6.829: Computer Networks and Mobile Systems

Lecture 16:

Data Center Network Architectures

Fall 2017

Some slides adapted from Mohammad Alizadeh, Albert Greenberg and Changhoon Kim

What are Data Centers?

Large facilities with 10s of thousands of networked servers

- Compute, storage, and networking working in concert
- "Warehouse-Scale Computers"
- Huge investment: ~ $0.5 billion for large datacenter

Data Center Costs

Amortized Cost*	Component	Sub-Components
~45%	Servers	CPU, memory, disk
~25%	Power infrastructure	UPS, cooling, power distribution
~15%	Power draw	Electrical utility costs
~15%	Network	Switches, links, transit

The Cost of a Cloud: Research Problems in Data Center Networks. Sigcomm CCR 2009. Greenberg, Hamilton, Maltz, Patel.

*3 yr amortization for servers, 15 yr for infrastructure; 5% cost of money

Server Costs

30% utilization considered “good” in most data centers!

Uneven application fit

- Each server has CPU, memory, disk: most applications exhaust one resource, stranding the others

Uncertainty in demand

- Demand for a new service can spike quickly

Risk management

- Not having spare servers to meet demand brings failure just when success is at hand

Goal: Agility – Any service, Any Server

Turn the servers into a single large fungible pool

- Dynamically expand and contract service footprint as needed

Benefits

- Lower cost (higher utilization)
- Increase developer productivity
- Achieve high performance and reliability

Achieving Agility

Workload management

- Means for rapidly installing a service's code on a server
- *Virtual machines, disk images, containers*

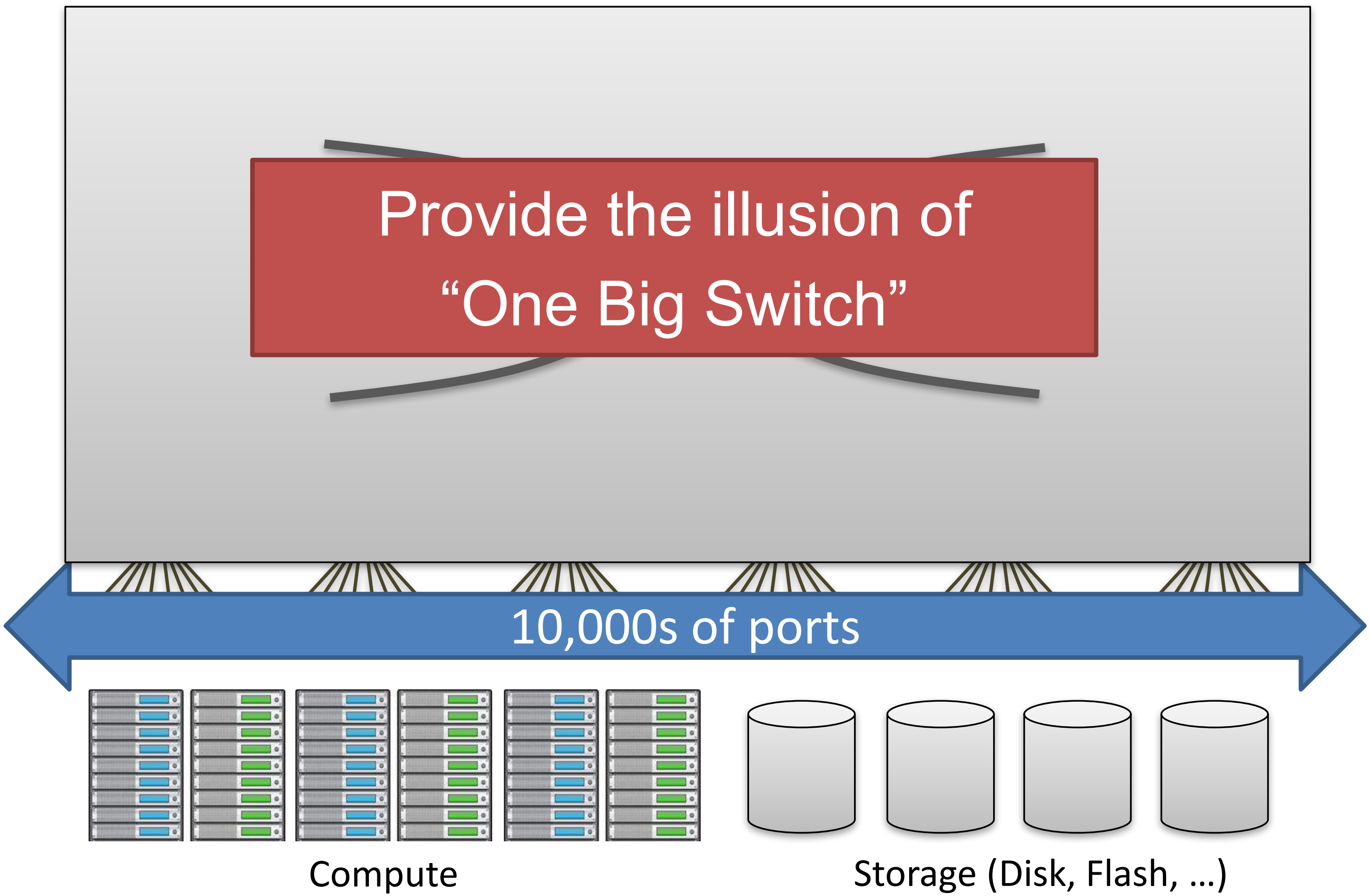
Storage Management

- Means for a server to access persistent data
- *Distributed filesystems (e.g., HDFS, blob stores)*

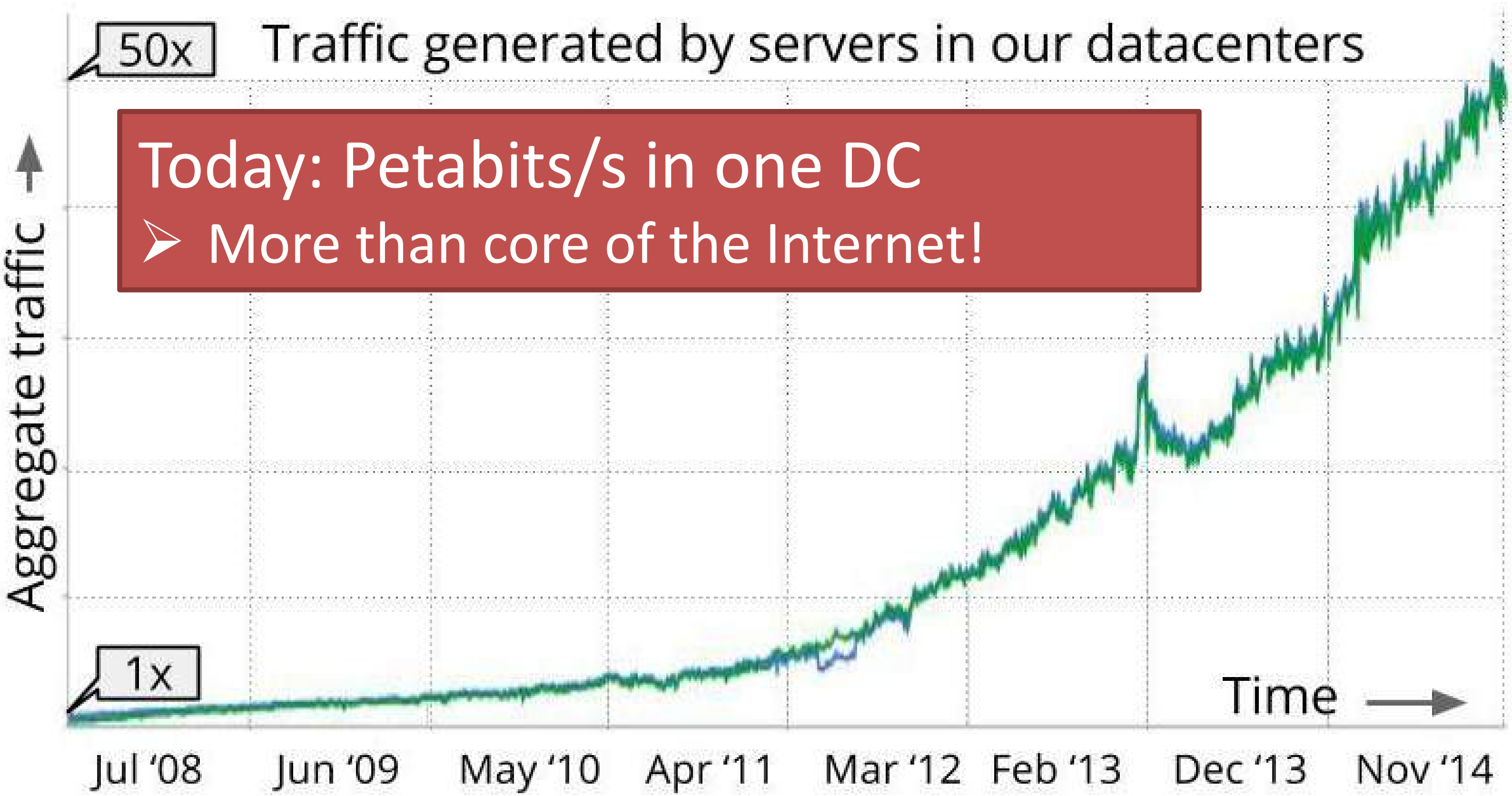
Network

- Means for communicating with other servers, regardless of where they are in the data center

Datacenter Networks



Datacenter Traffic Growth



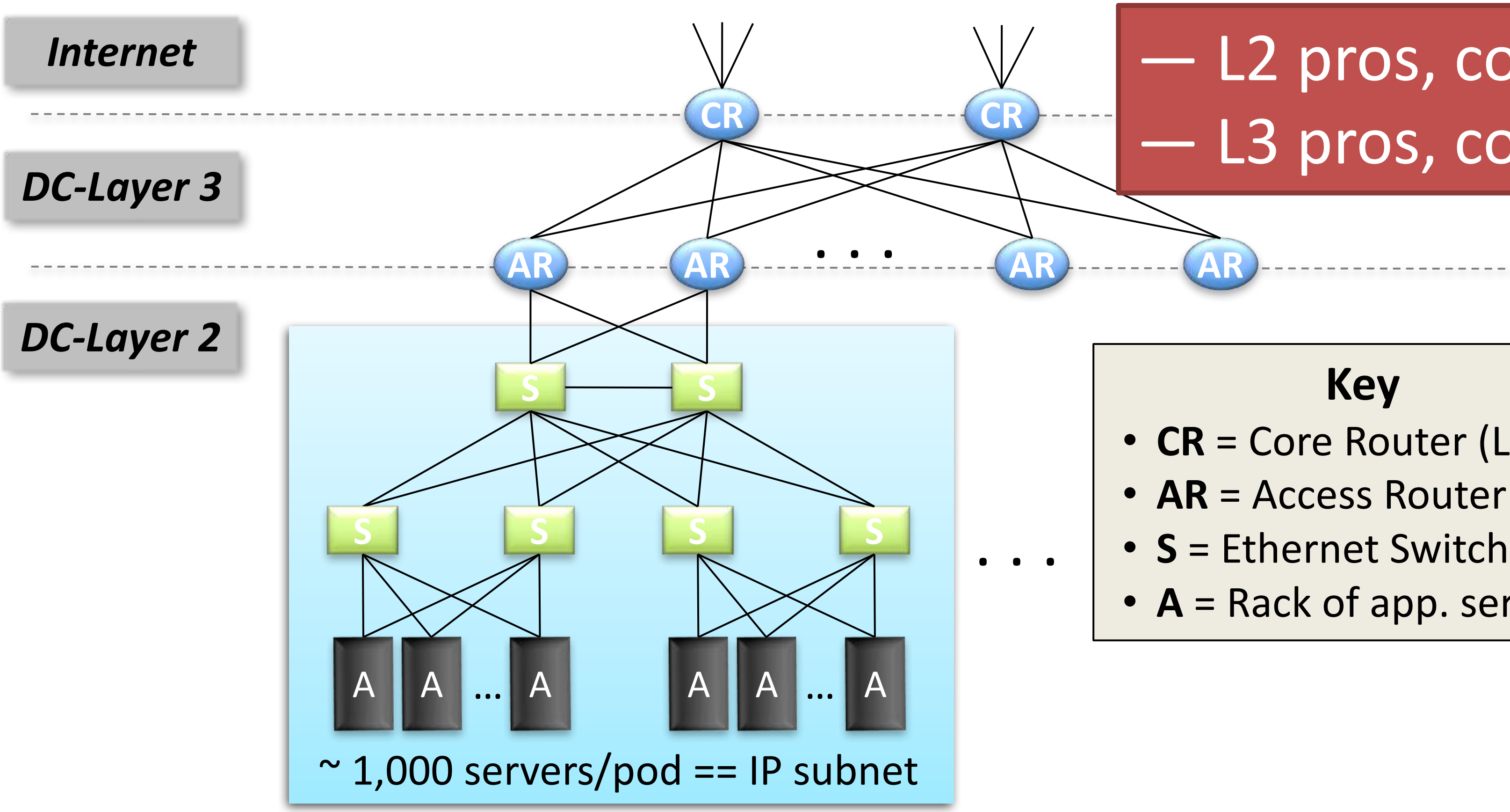
Today: Petabits/s in one DC

- More than core of the Internet!

✧ Source: "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", SIGCOMM 2015.

Conventional DC Network Problems

Conventional DC Network



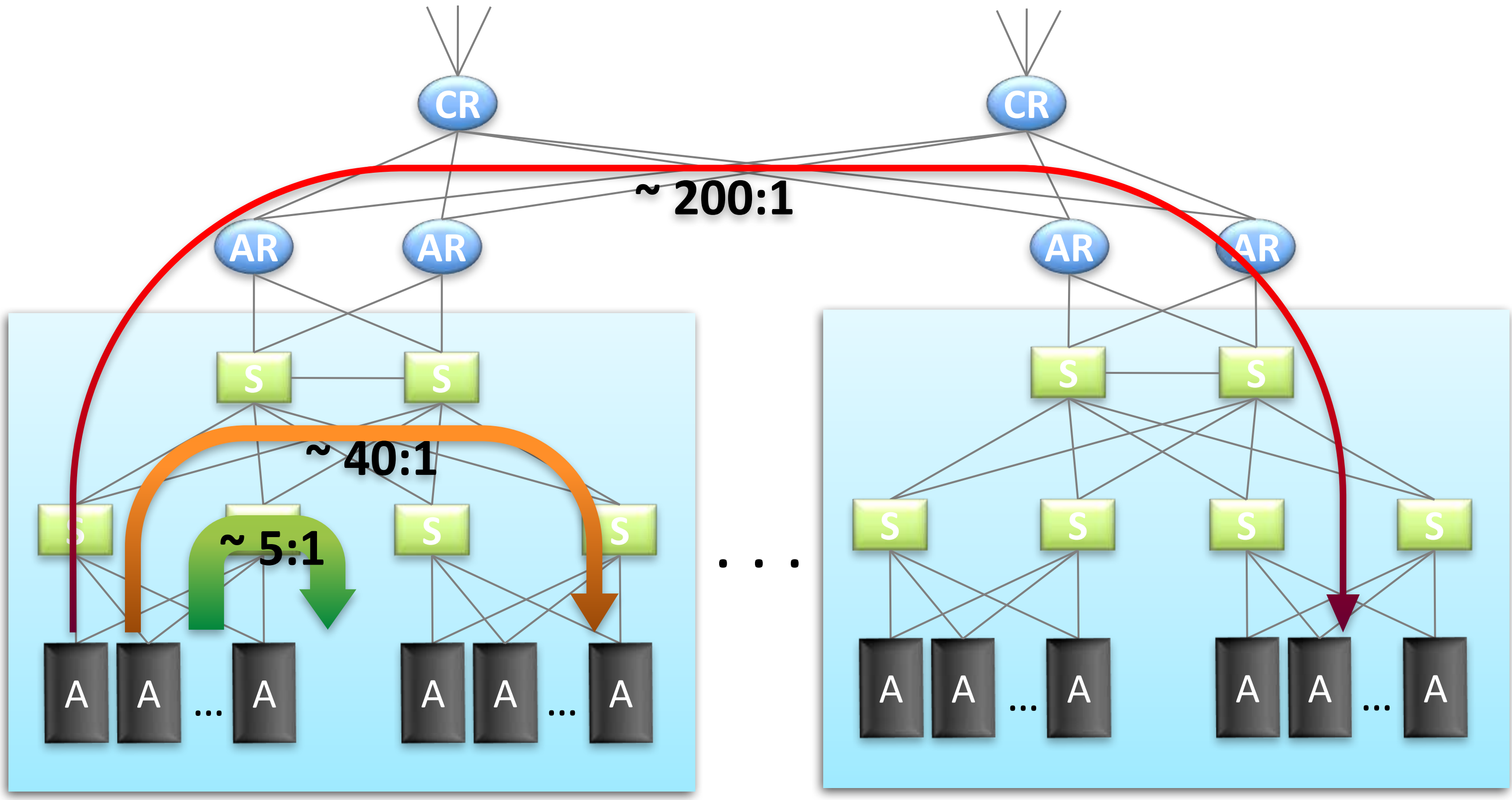
— L2 pros, cons?
— L3 pros, cons?

Key

- **CR** = Core Router (L3)
- **AR** = Access Router (L3)
- **S** = Ethernet Switch (L2)
- **A** = Rack of app. servers

Reference – "Data Center: Load balancing Data Center Services", Cisco 2004

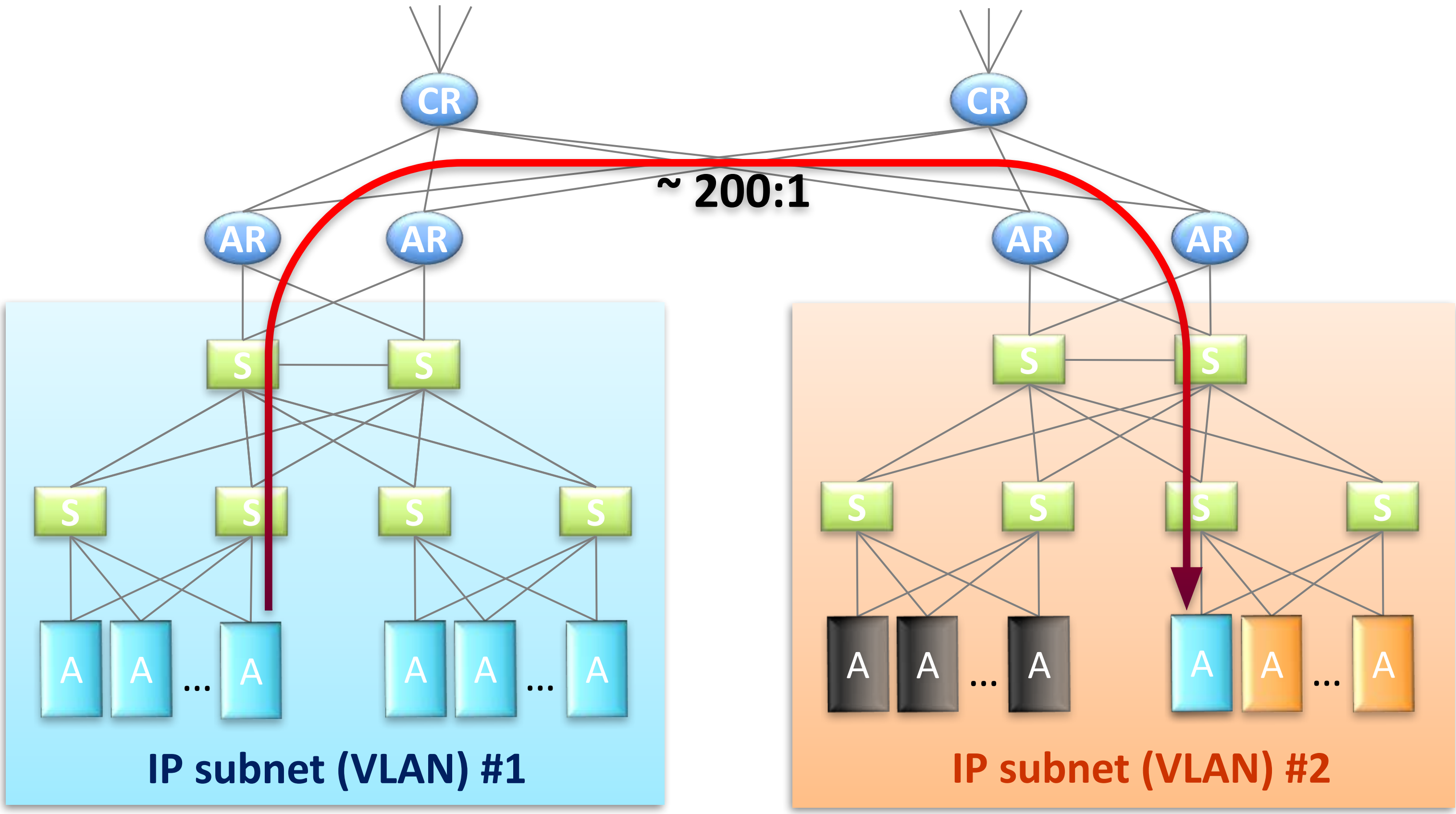
Conventional DC Network Problems



Dependence on high-cost proprietary routers

Extremely limited server-to-server capacity

Conventional DC Network Problems

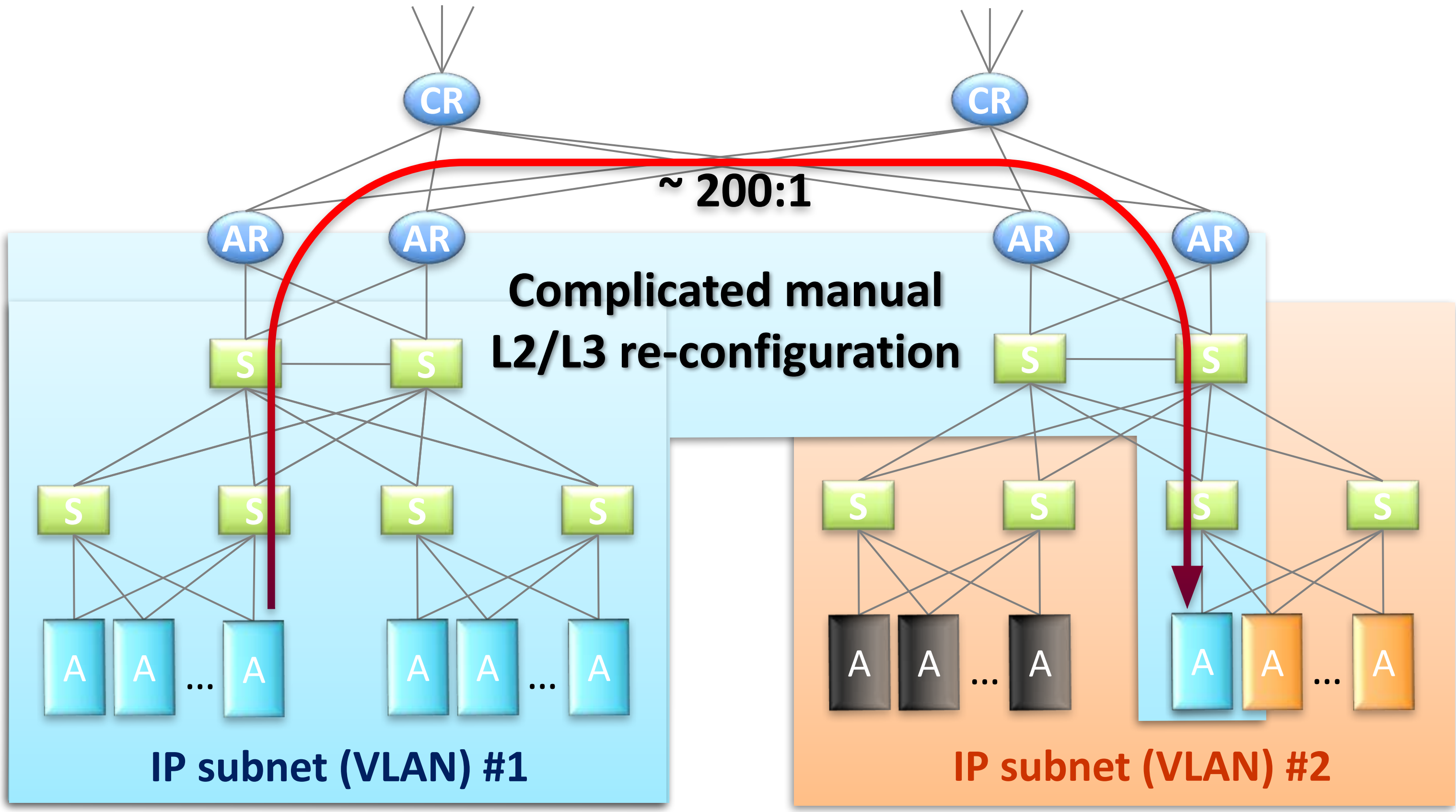


Dependence on high-cost proprietary routers

Extremely limited server-to-server capacity

Resource fragmentation

And More Problems …



Poor reliability

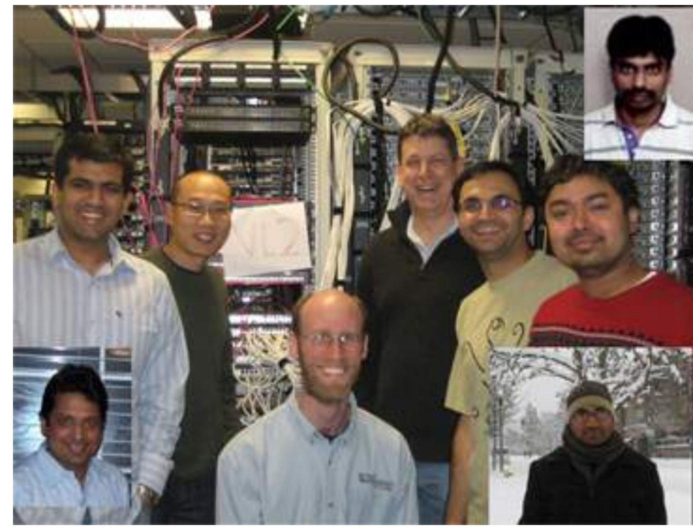
Lack of performance isolation

VL2 Paper

Measurements

VL2 Design

- Clos topology
- Valiant LB
- Name/location separation
  (precursor to network virtualization)



http://research.microsoft.com/en-US/news/features/datacenternetworking-081909.aspx

Measurements

DC Traffic Characteristics

Instrumented a large cluster used for data mining and identified distinctive traffic patterns

Traffic patterns are **highly volatile**

- A large number of distinctive patterns even in a day

Traffic patterns are **unpredictable**

- Correlation between patterns very weak

Traffic-aware optimization needs to be done frequently and rapidly

DC Opportunities

DC controller knows **everything** about **hosts**

Host OS's are easily **customizable**

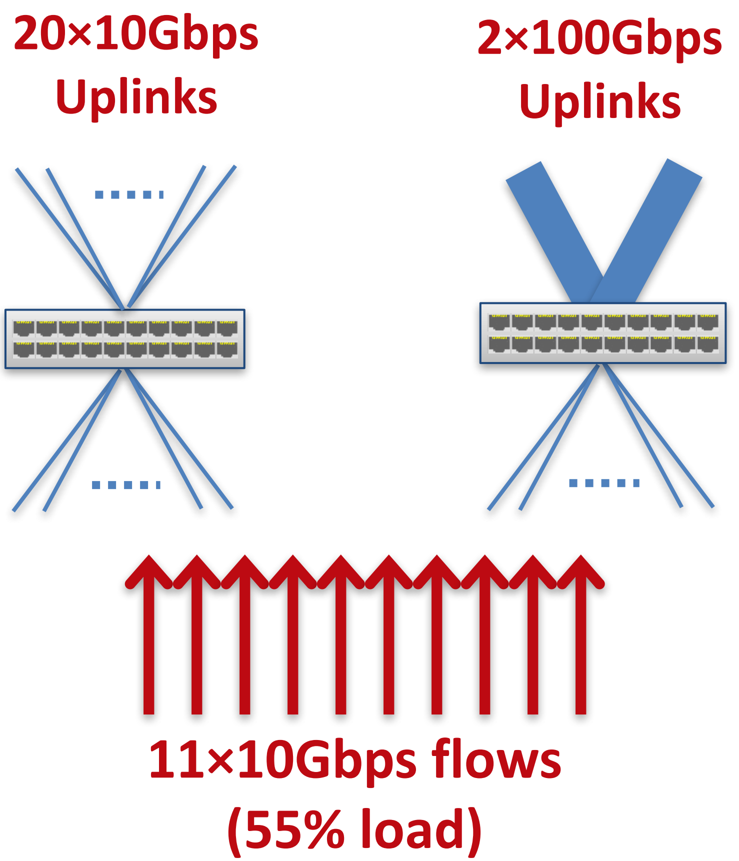
Probabilistic flow distribution would work well enough, because …

- Flows are numerous and not huge – no elephants
- Commodity switch-to-switch links are substantially thicker (~ 10x) than the maximum thickness of a flow

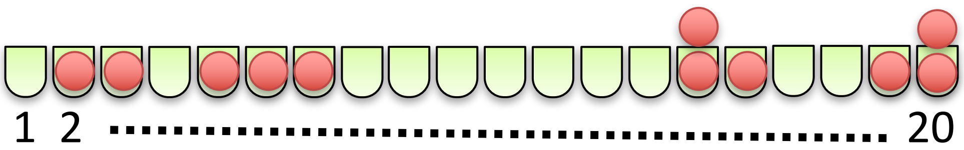
DC network can be made simple

Intuition

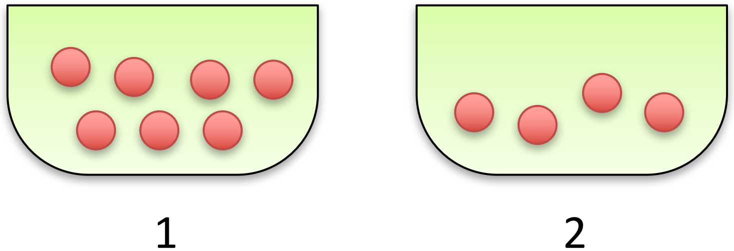
Higher speed links improve *flow-level* load balancing (ECMP)



Prob of 100% throughput = 3.27%

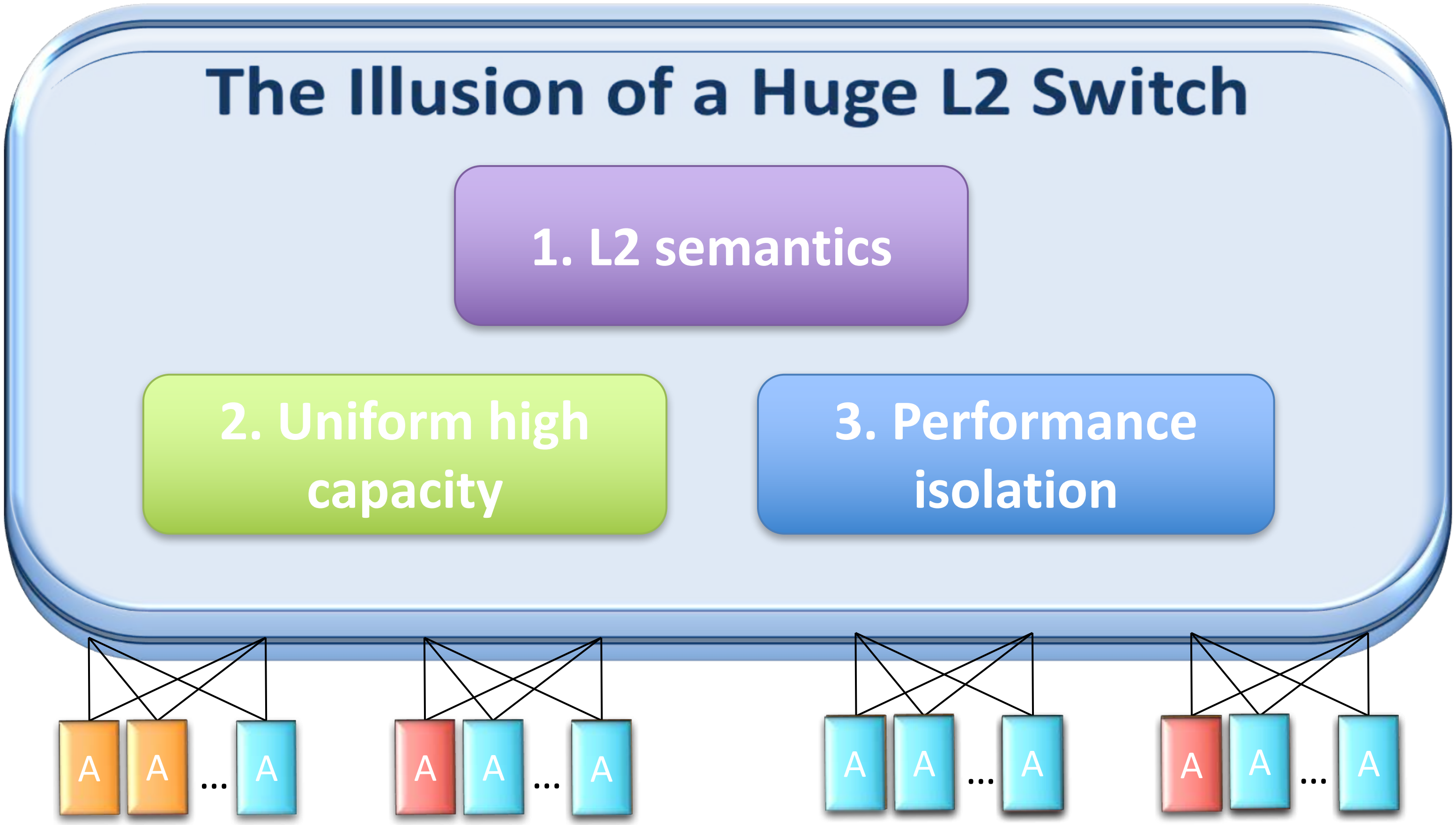


Prob of 100% throughput = 99.95%



Virtual Layer 2

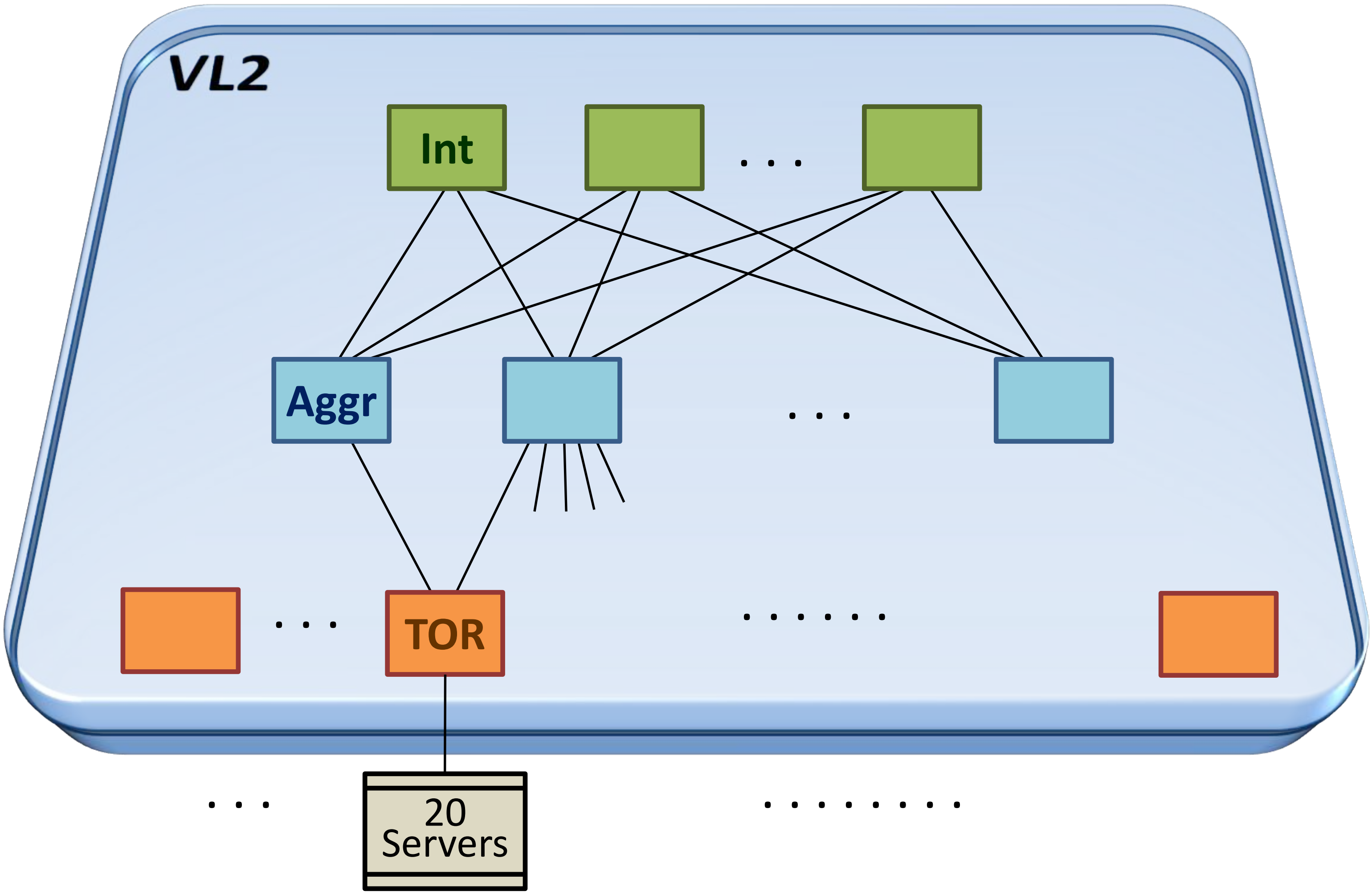
VL2 Goals



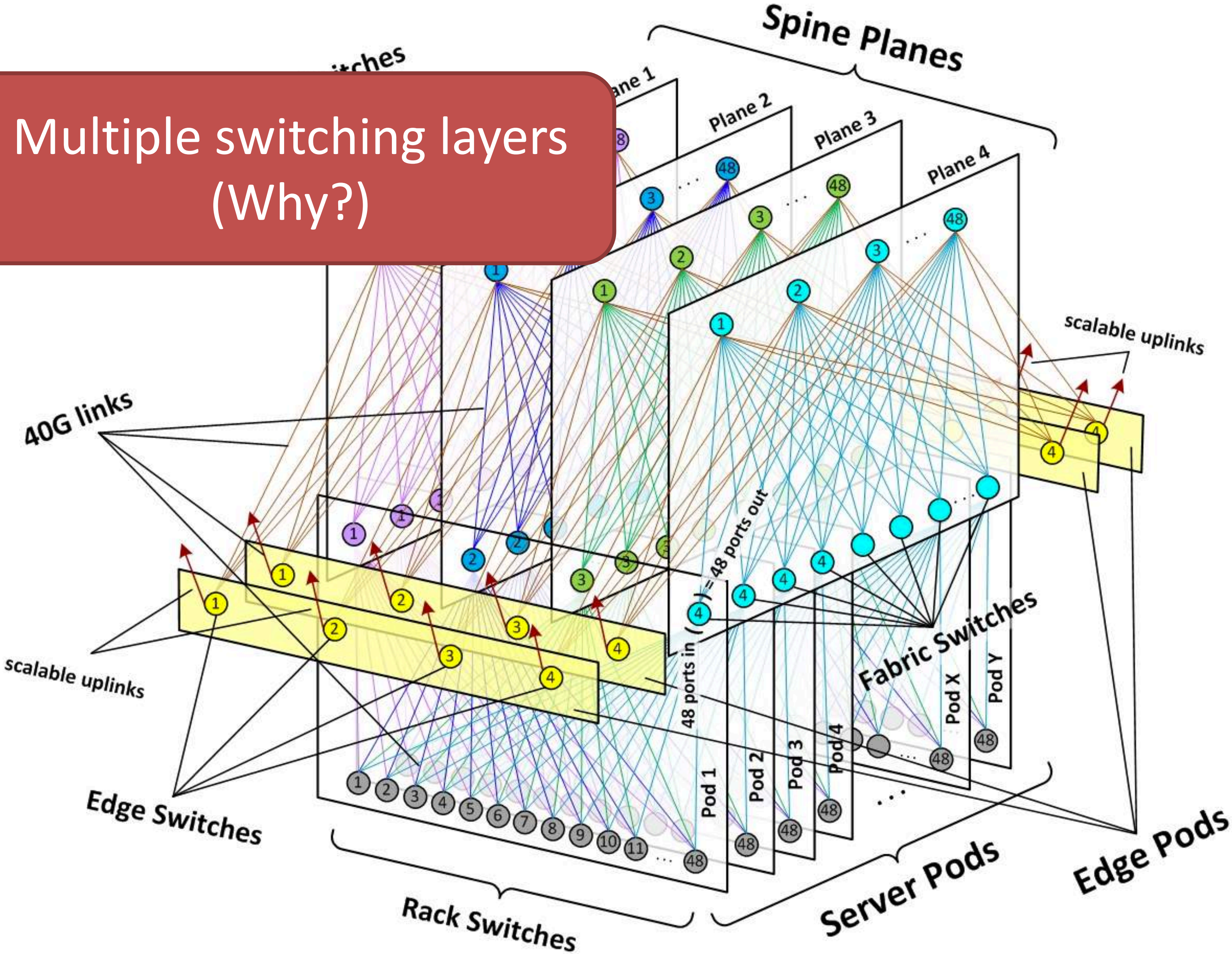
The Illusion of a Huge L2 Switch
1. L2 semantics
2. Uniform high capacity
3. Performance isolation
A
A
…
A
A
A
…
A
A
A
…
A
A
A
…
A

Clos Topology

Offer huge capacity via multiple paths (scale out, not up)



Multiple switching layers (Why?)



- https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

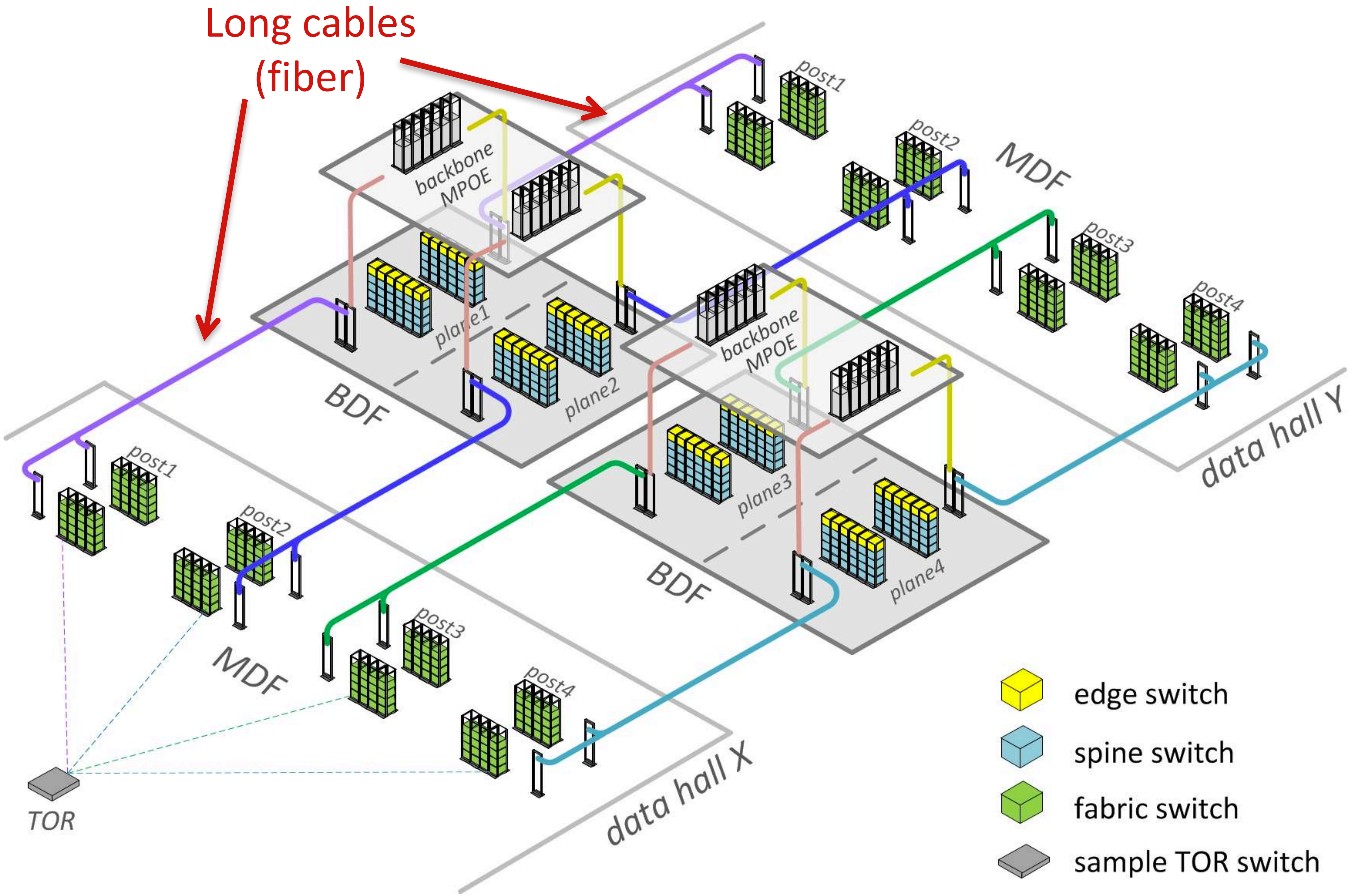
Building Block:
Merchant Silicon Switching Chips

Switch ASIC

Facebook Wedge

6 pack

- Image courtesy of Facebook



✧ https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

VL2 Design Principles

Randomizing to Cope with Volatility

– Tremendous variability in traffic matrices

Separating Names from Locations

– Any server, any service

Embracing End Systems

– Leverage the programmability & resources of servers
– Avoid changes to switches

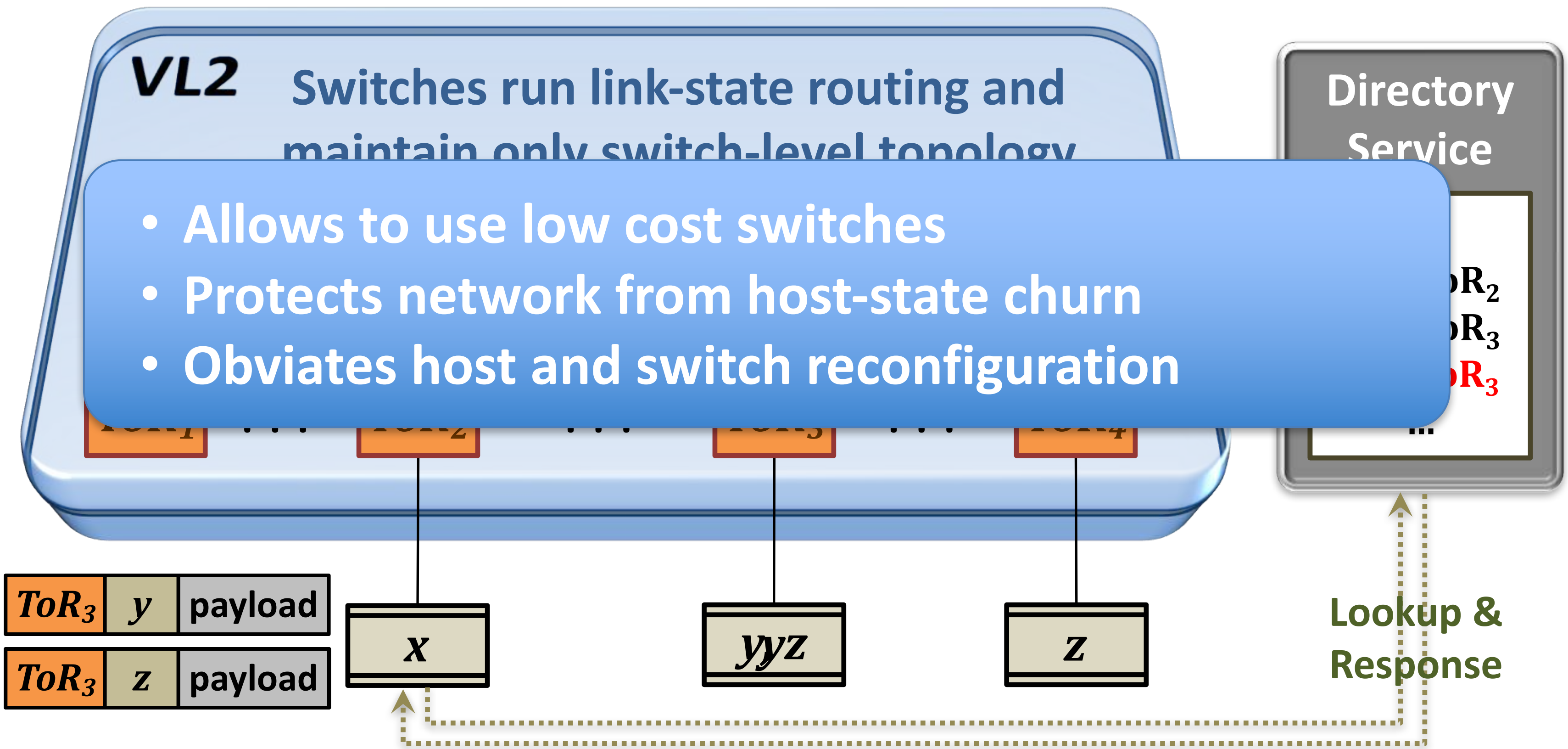
Building on Proven Networking Technology

– Build with parts shipping today
– Leverage low cost, powerful merchant silicon ASICs

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

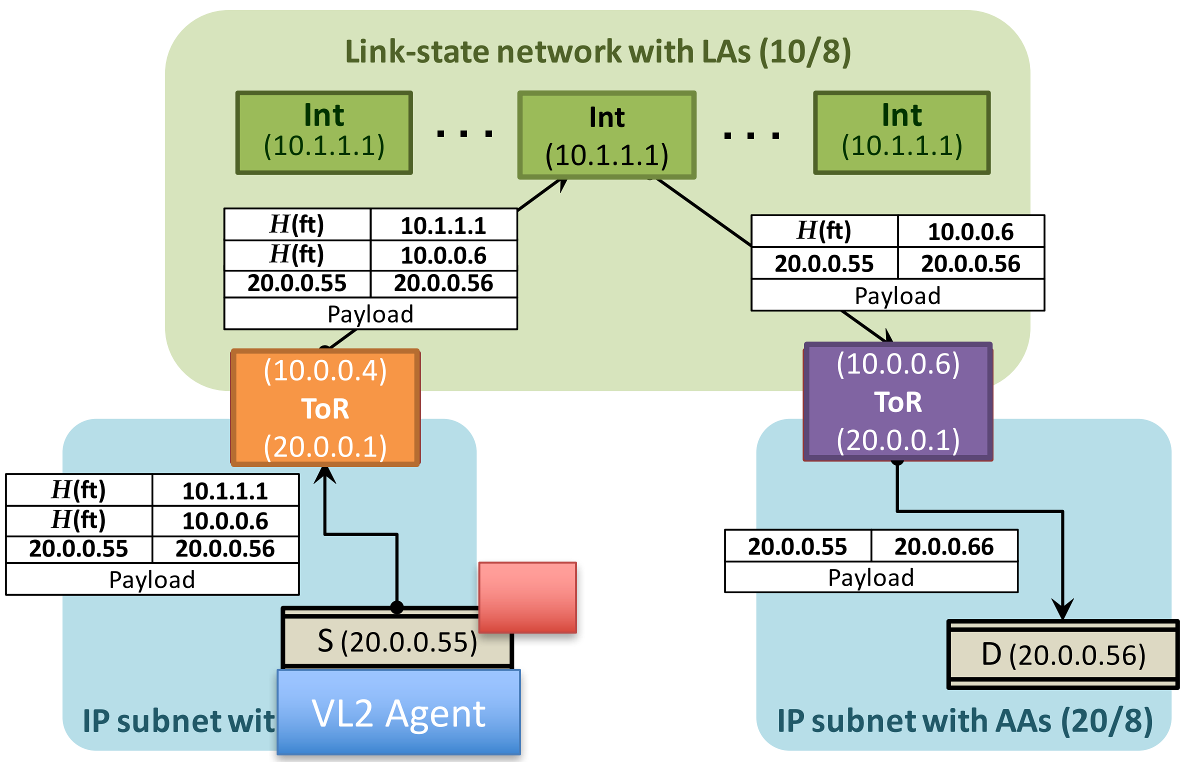
Addressing and Routing: Name-Location Separation

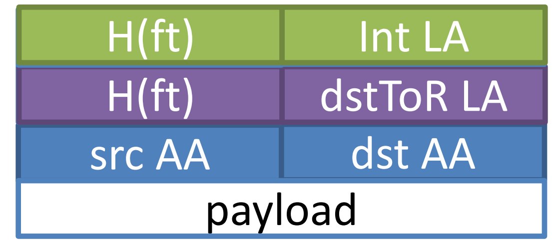


- Allows to use low cost switches
- Protects network from host-state churn
- Obviates host and switch reconfiguration

Servers use flat names

VL2 Agent in Action





VLB

ECMP

Why use hash for Src IP?
Why anycast & double encap?

Other details

How does L2 broadcast work?

- ARP requests is replaced by directory server (intercepted by VL2 agent)

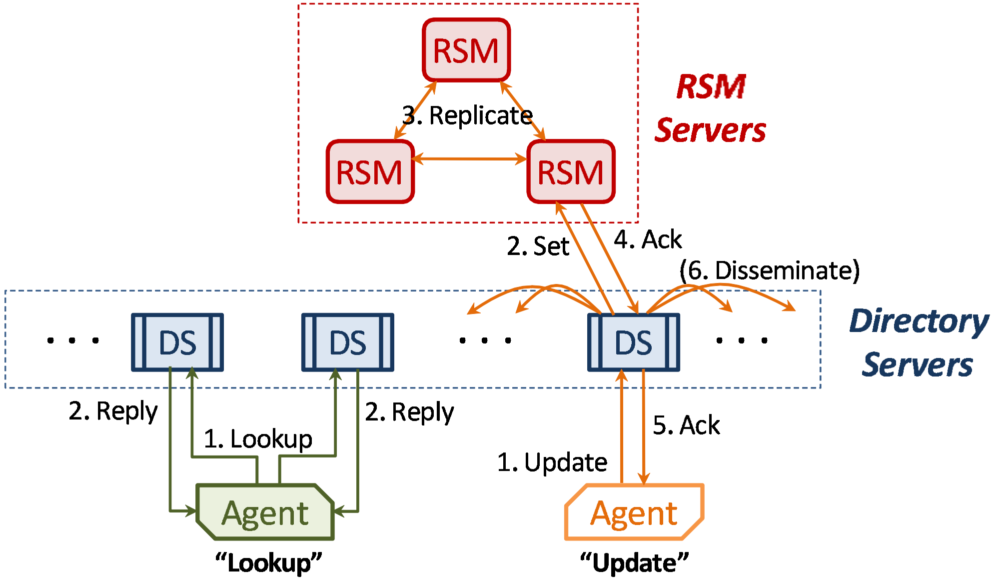
How does Internet communication work?

- Servers that need to be directly reachable from the Internet(e.g., front-end web servers) are assigned two addresses: an LA in addition to the AA used for intra-data-center communication with backend servers
- This LA is drawn from a pool that is announced via BGP and is externally reachable.

VL2 Directory System

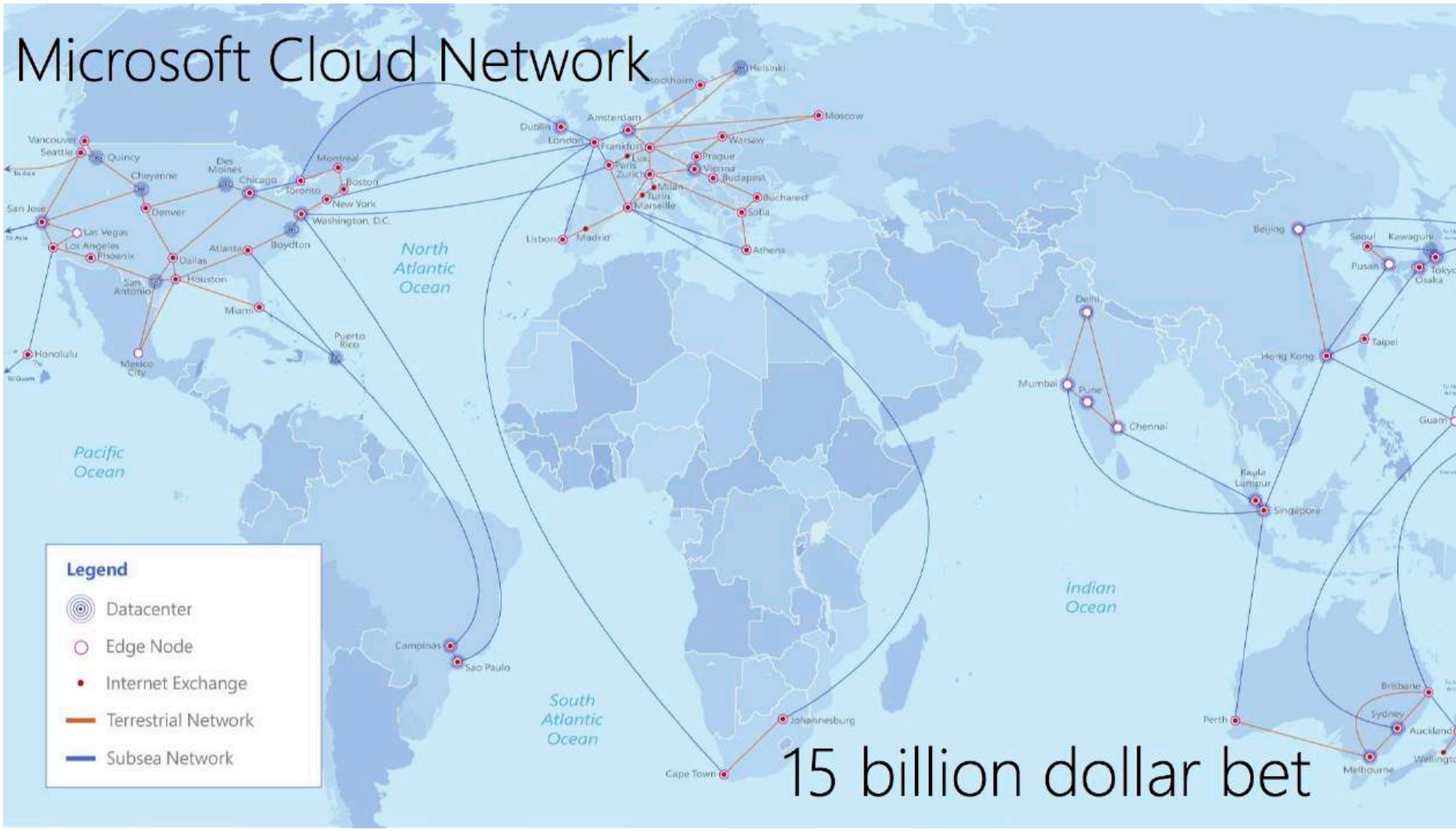
Read-optimized Directory Servers for lookups

Write-optimized
Replicated State
Machines for updates

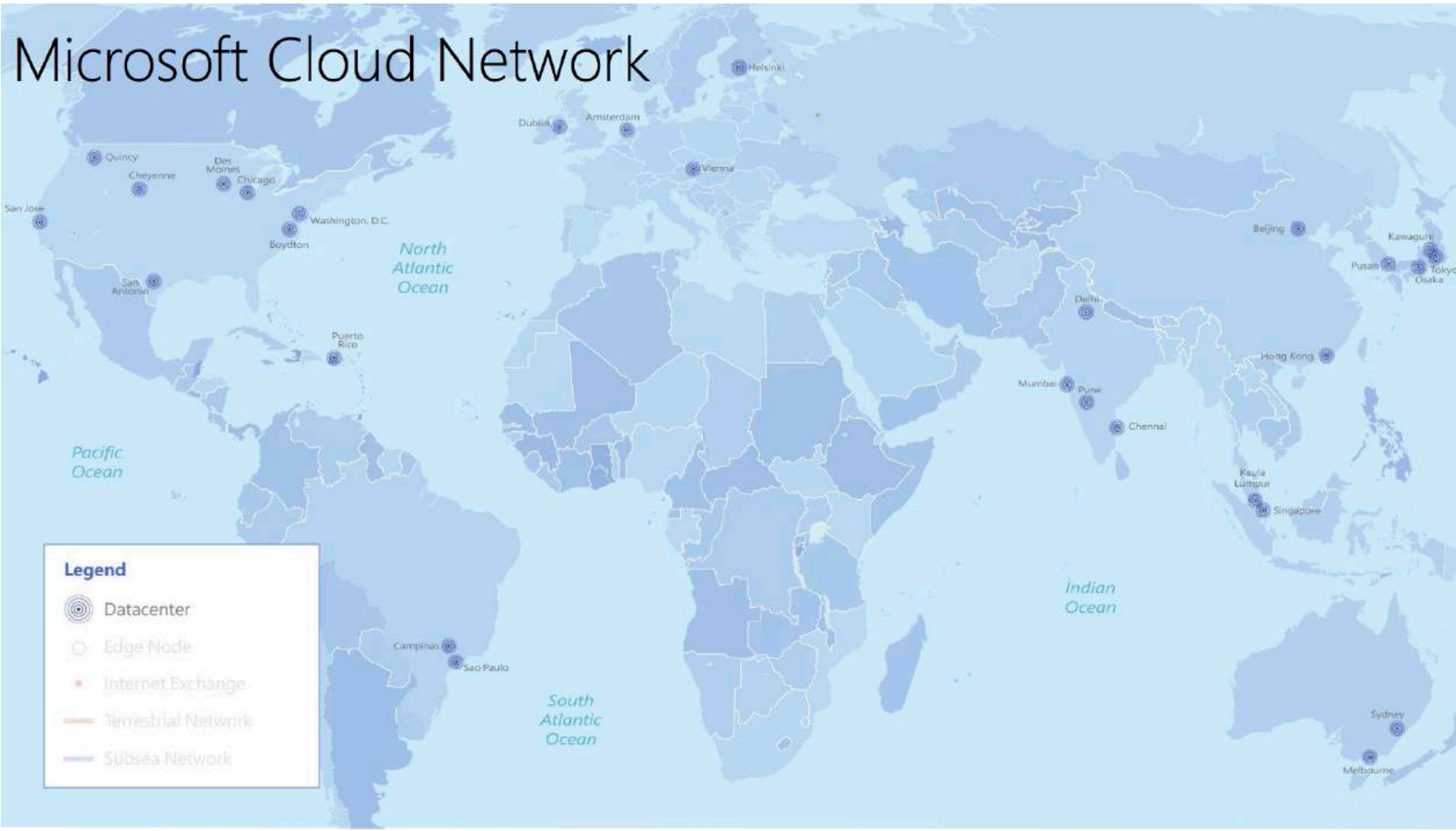


Stale mappings?

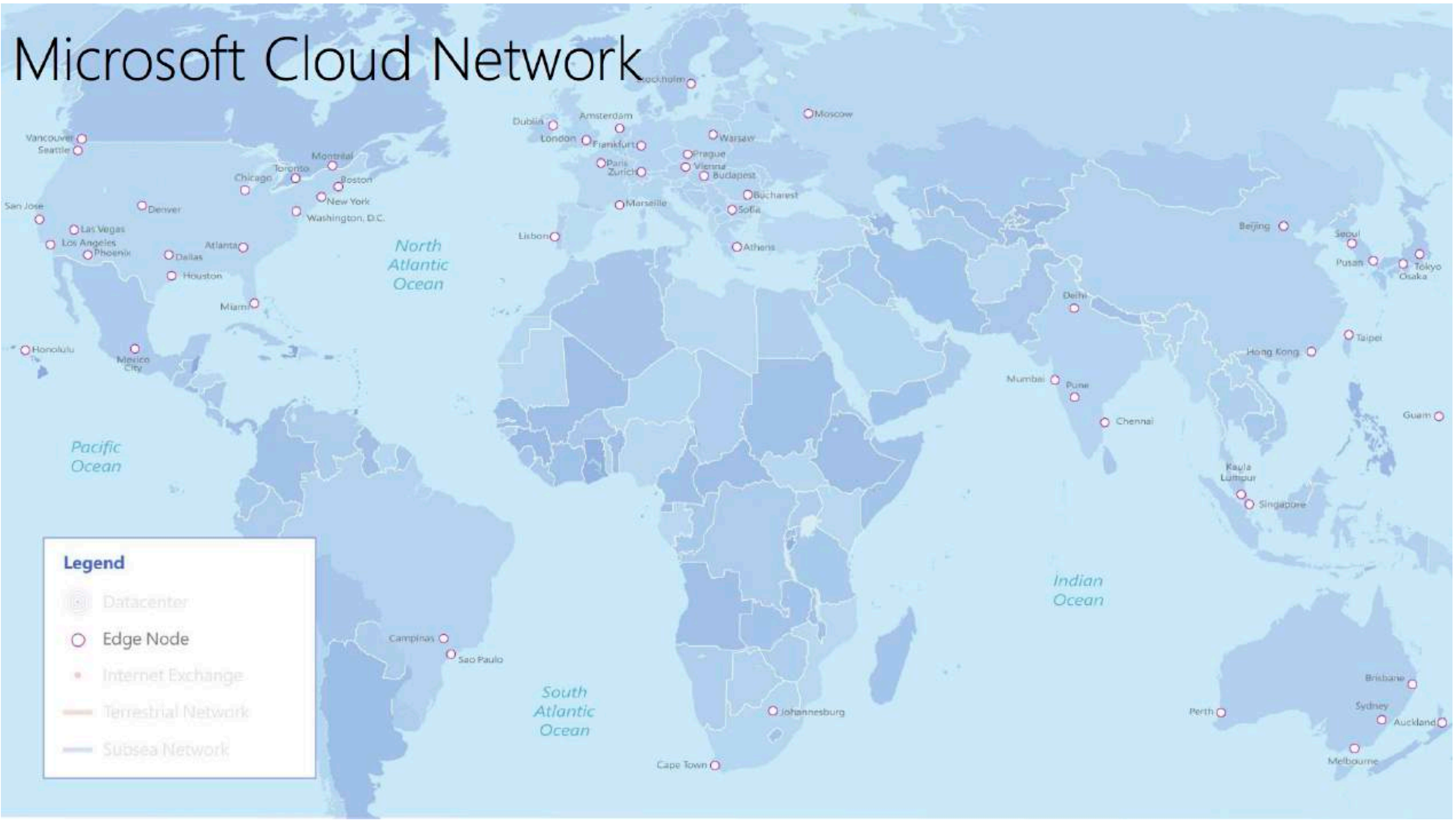
VL2 -> Azure


Microsoft Cloud Network
Legend
Datacenter
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network
North Atlantic Ocean
South Atlantic Ocean
Pacific Ocean
Indian Ocean
Vancouver
Seattle
Quincy
Cheyenne
Des Moines
Chicago
Toronto
Montreal
Boston
New York
Washington, D.C.
Boydton
San Jose
Las Vegas
Los Angeles
Phoenix
Denver
Dallas
Atlanta
San Antonio
Houston
Miami
Puerto Rico
Honolulu
Mexico City
Campinas
Sao Paulo
Dublin
London
Amsterdam
Stockholm
Helsinki
Moscow
Frankfurt
Paris
Zurich
Warsaw
Prague
Vienna
Budapest
Milan
Marseille
Bucharest
Sofia
Athens
Lisbon
Madrid
Johannesburg
Cape Town
Delhi
Mumbai
Pune
Chennai
Singapore
Beijing
Seoul
Pusan
Tokyo
Osaka
Taipei
Hong Kong
Guam
Perth
Brisbane
Sydney
Melbourne
Auckland
Wellington
15 billion dollar bet

VL2 -> Azure


Microsoft Cloud Network
Helsinki
Dublin
Amsterdam
Quincy
Cheyenne
Des Moines
Chicago
San Jose
Washington, D.C.
Boydton
San Antonio
Vienna
North Atlantic Ocean
Puerto Rico
Pacific Ocean
Beijing
Kawaguchi
Pusan
Tokyo
Osaka
Delhi
Hong Kong
Mumbai
Pune
Chennai
Kuala Lumpur
Singapore
Indian Ocean
Campinas
Sao Paulo
South Atlantic Ocean
Sydney
Melbourne
Legend
Datacenter
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network

VL2 -> Azure


Microsoft Cloud Network
Legend
Datacenter
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network
North Atlantic Ocean
South Atlantic Ocean
Pacific Ocean
Indian Ocean
Vancouver
Seattle
San Jose
Los Angeles
Las Vegas
Phoenix
Denver
Dallas
Houston
Chicago
Toronto
Montreal
Boston
New York
Washington, D.C.
Atlanta
Miami
Honolulu
Mexico City
Campinas
Sao Paulo
Dublin
London
Amsterdam
Frankfurt
Paris
Zurich
Marseille
Lisbon
Stockholm
Warsaw
Prague
Vienna
Budapest
Bucharest
Sofia
Athens
Moscow
Johannesburg
Cape Town
Delhi
Mumbai
Pune
Chennai
Beijing
Seoul
Pusan
Tokyo
Osaka
Taipei
Hong Kong
Guam
Kuala Lumpur
Singapore
Perth
Brisbane
Sydney
Melbourne
Auckland

VL2 -> Azure

	2010	2015
Compute Instances	100K	Millions
Azure Storage	10's of PB	Exabytes
Datacenter Network	10's of Tbps	Pbps

Microsoft