Topology Switching for Data Center Networks

Kevin Webb, Alex Snoeren, Ken Yocum

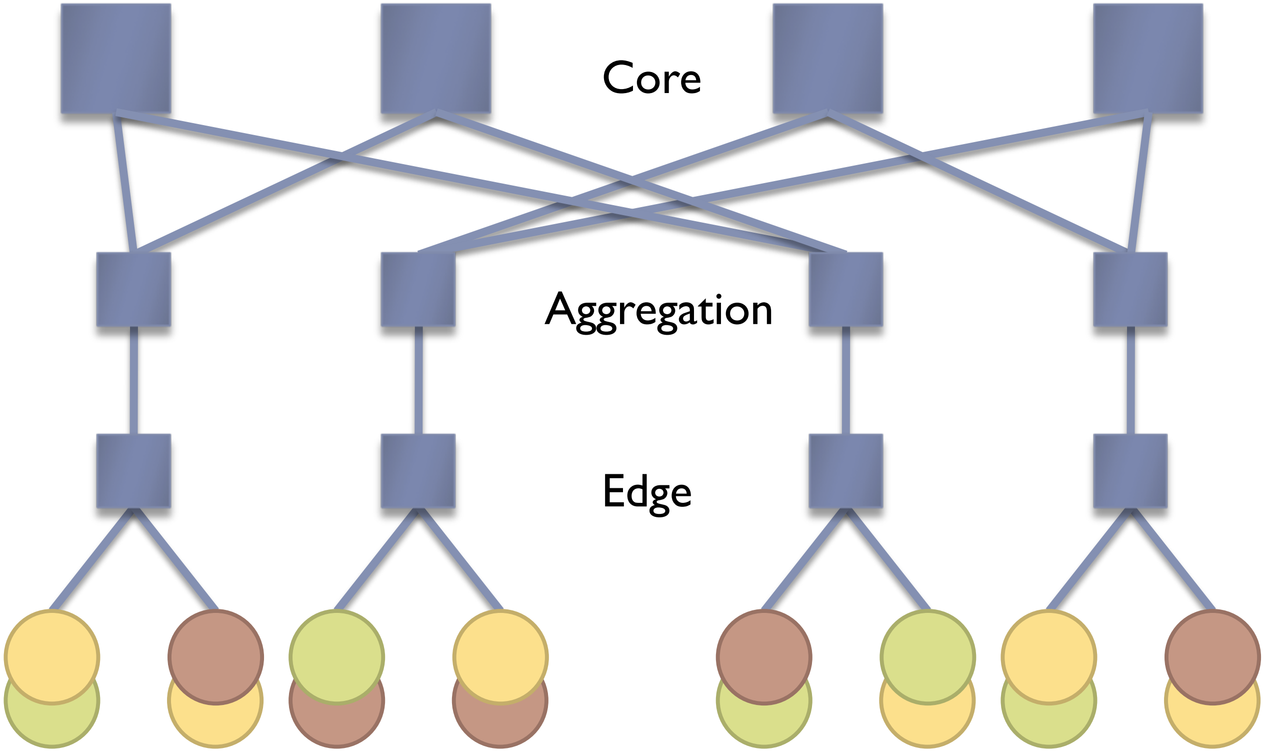
UC San Diego Computer Science

March 29, 2011
Hot-ICE 2011

UCSDCSE
Computer Science and Engineering

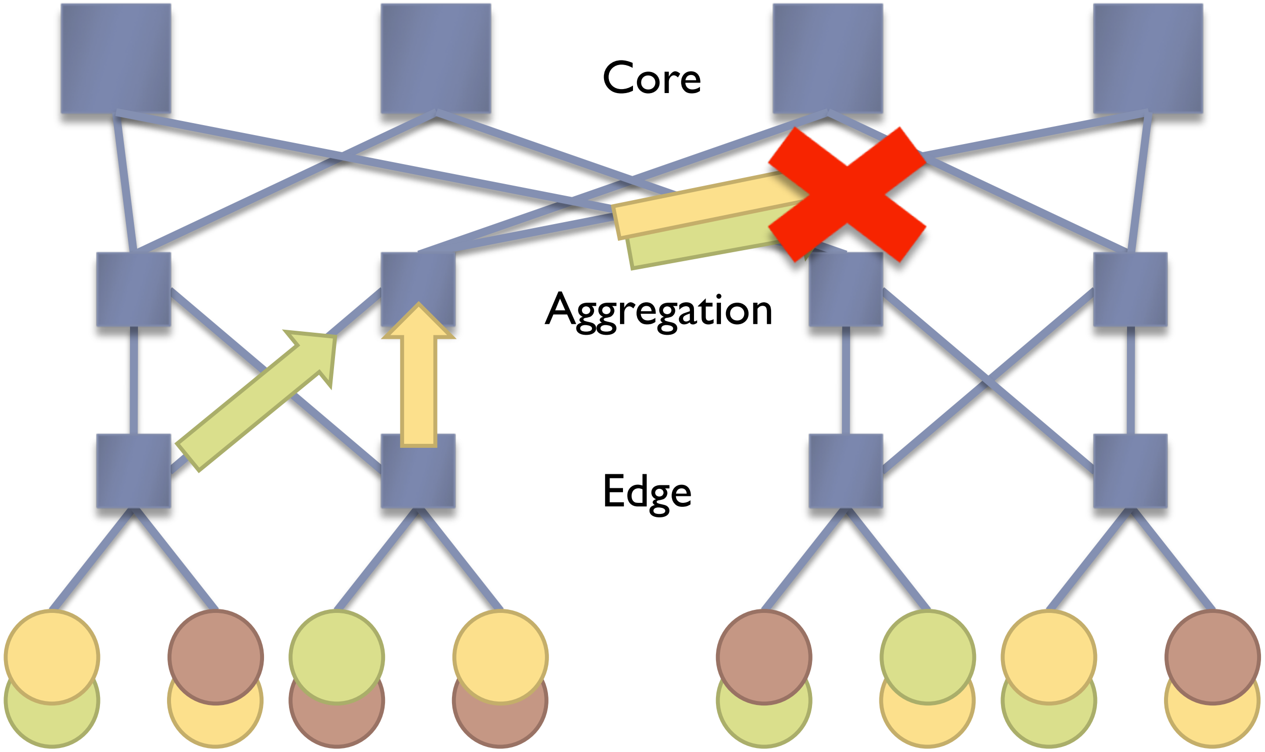
Data Center Networks

- Hosting myriad of applications:
  - Big data: MapReduce
  - Web services
  - HPC: MPI
  - DB, Storage
  - Many others!

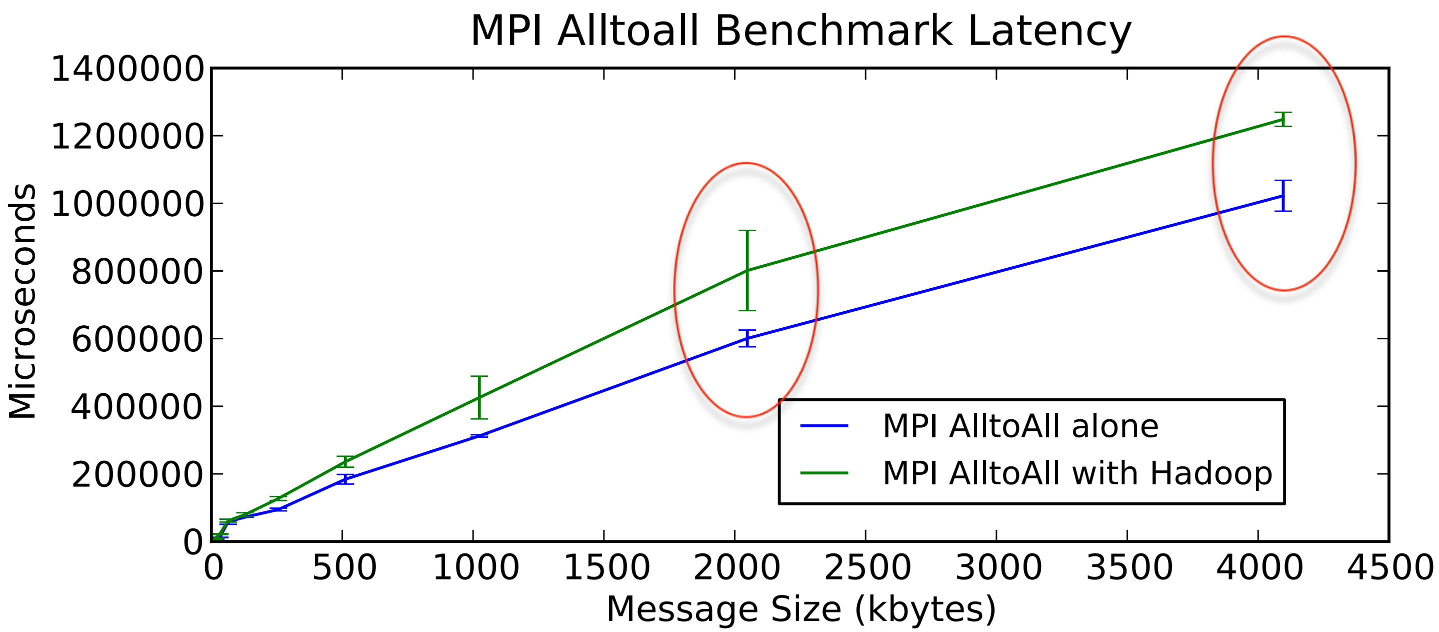


Data Center Networks

- DC engineers adding links
  - Applications need other important characteristics!



Inter-application Interference



- Experiment
  - All to all MPI and Hadoop data processing
  - Openflow ECMP network
  - Interference > 20% latency increase

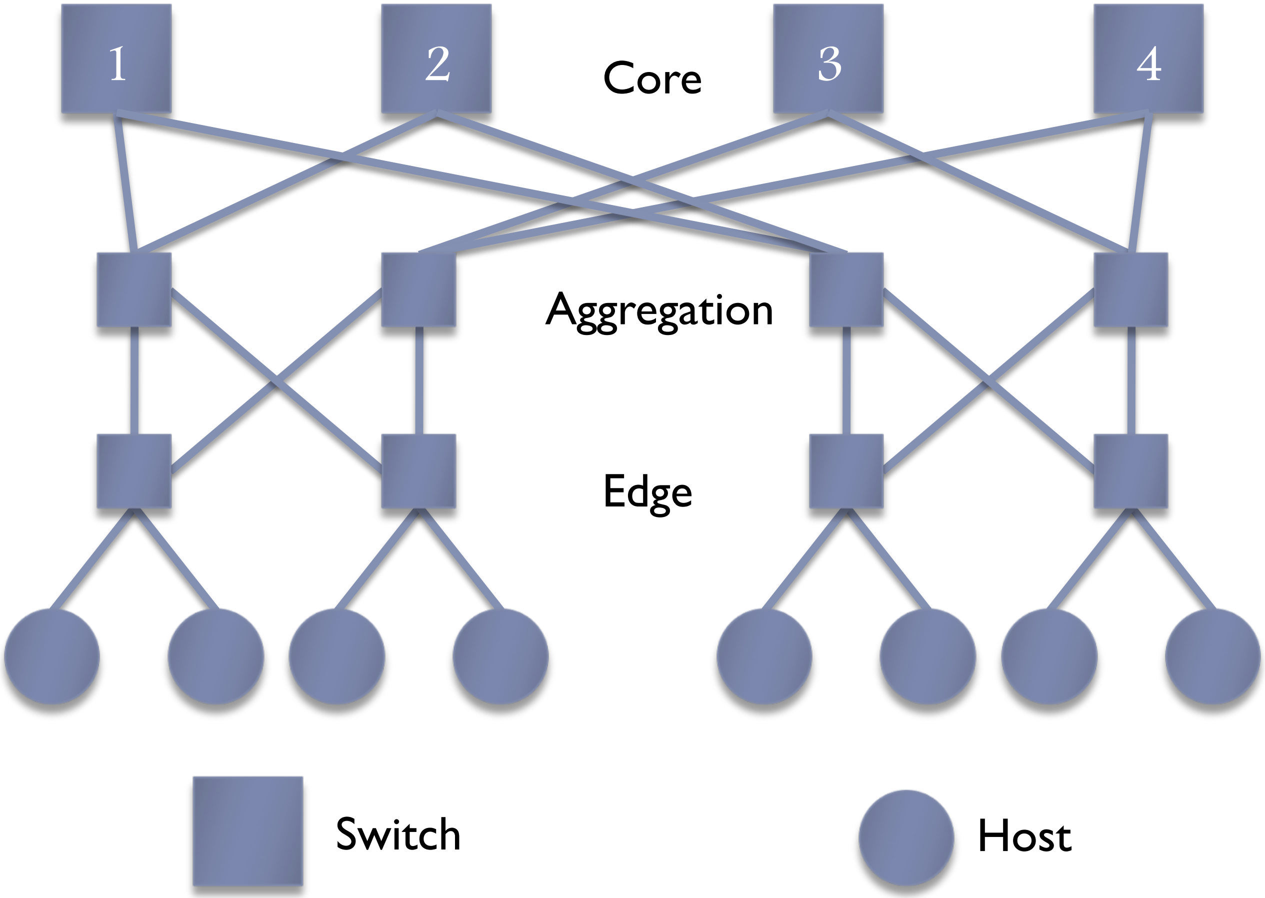
UCSDCSE
Computer Science and Engineering

Topology Switching Overview

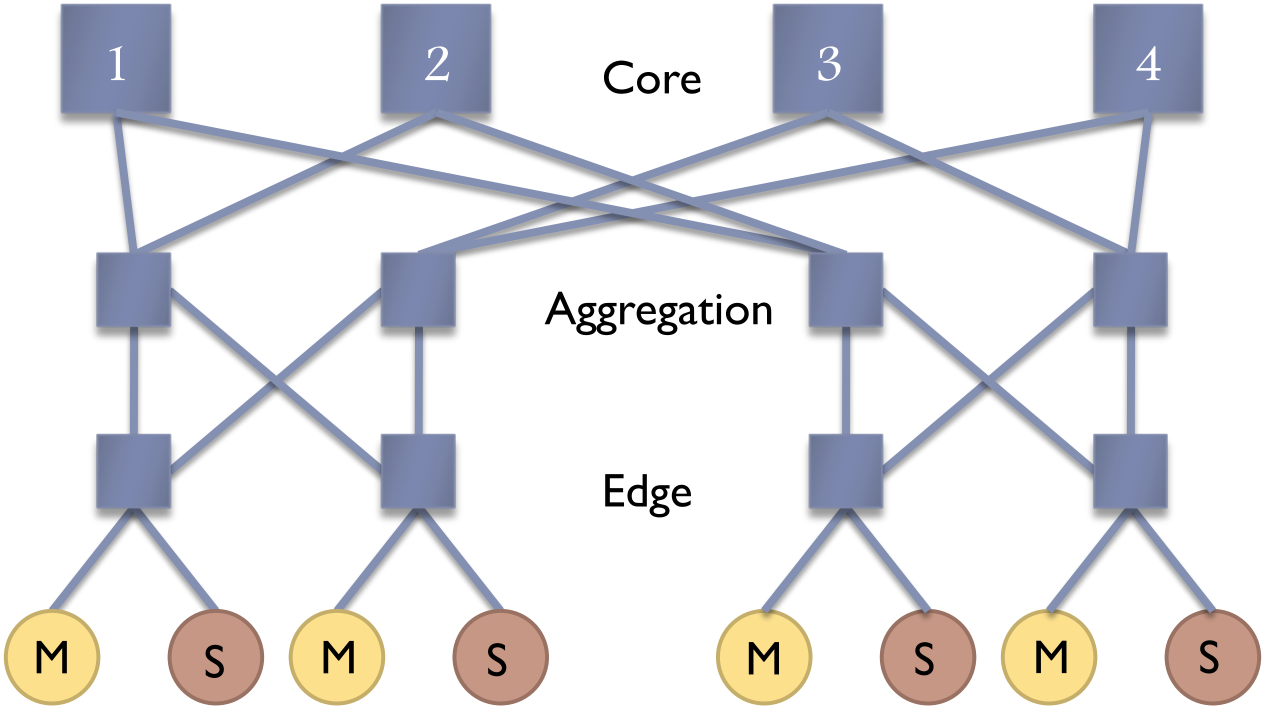
- Applications request specific characteristics
  - Bandwidth, Redundancy, Latency, Isolation, others…

- Idea: Create routes based on applications' needs
  - Per application instance
  - Per application phase: Hadoop shuffle vs. HDFS writes

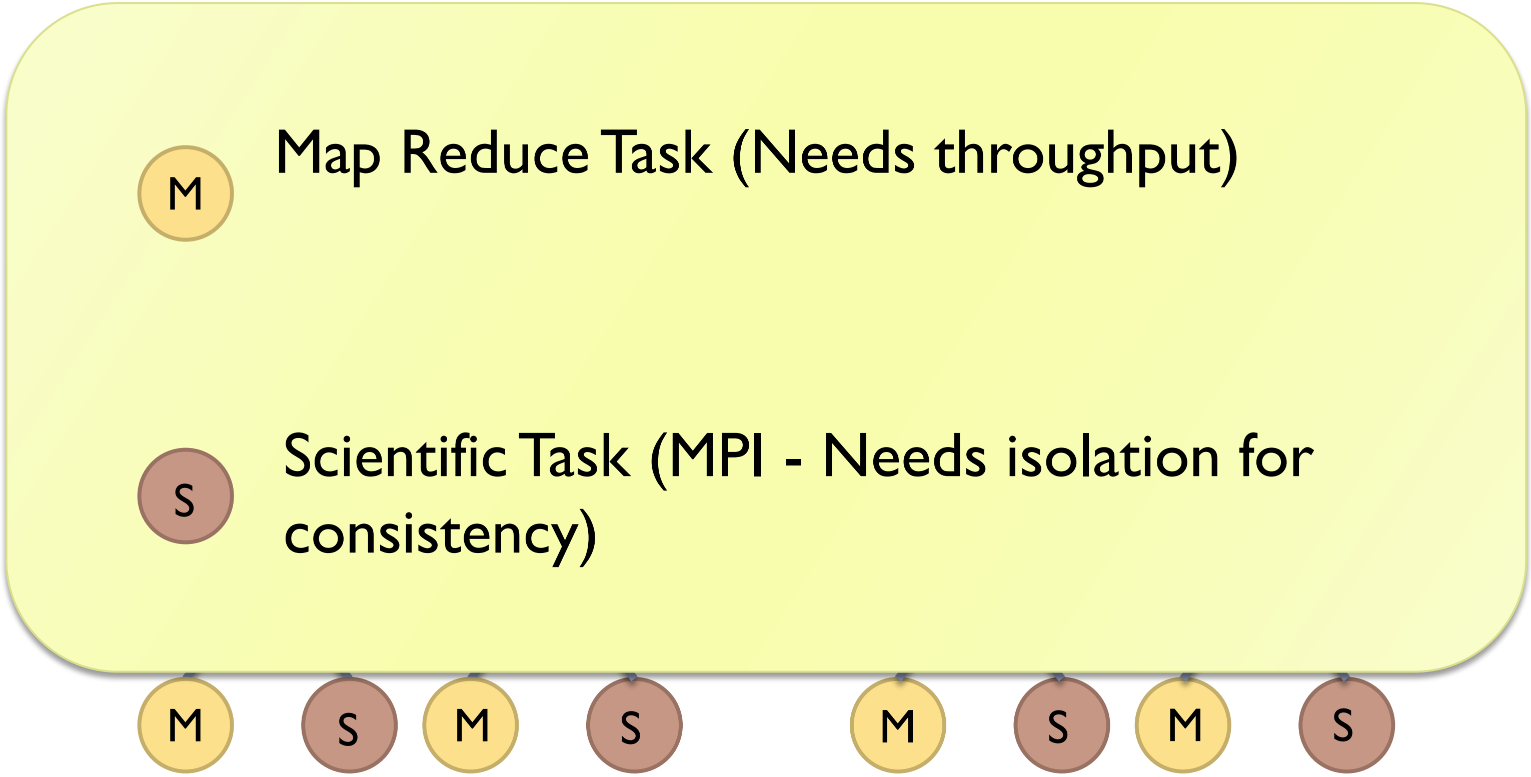
Example



Example



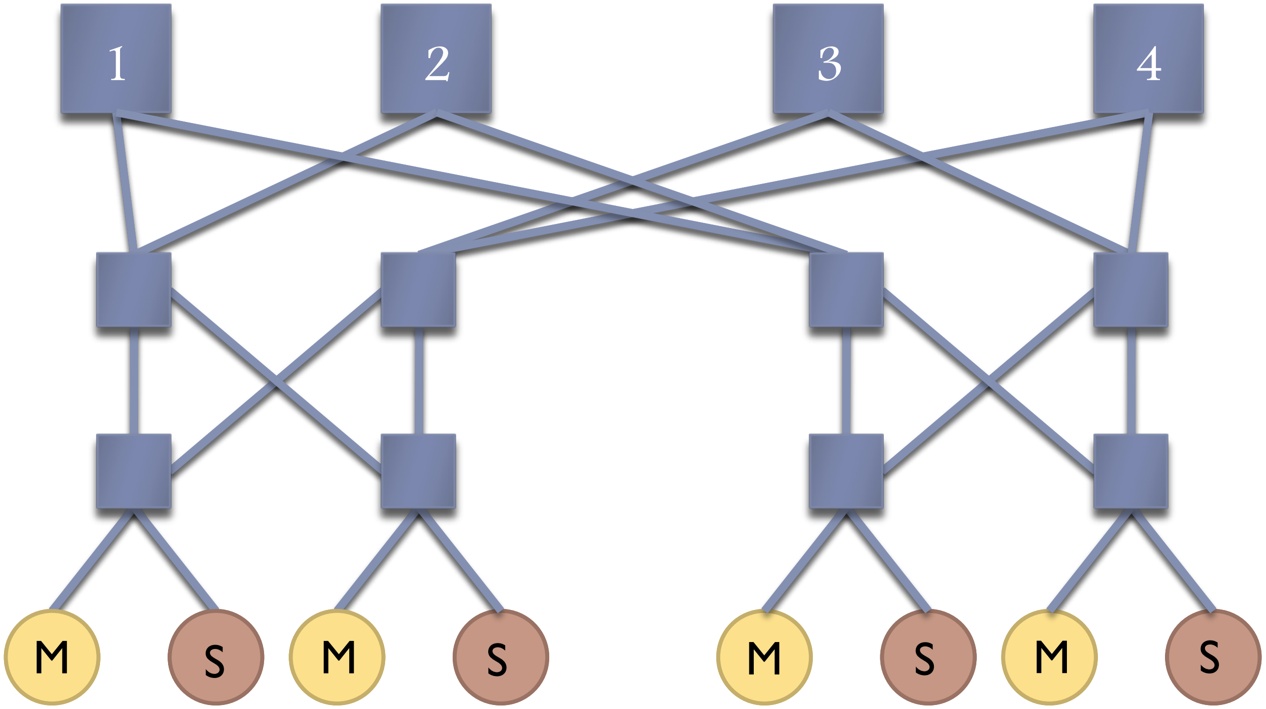
Example



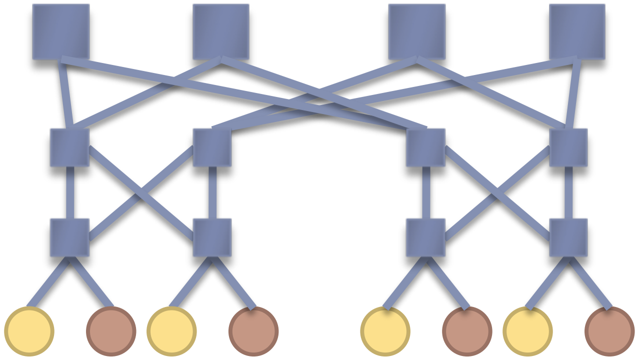
M Map Reduce Task (Needs throughput)

S Scientific Task (MPI - Needs isolation for consistency)

Example

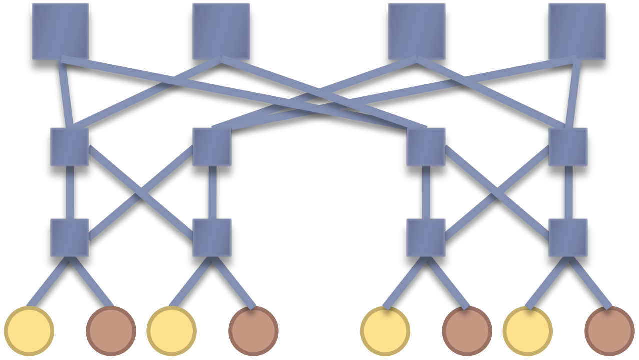


Example



Scientific Network
Exclusive
Free from interference

Map Reduce Network
Multiple paths - high capacity



Challenges

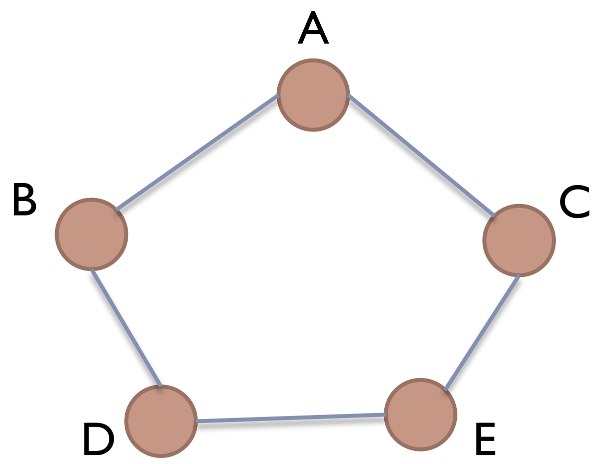
- Reconfigurable network infrastructure
  - Frequent allocations
  - Flexible routing rules
  - Openflow

- Allocation algorithms
  - Throughput, Reliability, Isolation, etc.
  - Evaluation metrics

- Cooperative online allocation of network resources
  - Limit conflict between allocations
  - Can’t take too long

Abstraction

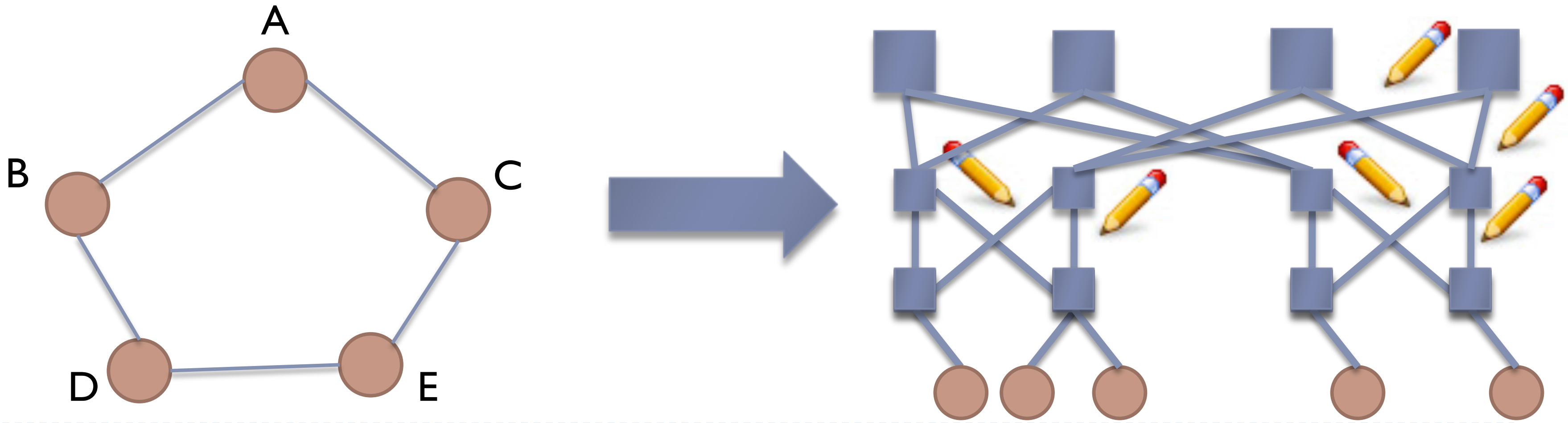
- Applications submit ***routing tasks***:
  - Set of communicating end hosts
  - Logical topology: mesh, ring, tree, custom



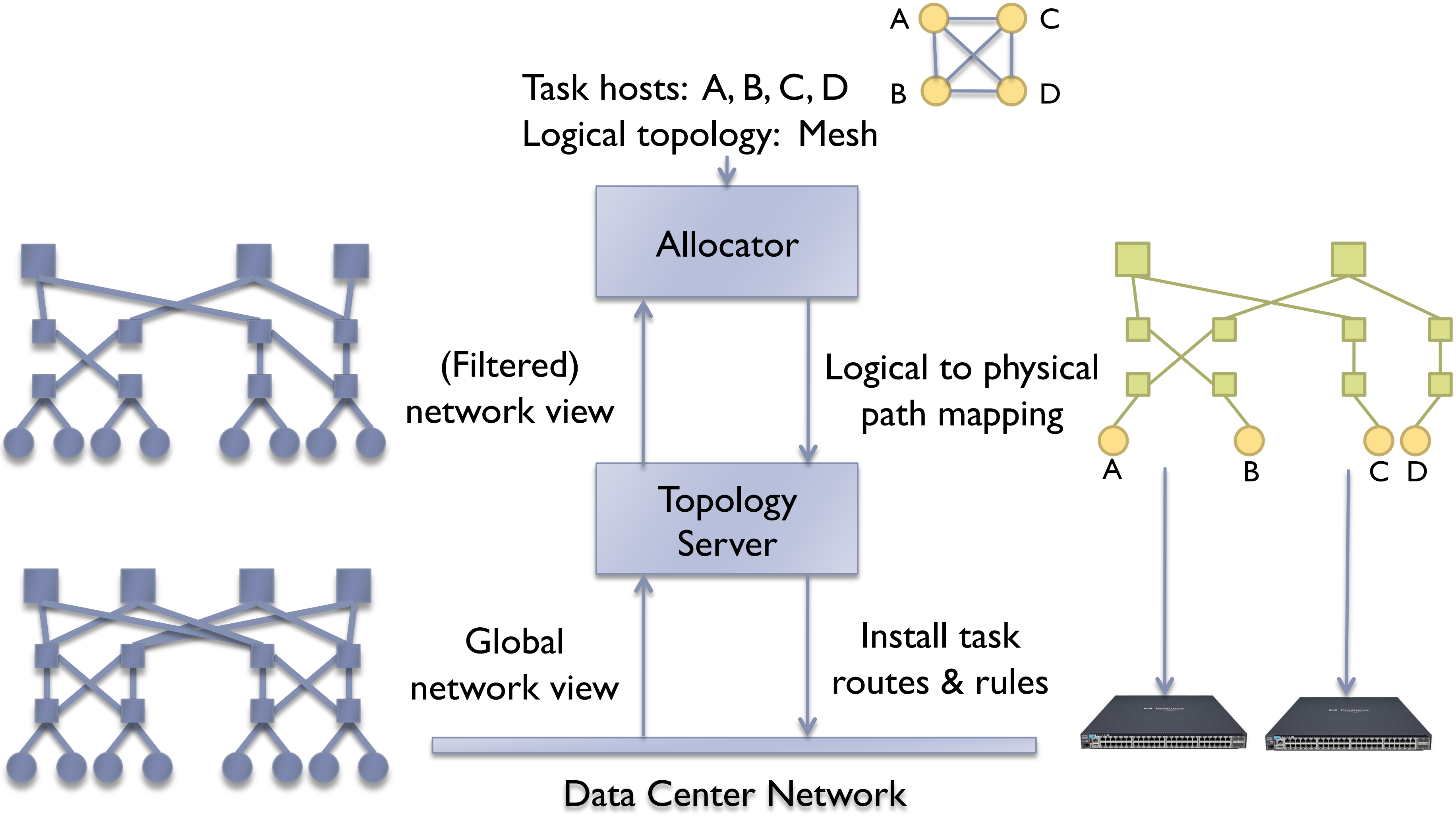
Abstraction

- Routing tasks utilize an ***allocator***:
  - Quantifiable metric
    - Guides allocation, indicates success
  - Allocation algorithm
    - Chooses physical paths
  - Graph annotation & filtering
    - Record allocation results to reduce conflict

# Tasks
# Routes



Architecture



Task hosts: A, B, C, D

Logical topology: Mesh

Allocator

(Filtered) network view

Logical to physical path mapping

Topology Server

Global network view

Install task routes & rules

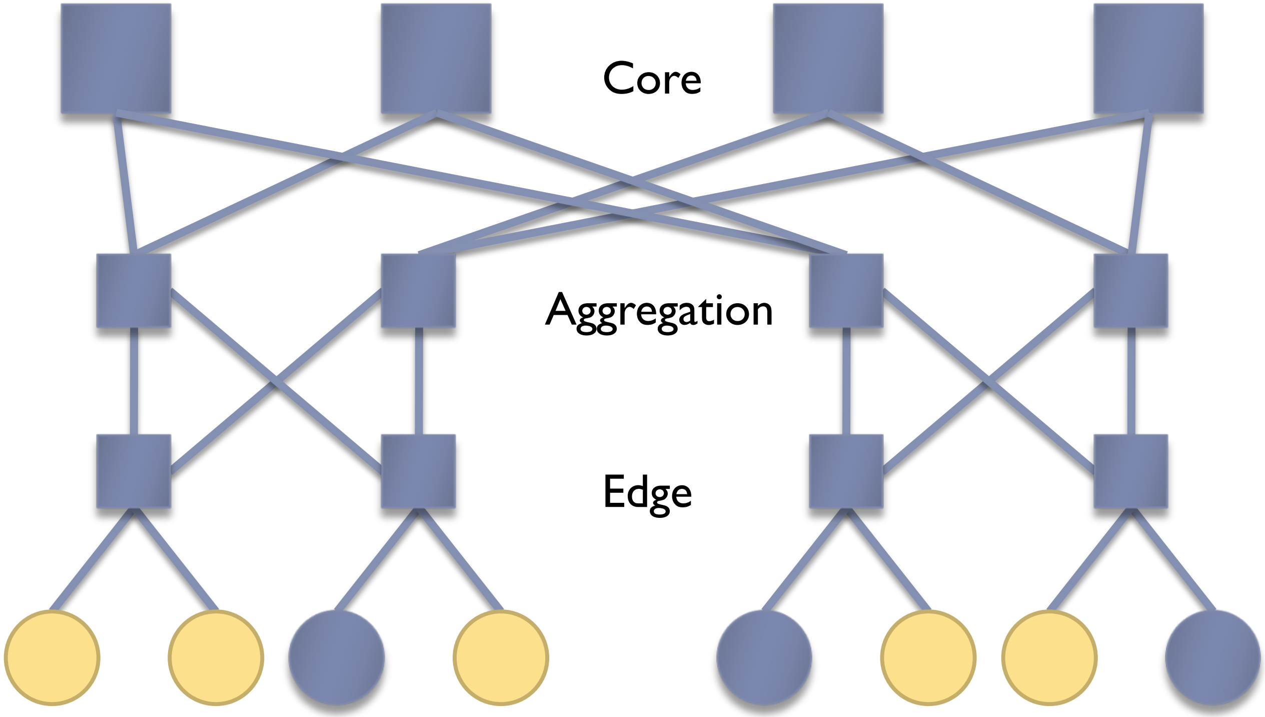
Data Center Network

Three Allocators

- Bandwidth
  - Finds least loaded paths to maximize capacity
- Resiliency
  - Allocates *N* disjoint paths between every host pair
- *K*-Isolation
  - Chooses paths with at most *k* other tasks
  - Reduces inter-task interference, more consistent

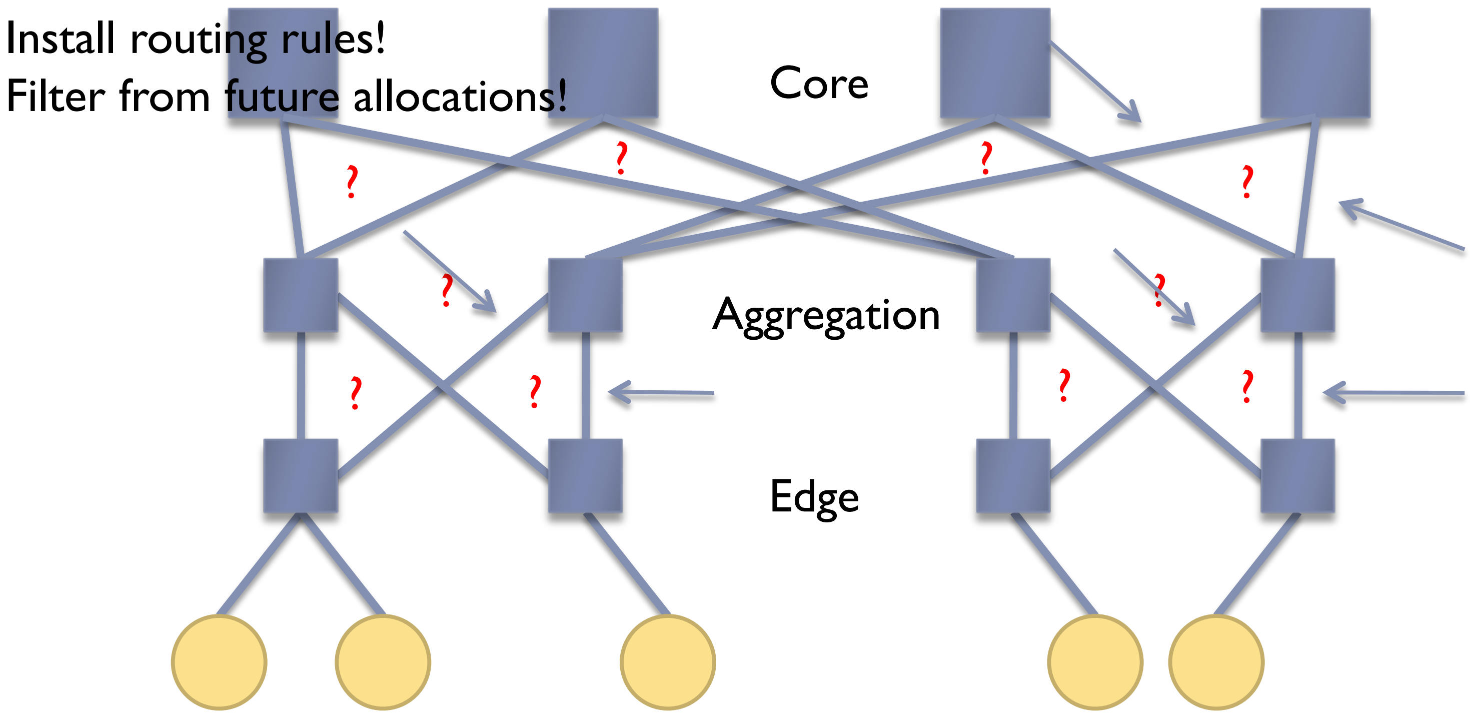
K-Isolation

- Goal: Don’t share links with more than *K* other tasks



K-Isolation

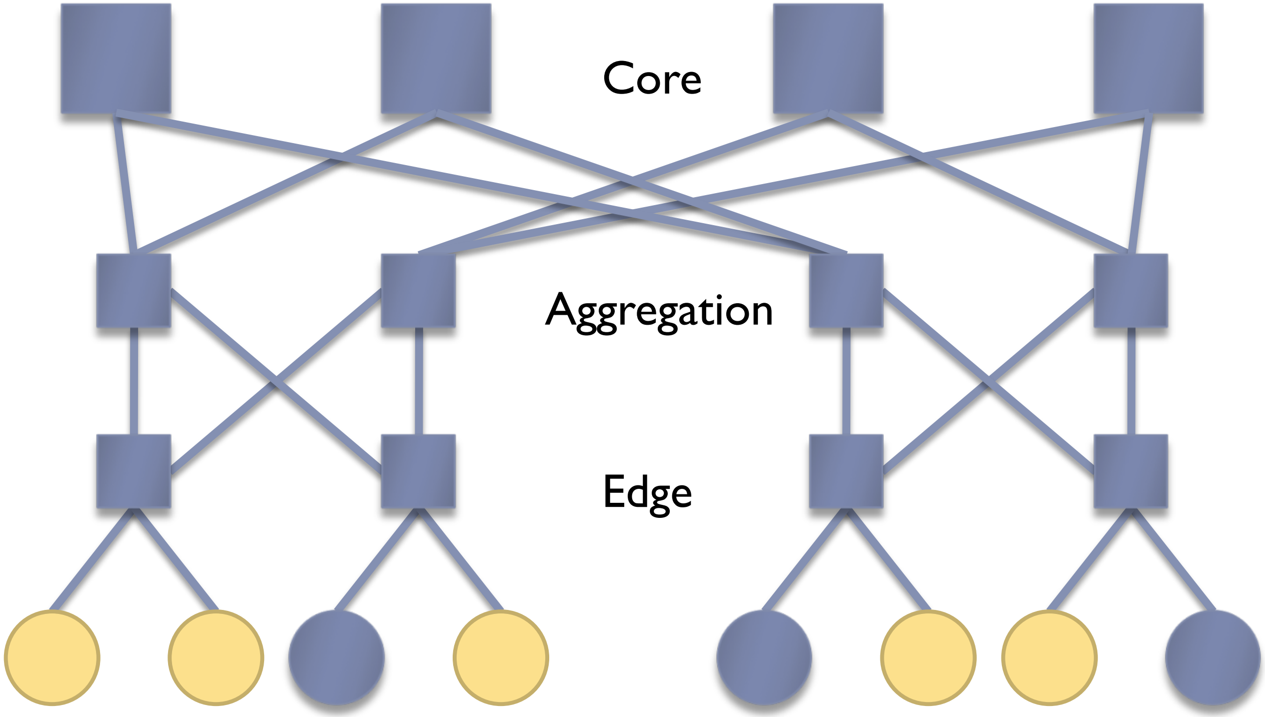
- Goal: Don’t share links with any other tasks



UCSDCSE
Computer Science and Engineering

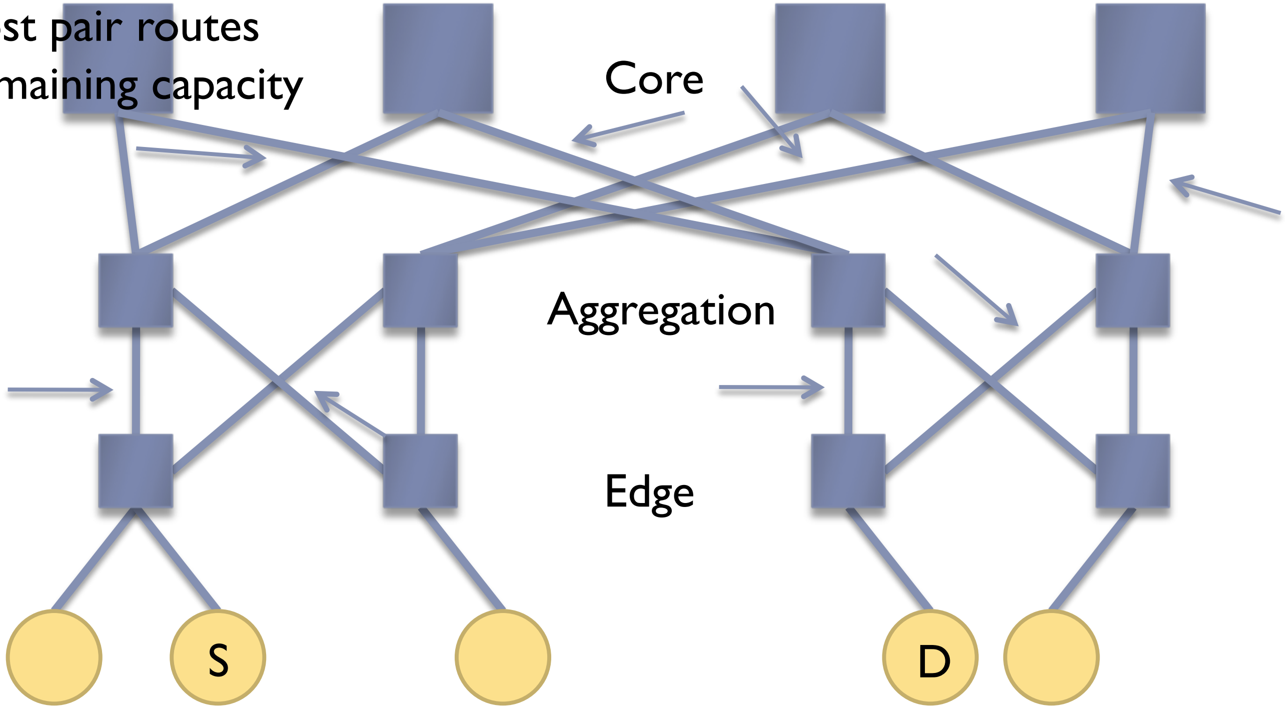
Resiliency

- Goal: Find *N* disjoint routes between end host pairs



Resiliency

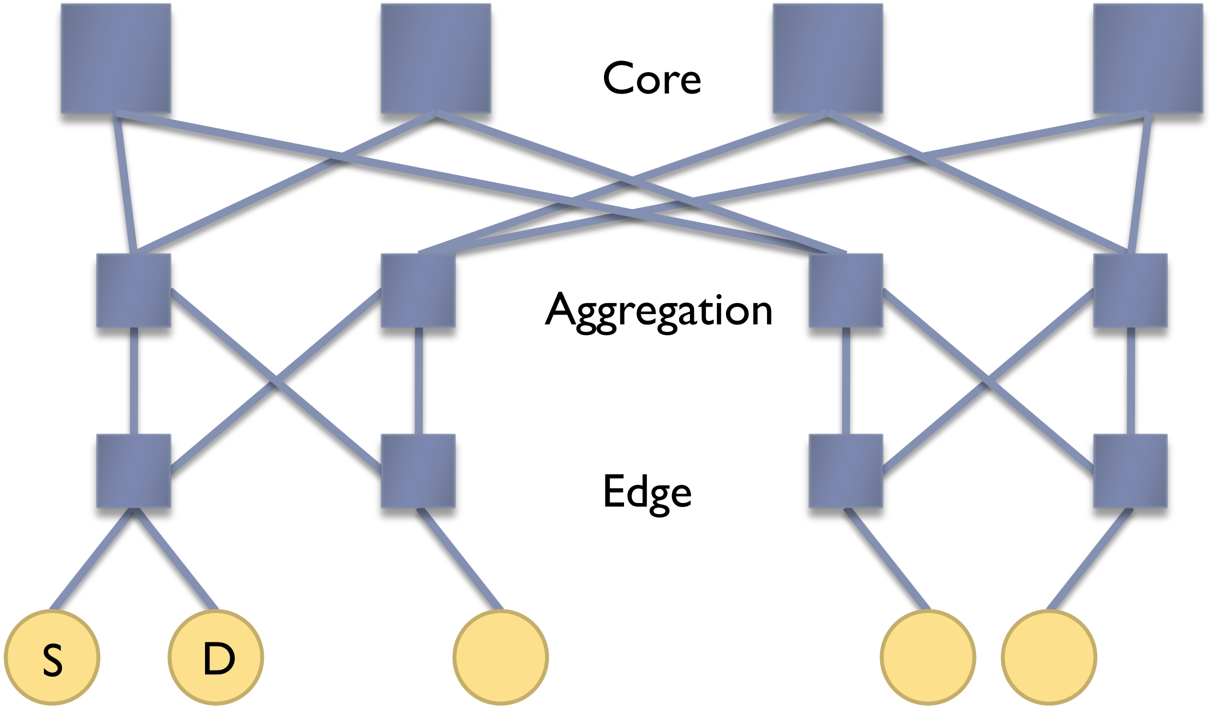
- Goal: Find 2 disjoint routes between end host pairs



Record host pair routes
Update remaining capacity

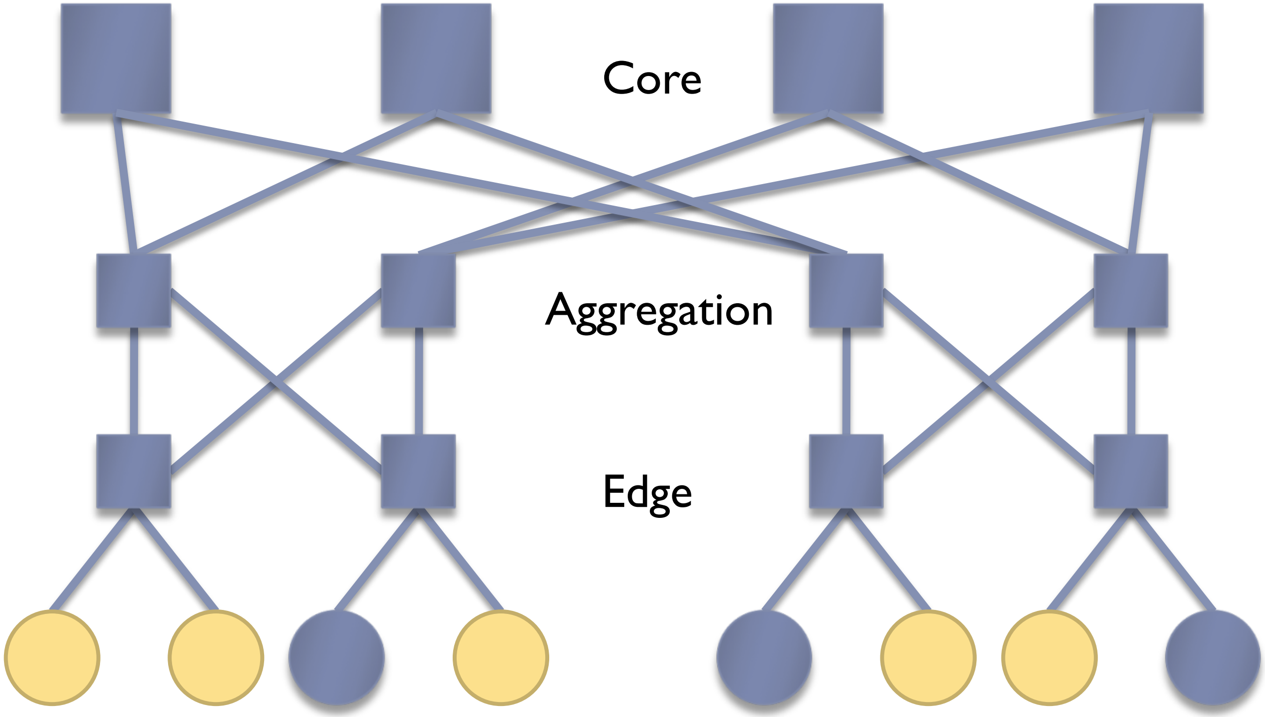
Resiliency

- Repeat for remaining source/destination pairs…



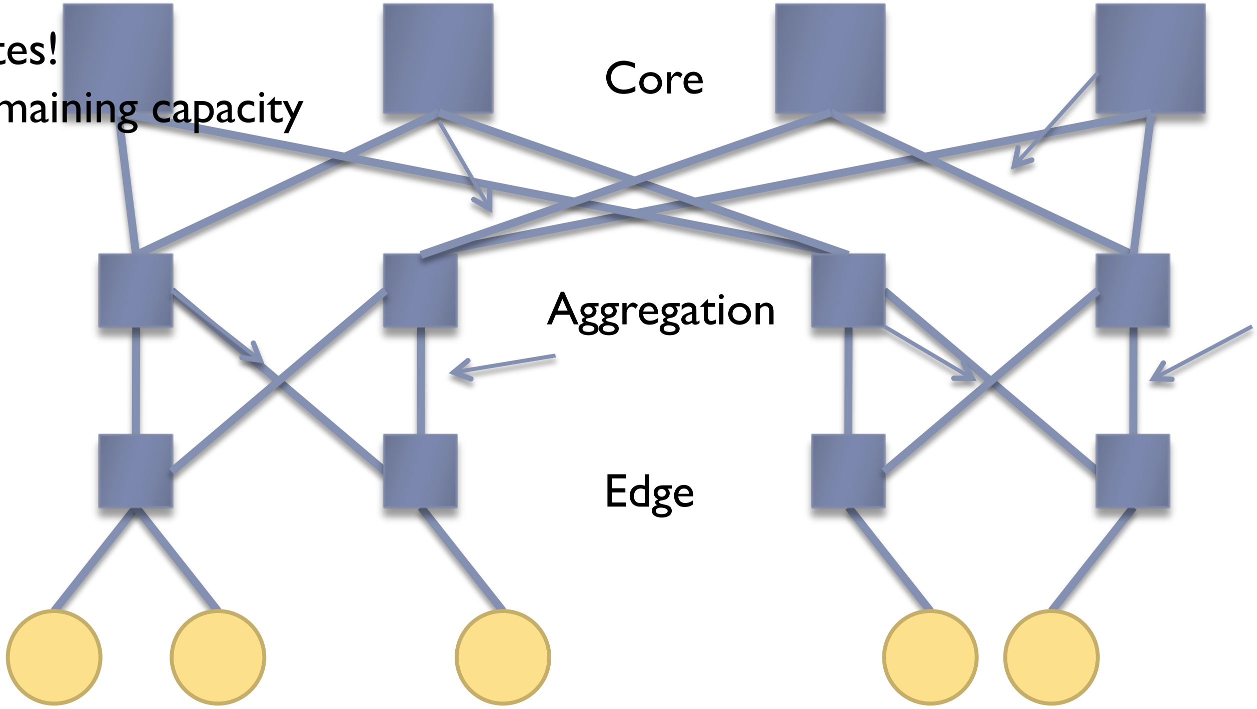
Bandwidth

- Goal: Maximize bisection bandwidth between hosts



Bandwidth

- Build max spanning tree over remaining capacity



Install routes!
Update remaining capacity

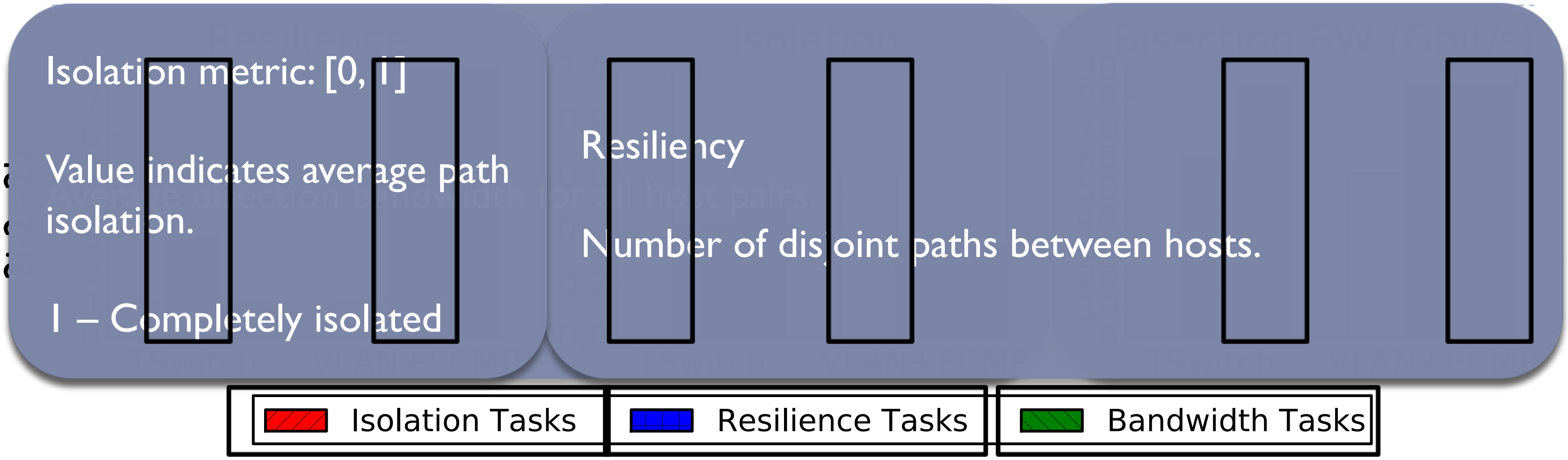
Simulations

- Does it work?

- Comparison against state of the art:
  - “Optimal” equal-cost paths for Resiliency/Bandwidth tasks
  - Spanning tree VLAN for isolation tasks

- Two distinct workloads:
  - Balanced - 6 tasks: 2 isolation, 2 resilience, 2 bandwidth
  - Stressed - 16 tasks: 7 isolation, 5 resilience, 4 bandwidth

Results



Isolation metric: [0, 1]

Value indicates average path isolation.

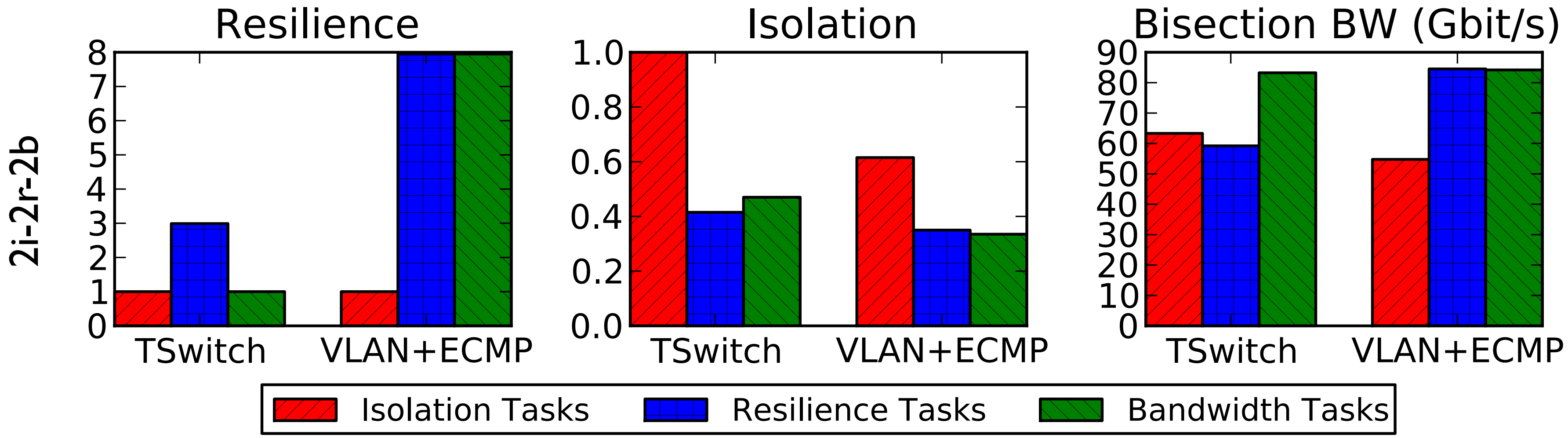
1 – Completely isolated

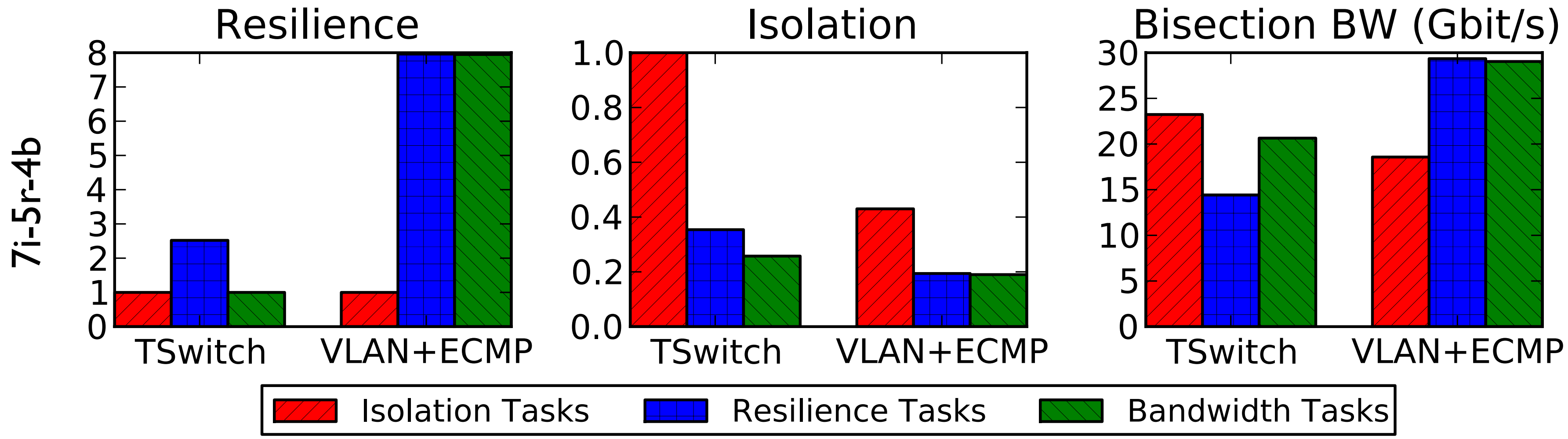
Resiliency

Number of disjoint paths between hosts.

Average bisection bandwidth for all host pairs.

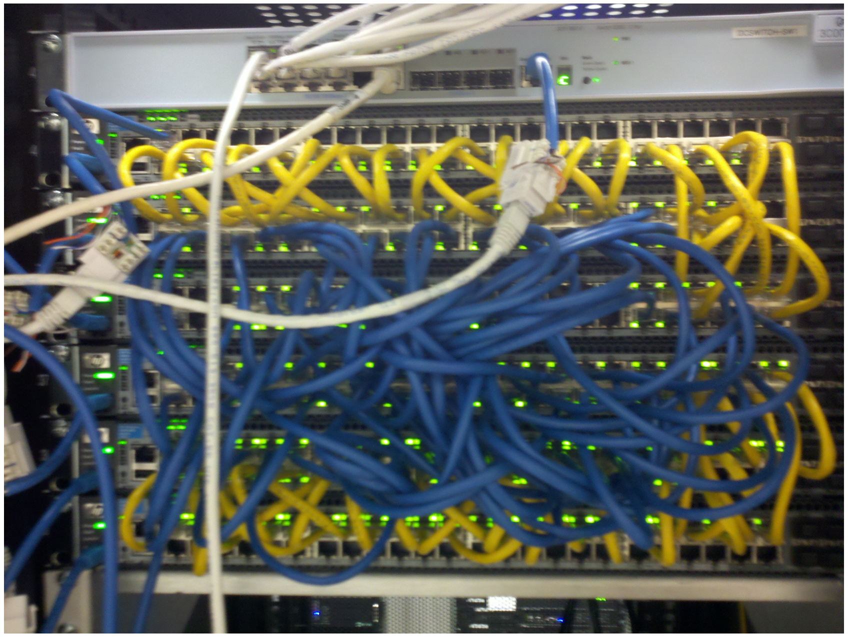
Results


2i-2r-2b
Resilience
0
1
2
3
4
5
6
7
8
TSwitch
VLAN+ECMP
Isolation
0.0
0.2
0.4
0.6
0.8
1.0
TSwitch
VLAN+ECMP
Bisection BW (Gbit/s)
0
10
20
30
40
50
60
70
80
90
TSwitch
VLAN+ECMP
Isolation Tasks
Resilience Tasks
Bandwidth Tasks


7i-5r-4b
Resilience
0
1
2
3
4
5
6
7
8
TSwitch
VLAN+ECMP
Isolation
0.0
0.2
0.4
0.6
0.8
1.0
TSwitch
VLAN+ECMP
Bisection BW (Gbit/s)
0
5
10
15
20
25
30
TSwitch
VLAN+ECMP
Isolation Tasks
Resilience Tasks
Bandwidth Tasks

Current Status

- Simulations promising!
- Ongoing work:
  - Quantifying interference
  - Refining allocation strategies
- Building architecture
  - Openflow-enabled switches
  - Routing rule instantiation
    - Limited TCAM size / speed



Thanks!

- Questions?