Data Center Network Topologies: VL2 (Virtual Layer 2)

Hakim Weatherspoon

Assistant Professor, Dept of Computer Science

CS 5413: High Performance Systems and Networking

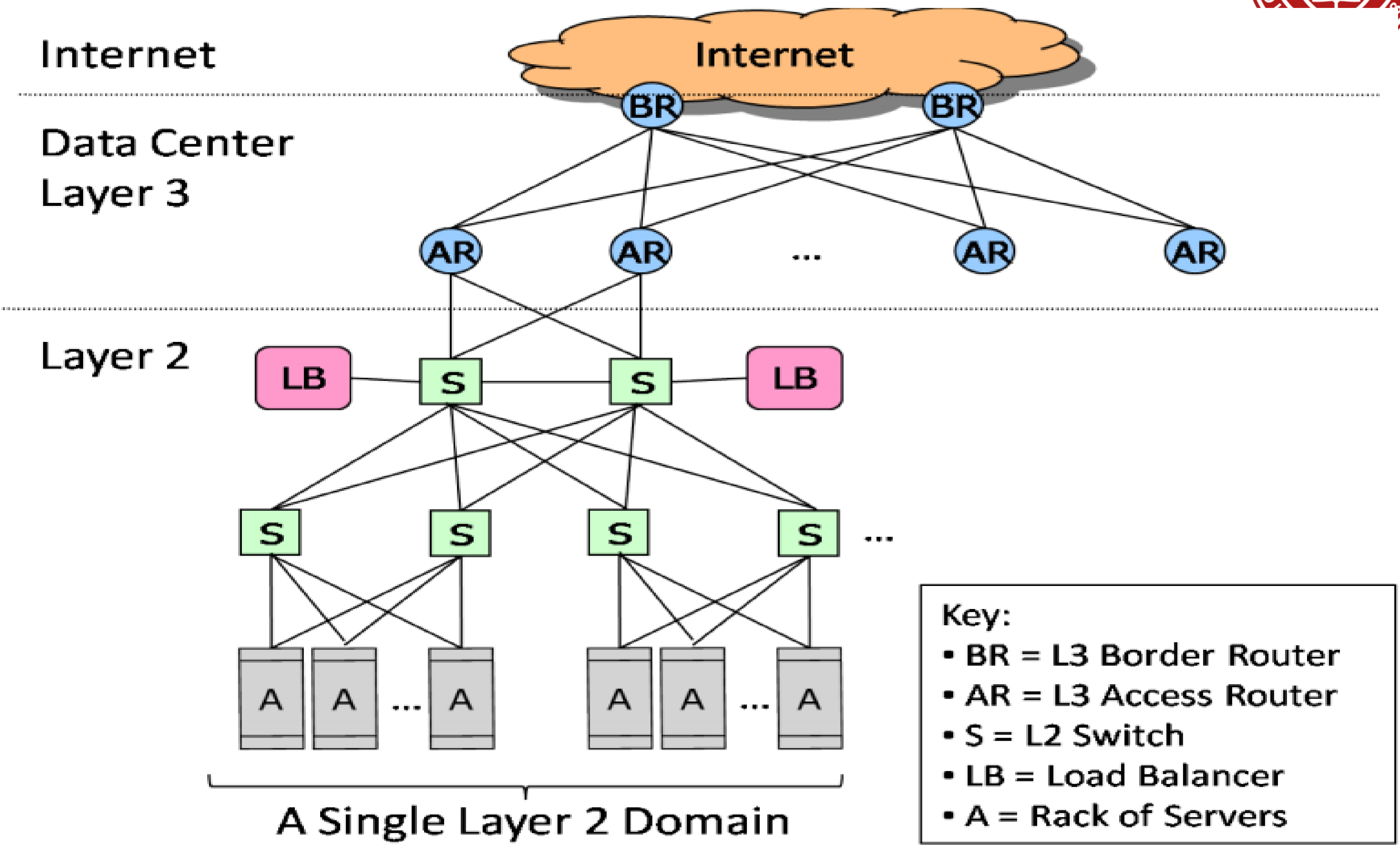
September 26, 2014

Slides used and adapted judiciously from COS-561, Advanced Computer Networks
At Princeton University

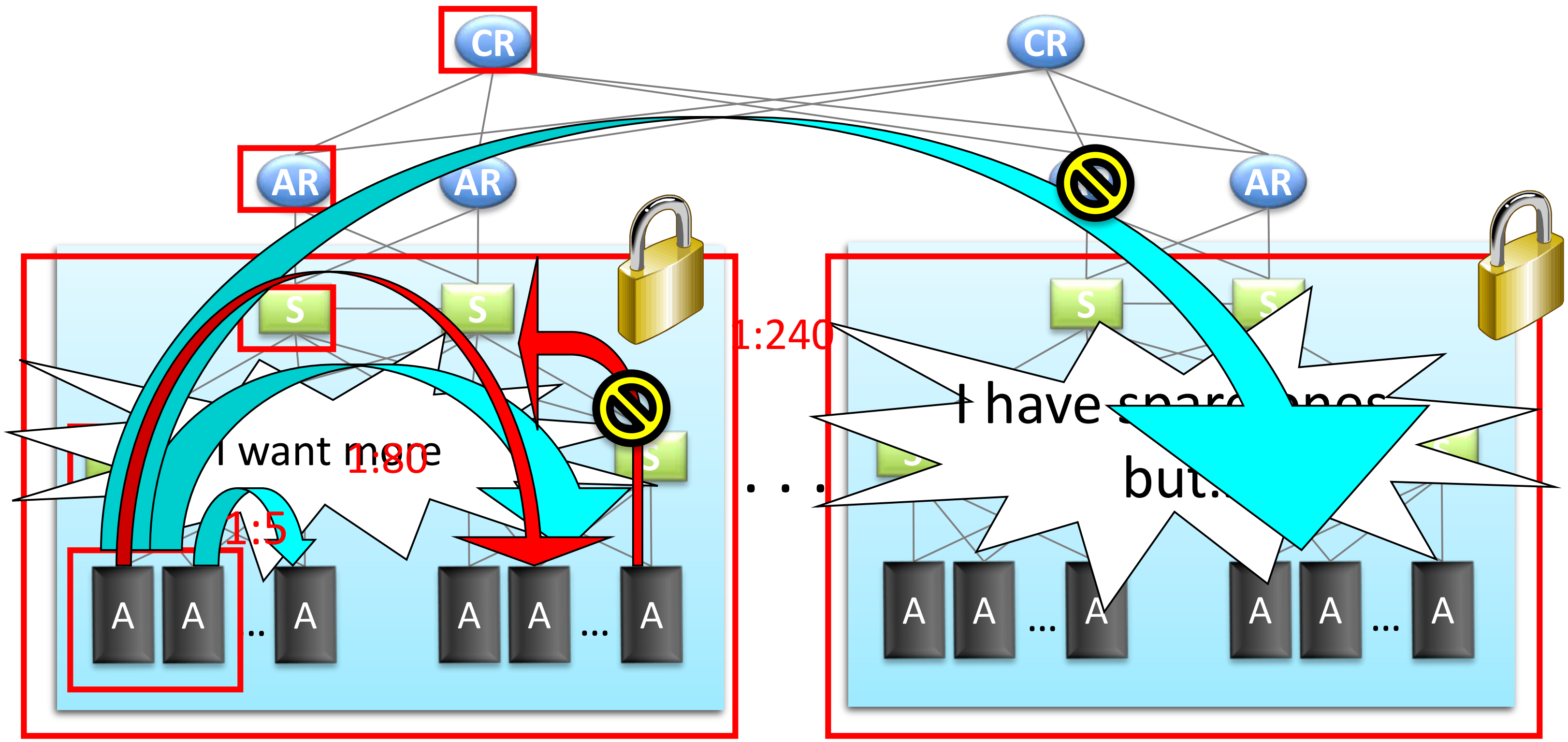
Goals for Today

- VL2: a scalable and flexible data center network
  - A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. ACM Computer Communication Review (CCR), August 2009, pages 51-62.

Architecture of Data Center Networks (DCN)



Conventional DCN Problems



- Static network assignment
- Fragmentation of resource

- Poor server to server connectivity
- Traffics affects each other
- Poor reliability and utilization

Objectives:

- Uniform high capacity:
  - Maximum rate of server to server traffic flow should be limited only by capacity on network cards
  - Assigning servers to service should be independent of network topology
- Performance isolation:
  - Traffic of one service should not be affected by traffic of other services
- Layer-2 semantics:
  - Easily assign any server to any service
  - Configure server with whatever IP address the service expects
  - VM keeps the same IP address even after migration

Measurements and Implications of DCN

- Data-Center traffic analysis:
  - Traffic volume between servers to entering/leaving data center is 4:1
  - Demand for bandwidth between servers growing faster
  - Network is the bottleneck of computation
- Flow distribution analysis:
  - Majority of flows are small, biggest flow size is 100MB
  - The distribution of internal flows is simpler and more uniform
  - 50% times of 10 concurrent flows, 5% greater than 80 concurrent flows

Measurements and Implications of DCN

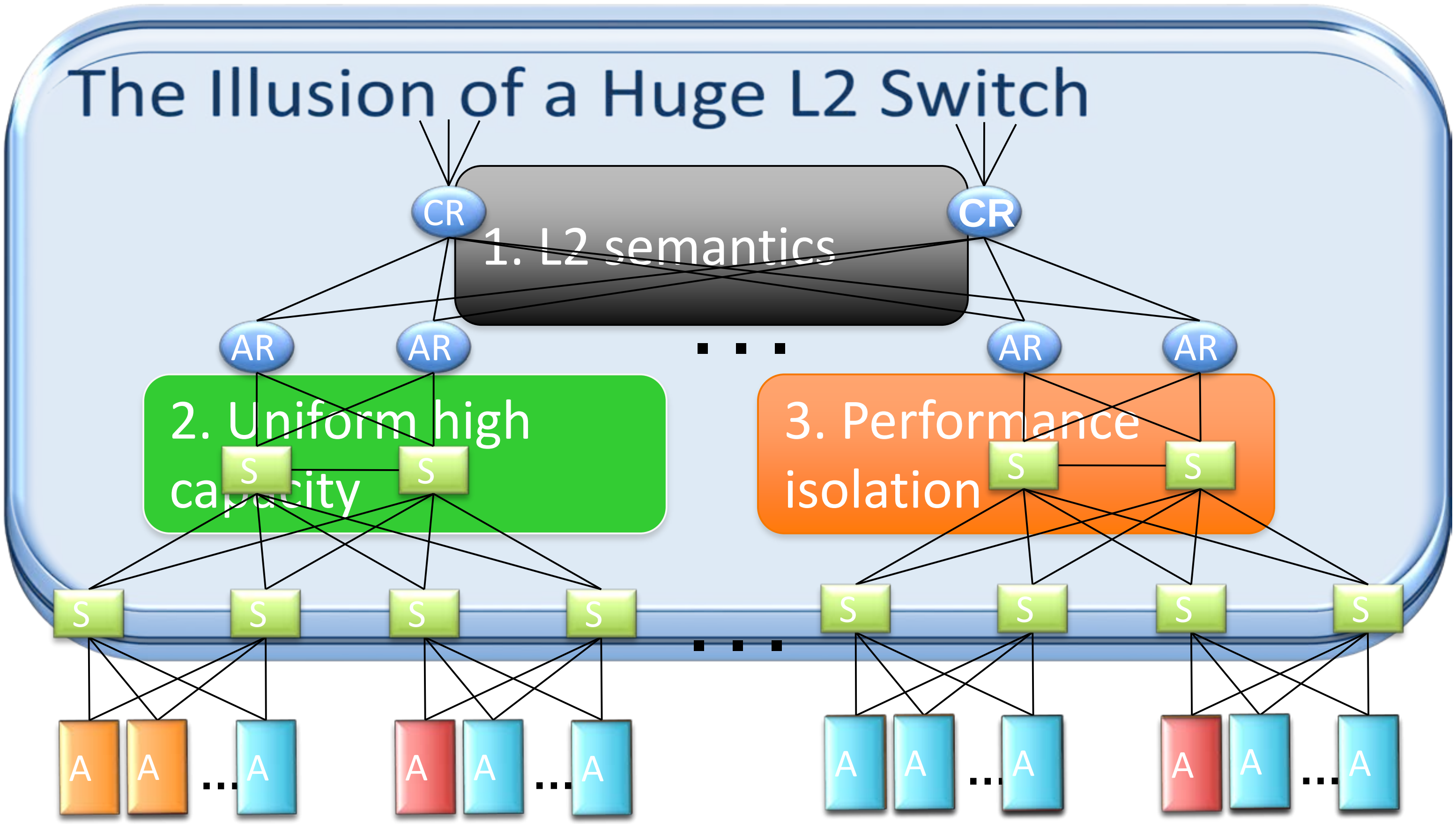
- Traffic matrix analysis:
  - Poor summarizing of traffic patterns
  - Instability of traffic patterns

- Failure characteristics:
  - Pattern of networking equipment failures: 95% < 1min, 98% < 1hr, 99.6% < 1 day, 0.09% > 10 days
  - No obvious way to eliminate all failures from the top of the hierarchy

Virtual Layer 2 Switch (VL2)

- Design principle:
  - Randomizing to cope with volatility:
    - Using Valiant Load Balancing (VLB) to do destination independent traffic spreading across multiple intermediate nodes
  - Building on proven networking technology:
    - Using IP routing and forwarding technologies available in commodity switches
  - Separating names from locators:
    - Using directory system to maintain the mapping between names and locations
  - Embracing end systems:
    - A VL2 agent at each server

Virtual Layer 2 Switch (VL2)



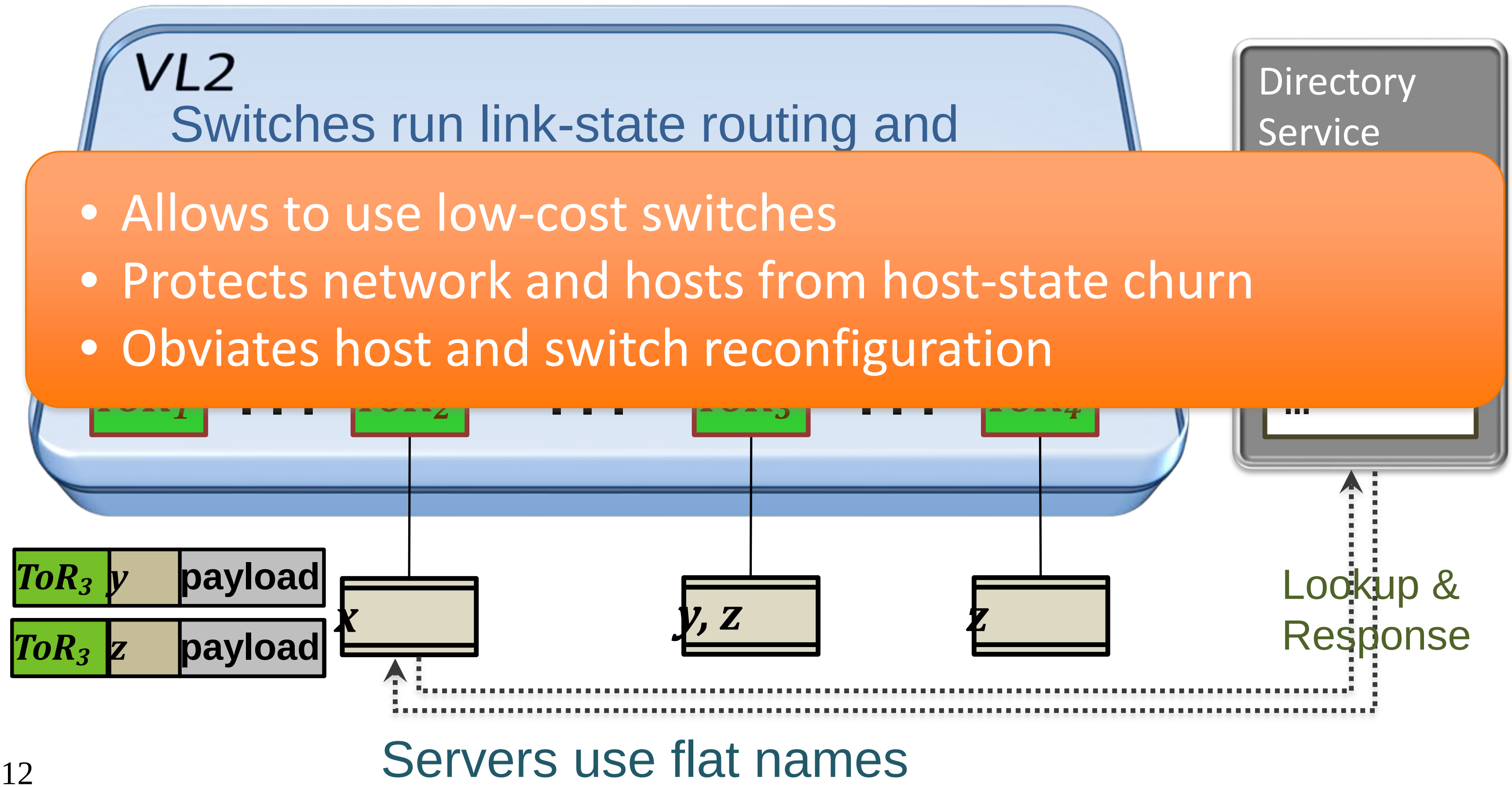
VL2 Goals and Solutions

Objective	Approach	Solution
1. Layer-2 semantics	**Employ flat addressing**	**Name-location separation & resolution service**
2. Uniform high capacity between servers	Guarantee bandwidth for hose-model traffic	Flow-based random traffic indirection (Valiant LB)
3. Performance Isolation	**Enforce hose model using existing mechanisms only**	TCP

"Hose": each node has ingress/egress bandwidth constraints

Name/Location Separation

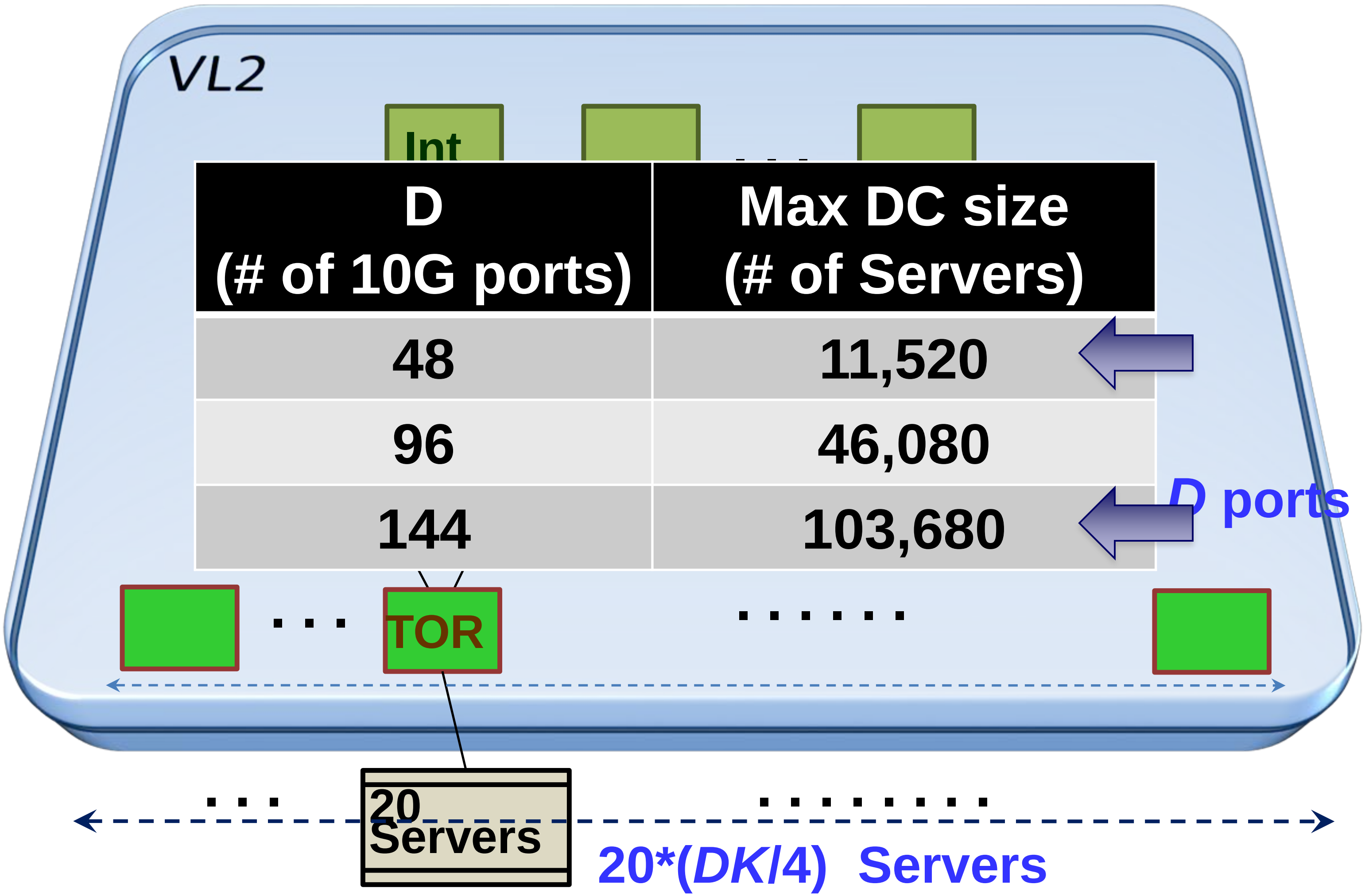
Cope with host churns with very little overhead



- Allows to use low-cost switches
- Protects network and hosts from host-state churn
- Obviates host and switch reconfiguration

Clos Network Topology

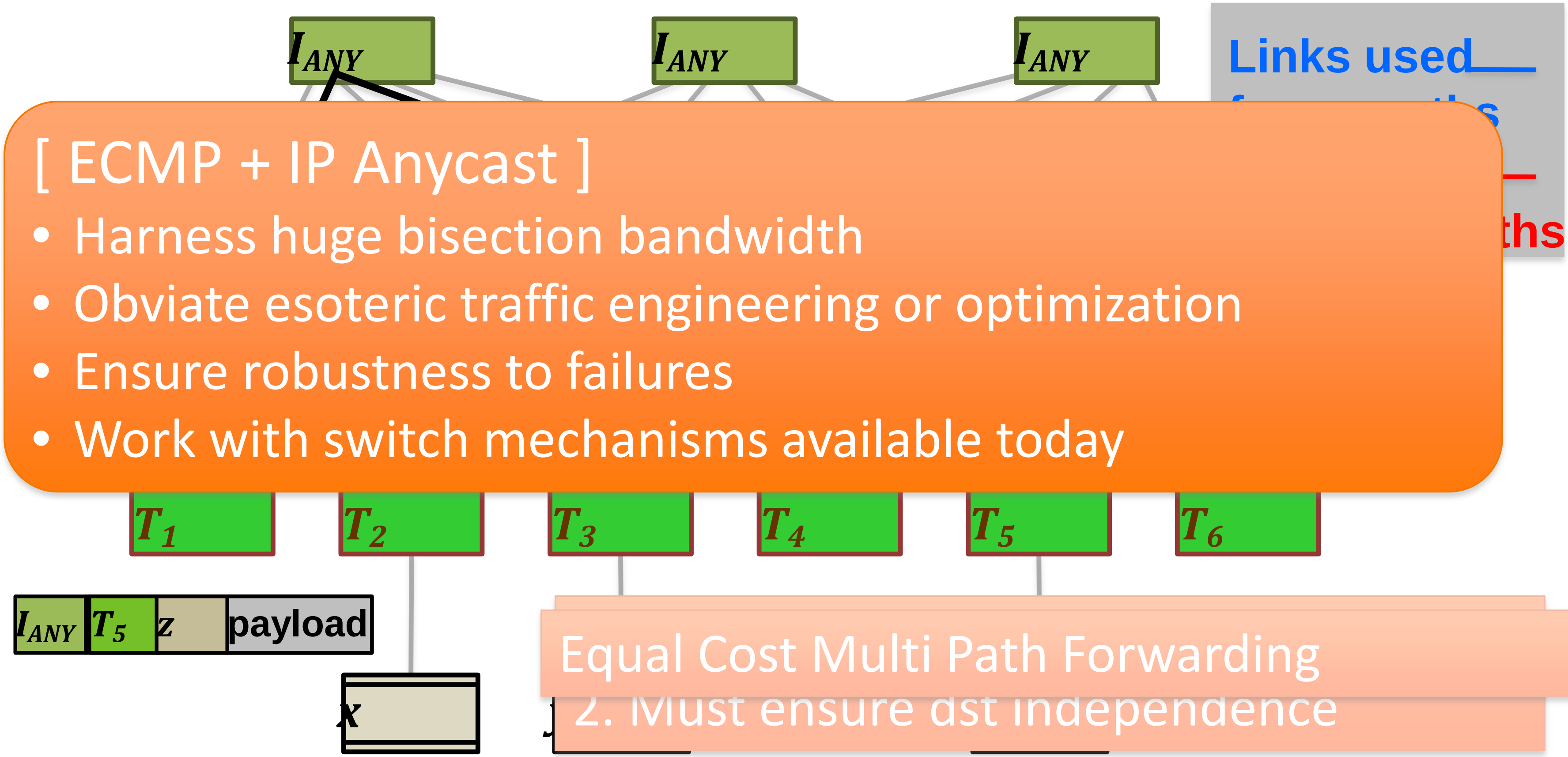
Offer huge aggr capacity & multi paths at modest cost



D (# of 10G ports)	Max DC size (# of Servers)
48	11,520
96	46,080
144	103,680

Valiant Load Balancing: Indirection

Cope with arbitrary TMs with very little overhead



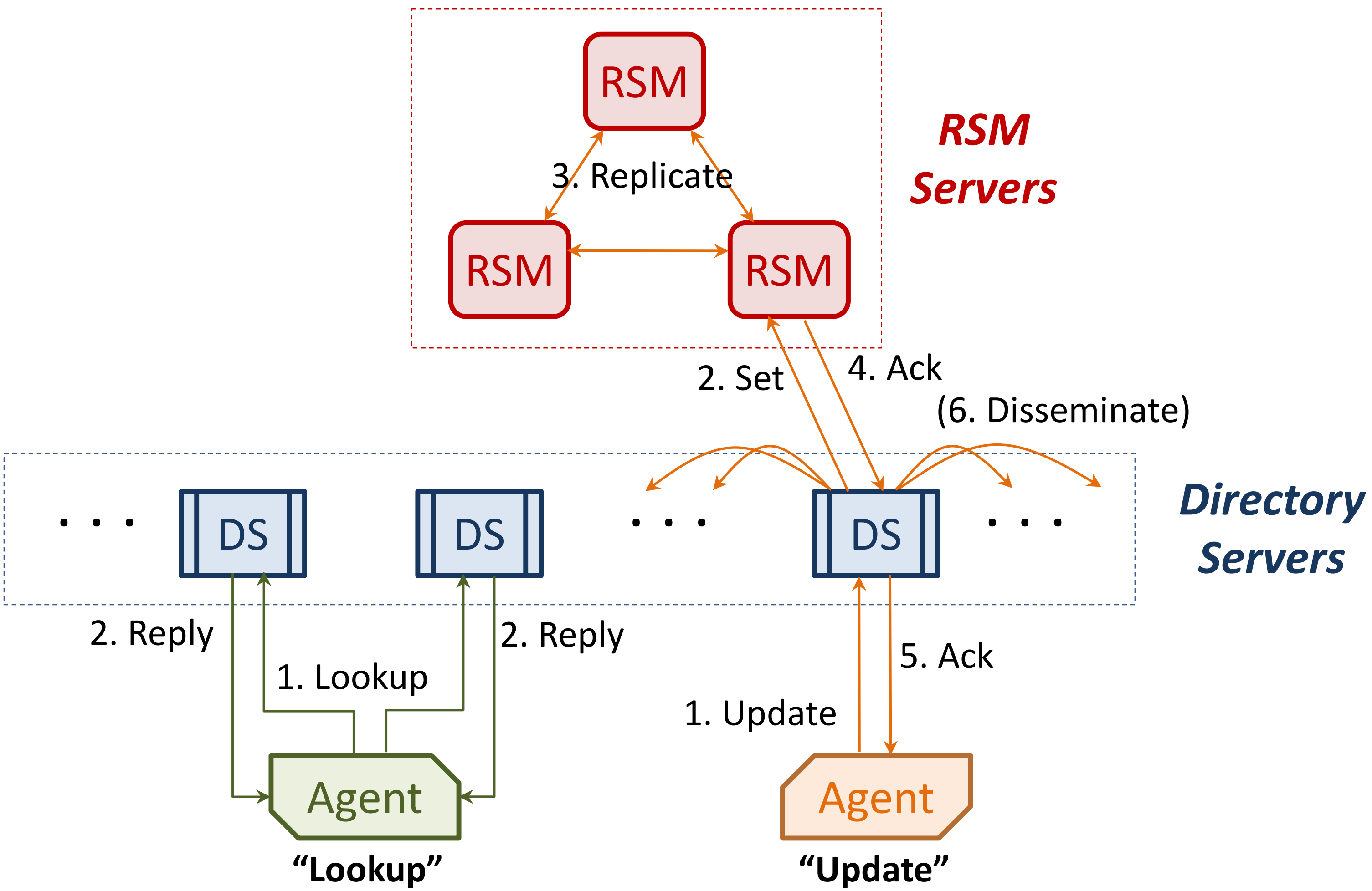
[ECMP + IP Anycast]

- Harness huge bisection bandwidth
- Obviate esoteric traffic engineering or optimization
- Ensure robustness to failures
- Work with switch mechanisms available today

Equal Cost Multi Path Forwarding

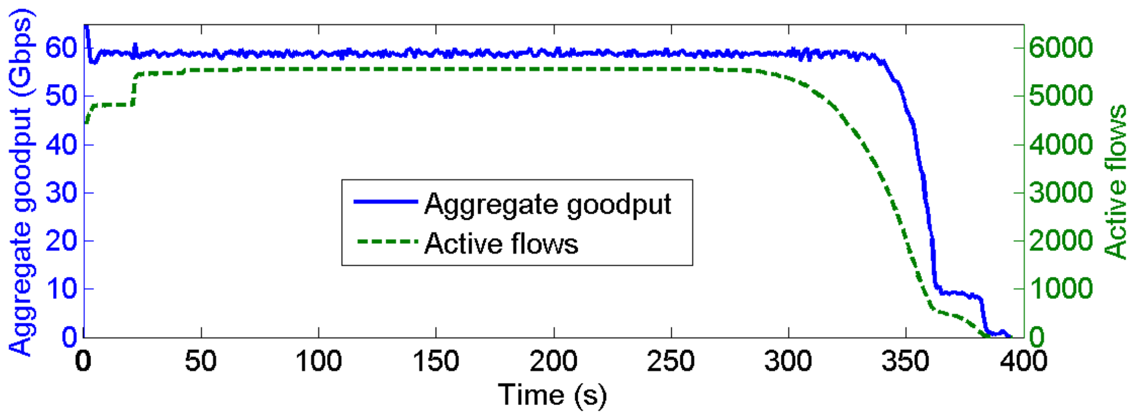
2. Must ensure dst independence

VL2 Directory System



Evaluation

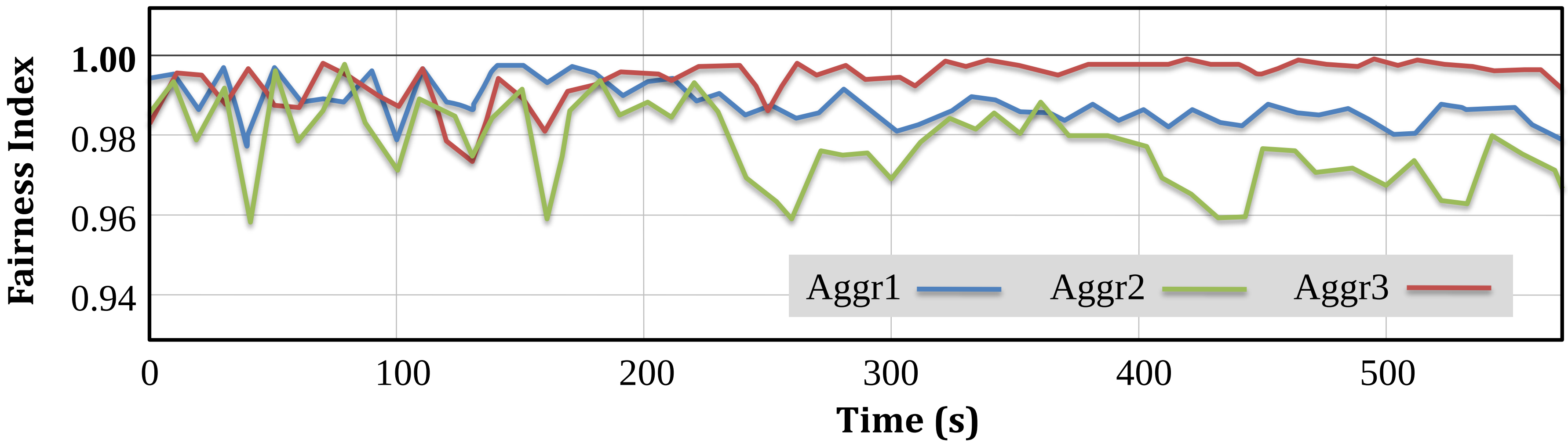
- Uniform high capacity:
  - All-to-all data shuffle stress test:
    - 75 servers, deliver 500MB



    - Maximal achievable goodput is 62.3
    - VL2 network efficiency as 58.8/62.3 = 94%

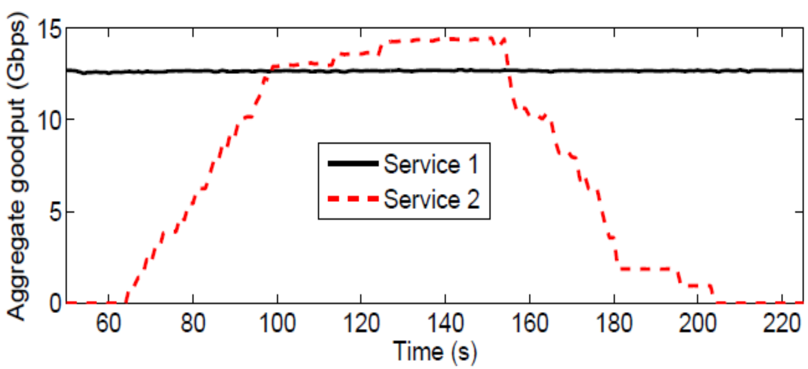
Evaluation

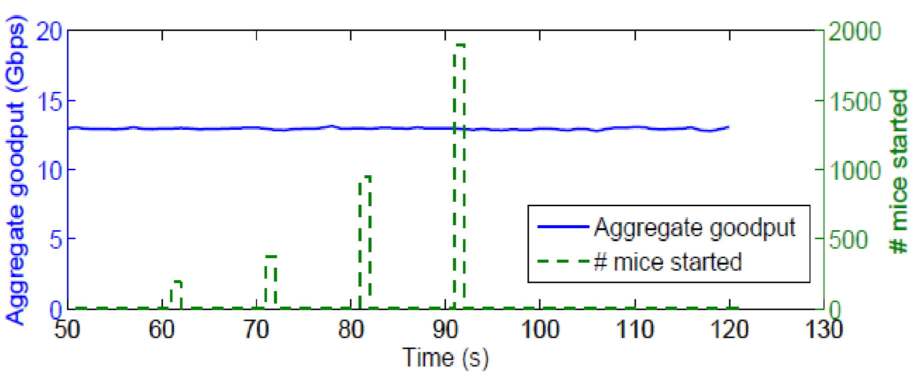
- Fairness:
  - 75 nodes
  - Real data center workload
  - Plot Jain's fairness index for traffics to intermediate switches



Evaluation

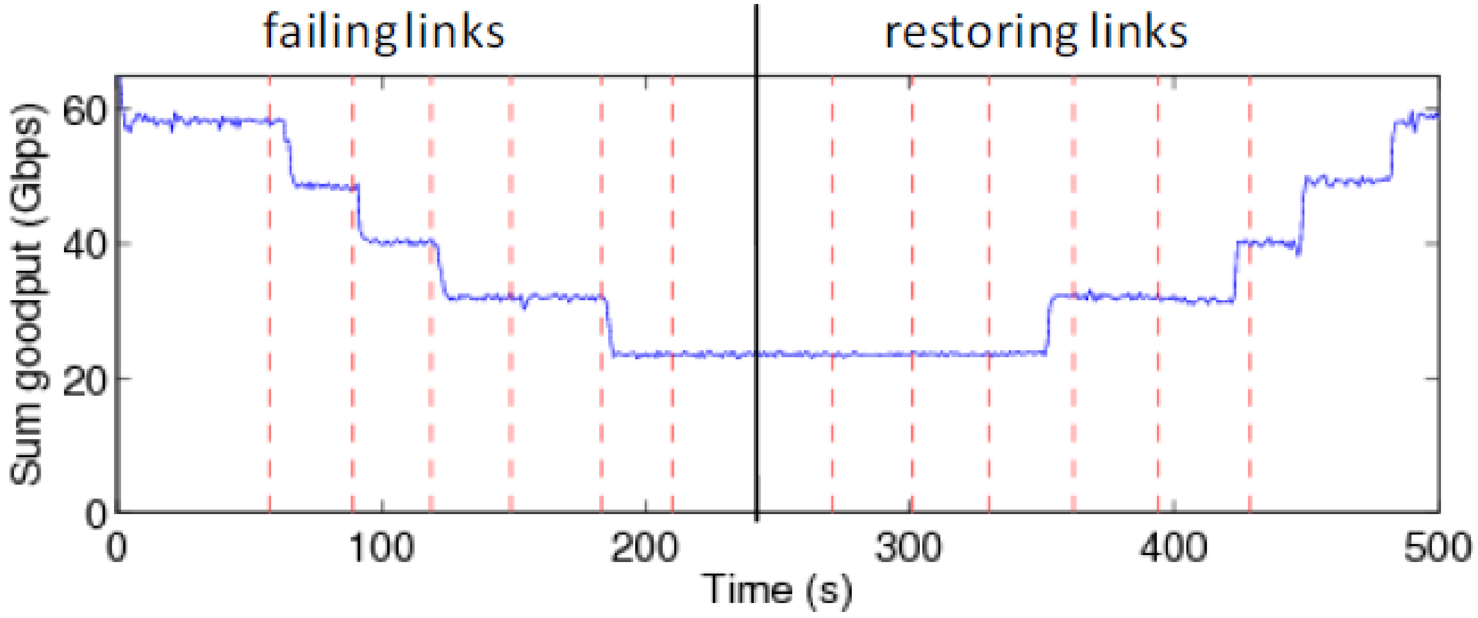
- Performance isolation:
  - Two types of services:
    - Service one: 18 servers do single TCP transfer all the time
    - Service two: 19 servers starts a 8GB transfer over TCP every 2 seconds





Evaluation

- Convergence after link failures
  - 75 servers
  - All-to-all data shuffle
  - Disconnect links between intermediate and aggregation switches



Perspective

- Studied the traffic pattern in a production data center and find the traffic patterns
- Design, build and deploy every component of VL2 in an 80 server testbed
- Apply VLB to randomly spreading traffics over multiple flows
- Using flat address to split IP addresses and server names

Critique

- The extra servers are needed to support the VL2 directory system,:
  - Brings more cost on devices
  - Hard to be implemented for data centers with tens of thousands of servers.
- All links and switches are working all the times, not power efficient
- No evaluation of real time performance.

VL2 vs. SEATTLE

- Similar "virtual layer 2" abstraction
  – Flat end-point addresses
  – Indirection through intermediate node
- Enterprise networks (Seattle)
  – Hard to change hosts → directory on the switches
  – Sparse traffic patterns → effectiveness of caching
  – Predictable traffic patterns → no emphasis on TE
- Data center networks (VL2)
  – Easy to change hosts → move functionality to hosts
  – Dense traffic matrix → reduce dependency on caching
  – Unpredictable traffic patterns → ECMP and VLB for TE

Other Data Center Architectures

- VL2: A Scalable and Flexible Data Center Network
  - consolidate layer-2/layer-3 into a “virtual layer 2”
  - separating “naming” and “addressing”, also deal with dynamic load-balancing issues
- A Scalable, Commodity Data Center Network Architecture
  - a new Fat-tree “inter-connection” structure (topology) to increases “bi-section” bandwidth
    - needs “new” addressing, forwarding/routing

Other Approaches:

- PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric
- BCube: A High-Performance, Server-centric Network Architecture for Modular Data Centers

Ongoing Research

Research Questions

- What topology to use in data centers?
  - Reducing wiring complexity
  - Achieving high bisection bandwidth
  - Exploiting capabilities of optics and wireless
- Routing architecture?
  - Flat layer-2 network vs. hybrid switch/router
  - Flat vs. hierarchical addressing
- How to perform traffic engineering?
  - Over-engineering vs. adapting to load
  - Server selection, VM placement, or optimizing routing
- Virtualization of NICs, servers, switches, …

Research Questions

- Rethinking TCP congestion control?
  - Low propagation delay and high bandwidth
  - “Incast” problem leading to bursty packet loss
- Division of labor for TE, access control, …
  - VM, hypervisor, ToR, and core switches/routers
- Reducing energy consumption
  - Better load balancing vs. selective shutting down
- Wide-area traffic engineering
  - Selecting the least-loaded or closest data center
- Security
  - Preventing information leakage and attacks

Before Next time

- Project Progress
  - **Need to setup environment as soon as possible**
  - And meet with groups, TA, and professor
- Lab0b – Getting Started with Fractus
  - Use Fractus instead of Red Cloud
    - Red Cloud instances will be terminated and state lost
  - **Due Monday, Sept 29**

- ***Required review and reading for Friday, September 26***
  - "The Click Modular Router", E. Kohler, R. Morris, B. Chen, and M. F. Kaashoek. *ACM Symposium on Operating Systems Principles (SOSP)*, December 1999, pages 217-231.
  - http://dl.acm.org/citation.cfm?id=319166
  - http://www.pdos.lcs.mit.edu/papers/click:sosp99/paper.pdf

- Check piazza: http://piazza.com/cornell/fall2014/cs5413
- Check website for updated schedule