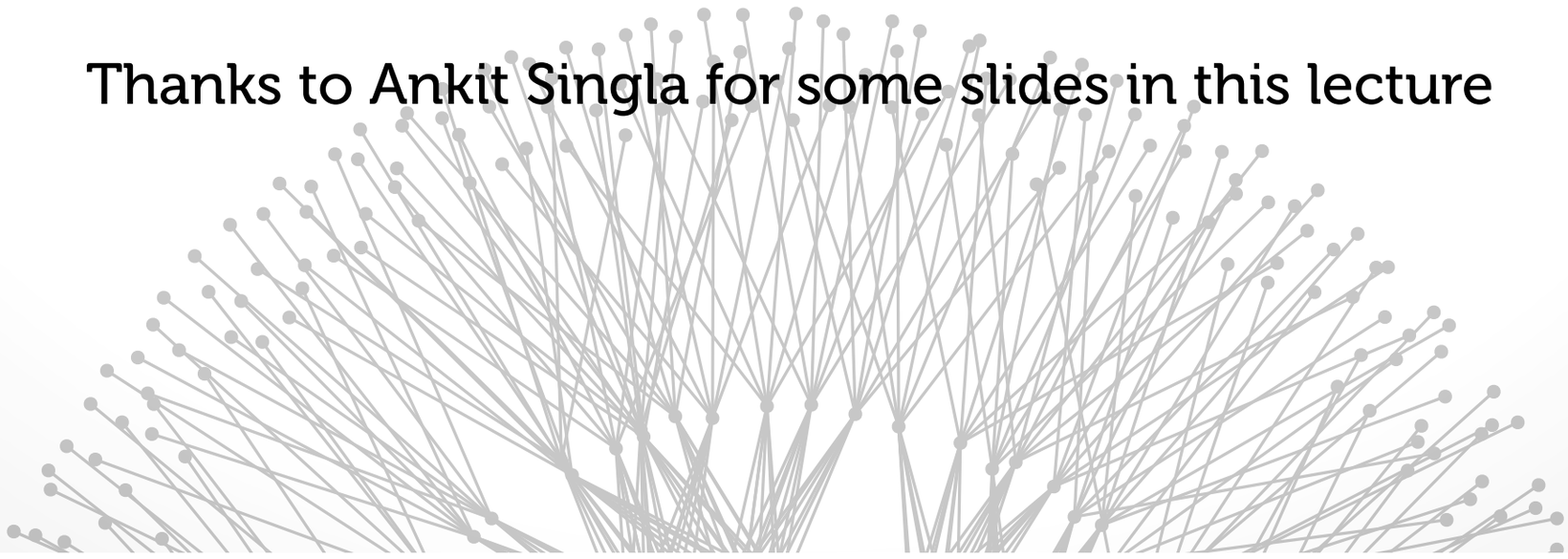
Data Center Networks

Brighten Godfrey
CS 538 April 4 2018

Thanks to Ankit Singla for some slides in this lecture



Introduction:
The Driving Trends

Cloud Computing

Computing as a utility

- Purchase however much you need, whenever you need it
- Service ranges from access to raw (virtual) machines, to higher level: distributed storage, web services

Implications

- Reduces barrier to entry to building large service
  - No need for up-front capital investment
  - No need to plan ahead
- May reduce cost
- Compute and storage becomes more centralized

The physical cloud: Data centers

Facebook data center, North Carolina

National Petascale Computing Facility, UIUC


google.com/datacenters

Weather Condition	Frequency of Occurrence
ASHRAE 50 Year	Combines worst recorded wetbulb (from 1972 to 2001) along with hottest drybulb in 50 years (these generally do not occur at the same time)
Extreme DB	This state point assumes the drybulb is even higher than worst in 50 years.
ASHRAE 0.4% (Evaporative)	This state point is exceeded only 35.0 hours in a statistically "typical" year
ASHRAE 1% (Evaporative)	This state point is exceeded only 87.6 hours in a statistically "typical" year
ASHRAE 2% (Evaporative)	This state point is exceeded only 175.2 hours in a statistically "typical" year
ASHRAE Winter Peak	This represents 50% RH at the coldest drybulb in 50 years

VITESSE COLD AISLE CRITERIA
65°F TO 80°F DB
41.9°F TO 59.0°F DP
MAX 65% RH

ASHRAE GUIDELINES
64.4°F TO 80.6°F DB
41.9°F TO 59.0°F DP
MAX 60%RH

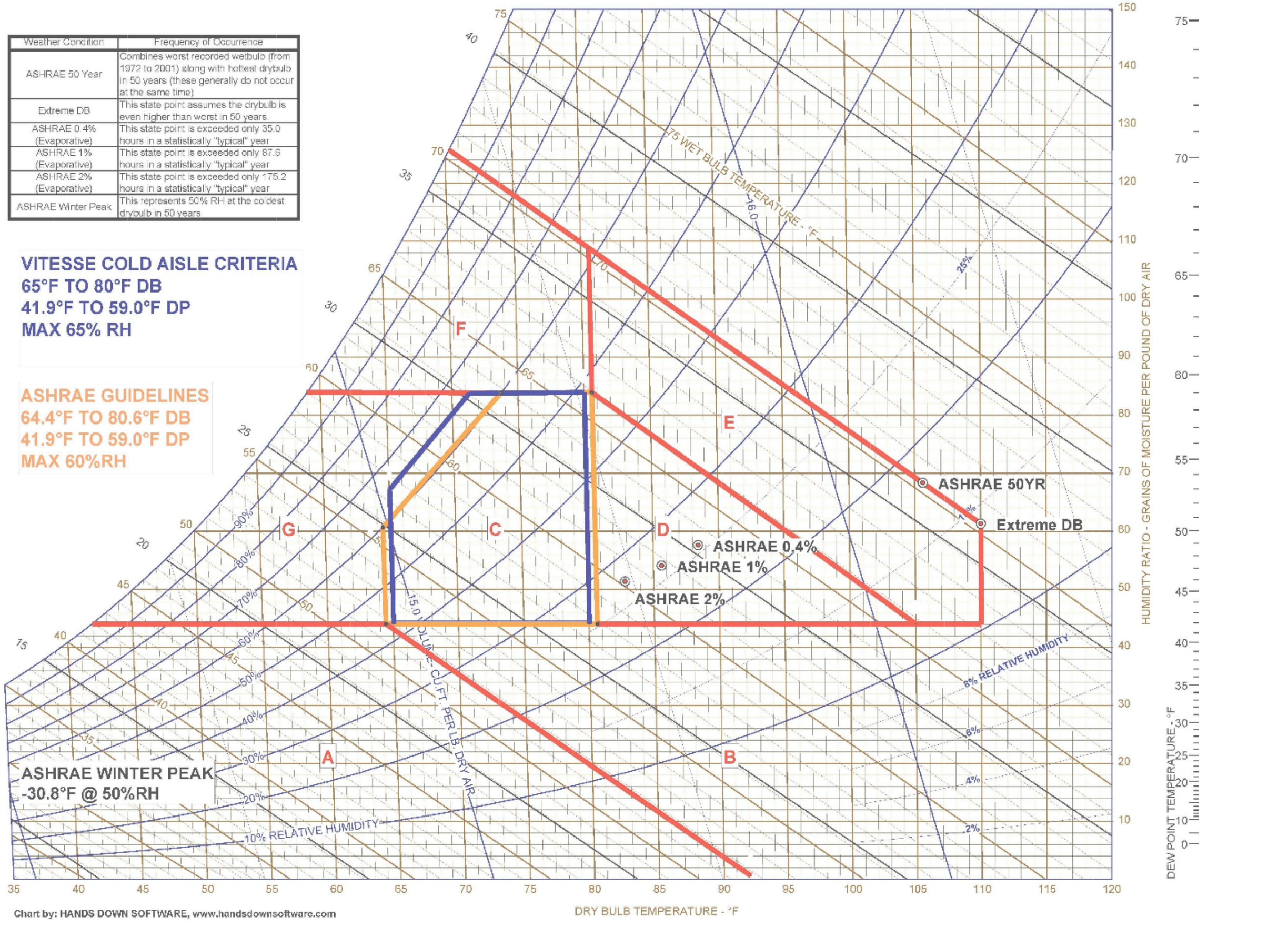


Chart by: HANDS DOWN SOFTWARE, www.handsdownsoftware.com

[Data Center v1.0, Open Compute Projec

Key advantage: economy of scale

One technician for each 15,000 servers [Facebook]

Facility / power infrastructure operated in bulk

- Power usage efficiency (PuE) ~ 1.8 in average DCs
- Pushed down to ~ 1.1 in large cloud DCs

Ability to custom-design equipment

- Servers, switches, NICs…

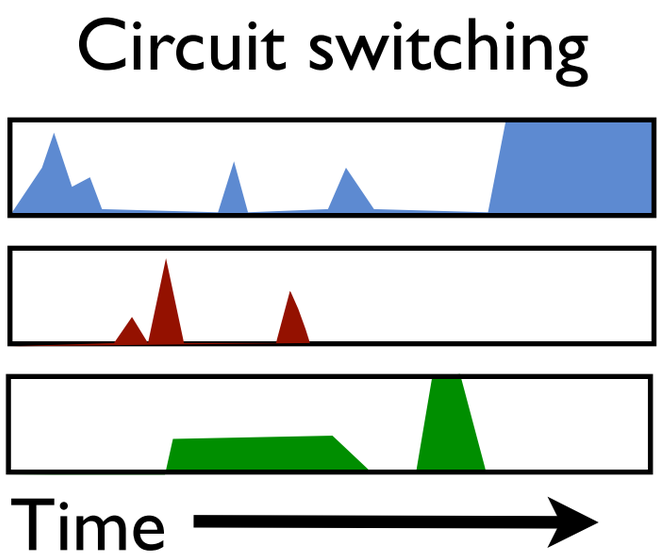
Statistical multiplexing

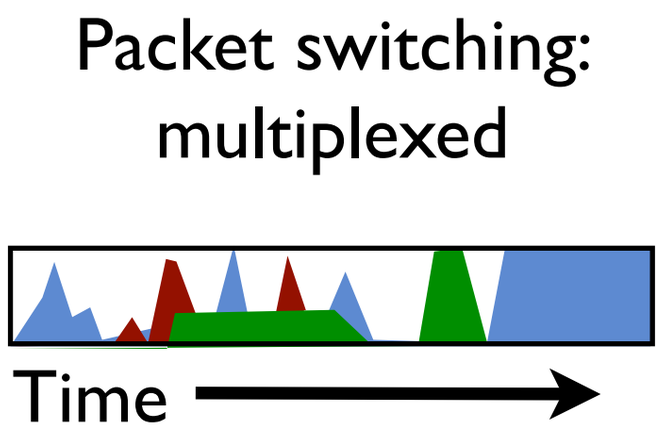
- Must provision for peak load
- Many users sharing a resource are unlikely to have their peaks all at the same time

Key advantage: economy of scale

Statistical multiplexing

- Must provision for peak load
- Many users sharing a resource are unlikely to have their peaks all at the same time
- Just as in packet switching





Challenges for Cloud Computing

Challenges

- Confidentiality of data and computation
- Isolation of resources
- Integration with existing systems
- Robustness
- Latency
- Bandwidth
- Programmability
- ...

Opportunities

- New systems and architectures
- Optimizations matter

Costs in a data center

Servers are expensive!

Amortized Cost	Component	Sub-Components
∼45%	Servers	CPU, memory, storage systems
∼25%	Infrastructure	Power distribution and cooling
∼15%	Power draw	Electrical utility costs
∼15%	Network	Links, transit, equipment

[Greenberg, CCR Jan. 2009]

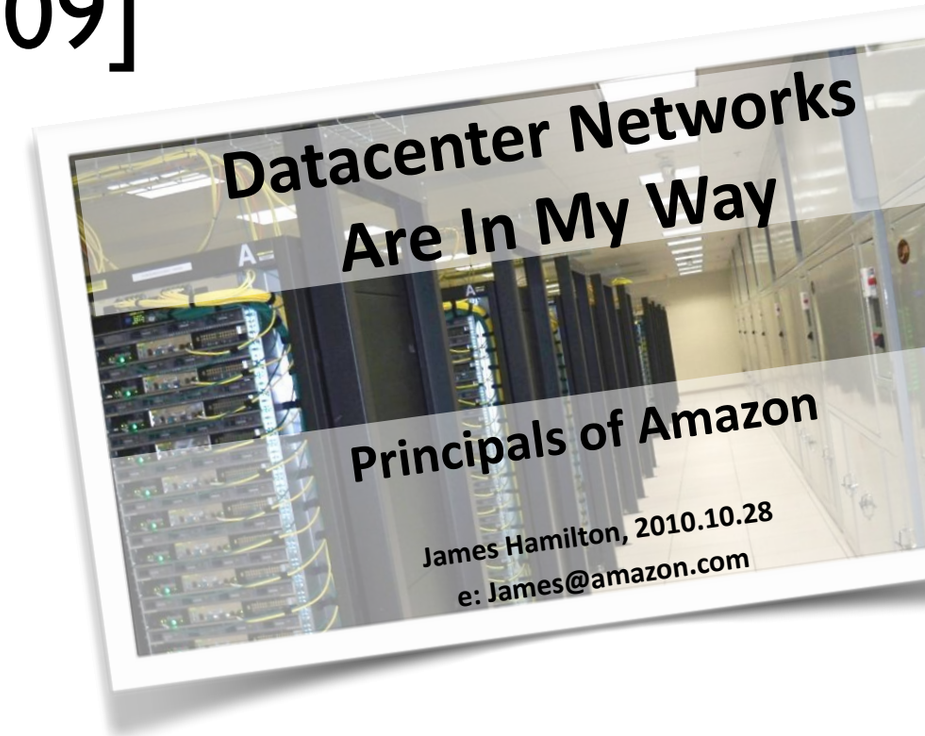
A key goal: Agility

Agility: Use any server for any service at any time

- Increase utilization of servers
- Reduce costs, increase reliability

What we need [Greenberg, ICDCS'09]

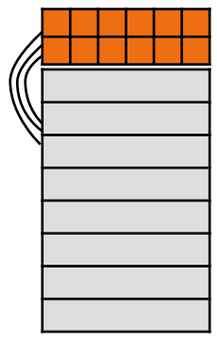
- Rapid installation of service's code
  - Solution: virtual machines



- Access to data from anywhere
  - Solution: distributed filesystems

- Ability to communicate between servers quickly, regardless of where they are in the data center

A server rack

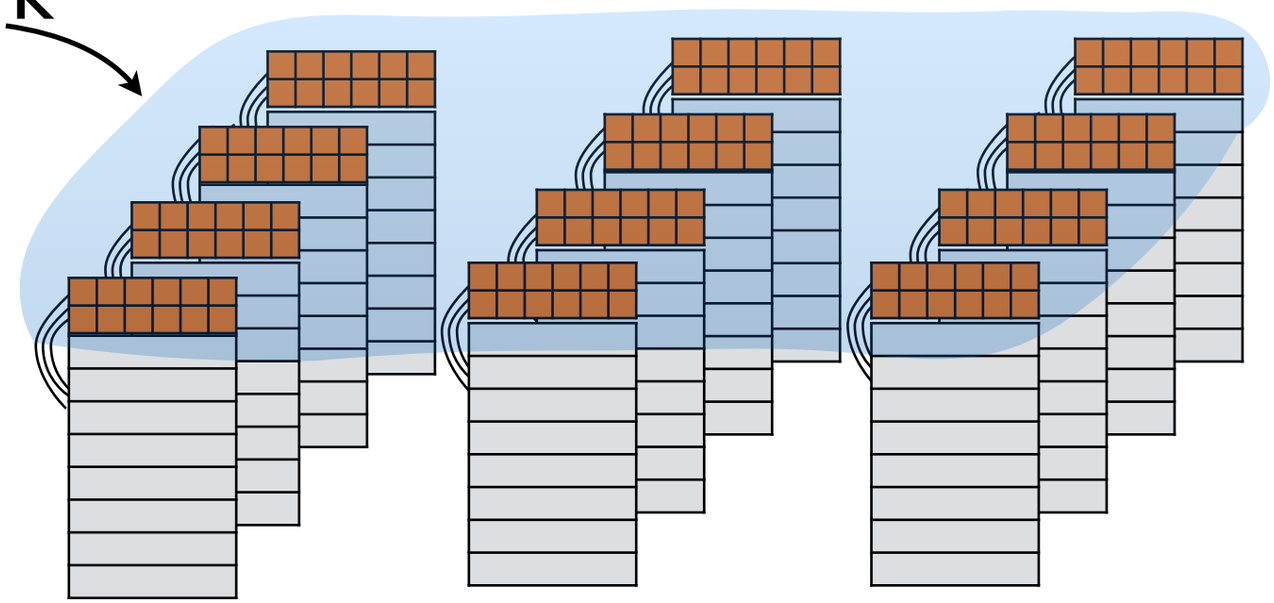


A top-of-rack switch

A rack of servers

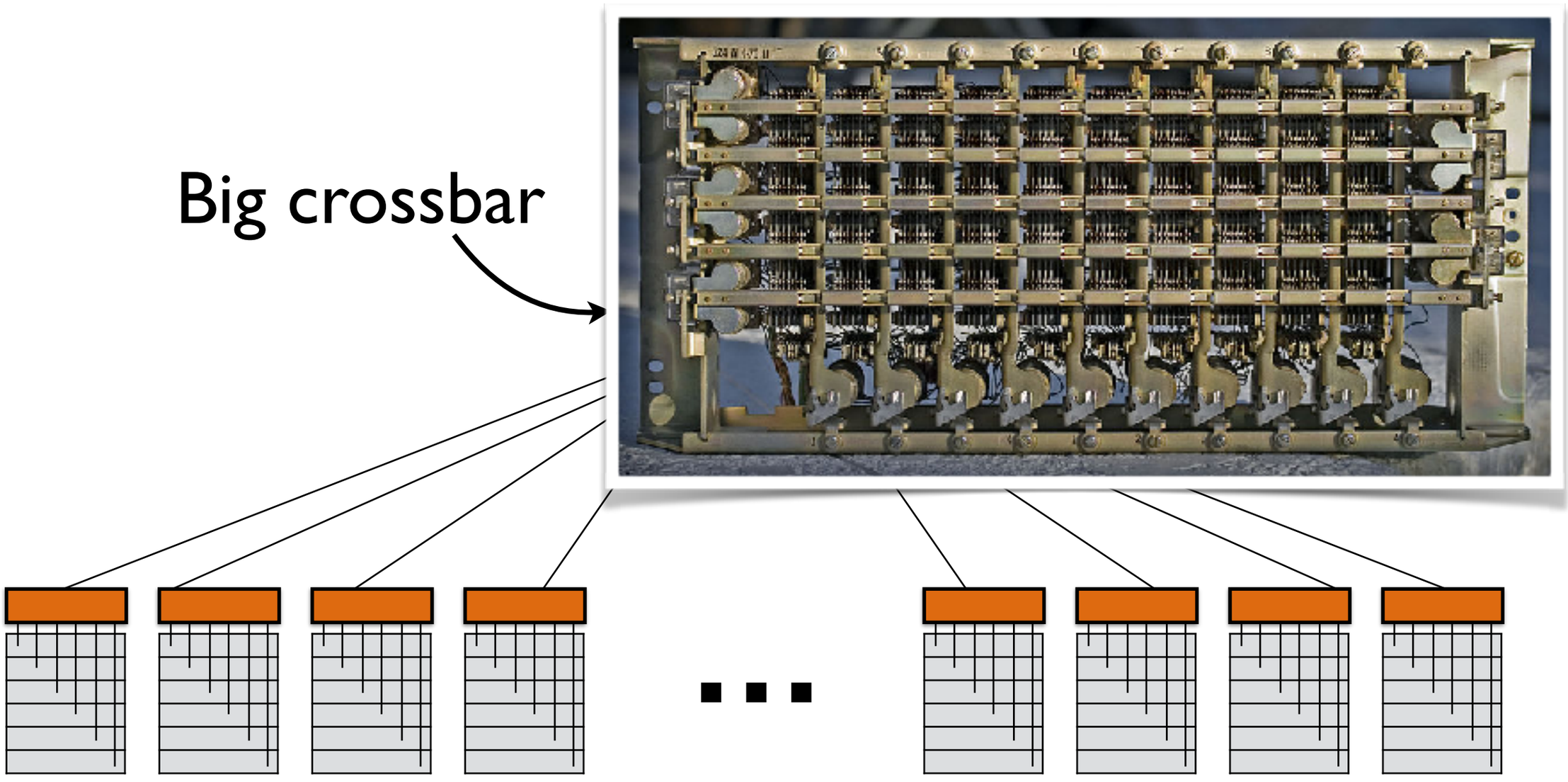
Lots of racks

How to network the racks?

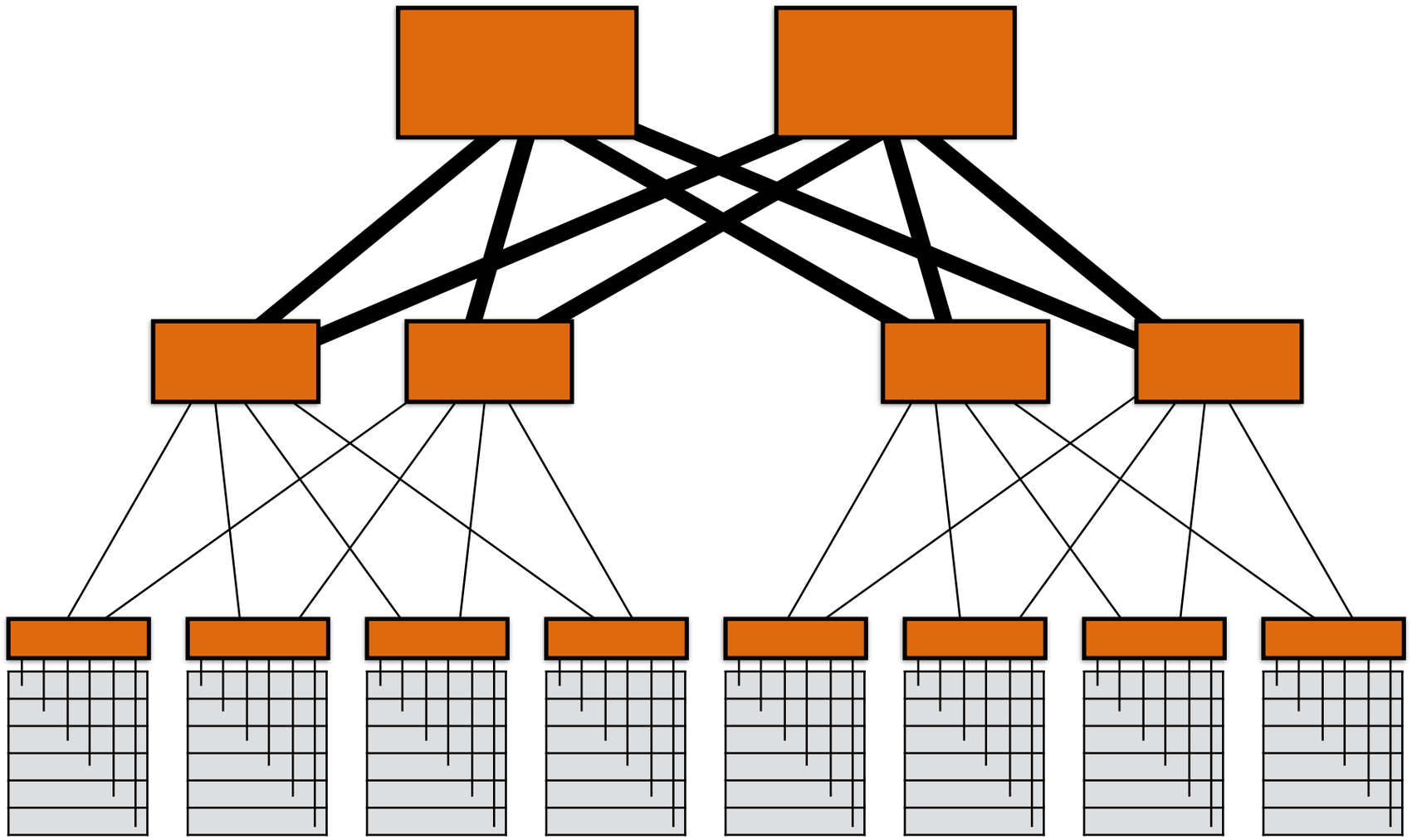


Facebook: machine-machine traffic "doubling at an interval of less than a year"

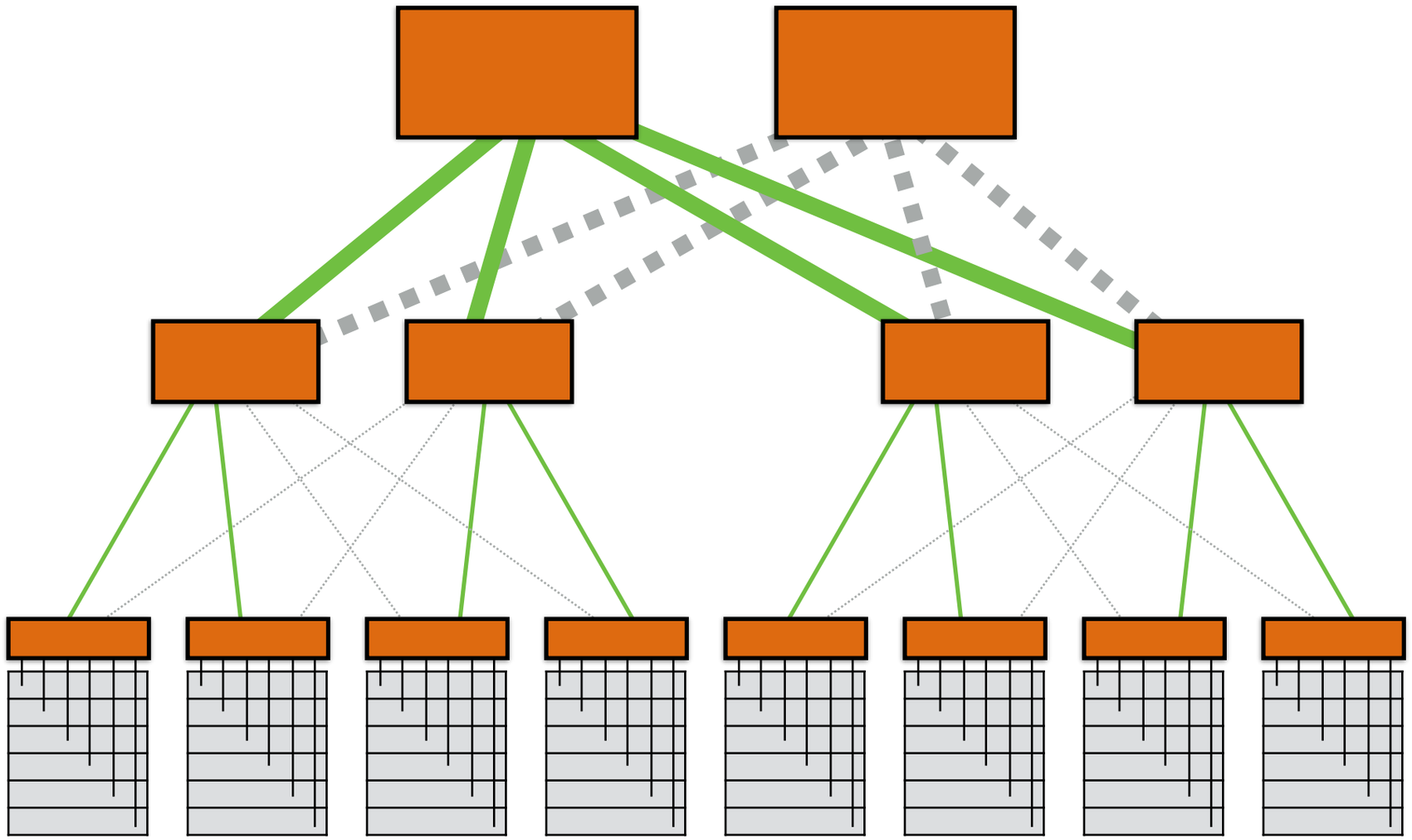
“Big switch” approach



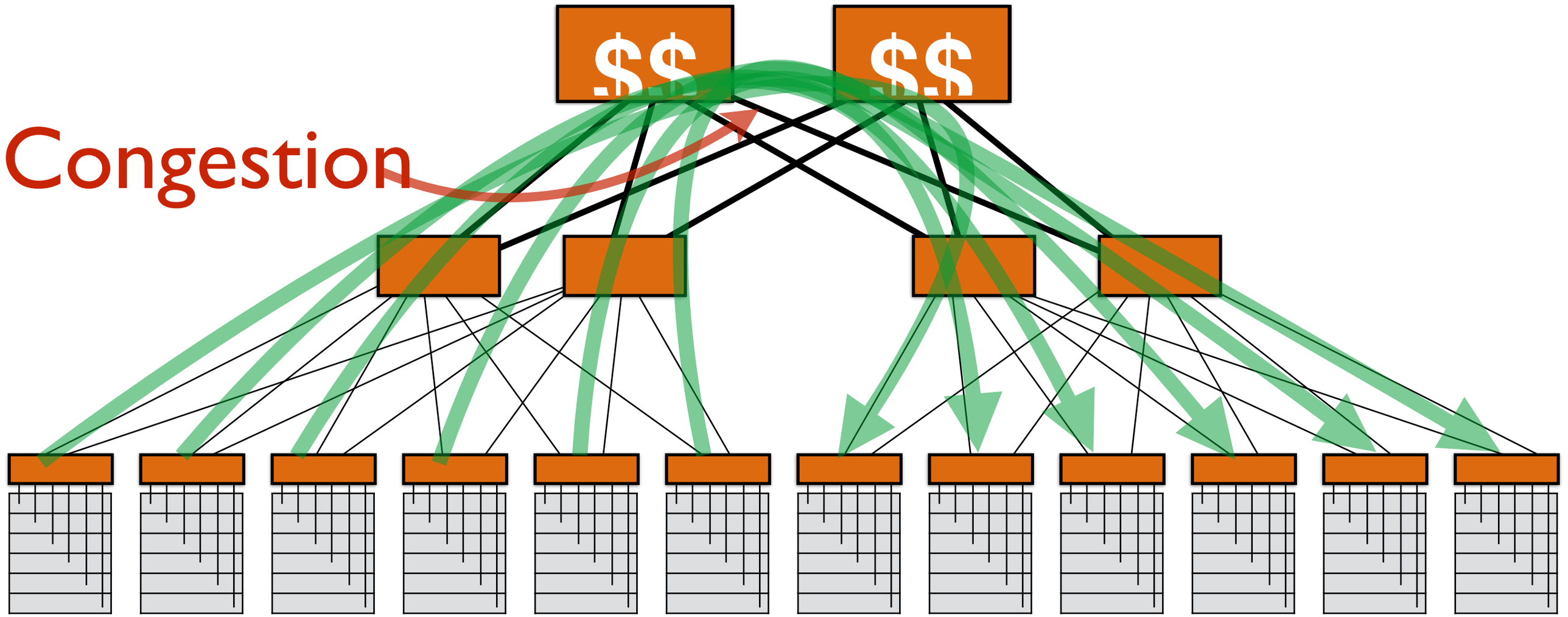
Alternative: tree network



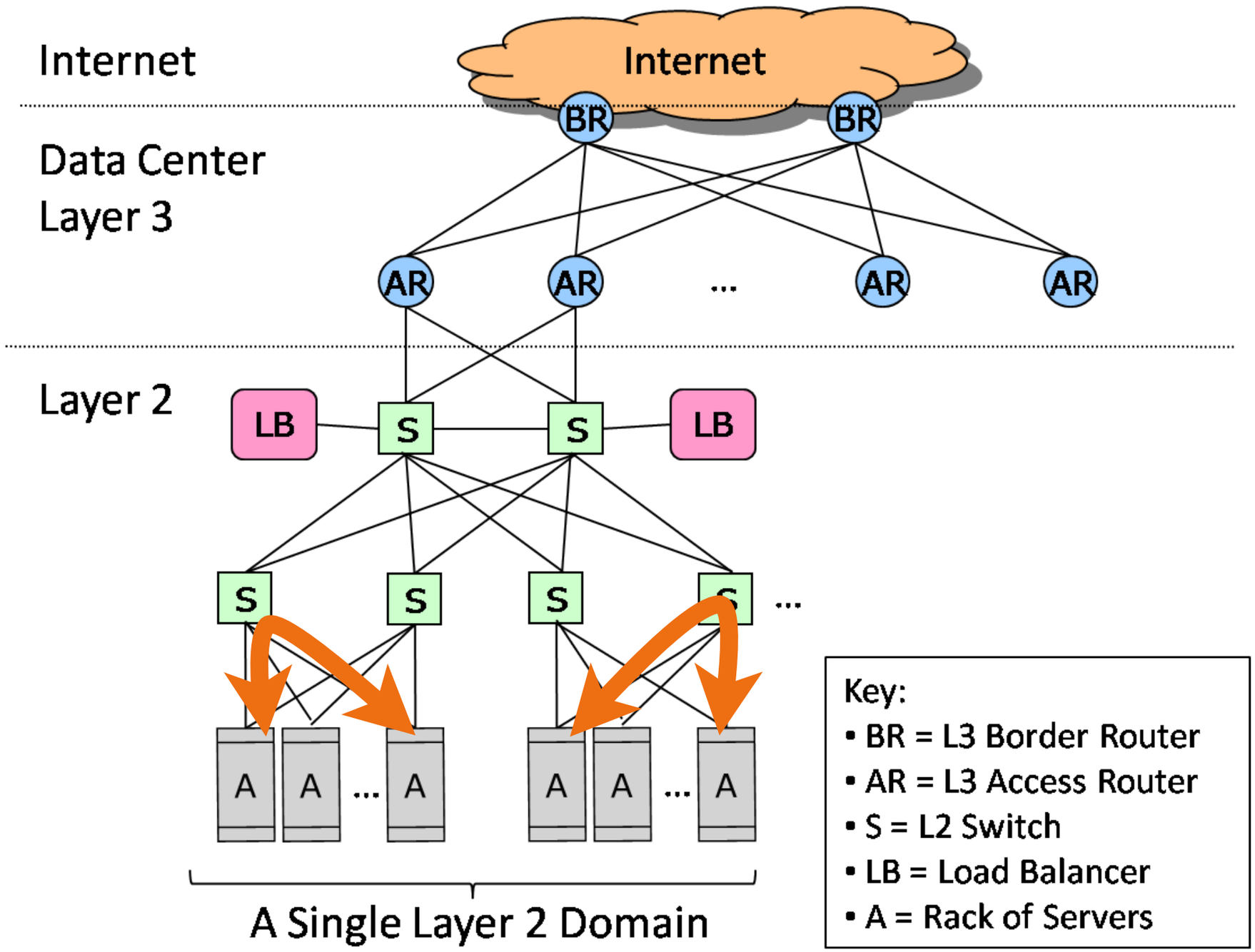
Alternative: tree network



Alternative: tree network

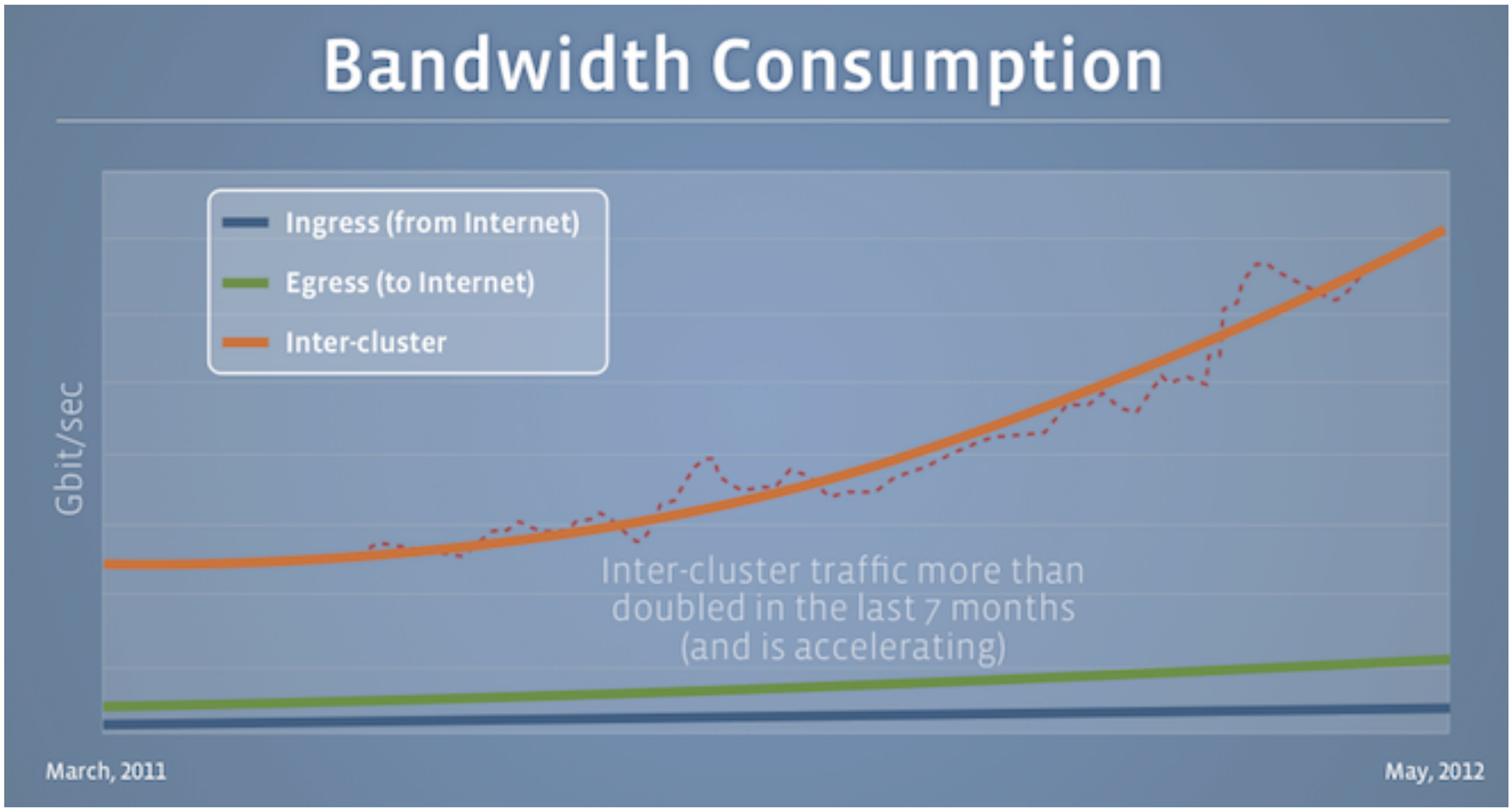


Traditional data center network



[Greenberg et al, CCR Jan. 2009]

The need for performance



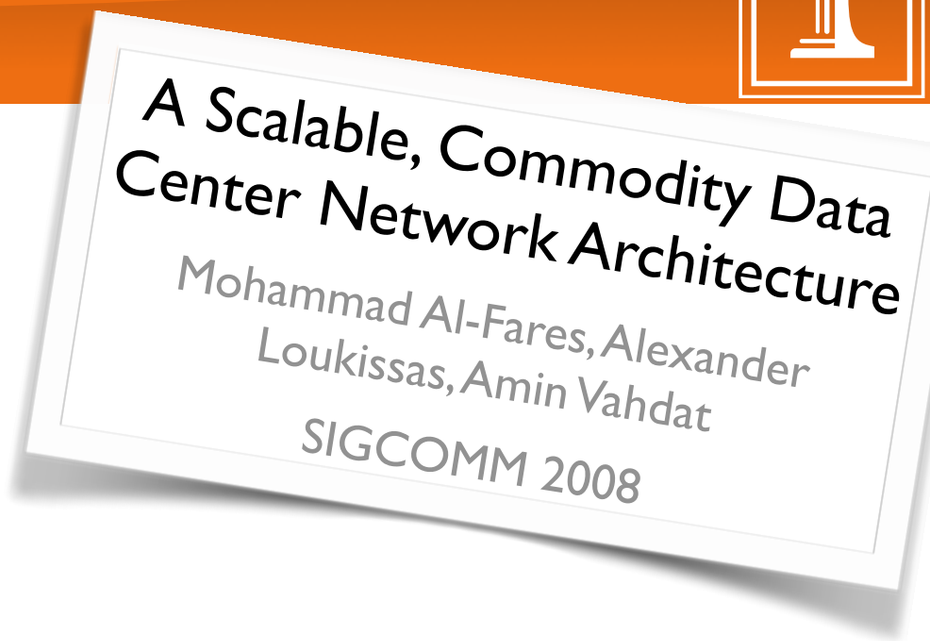
March 2011

May 2012

[Facebook, via Wired]

Modern Data Center Networks

Fat trees in data centers



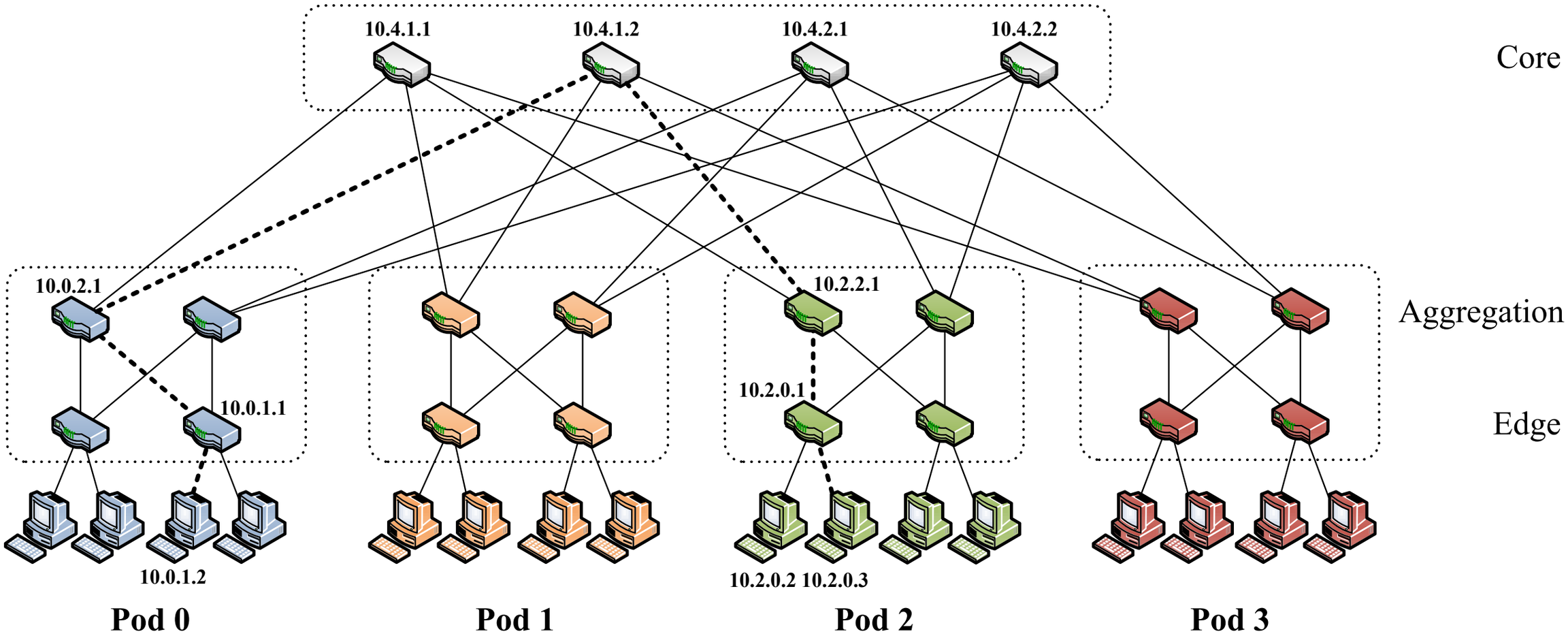
Argued for nonblocking bandwidth

- Servers limited only by their network card's speed, regardless of communication pattern between servers
- Also known as full throughput in the "hose model"
  - Maximum rate input from each "hose" (host)
  - Maximum rate output to each "hose"
  - Subject to those constraints, any traffic pattern is OK

Design

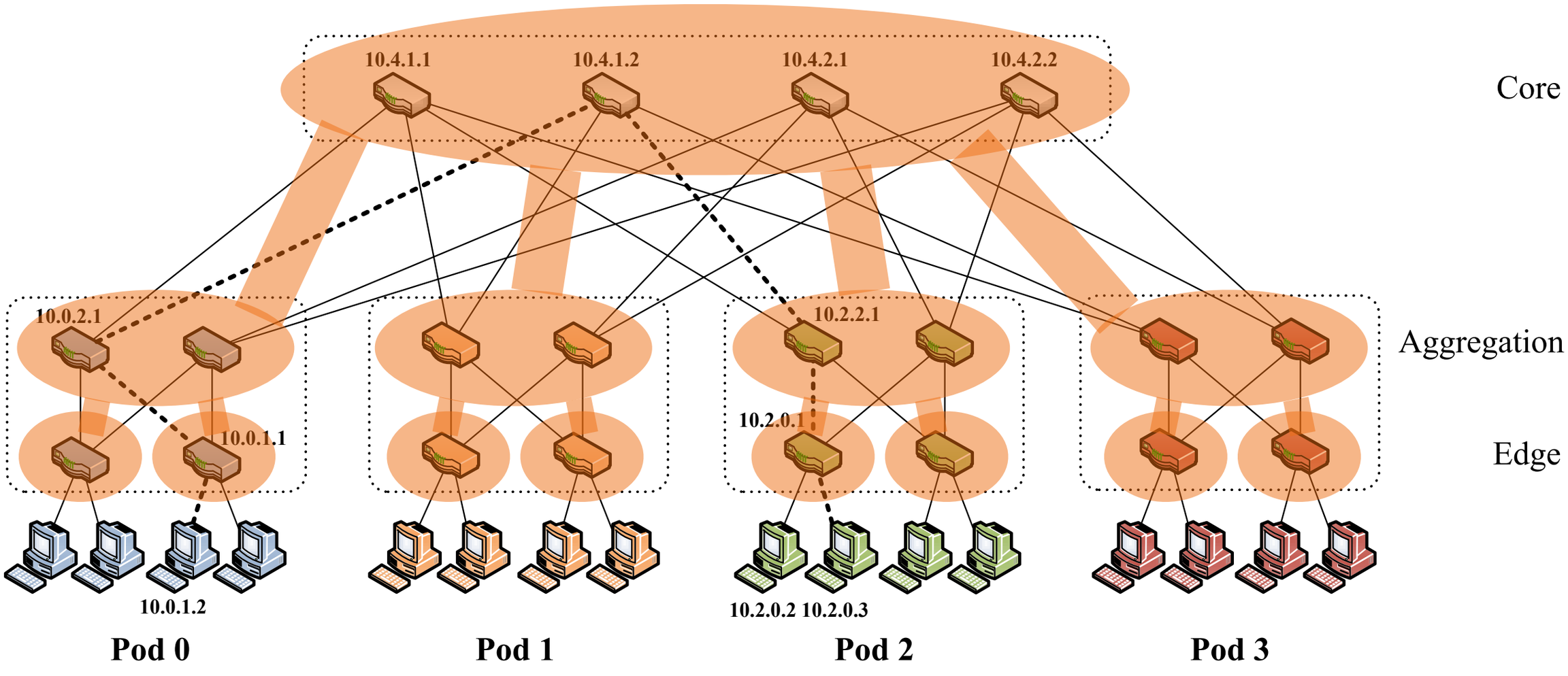
- Employed large number of commodity switches rather than "big iron"
- Arranged in Clos topology, and specifically a "fat tree"

Fat tree network



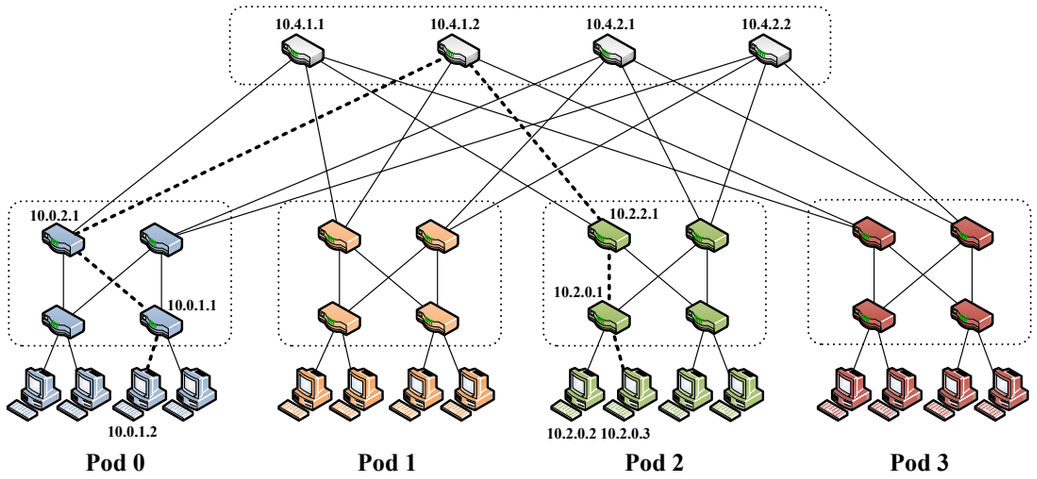
[Al-Fares, Loukissas, Vahdat, SIGCOMM '08]

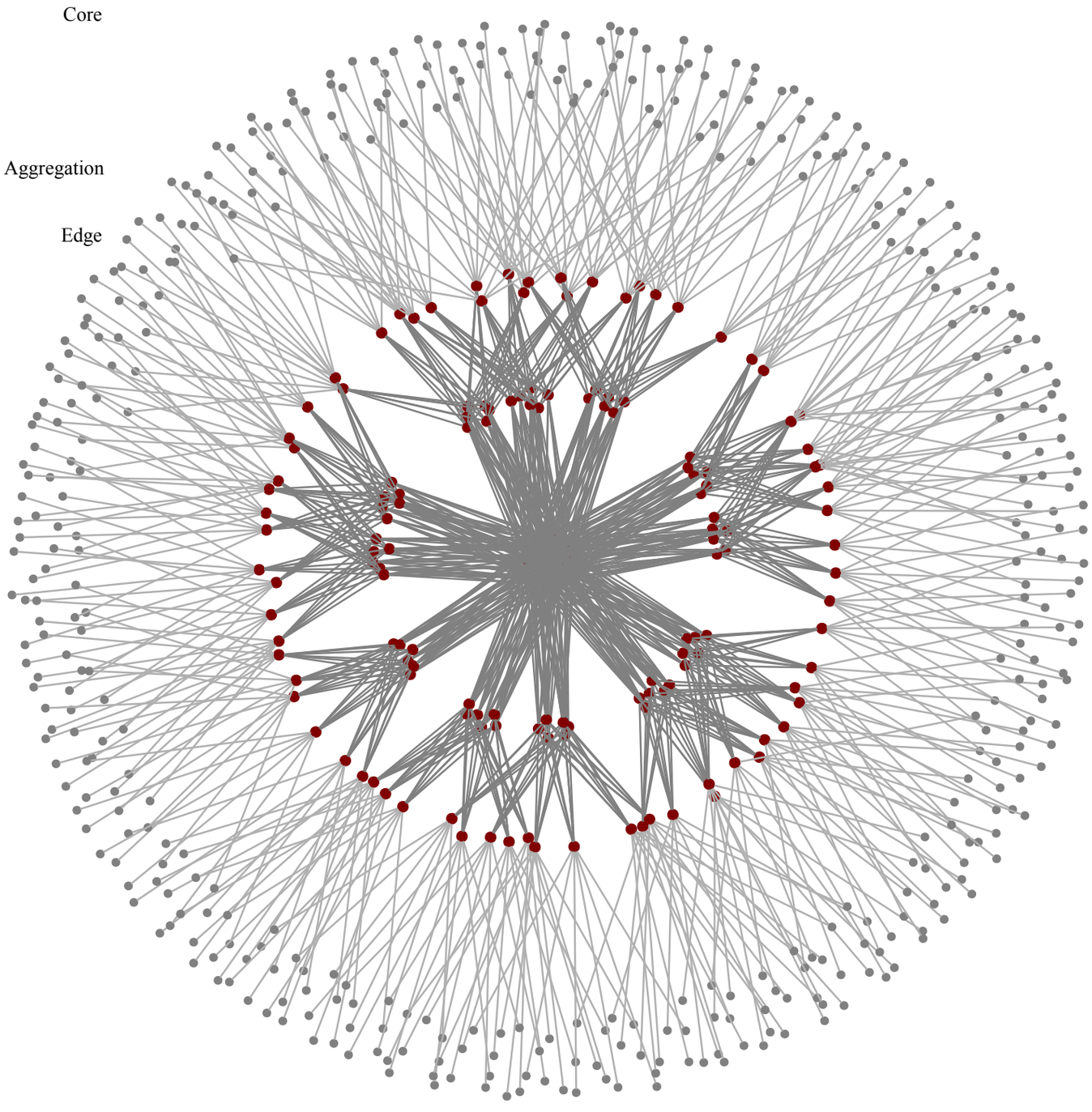
Fat tree network



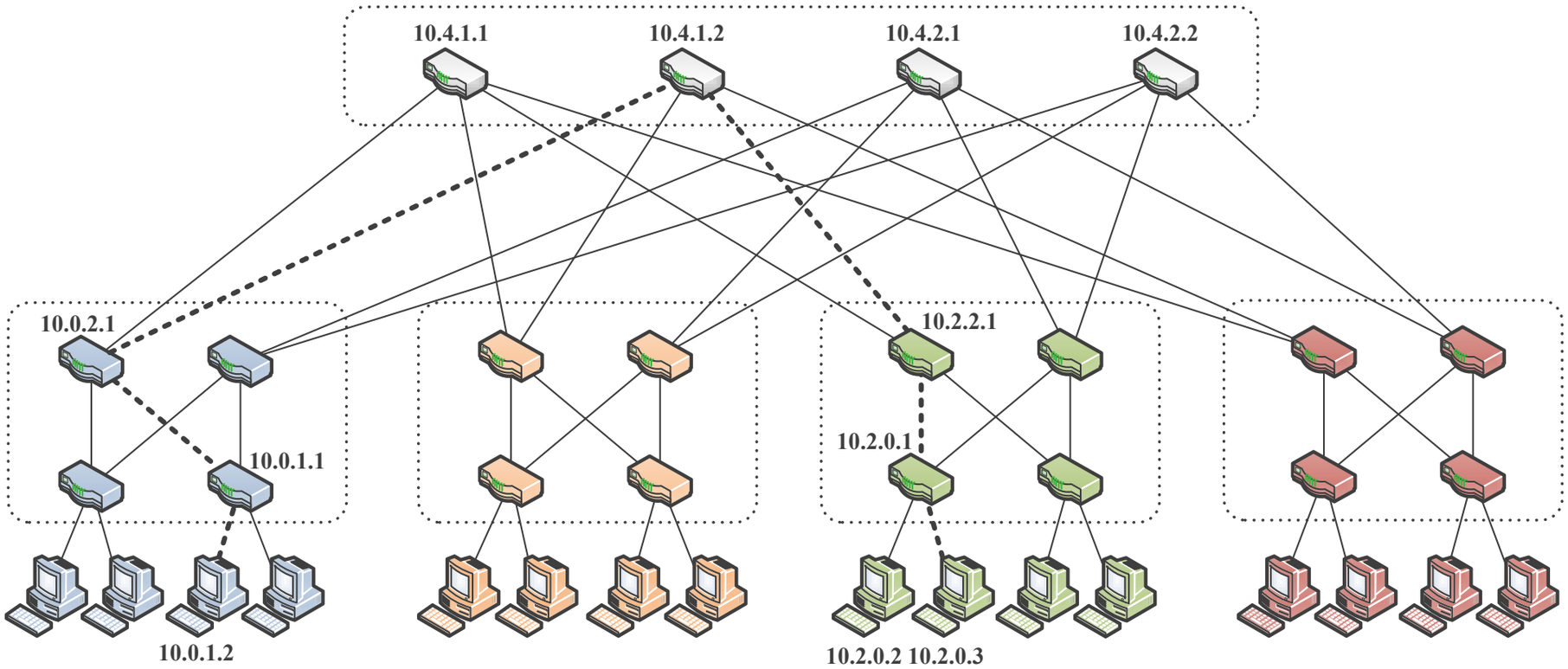
[Al-Fares, Loukissas, Vahdat, SIGCOMM '08]

Fat tree network





Fat tree network

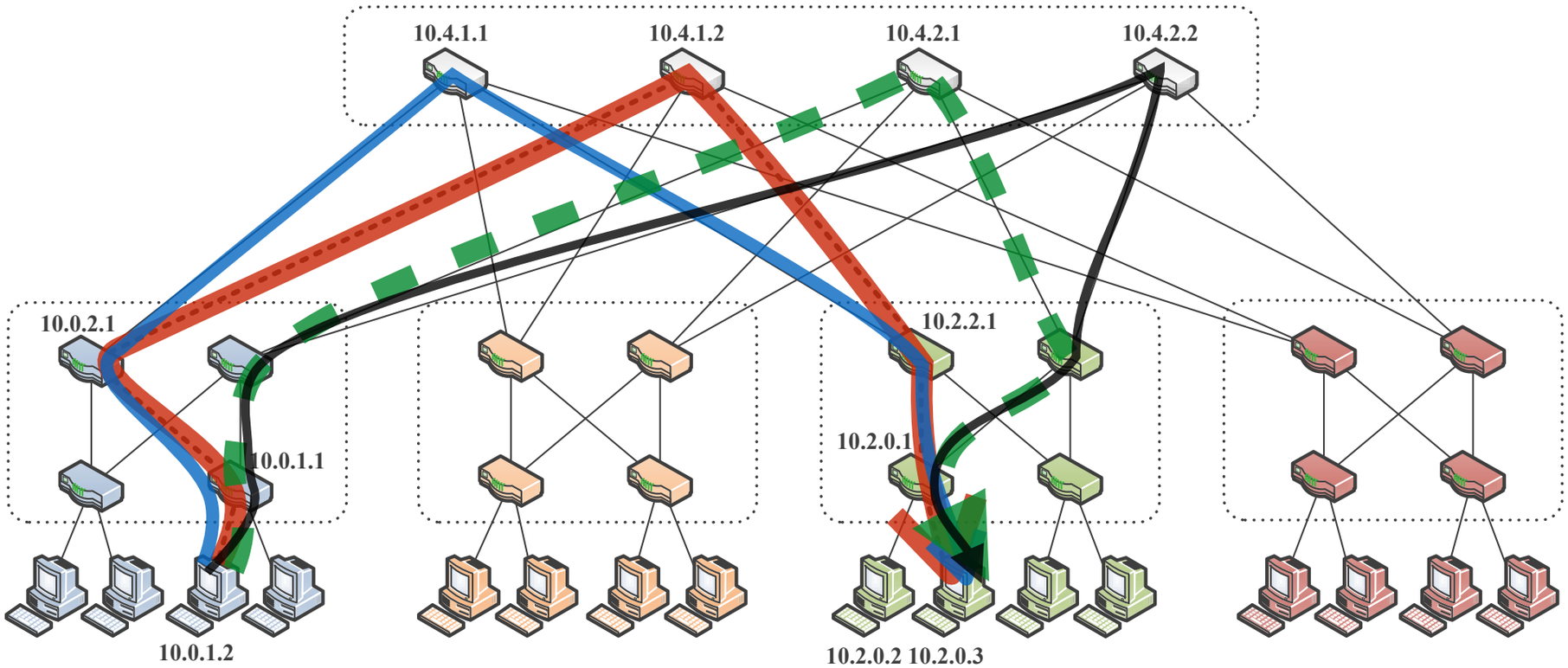


ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Fat tree network

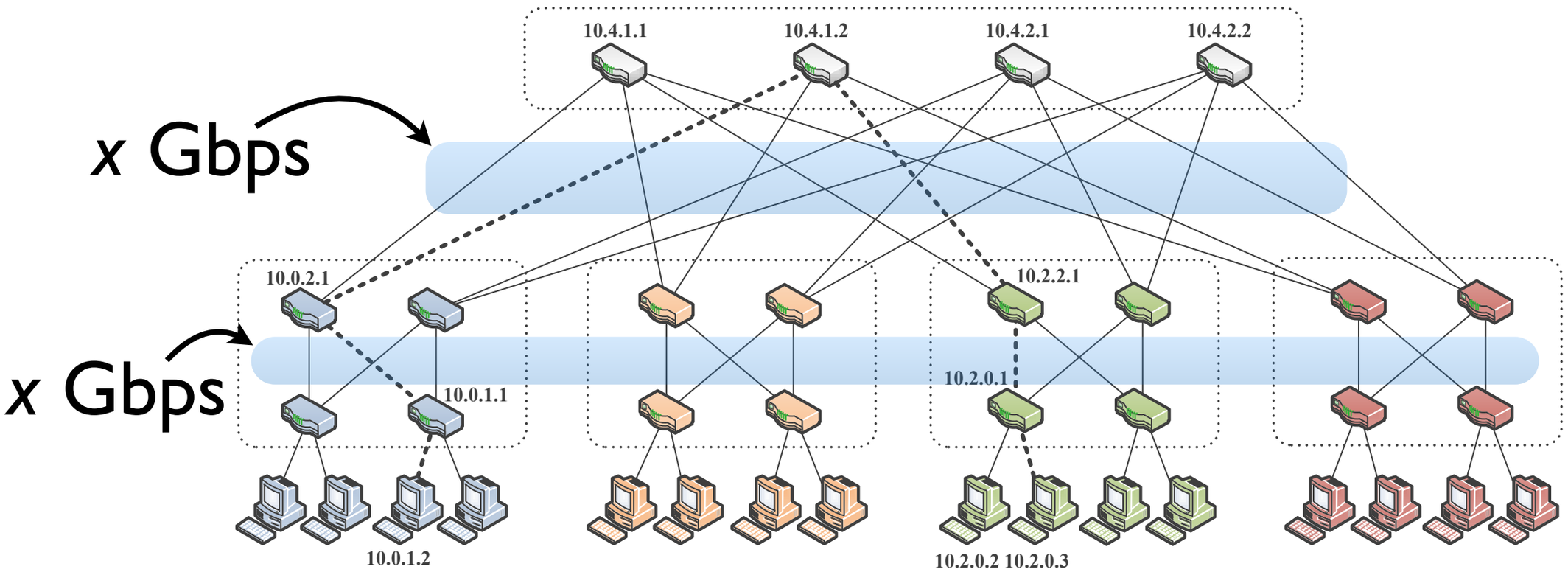


ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Fat tree network

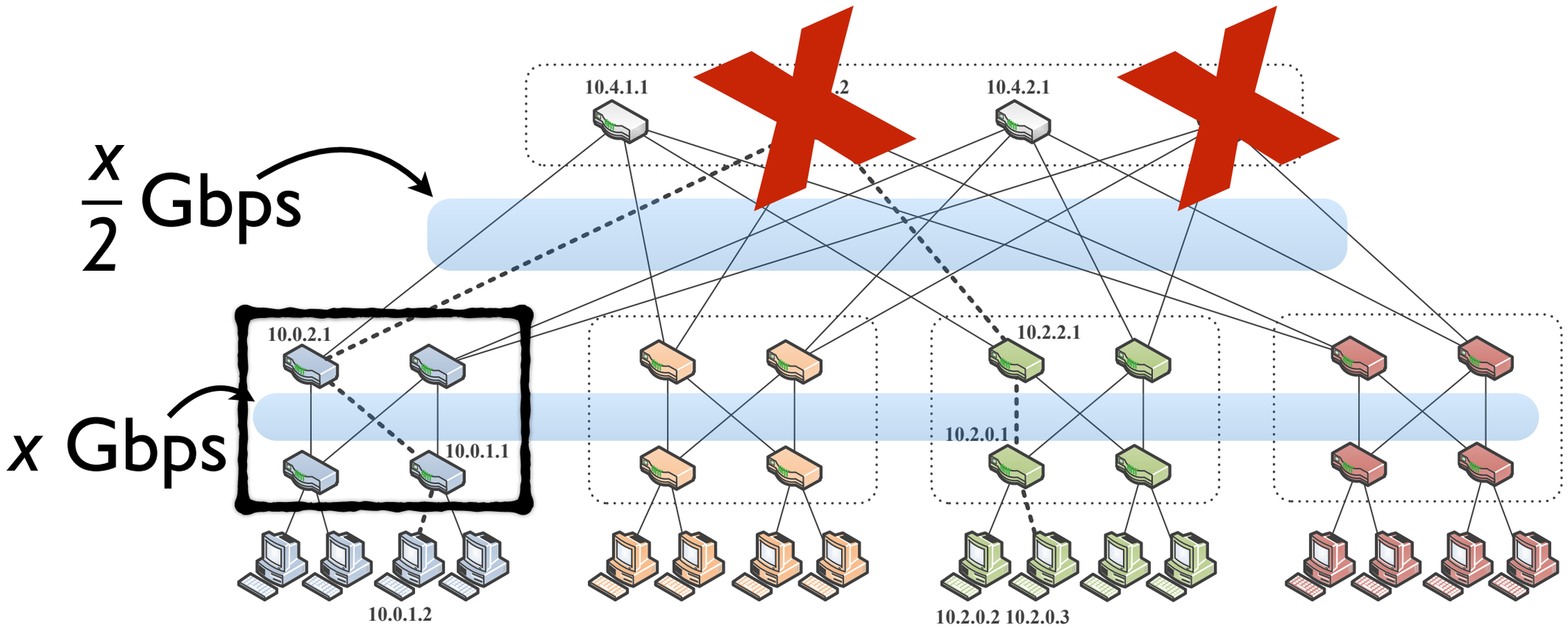


ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Oversubscribed fat tree



ACM SIGCOMM, 2008

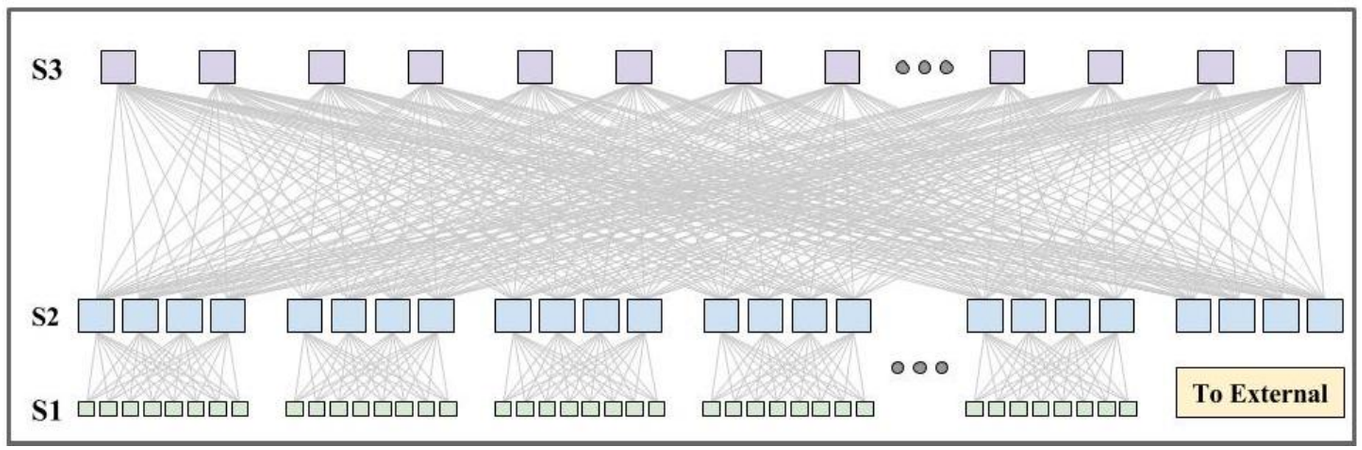
A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

ACM SIGCOMM, 2015

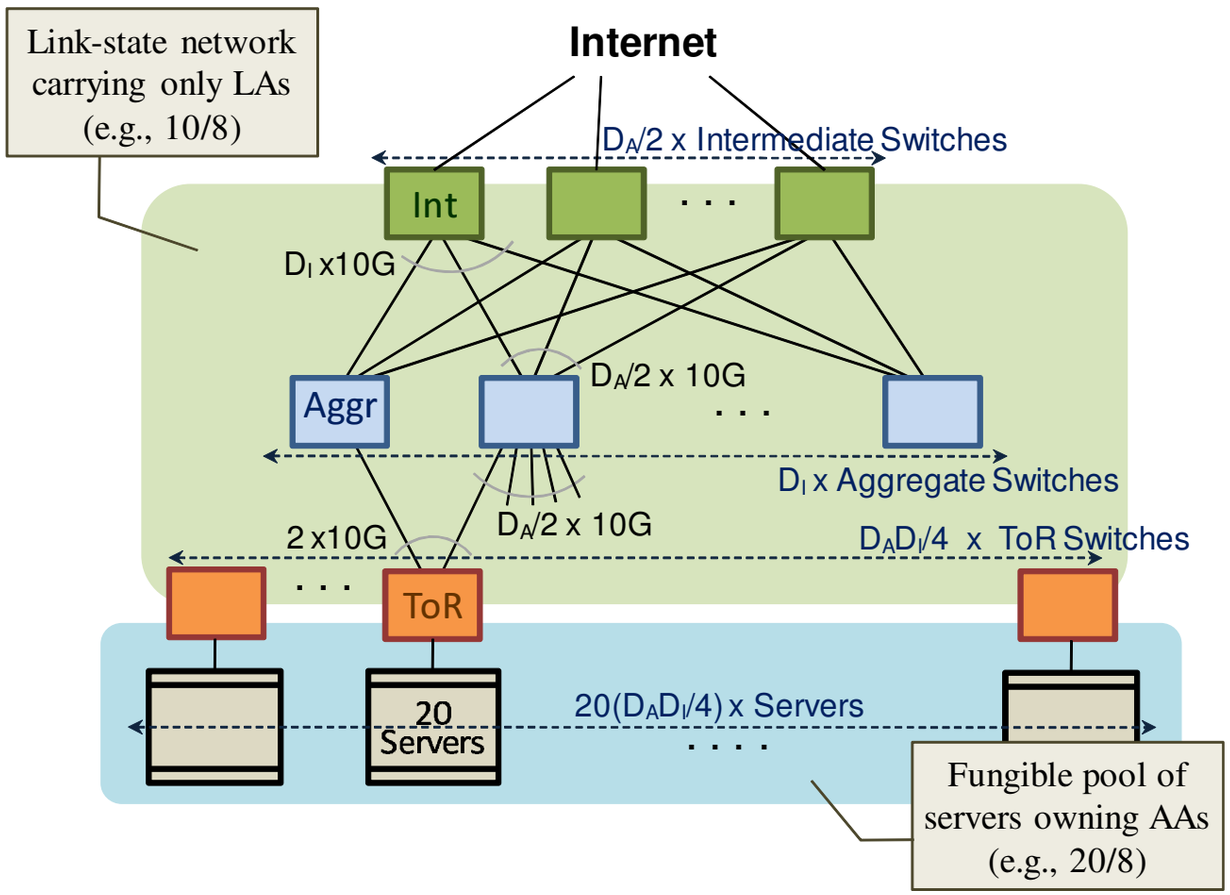
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network

Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, and Amin Vahdat
Google, Inc.

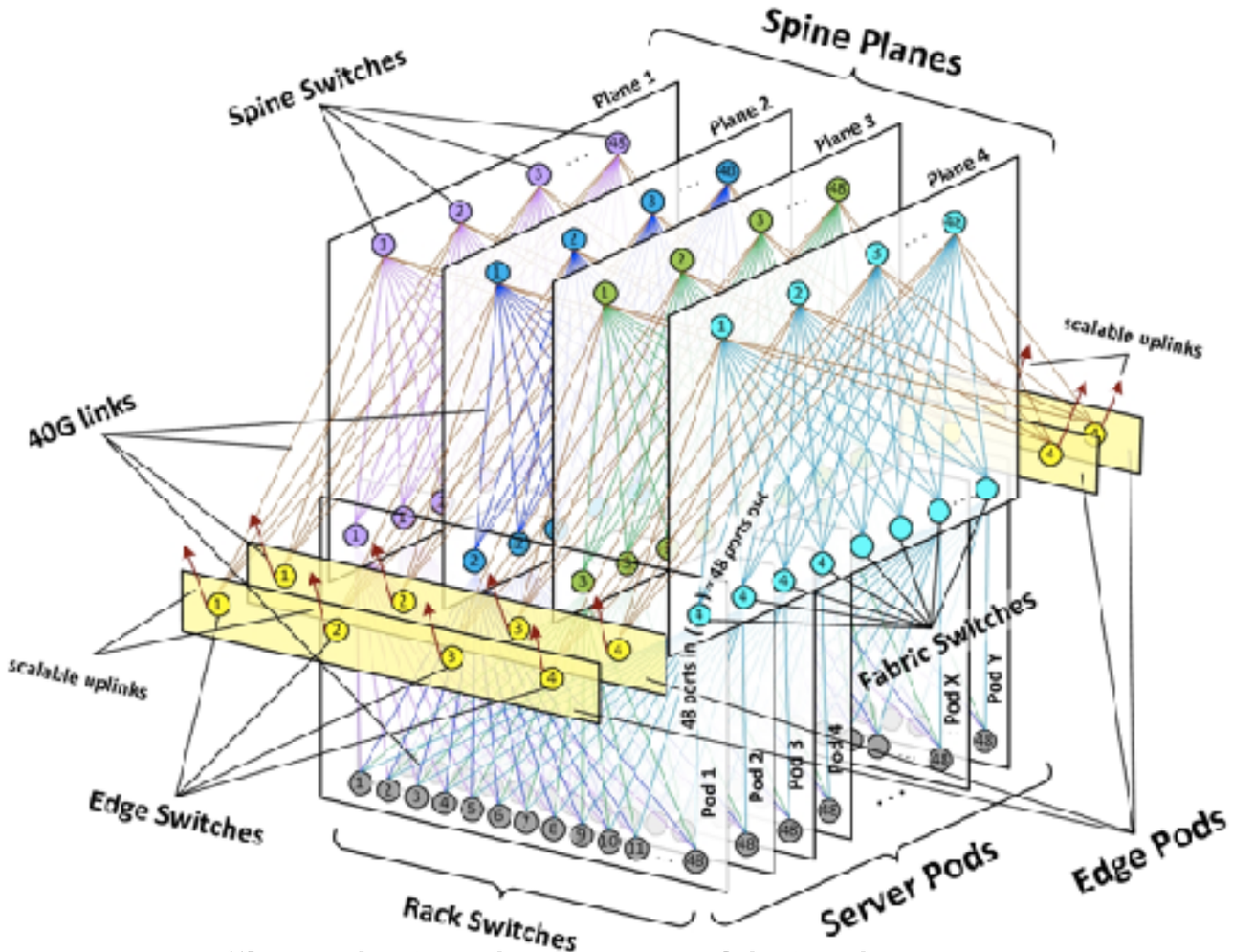


age: Robert Harker]

Variants of this design are common



VL2 @ **Microsoft**, ACM SIGCOMM'09
Greenburg, Hamilton, Jain, Kandula, Kim, Lahiri, Maltz, Patel, Sengupta



"Introducing data center fabric, the next-generation **Facebook** data center network", Alexey Andreyev, 2015

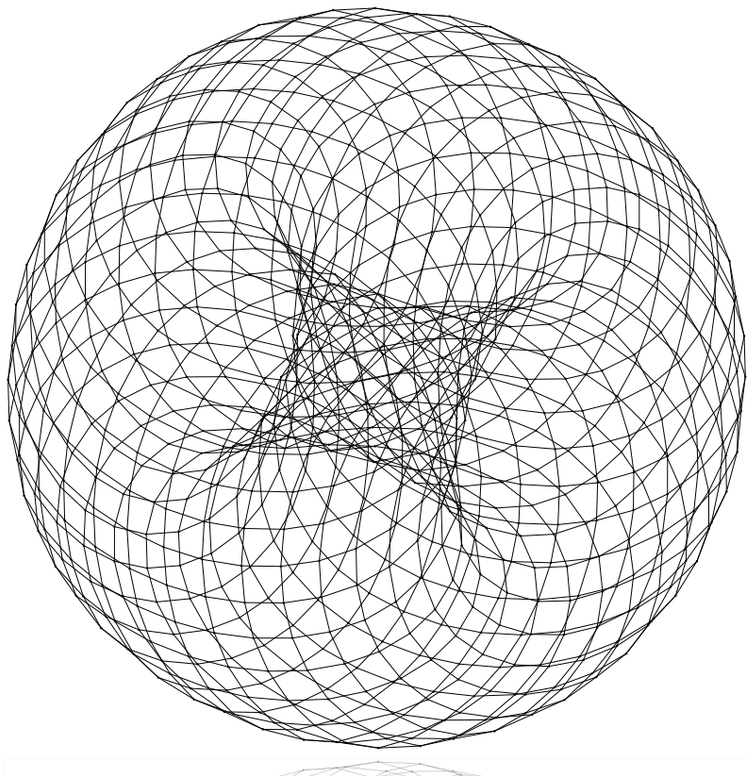
VL2

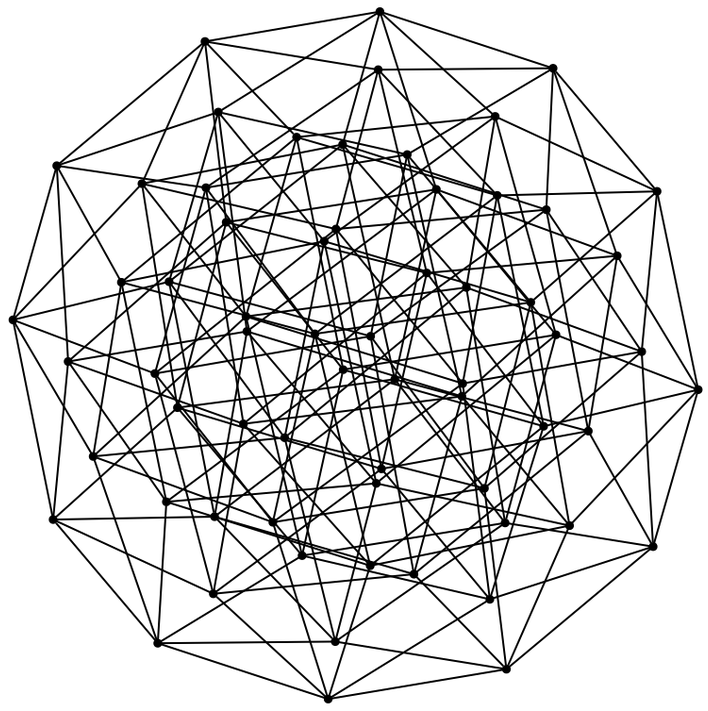
[Greenberg, Hamilton, Jain, Kandula, Kim, Lahiri, Maltz, Patel, Sengupta, SIGCOMM 2009]

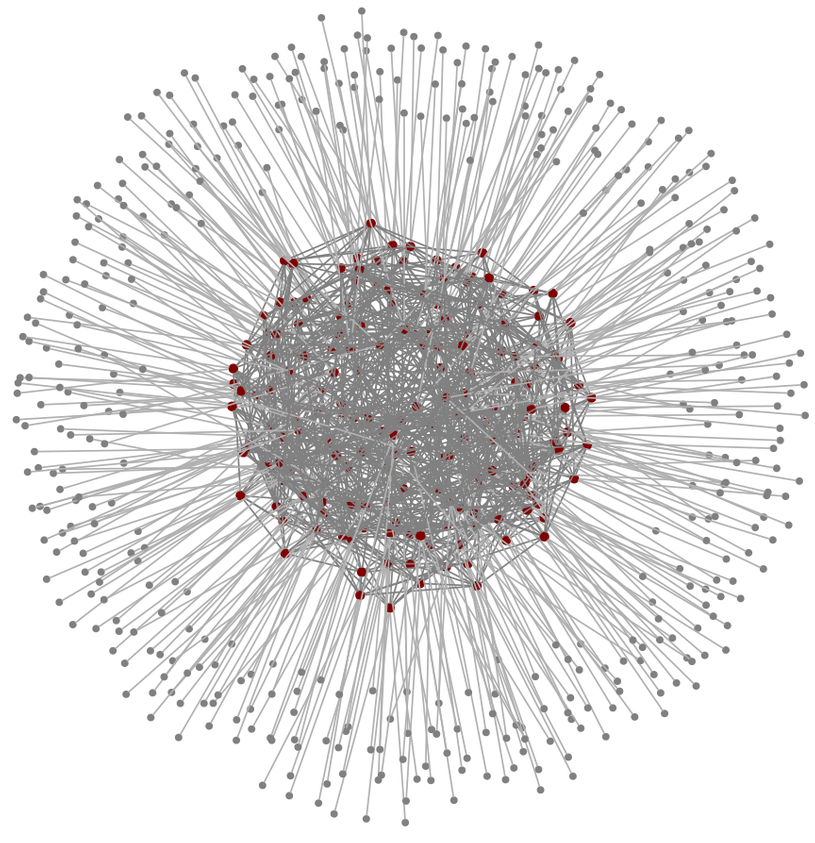
Key features

- High bandwidth network
  - Another folded Clos network
  - Slightly different than fat tree (e.g., uses faster 10 Gbps links at higher layers)
- Randomized (Valiant) load balancing
  - Makes better use of network resources
- Flat addressing
  - Ethernet-style (layer 2) addresses to forward data, rather than IP addresses
  - Separates names from locations

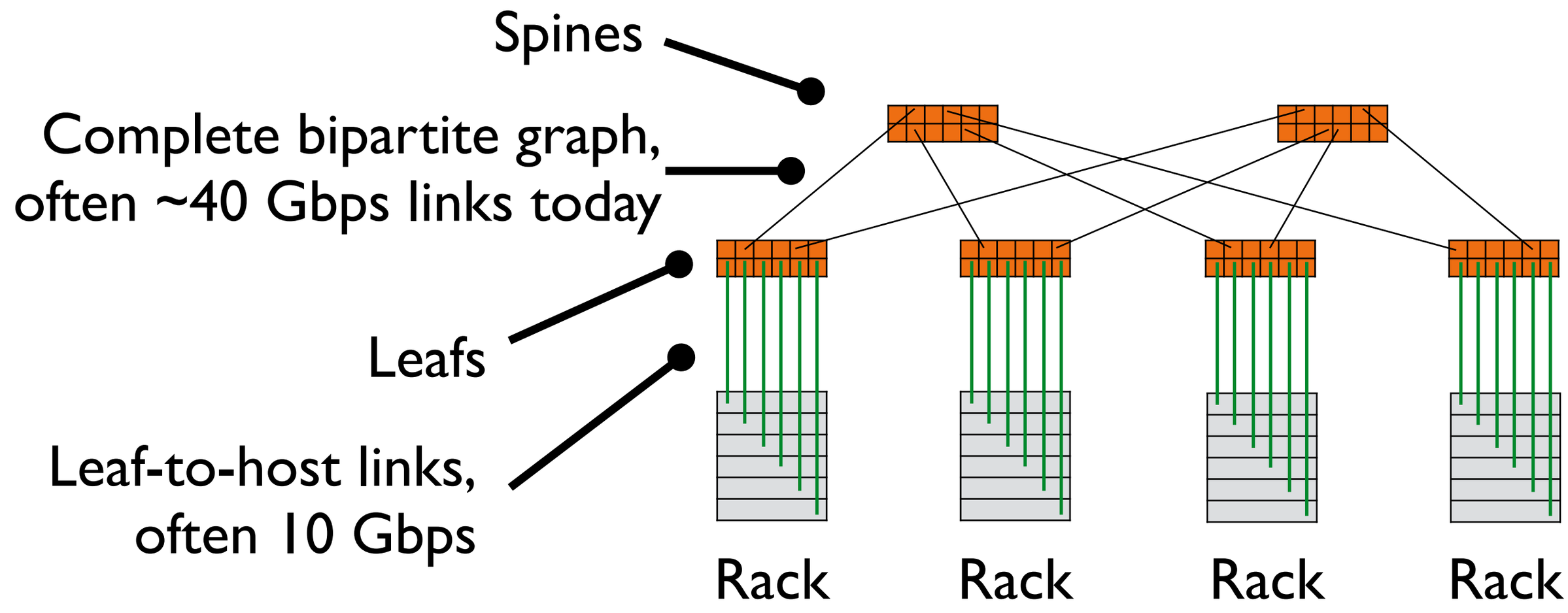
Many other proposed topologies







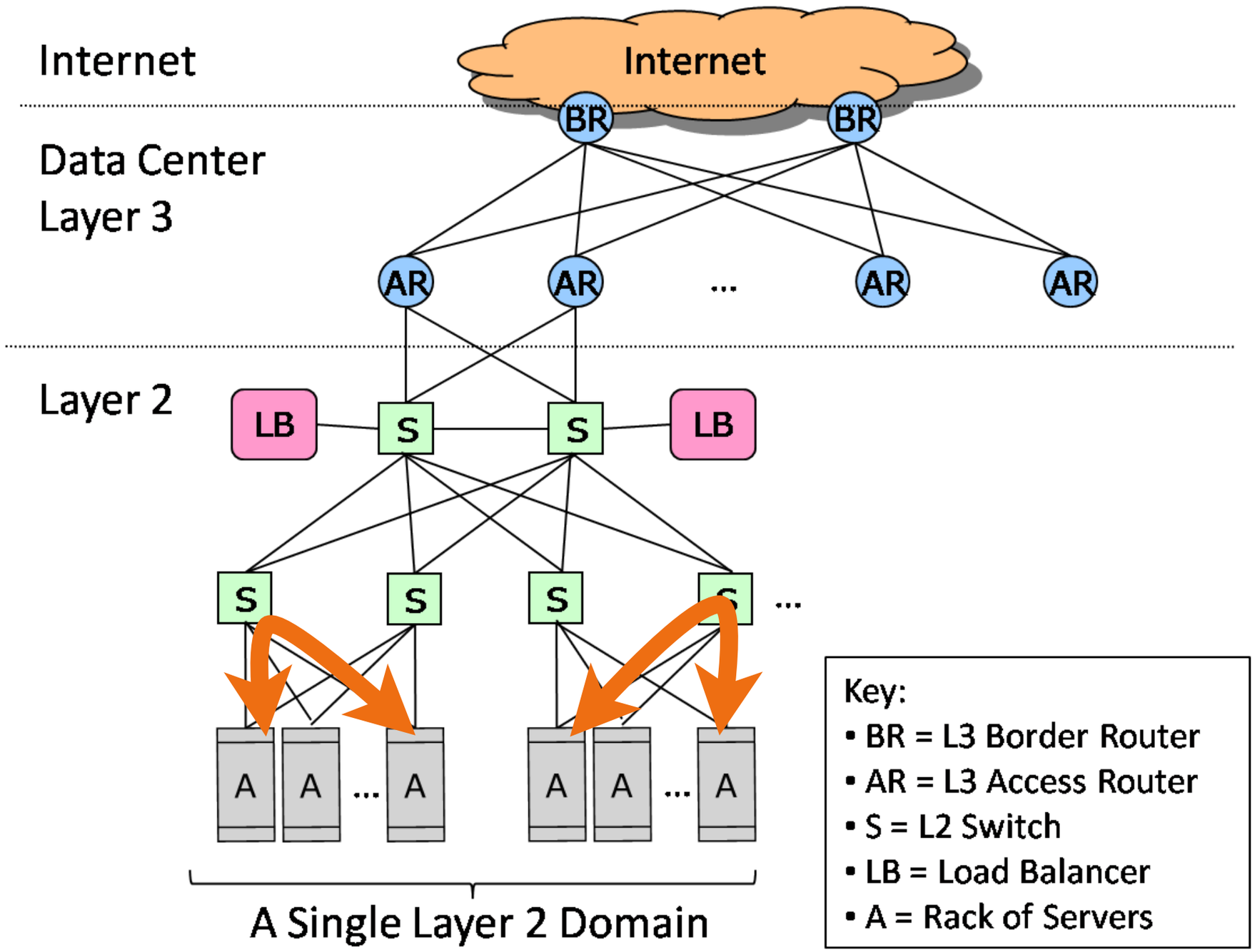
Leaf-spine for smaller networks



Outside of the hyper scale cloud providers, this 2-tier design is typically scalable enough and is now common.

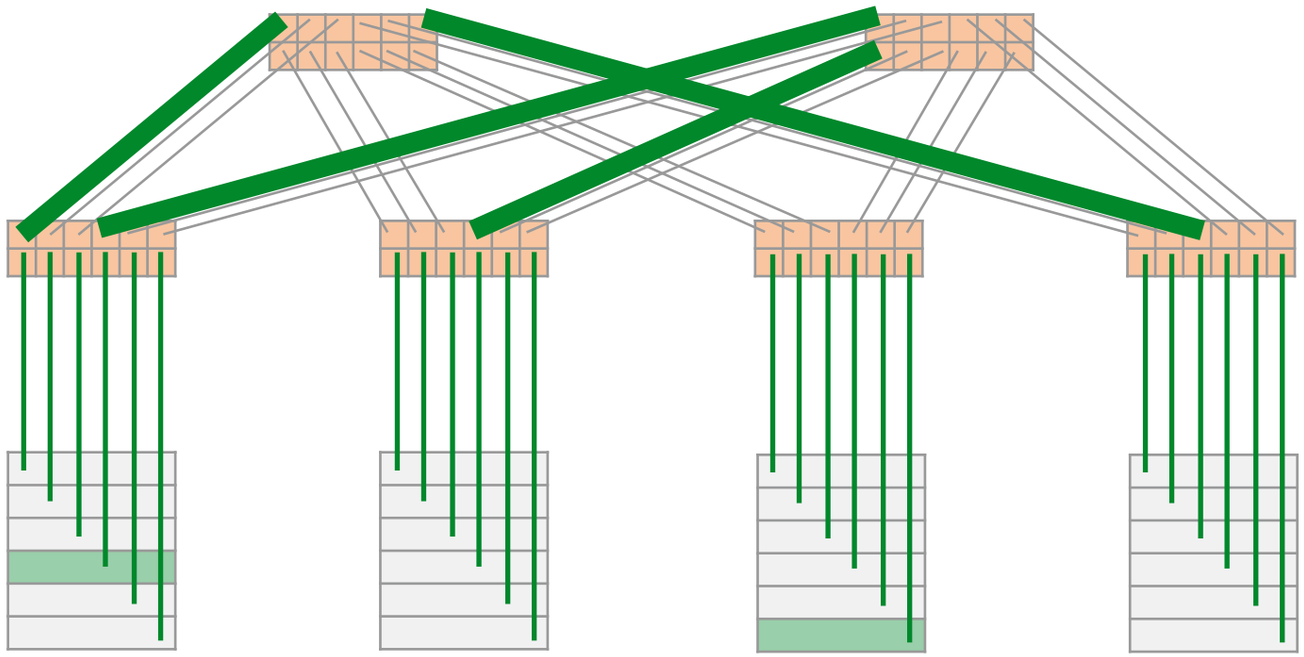
Routing in Data Center Networks

Traditional data center network



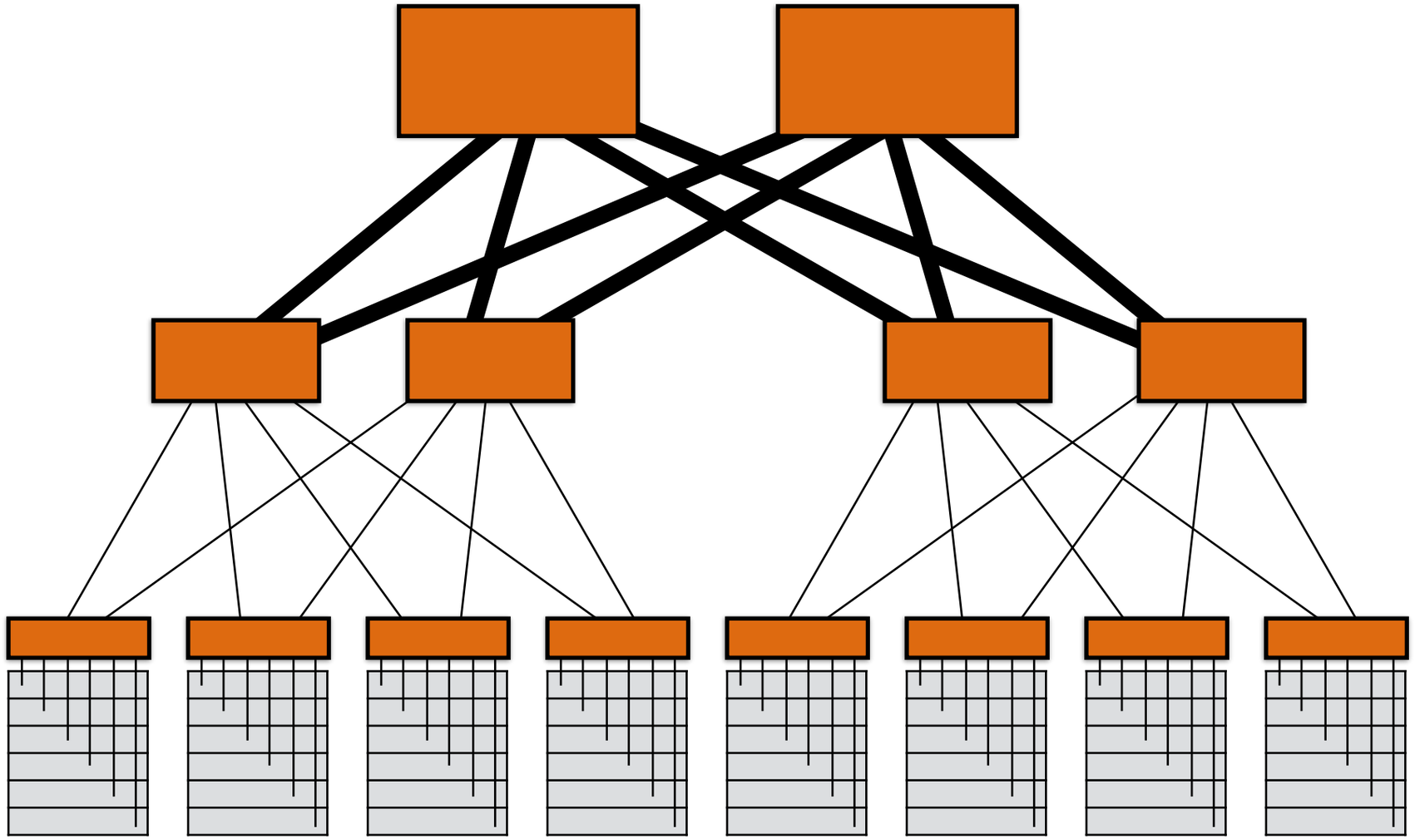
[Greenberg et al, CCR Jan. 2009]

Spanning tree

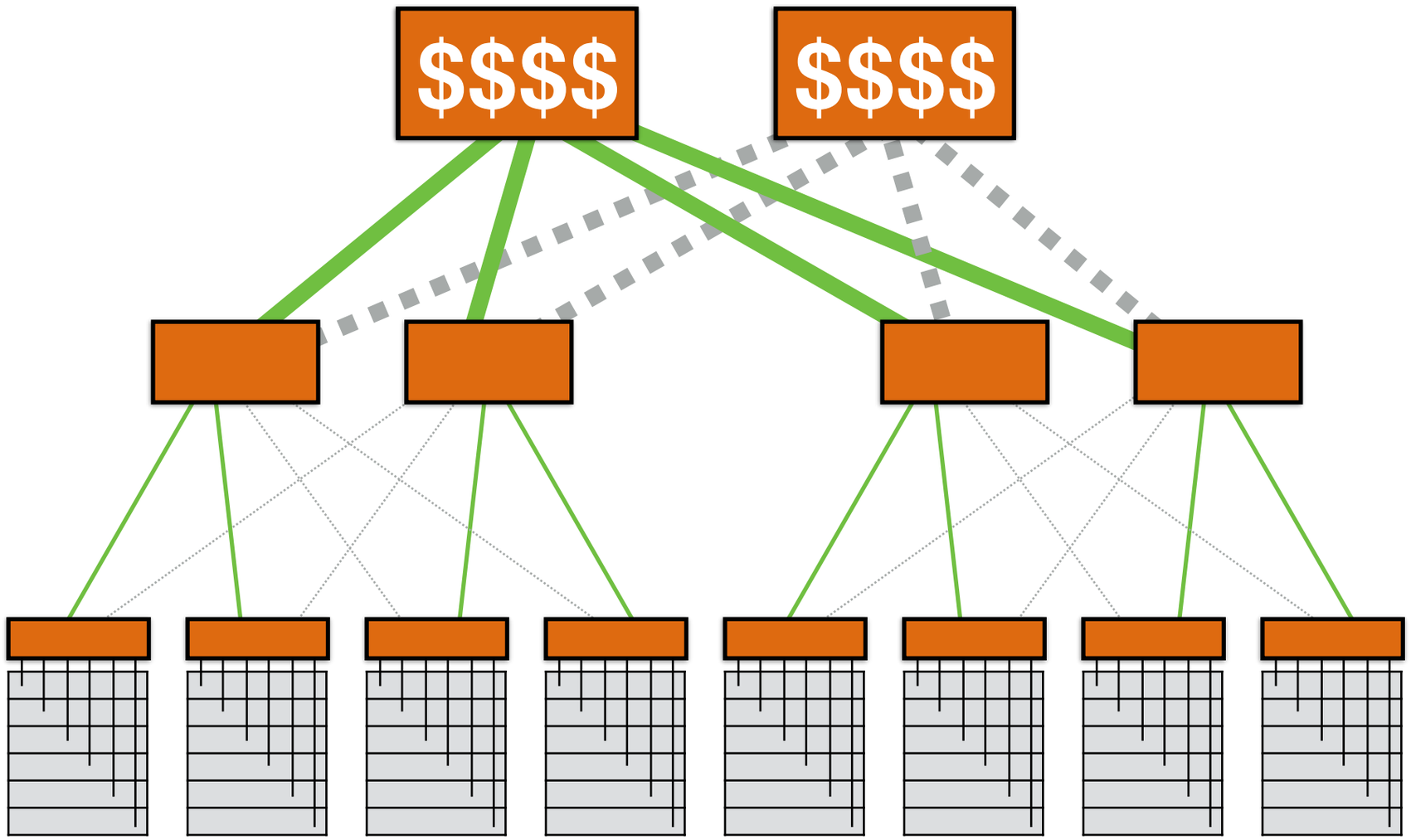


Disable all other links!

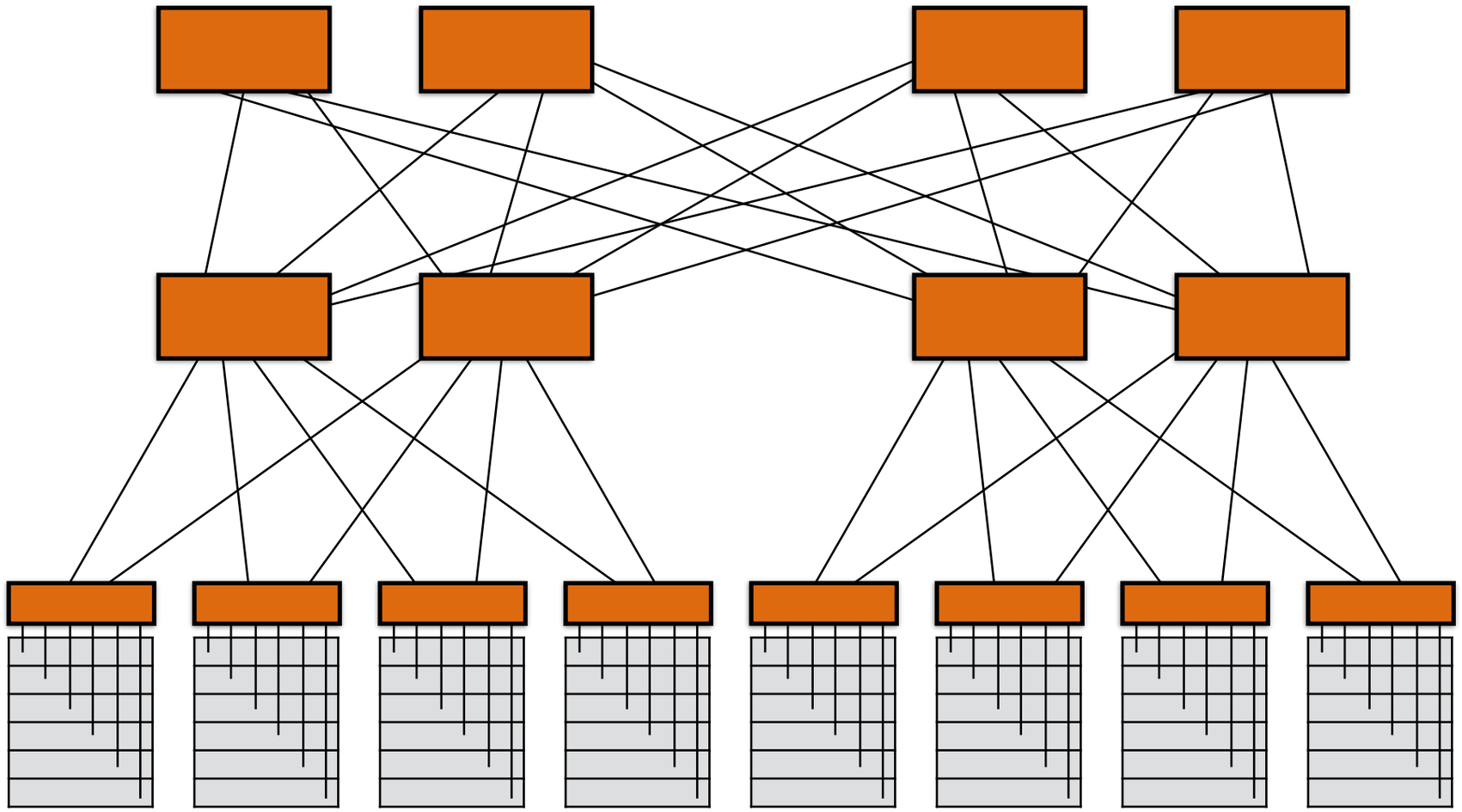
STP works for tree-like networks



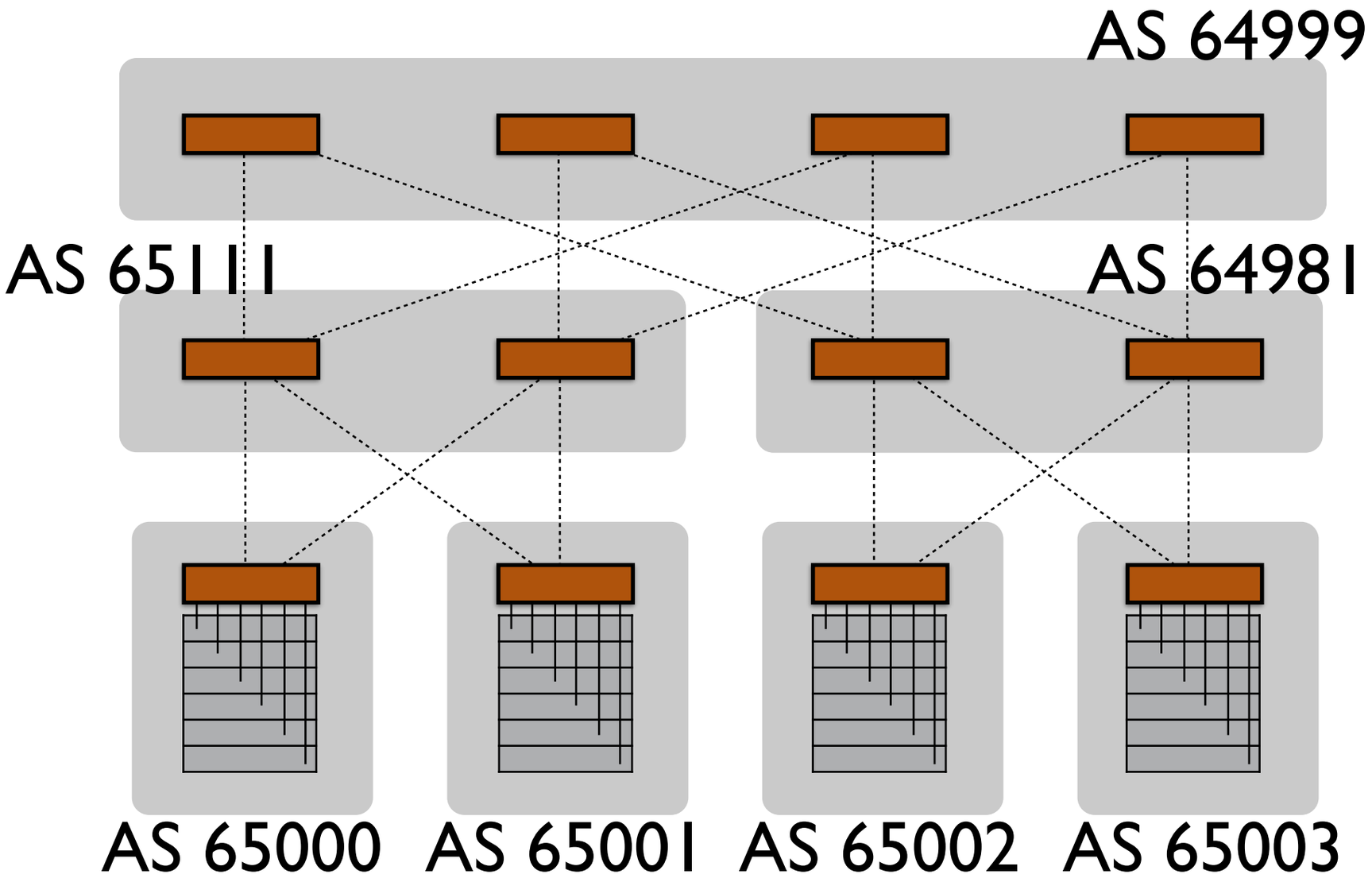
STP works for tree-like networks



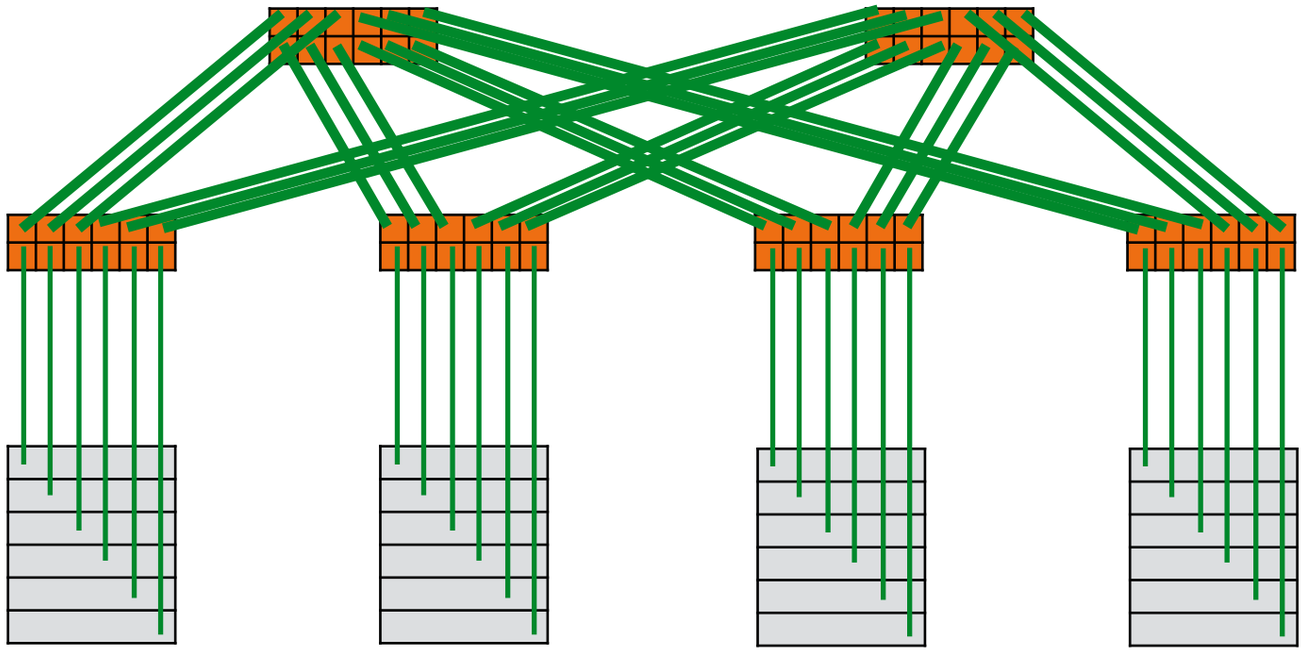
STP works for tree-like networks



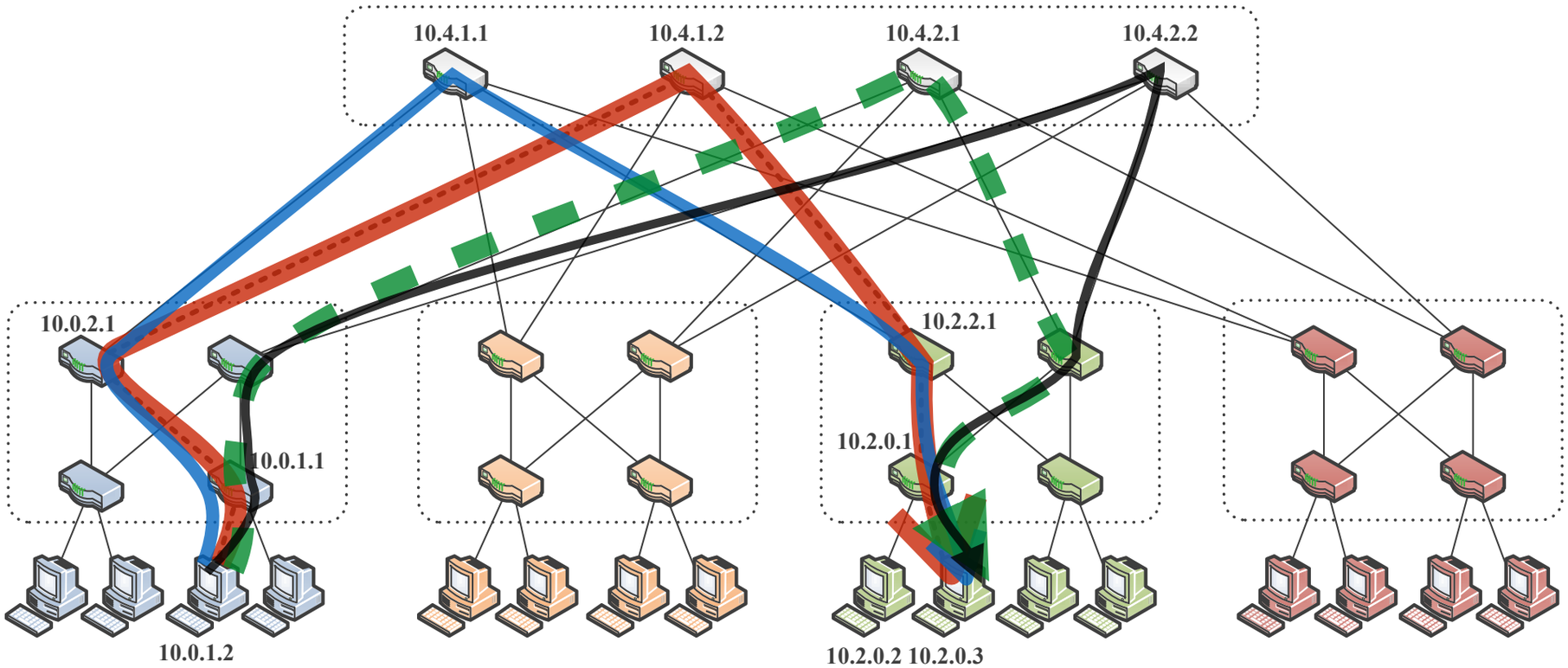
BGP in the data center



Routing



Multipath routing



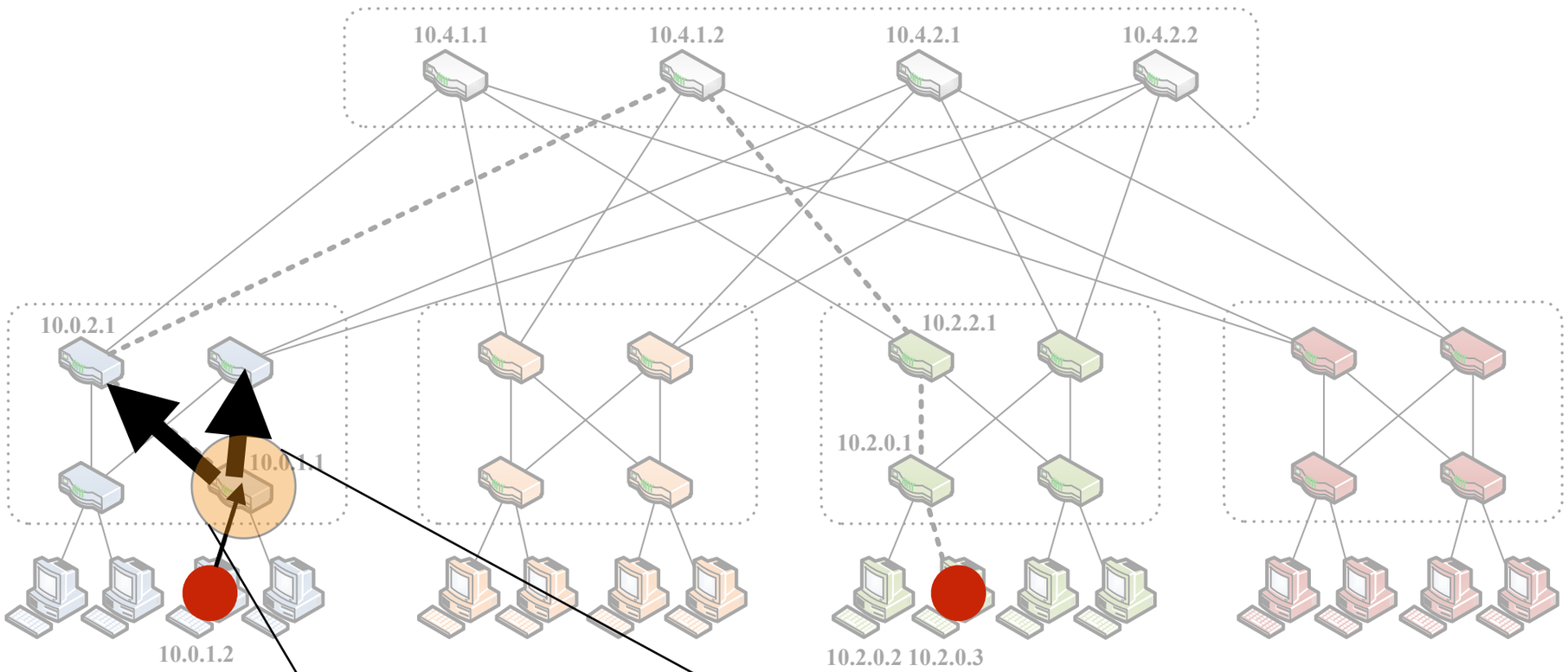
A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares
malfares@cs.ucsd.edu

Alexander Loukissas
aloukiss@cs.ucsd.edu

Amin Vahdat
vahdat@cs.ucsd.edu

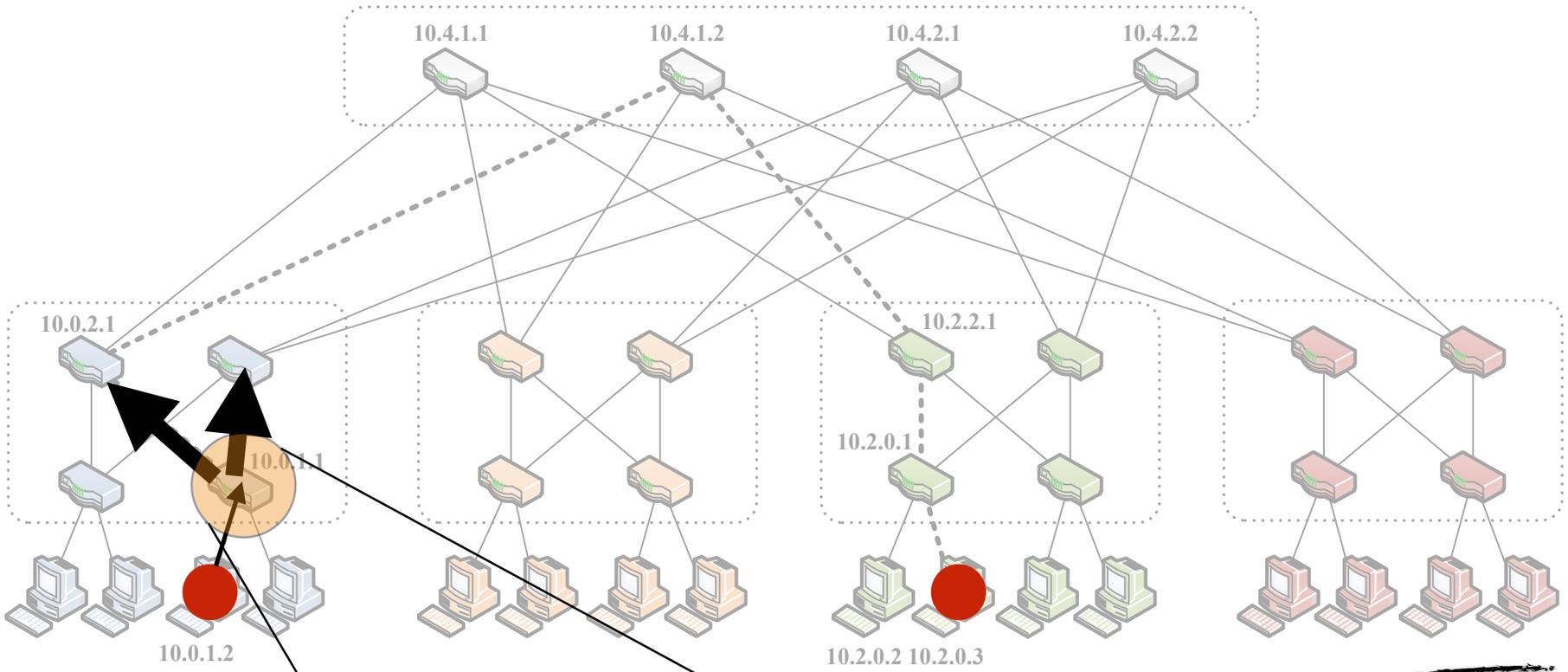
Equal cost multi-path (ECMP)



10.2.0.3	port 1
10.2.0.3	port 2

Choose uniformly at random!

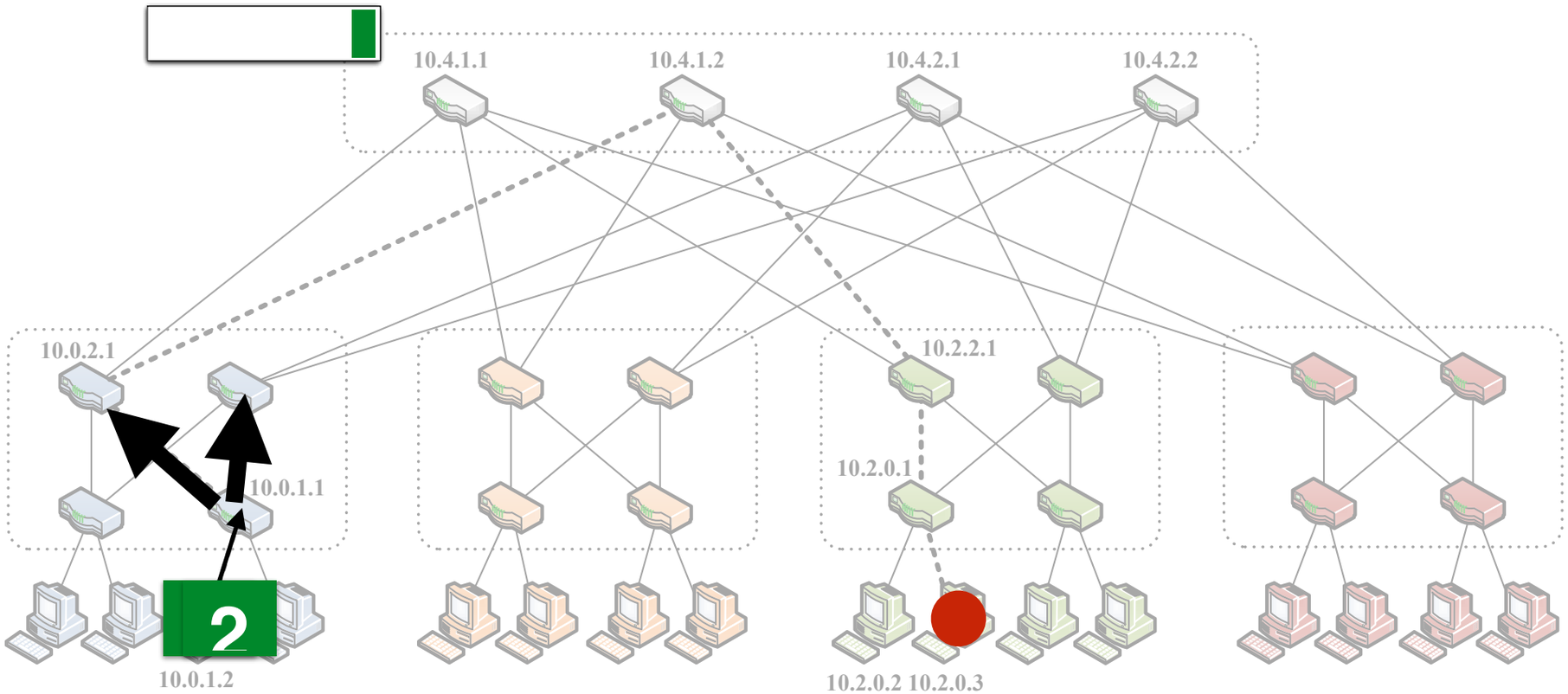
Equal cost multi-path (ECMP)



10.2.0.3	port 1
10.2.0.3	port 2

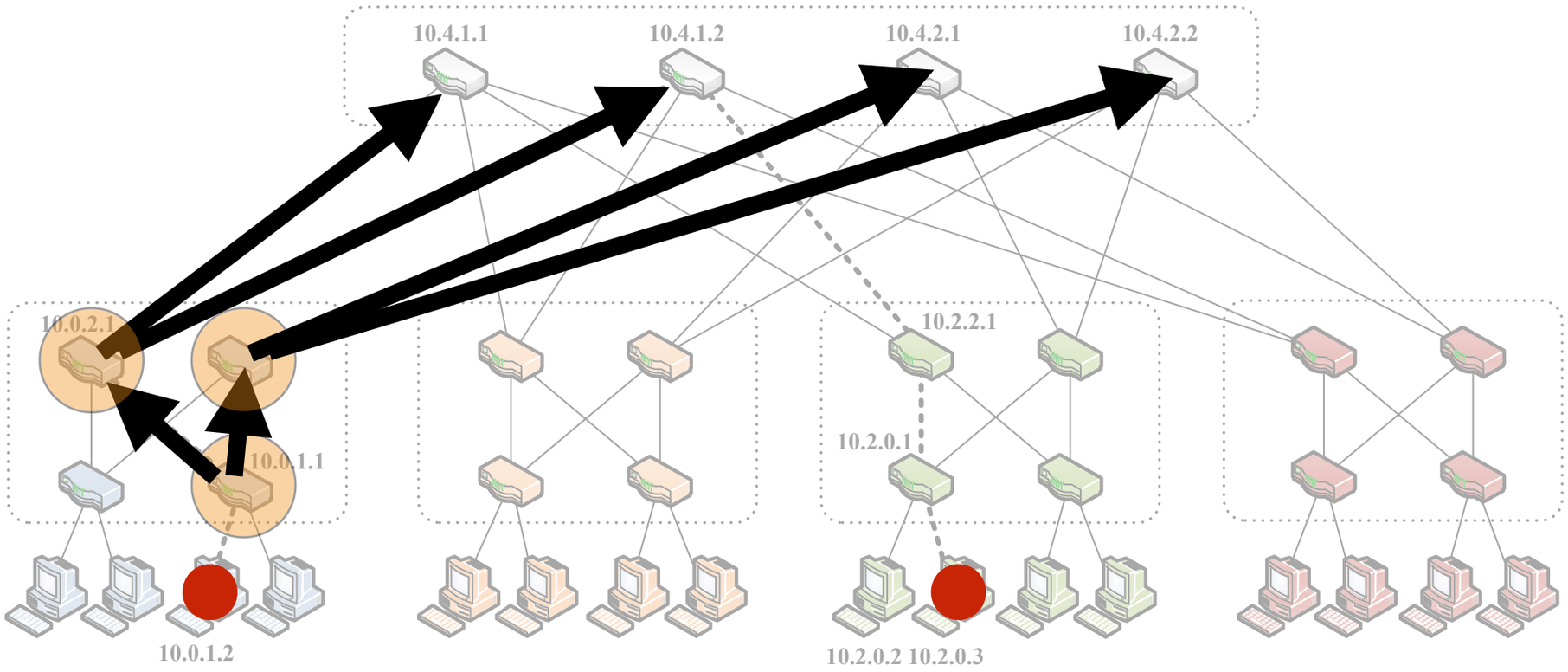
Choose uniformly at random!

Equal cost multi-path (ECMP)

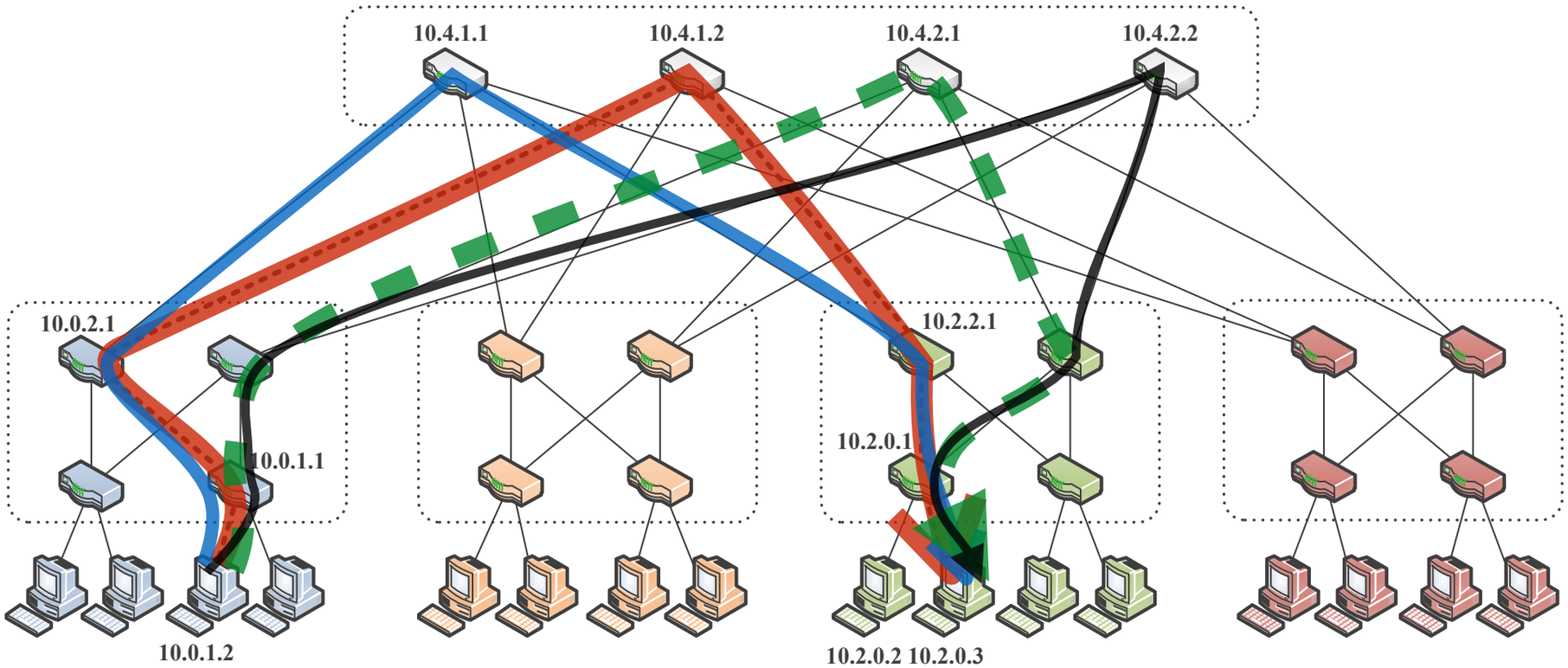


Output port = **hash** (packet header)

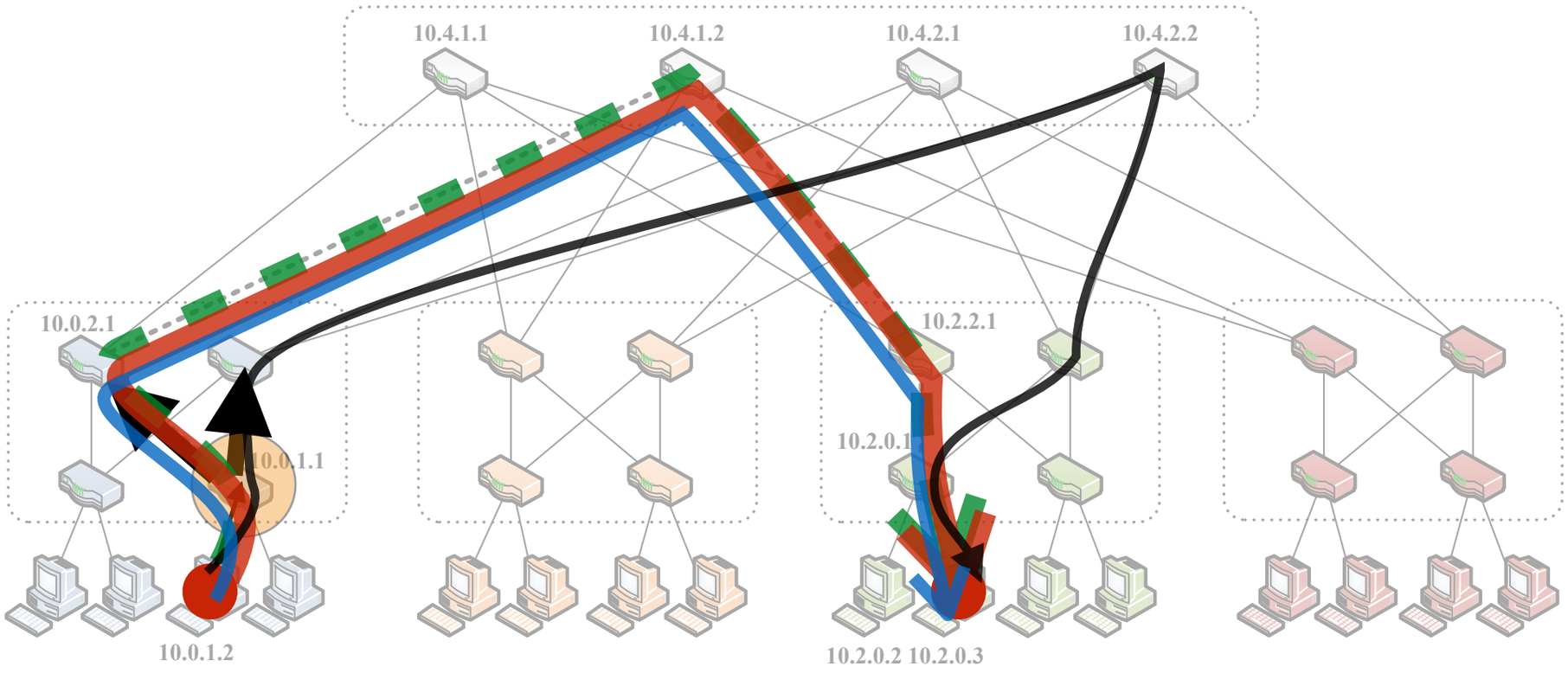
Equal cost multi-path (ECMP)



Equal cost multi-path (ECMP)



ECMP: traffic imbalance

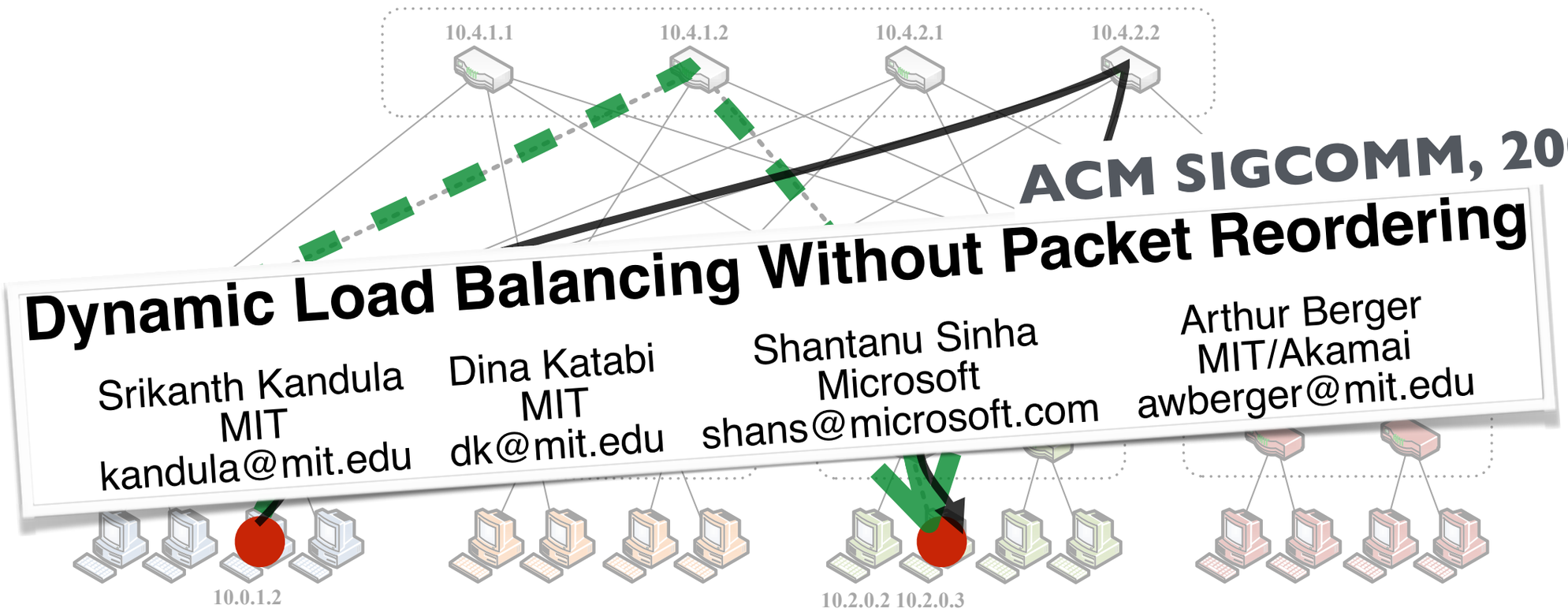


Output port = **hash** (packet header "5-tuple")

IP src & dst.
Protocol number
Protocol port src & dst.

Result: each TCP connection (a "flow") stays on one path

Flowlets



ACM SIGCOMM, 2007

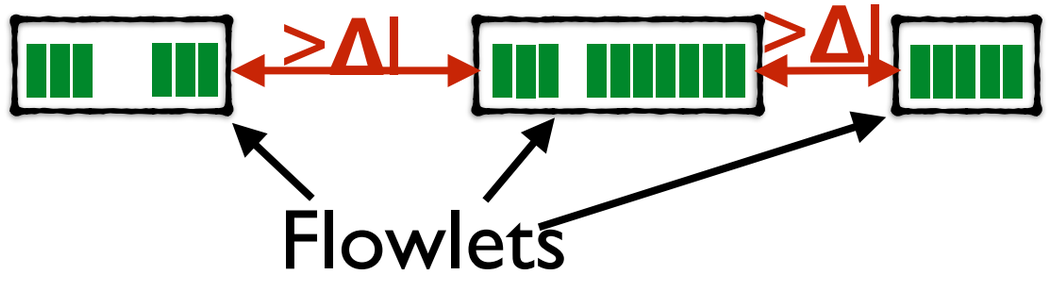
Dynamic Load Balancing Without Packet Reordering

Srikanth Kandula
MIT
kandula@mit.edu

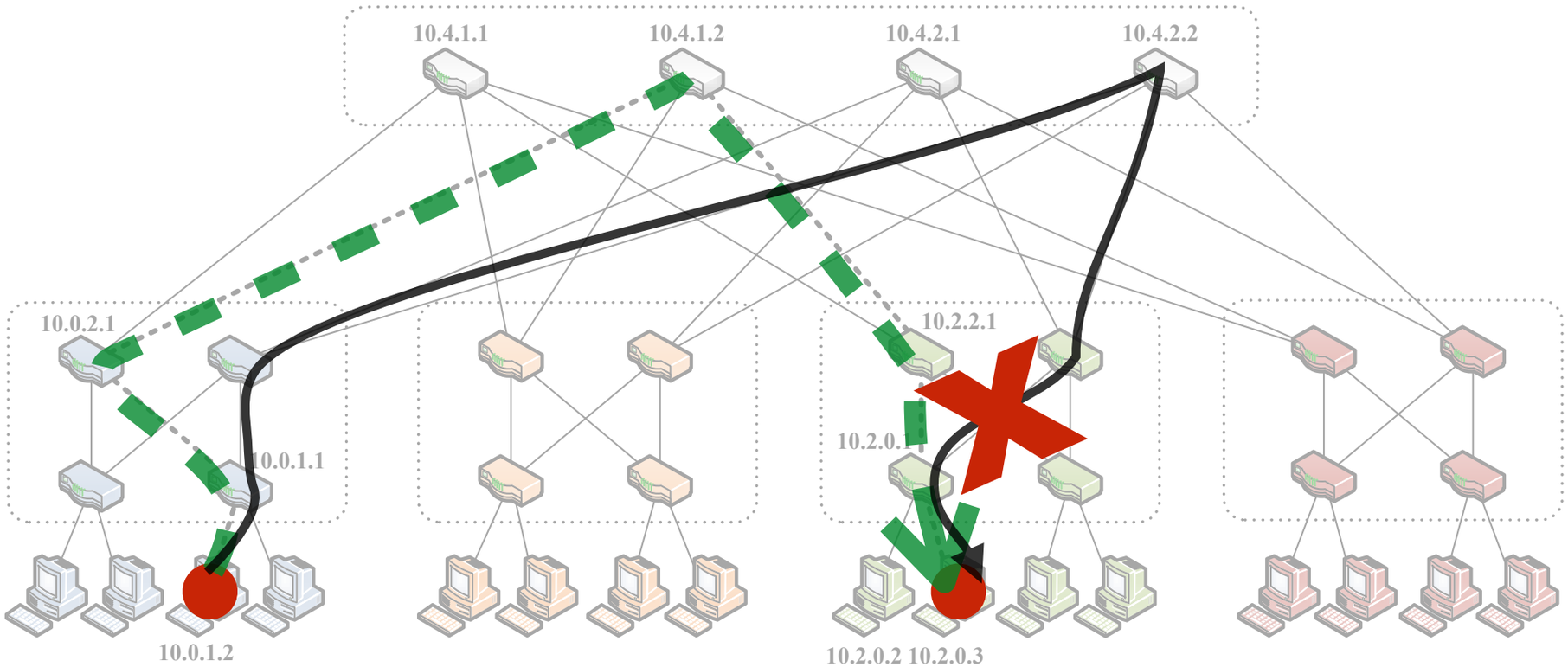
Dina Katabi
MIT
dk@mit.edu

Shantanu Sinha
Microsoft
shans@microsoft.com

Arthur Berger
MIT/Akamai
awberger@mit.edu

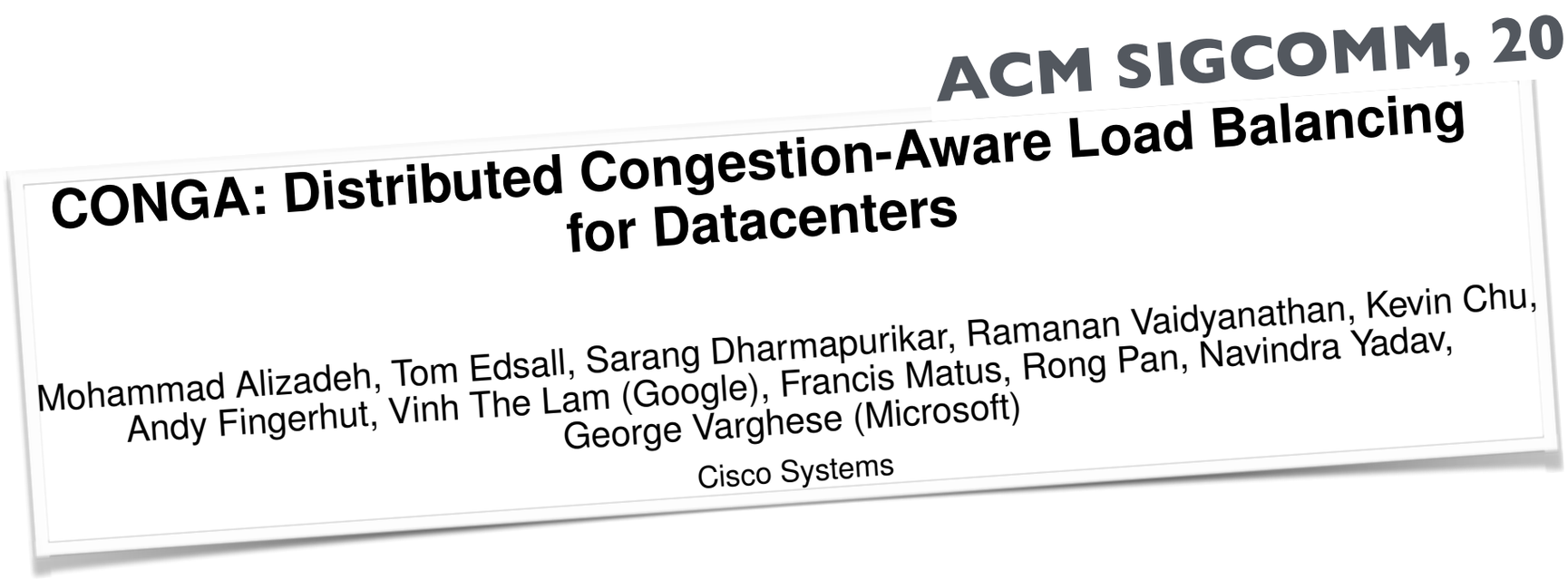


ECMP: local, oblivious choices

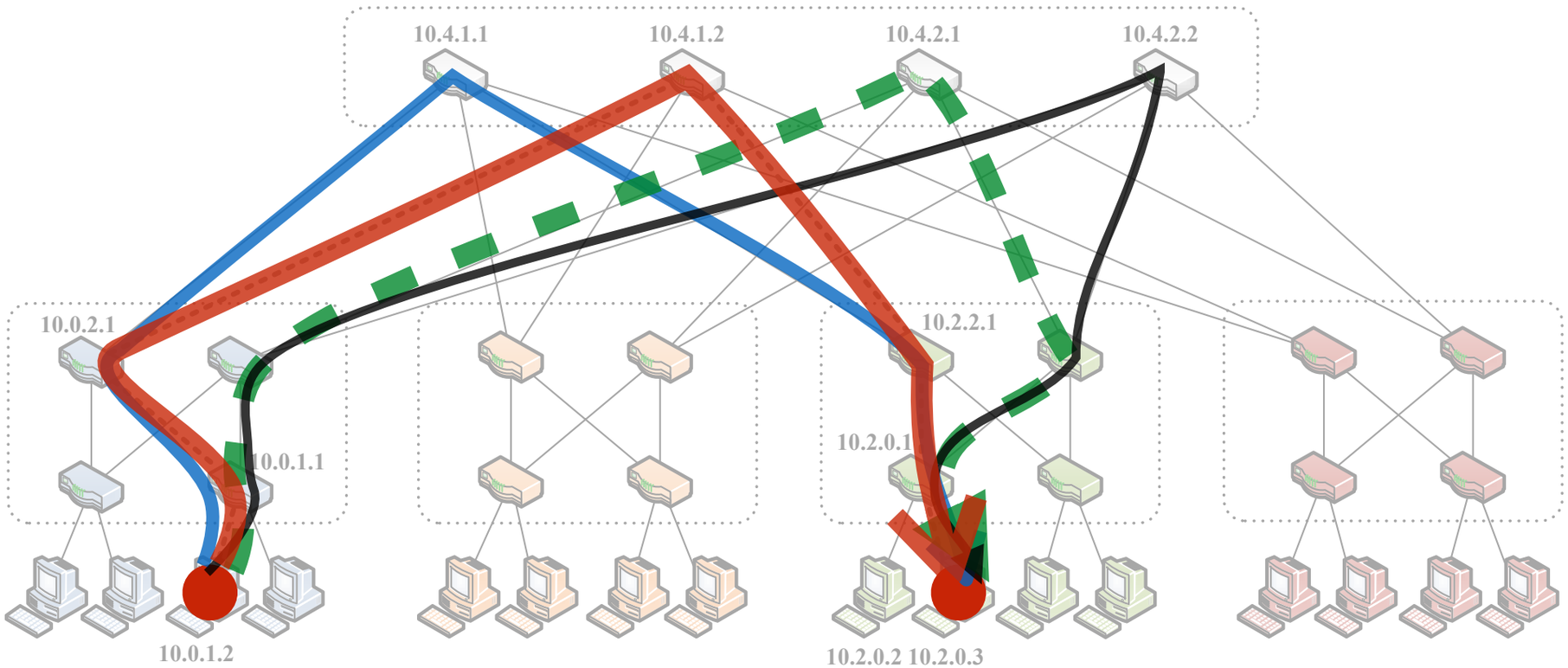


Cannot avoid distant failures!

CONGA: edge based monitoring



CONGA: edge based monitoring



Leaf switch monitors path performance in real time:

	Path 1	Path 2	Path 3	Path 4
Dest. switch	3	7	2	1

Research directions in DC nets

Physical layer

Topology

Load balancing

Transport

Control & SDN