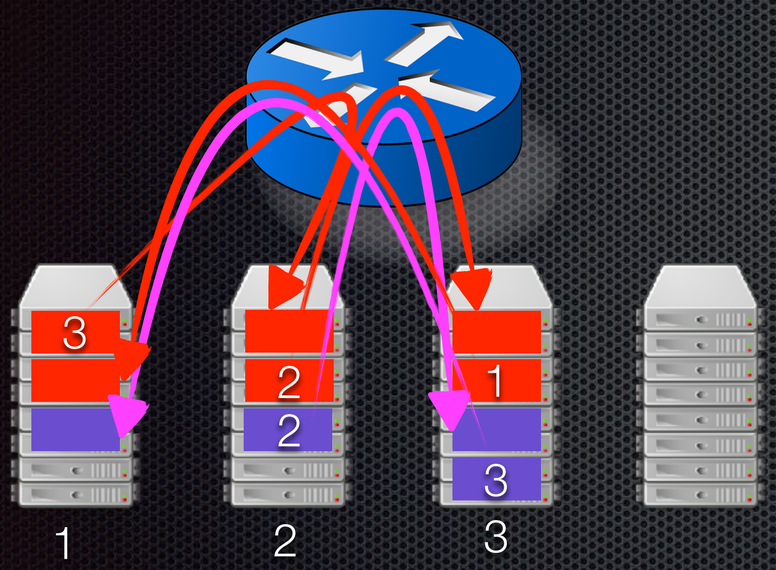
- Objective: minimizing average coflow completion time

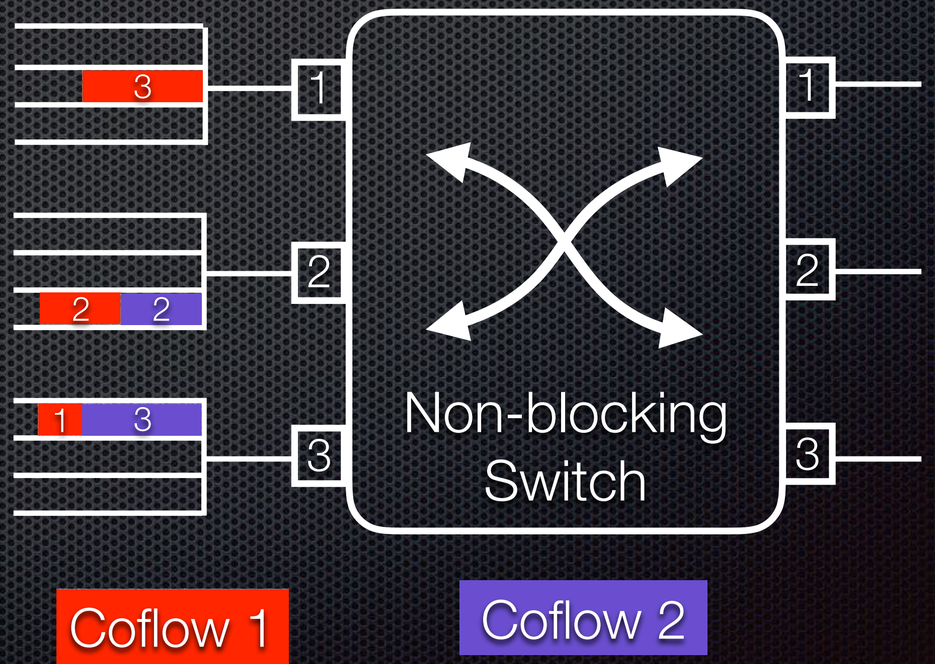




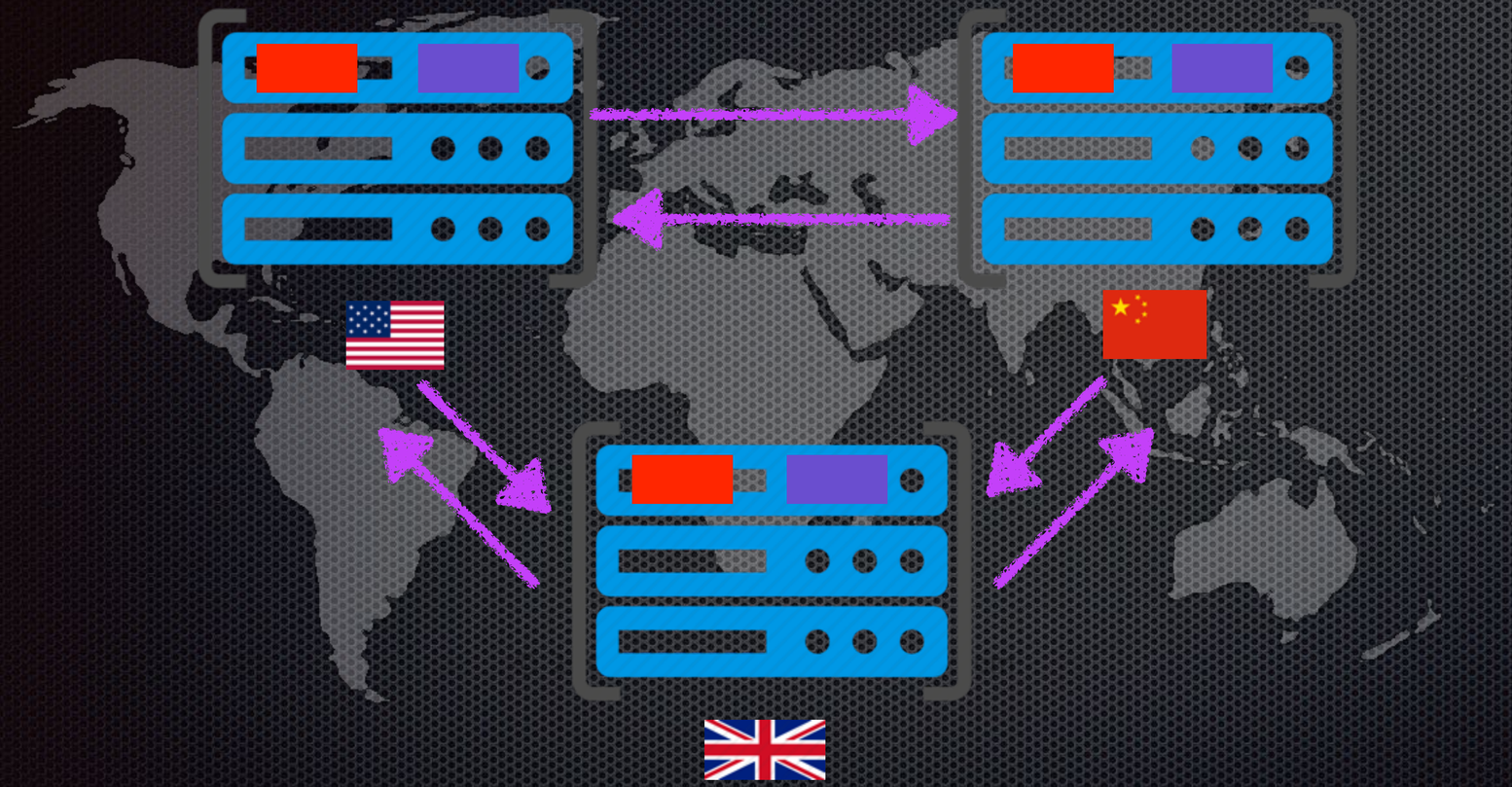
Coflow Scheduling

- Objective: minimizing average coflow completion time

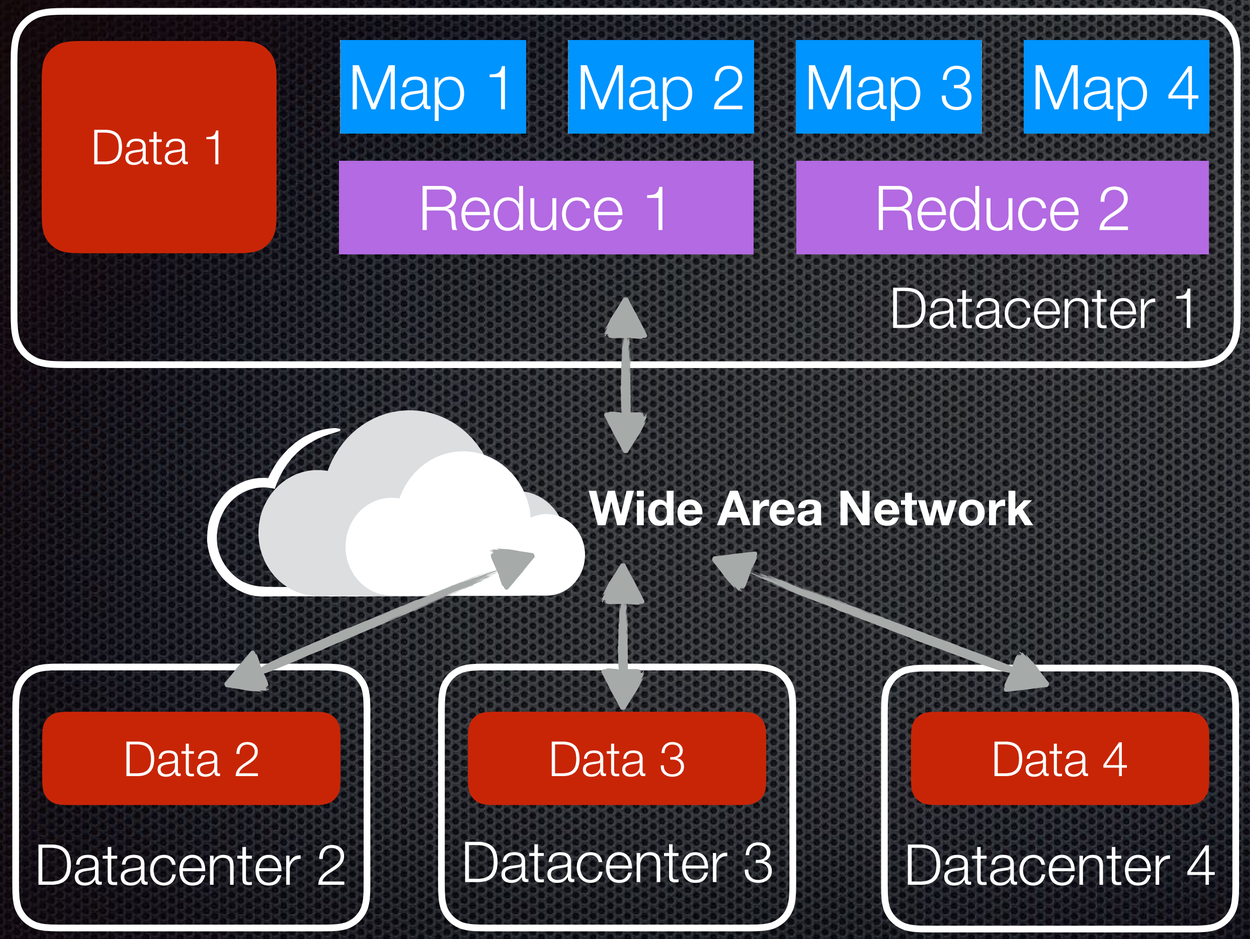




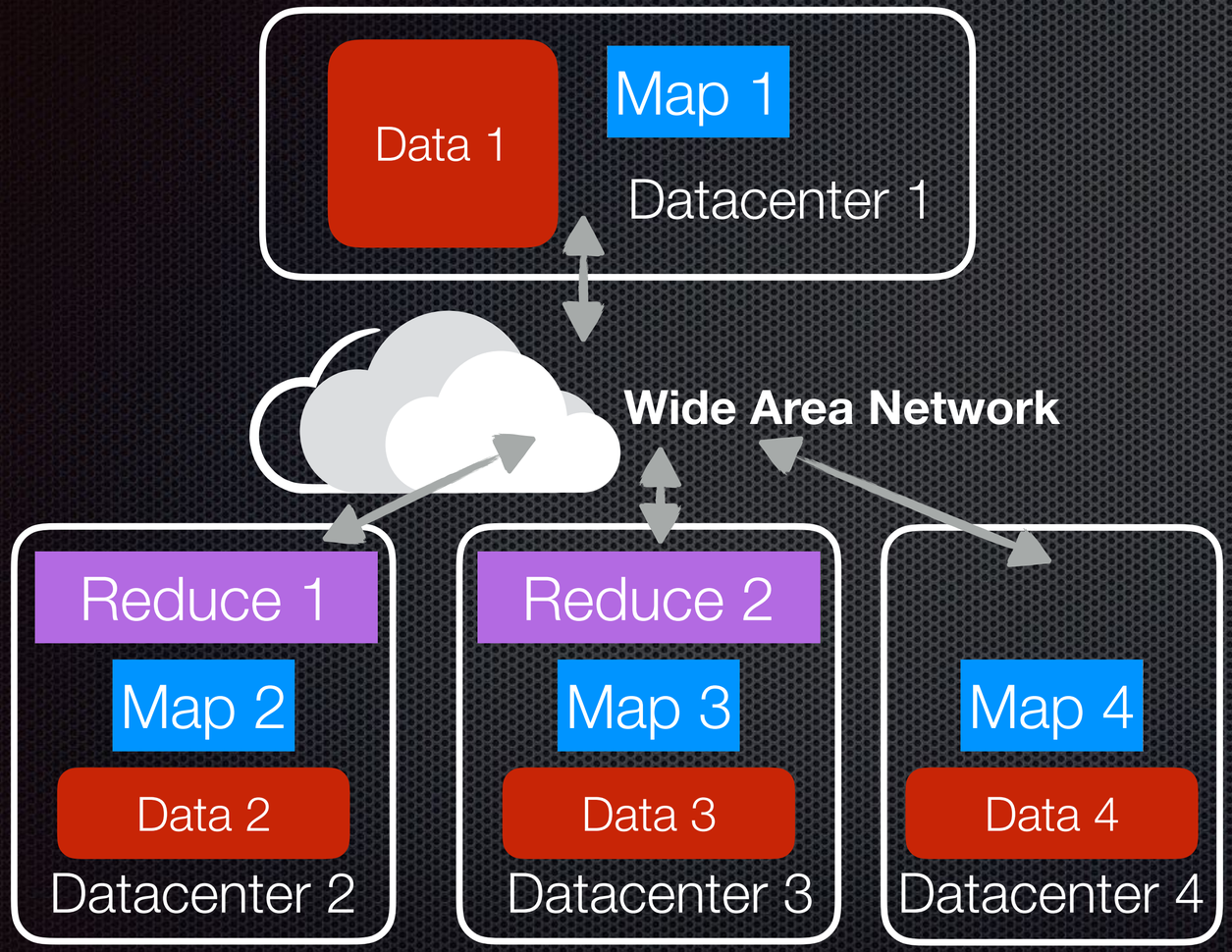
Wide-Area Data Analytics



Wide-Area Data Analytics



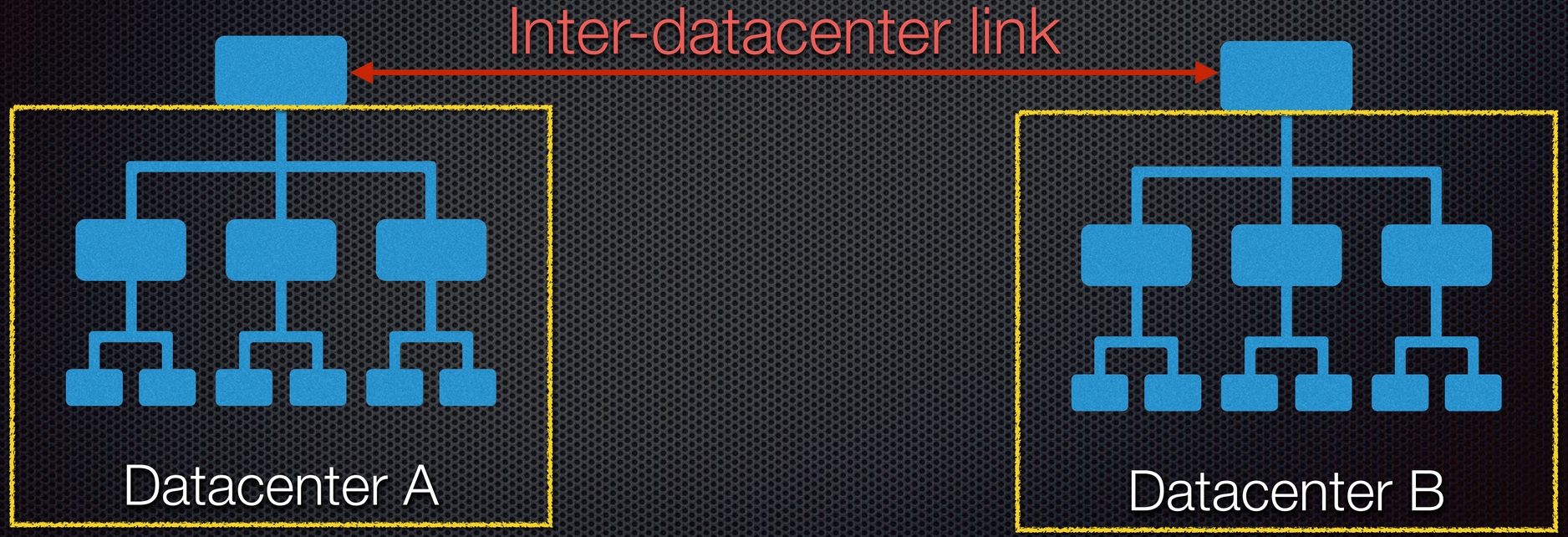
Wide-Area Data Analytics



With tasks placed in different datacenter, what about their generated **inter-datacenter coflows**?

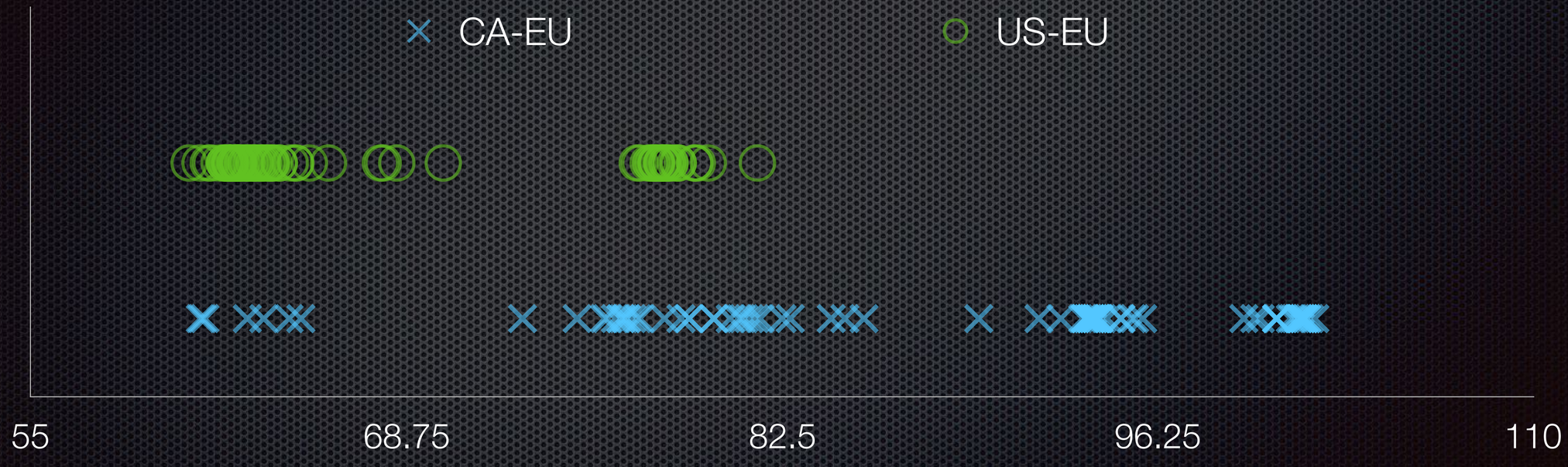
Challenges

- Dumb bell network model: inter-datacenter links are the **only** bottleneck

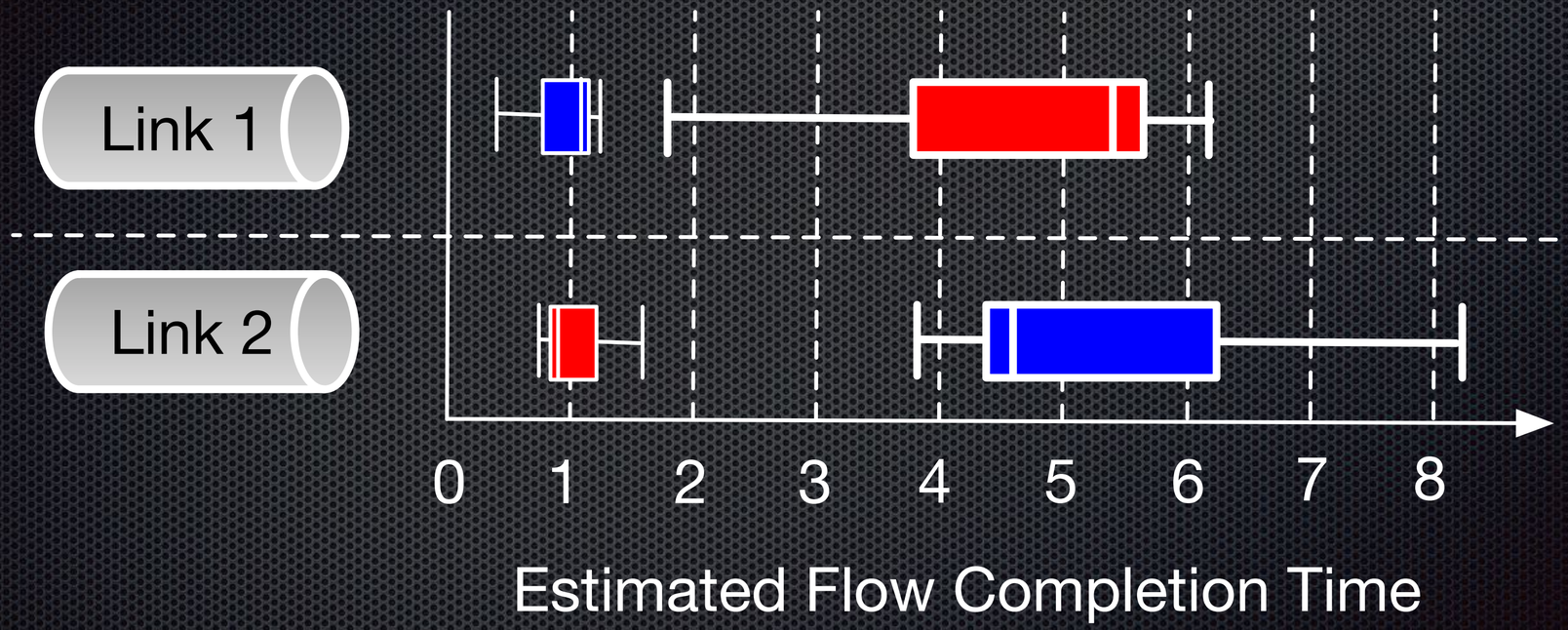


Challenges

- Constantly **changing** available bandwidth



Can existing heuristics work?



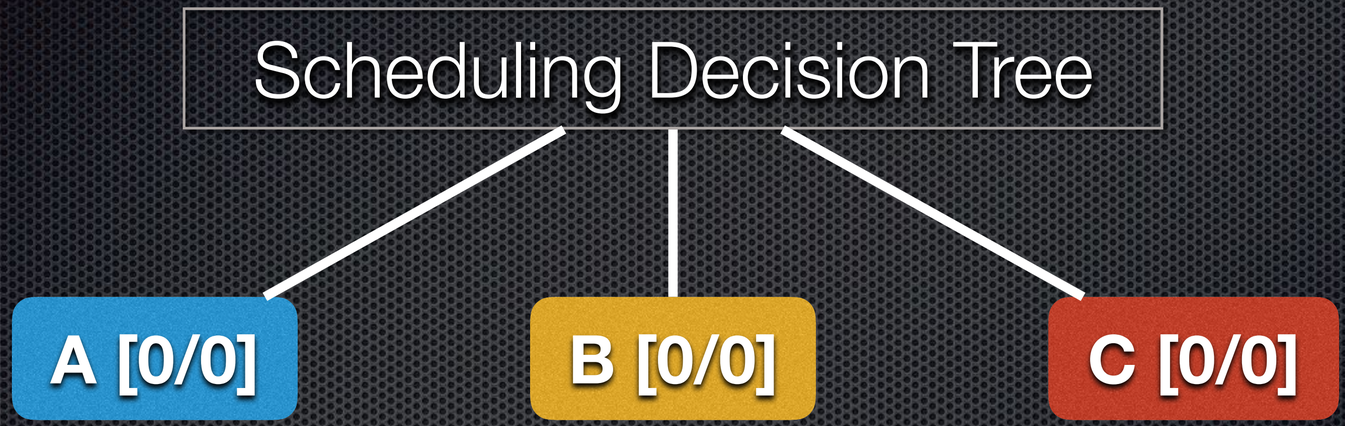
Coflow scheduling should consider the distribution of available bandwidth.

Monte Carlo Simulation

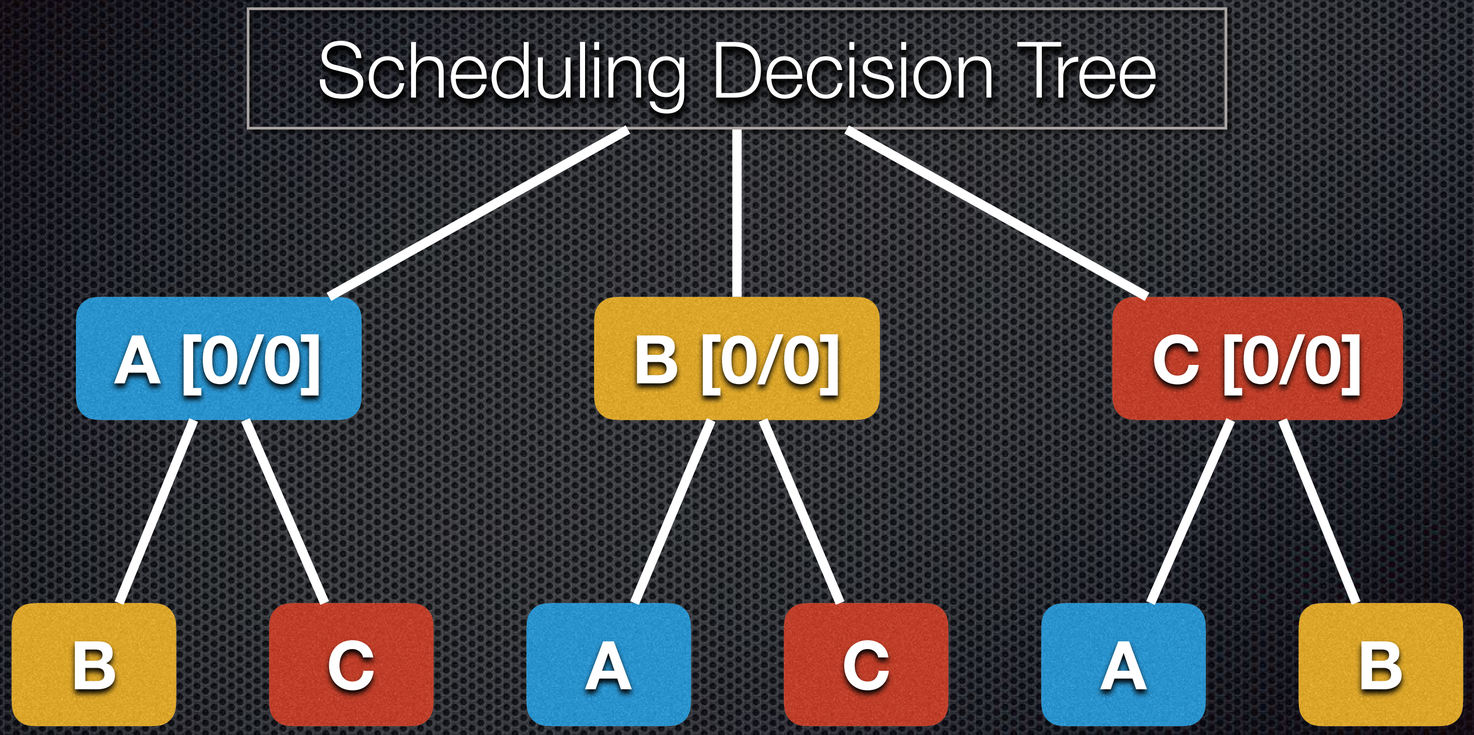
Monte Carlo Simulation

Scheduling Decision Tree

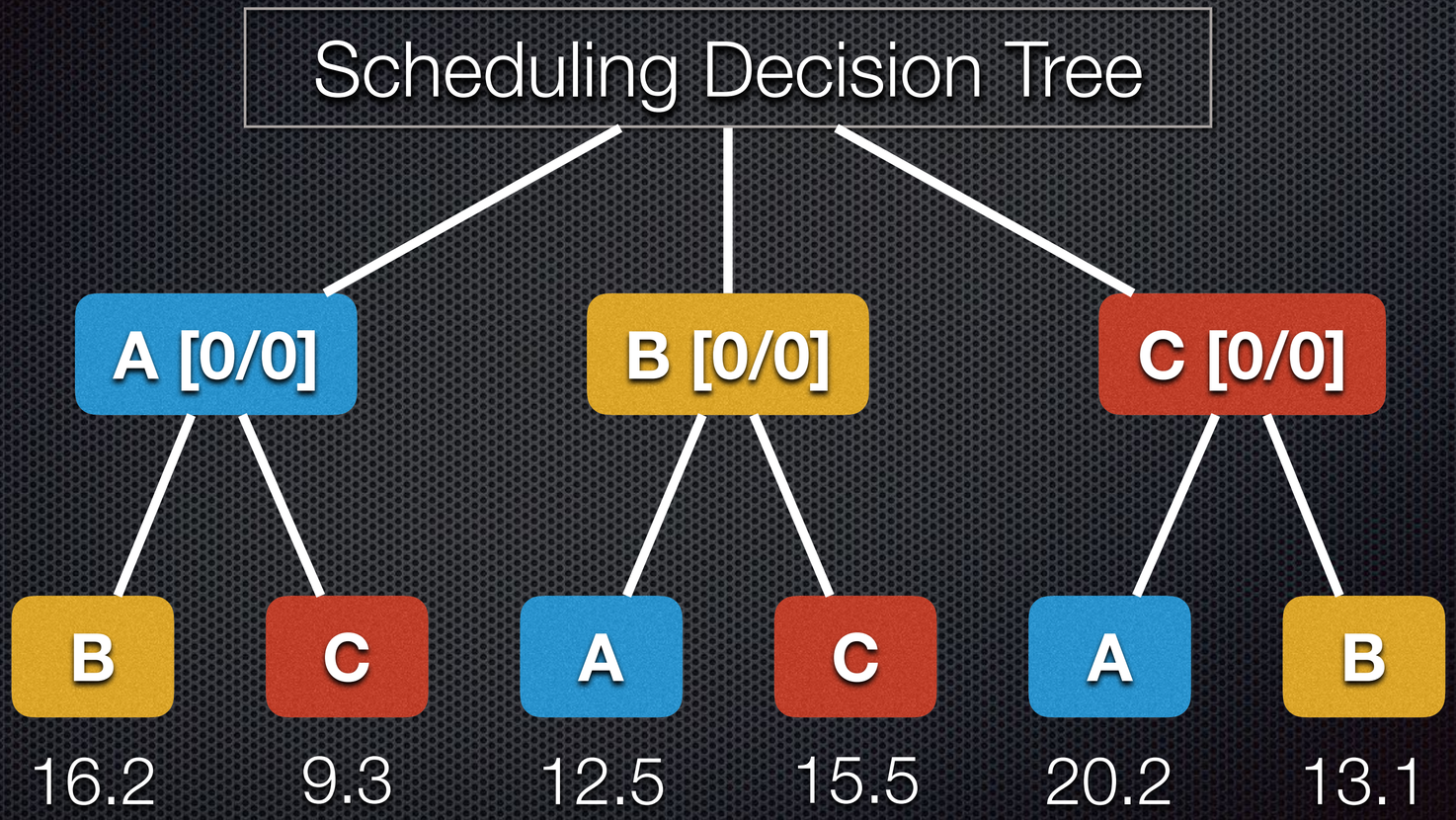
Monte Carlo Simulation



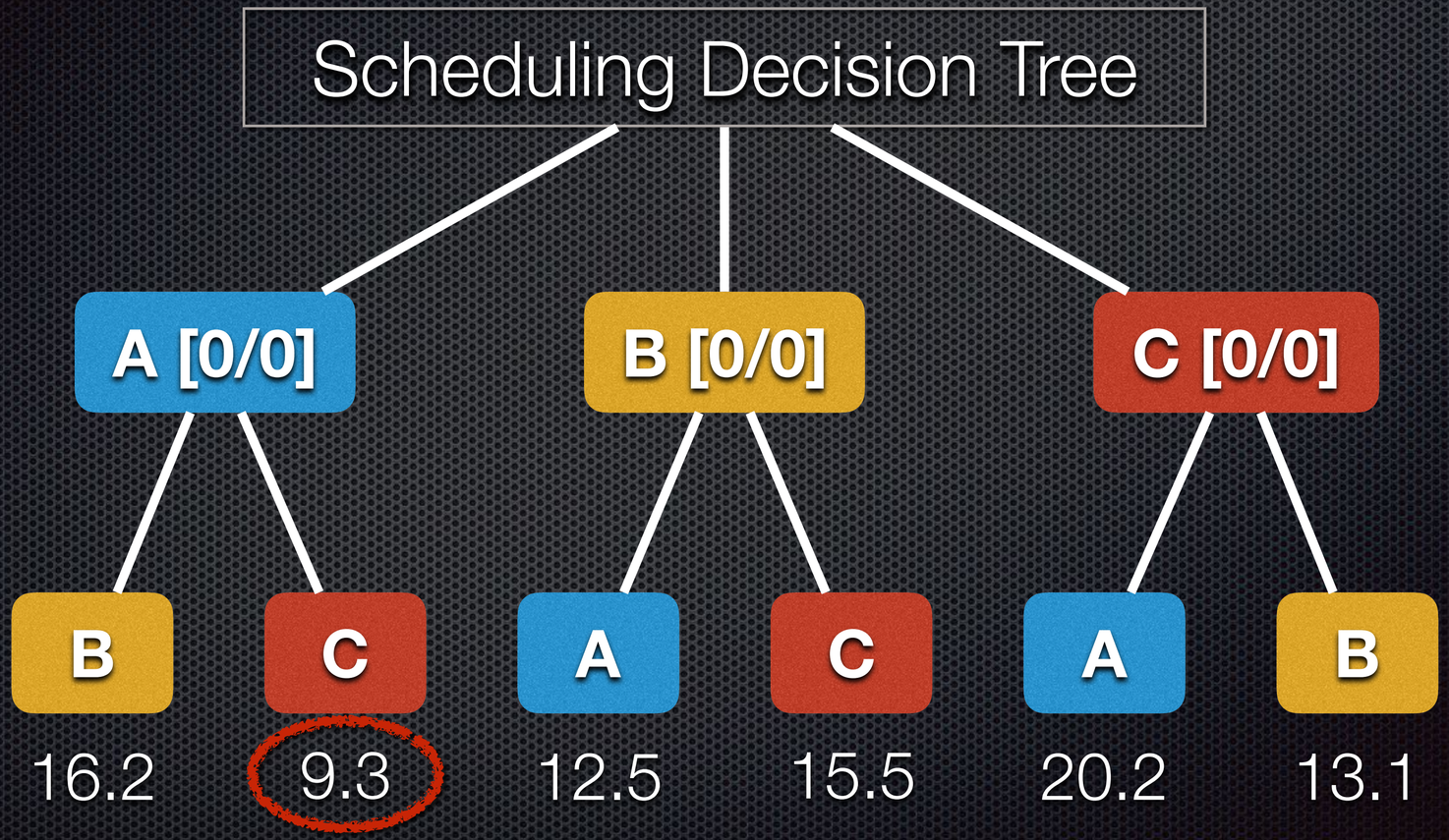
Monte Carlo Simulation



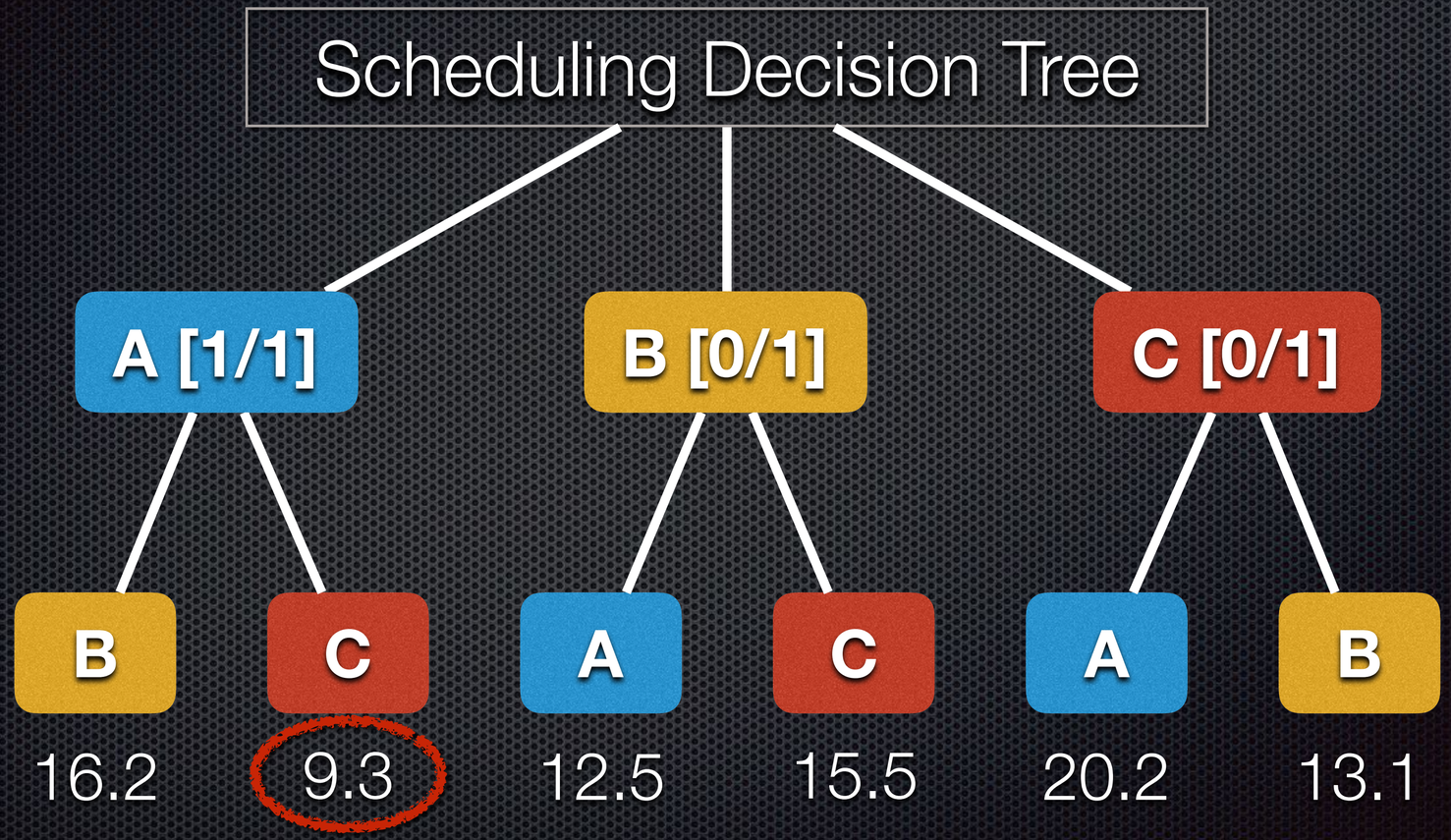
Monte Carlo Simulation



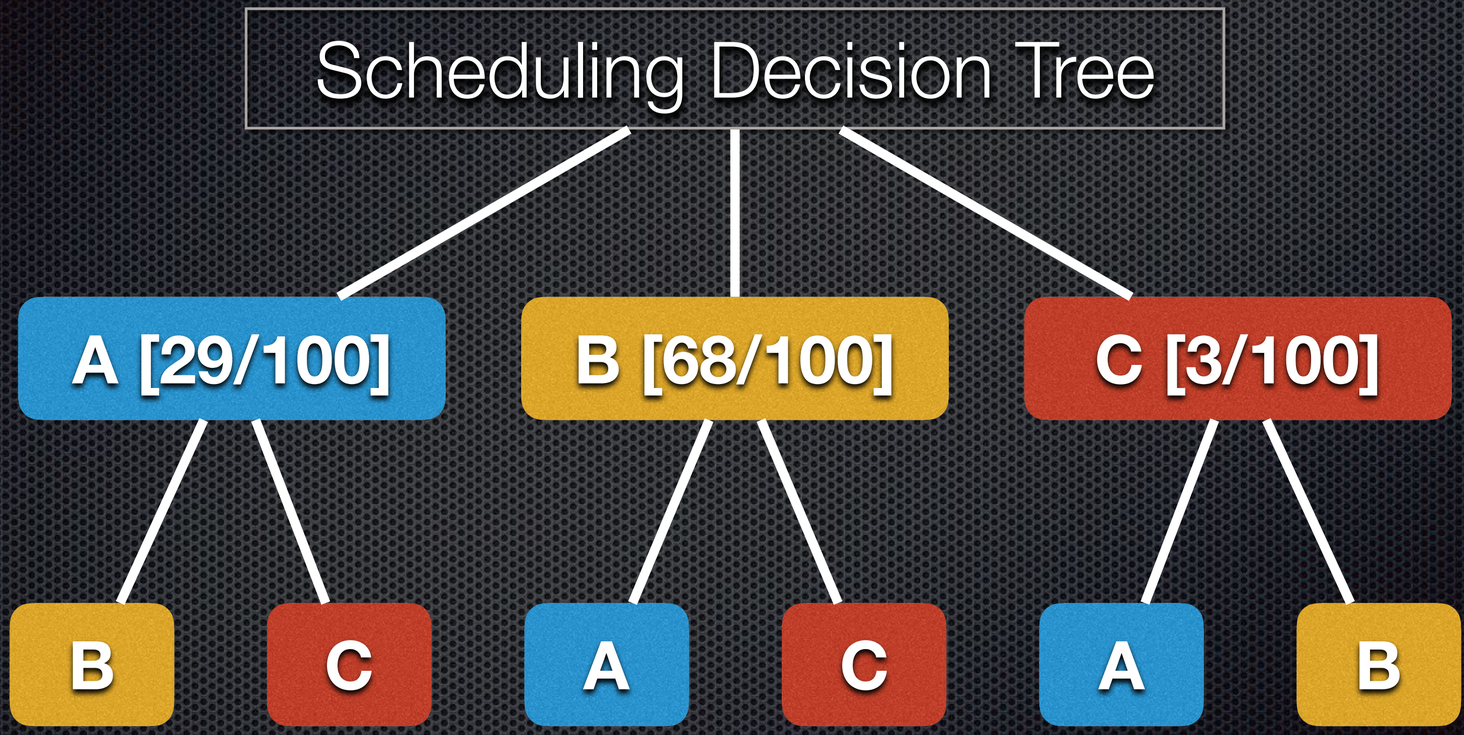
Monte Carlo Simulation



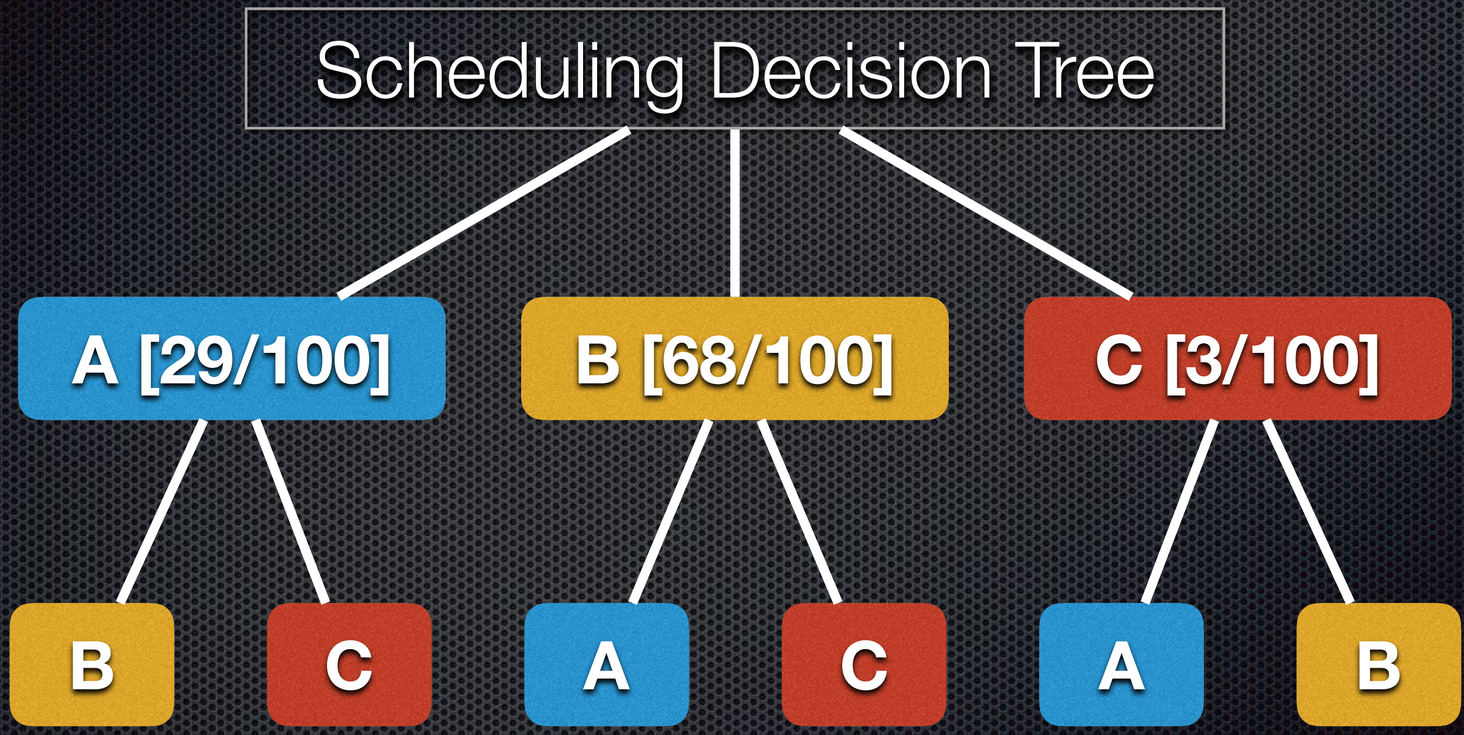
Monte Carlo Simulation



Monte Carlo Simulation



Monte Carlo Simulation



Complexity? 100 * O(n!)

Reduced Simulation Complexity

Reduced Simulation Complexity

Bounded Search Depth

$$\Theta(t \times n^d)$$

Reduced Simulation Complexity

Bounded Search Depth

$$\Theta(t \times n^d)$$

Reduced Search Breath (Early termination)

Reduced Simulation Complexity

Bounded Search Depth

$$\Theta(t \times n^d)$$

Reduced Search Breath (Early termination)

$$O(t \times n^d)$$

Reduced Simulation Complexity

Bounded Search Depth $\Theta(t \times n^d)$

Reduced Search Breath (Early termination) $O(t \times n^d)$

Online Incremental Search

Reduced Simulation Complexity

Bounded Search Depth $\Theta(t \times n^d)$

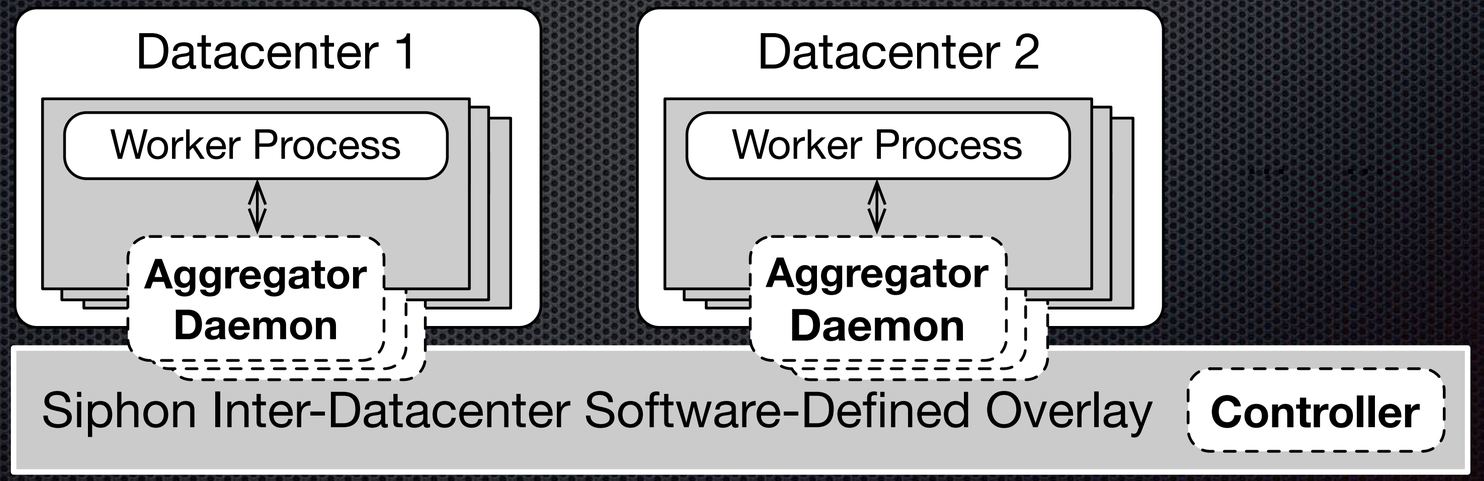
Reduced Search Breath (Early termination) $O(t \times n^d)$

Online Incremental Search $O(t \times n^{d-1})$

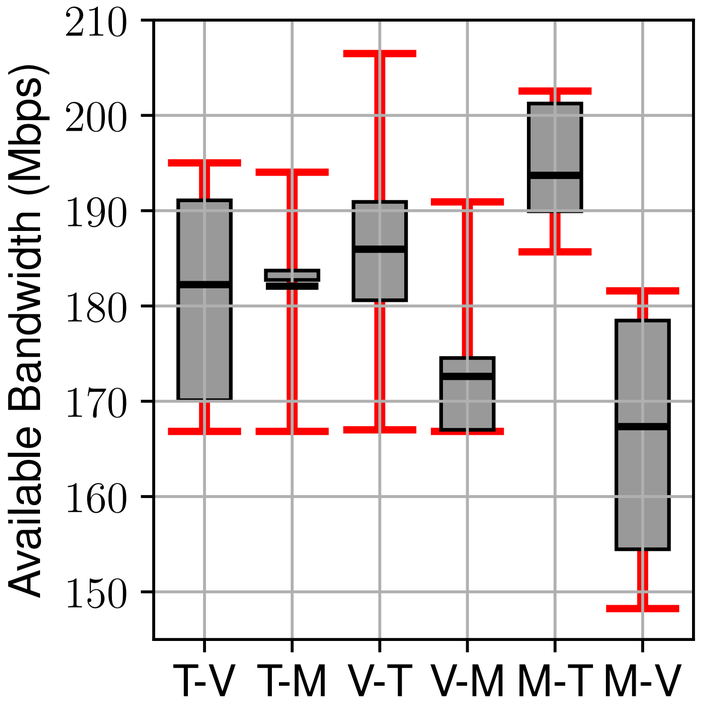
How to enforce the scheduling decisions?

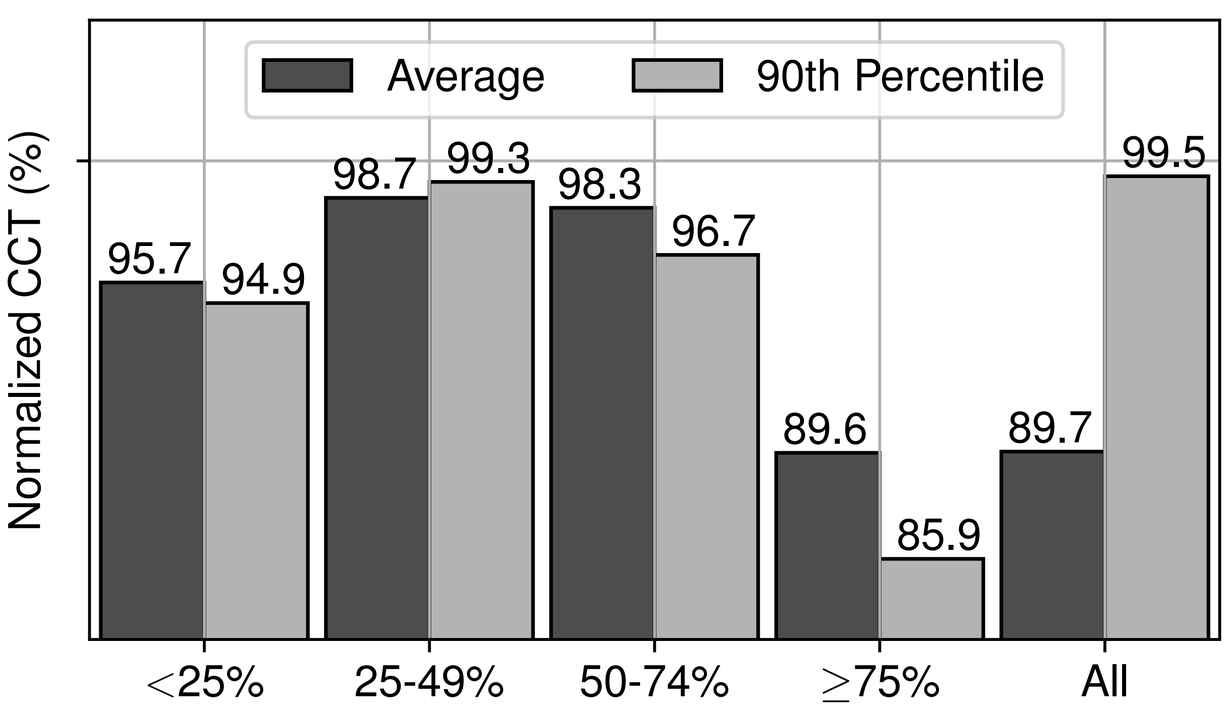
Siphon: System Overview

- Form a software-defined overlay network
  - Aggregators: measure bandwidth, schedule coflows based on priority assignment
  - Controller: compute priority based on Monte Carlo simulation



Performance: Coflow Scheduling

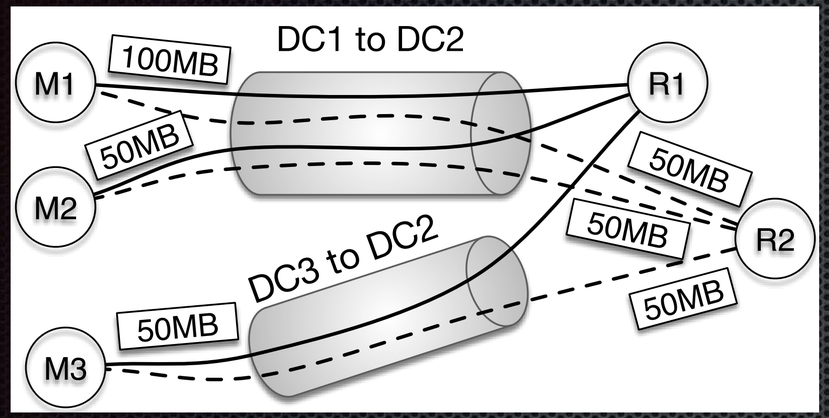


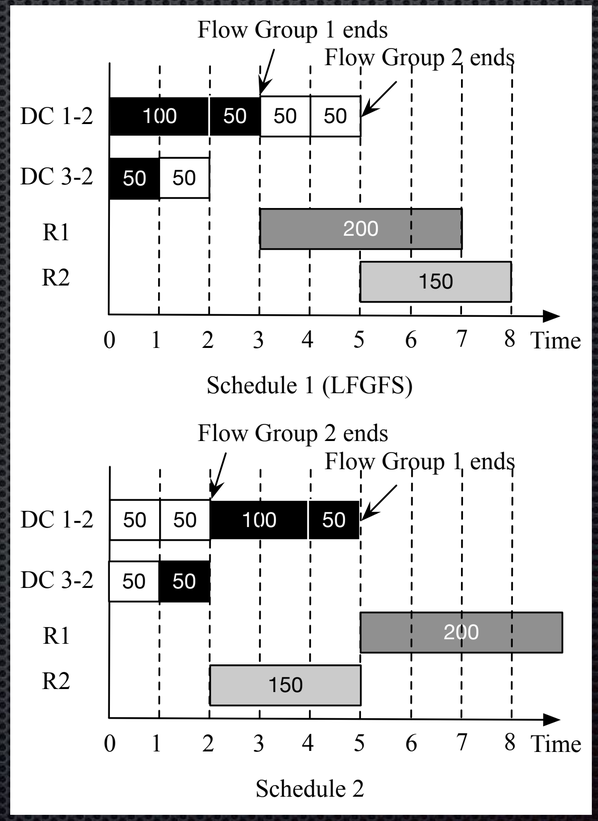


With Siphon, we can do more…

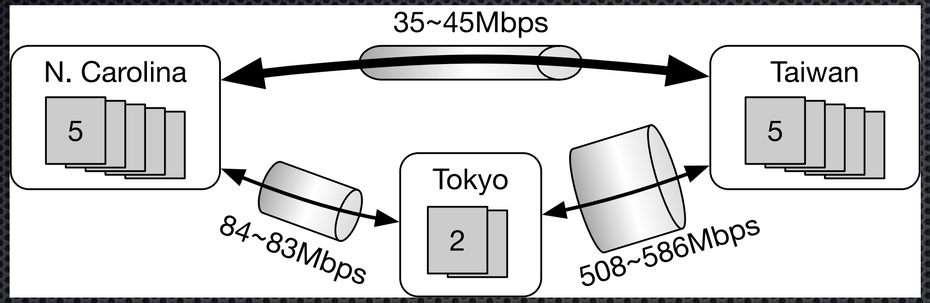
Intra-Coflow Scheduling

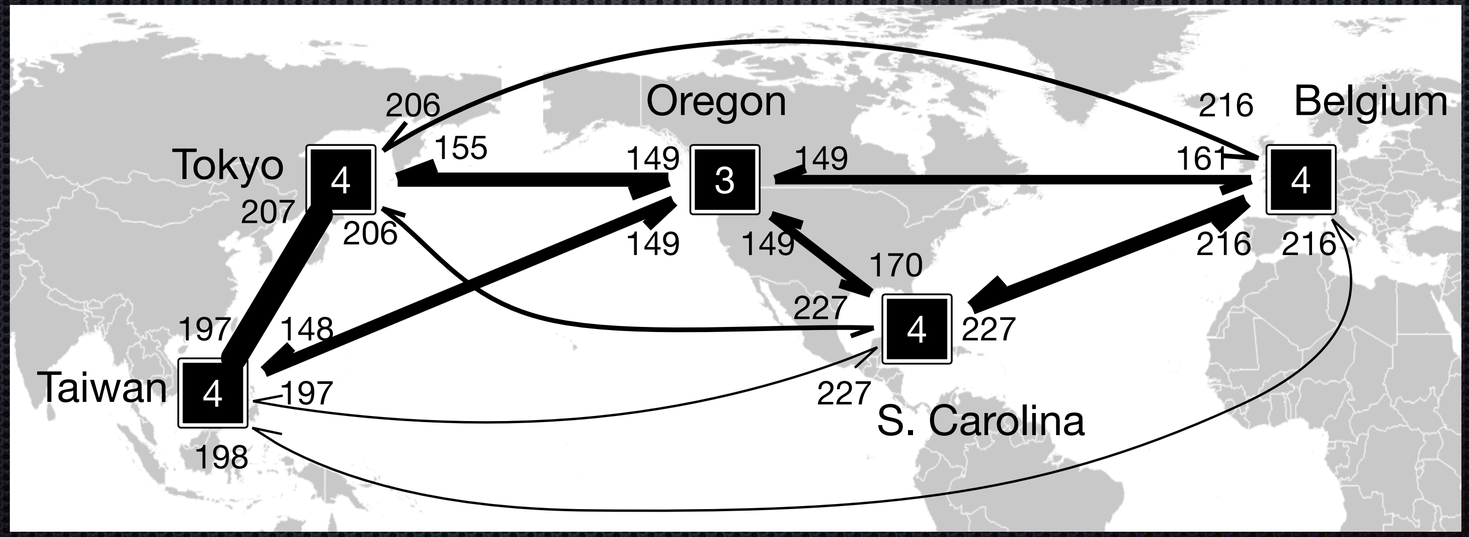
- Largest Flow Group First



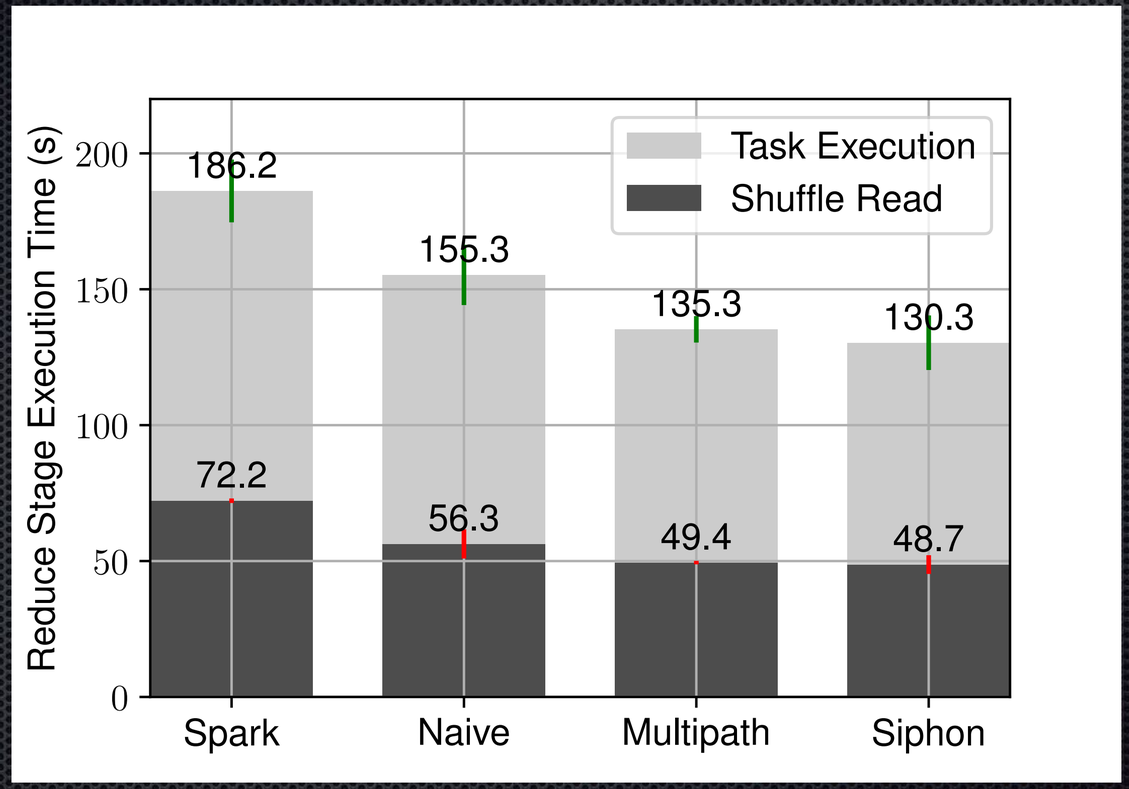


Coflow Multipath Routing

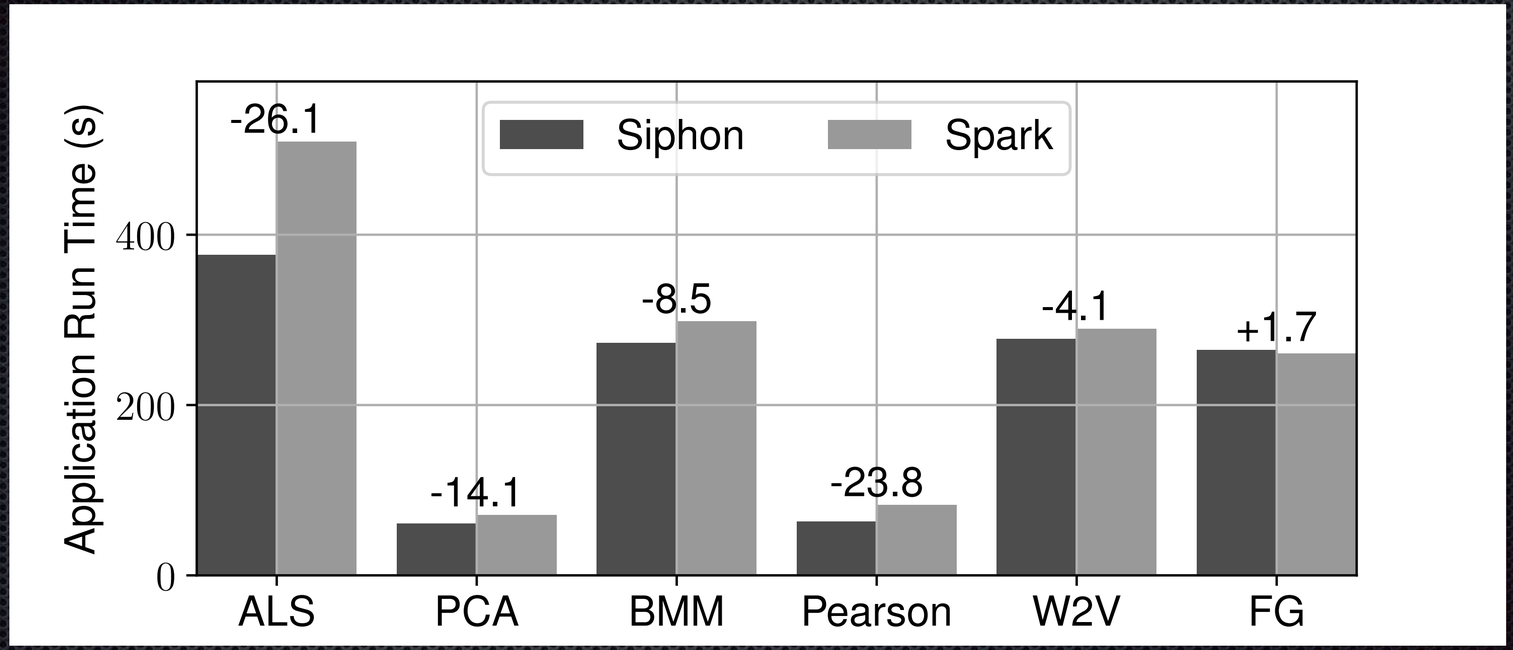




Performance: Intra-Coflow Scheduling



Performance: Benchmark Workloads



Takeaway

- Siphon is a software-defined inter-datacenter overlay that realizes:
  - Coflow scheduling in wide-area data analytics: Monte Carlo Simulation
  - Intra-coflow scheduling: Largest Flow Group First
  - Coflow multipath rerouting
- Shorter coflow completion time leads to better job-level performance

Thank you!