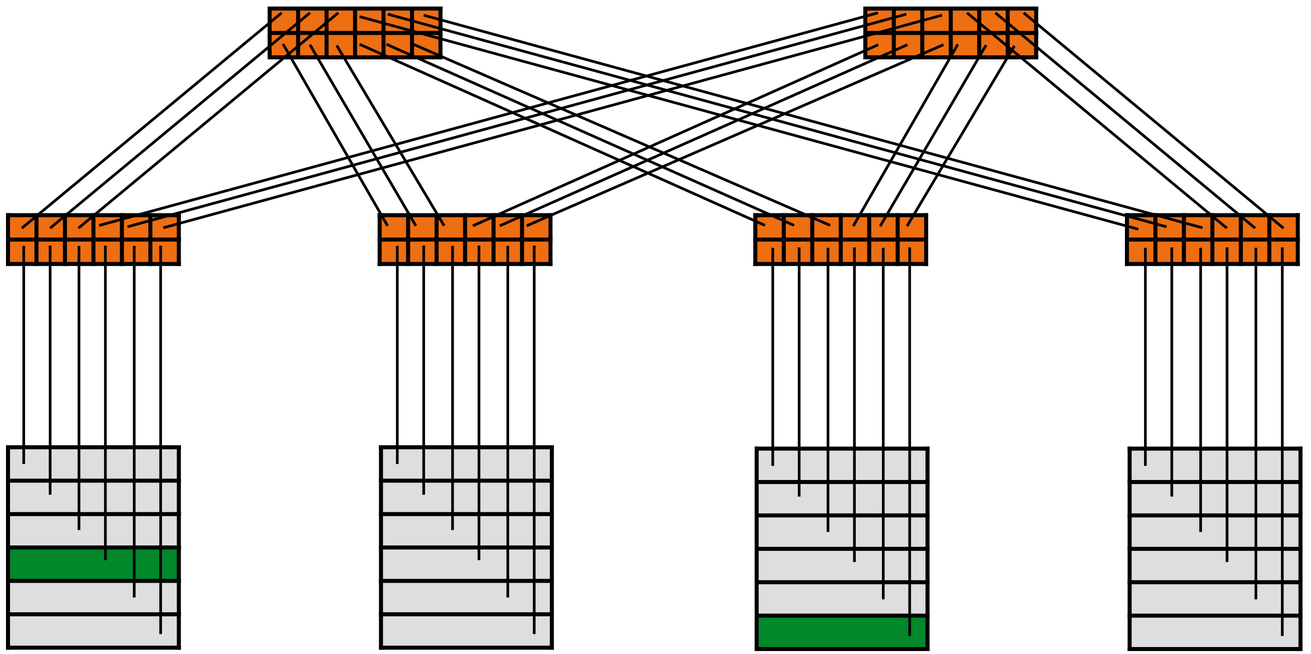
Data centers: routing

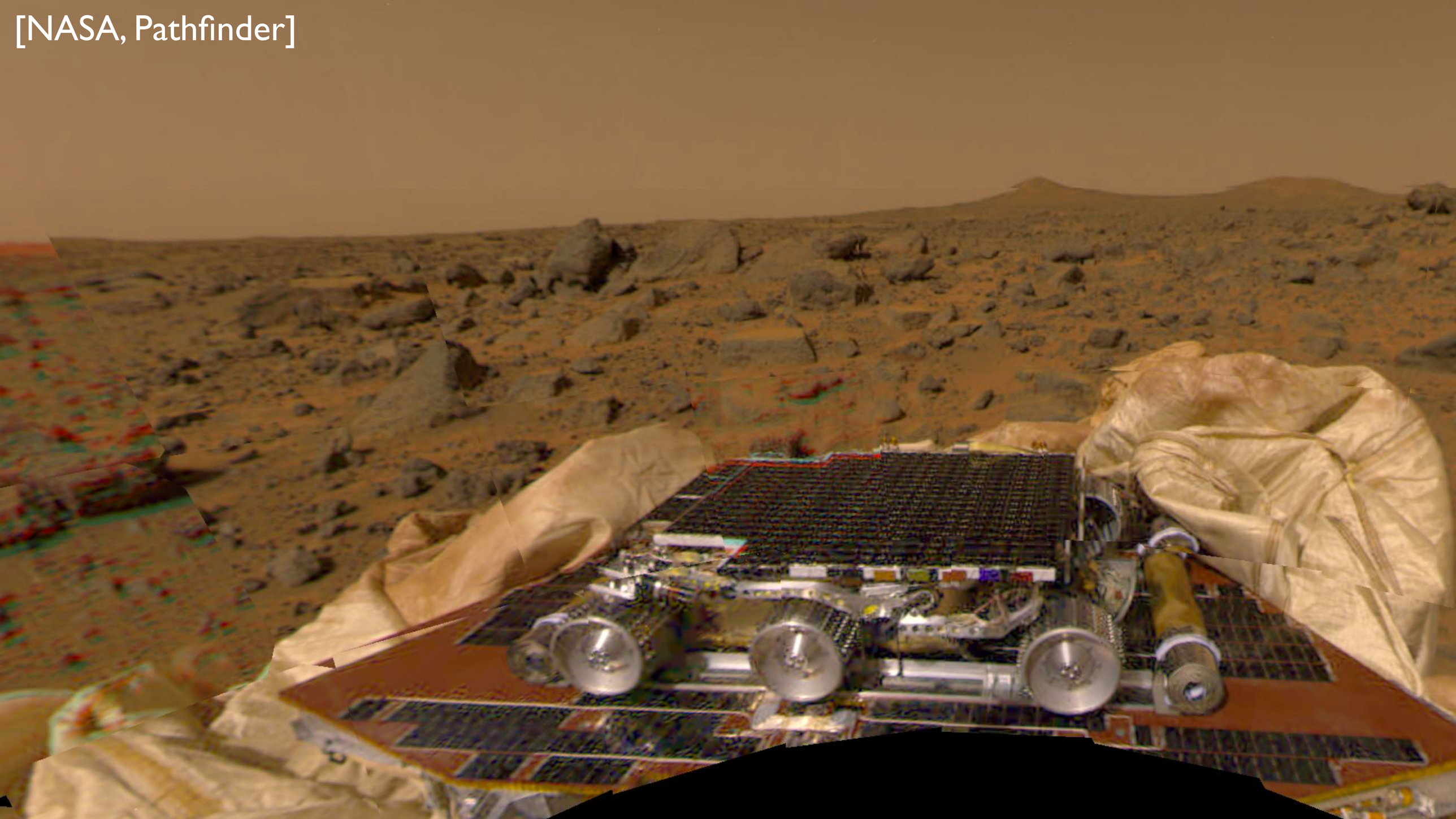
Ankit Singla

ETH Zürich Spring 2017

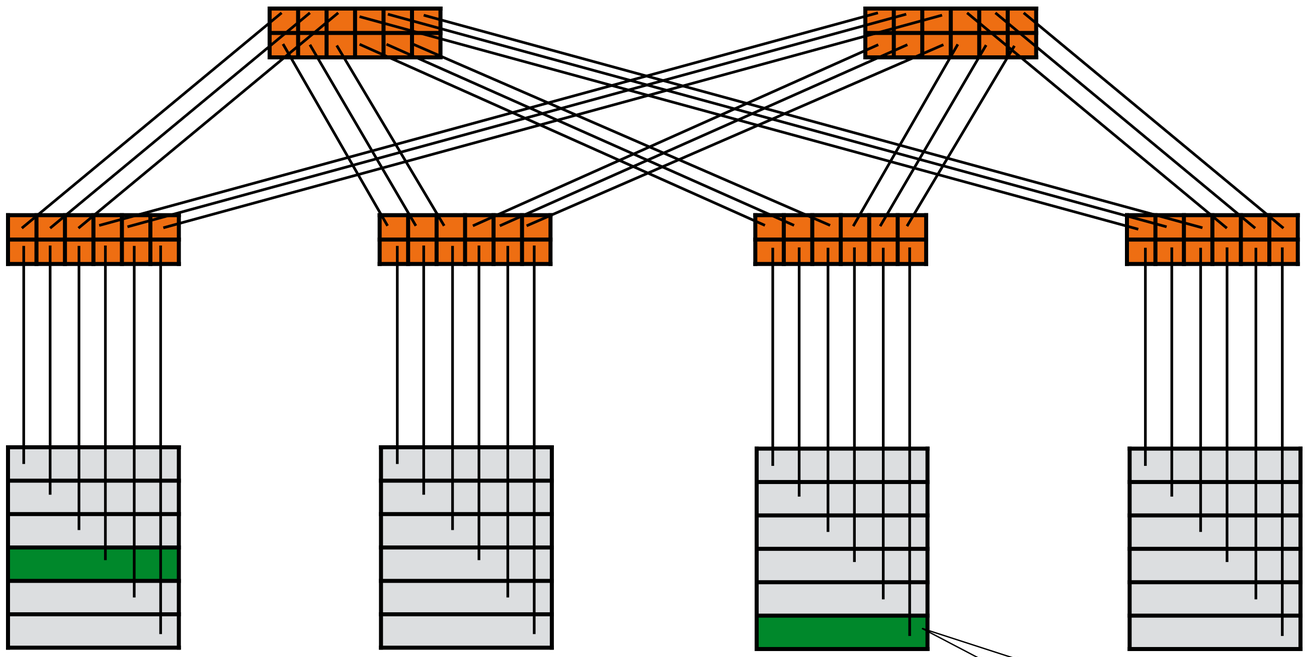
Goal: find paths through the network



How do I reach the destination server?


[NASA, Pathfinder]

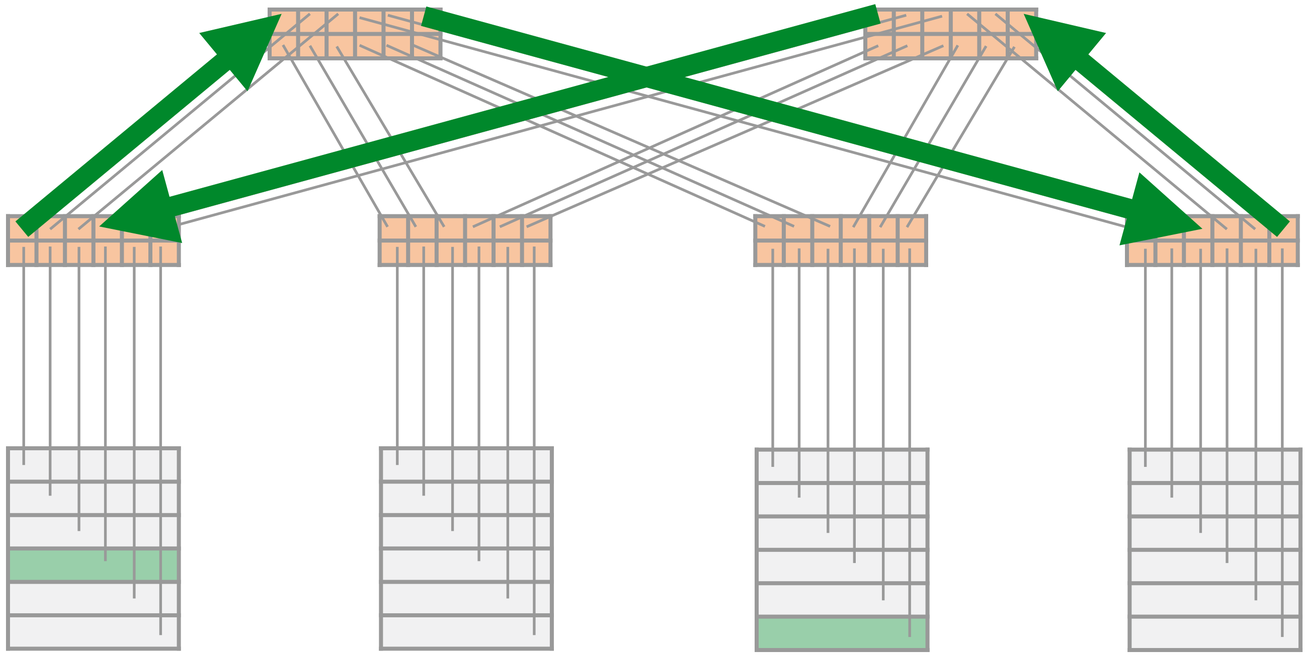
Plug-and-play at Layer 2



Send to a.b.c.d

(ARP) a.b.c.d =
A::B::C::D::E::F

Loop-freedom in forwarding

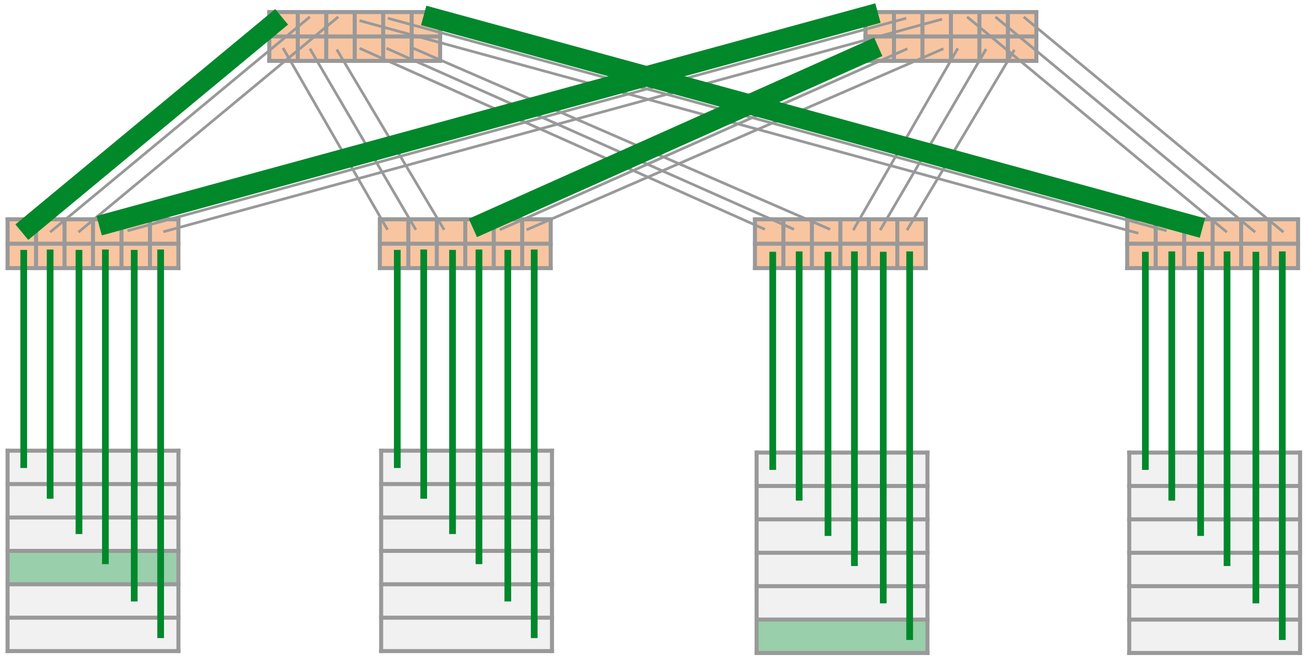


Radia Perlman: Inventor of STP

I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.

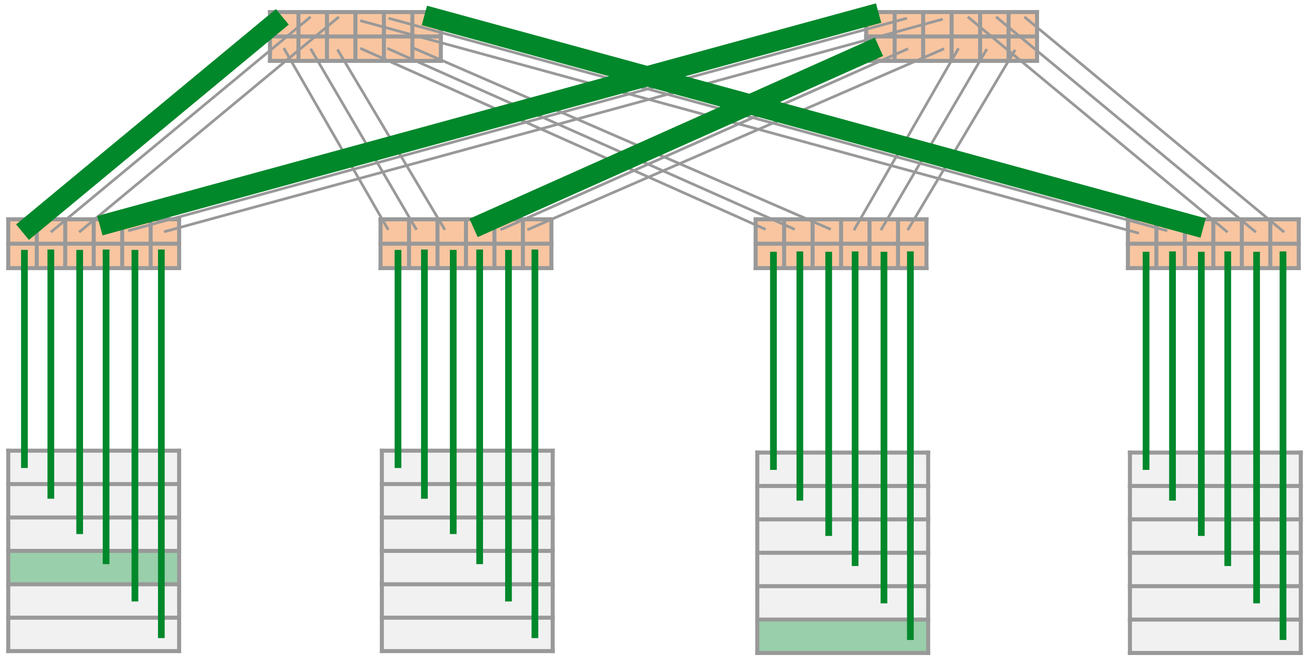
[Excerpt from "Algorhyme" by Radia Perlman]

Spanning tree



Disable all other links!

Spanning tree

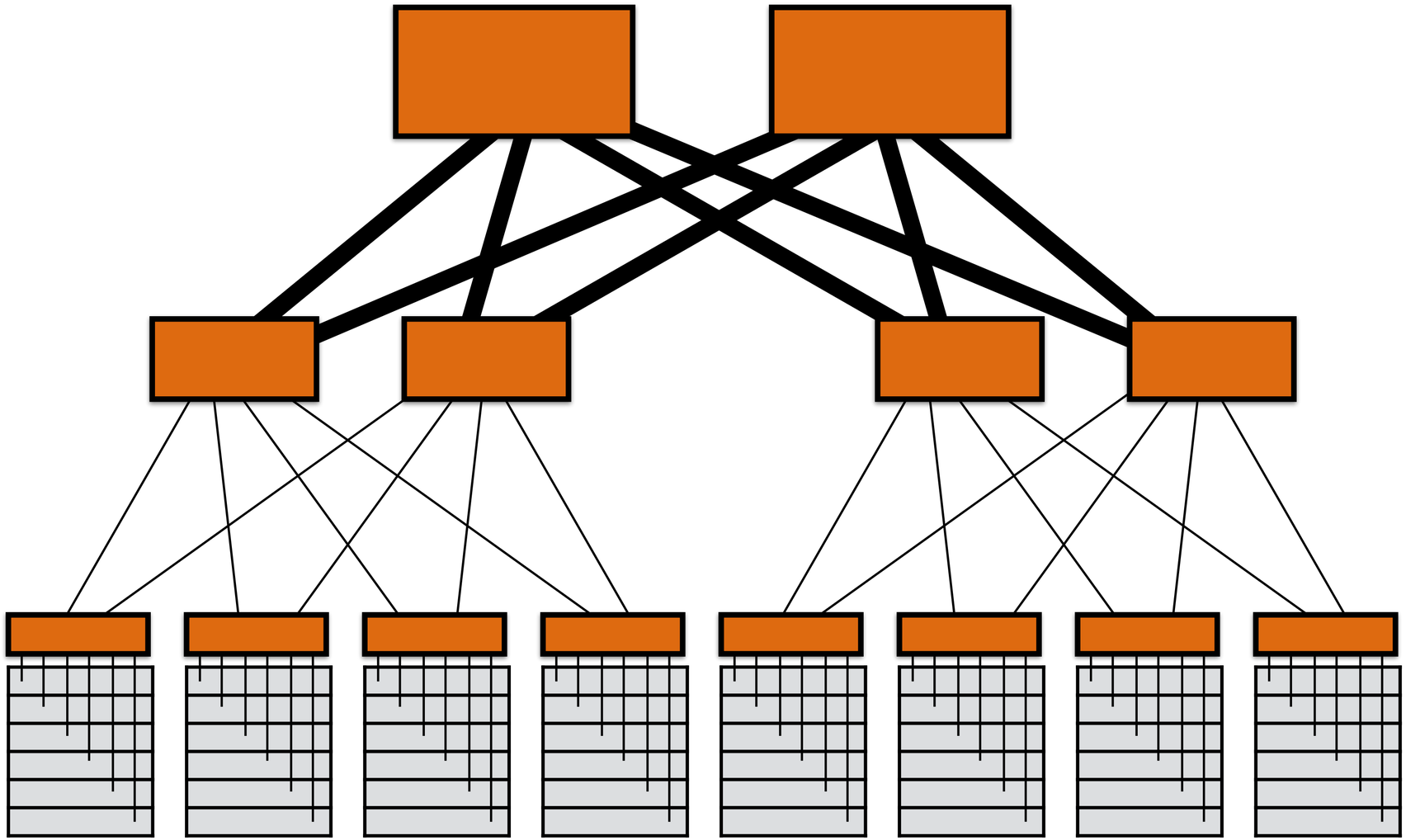


The Atlantic Popular Latest Sections Magazine

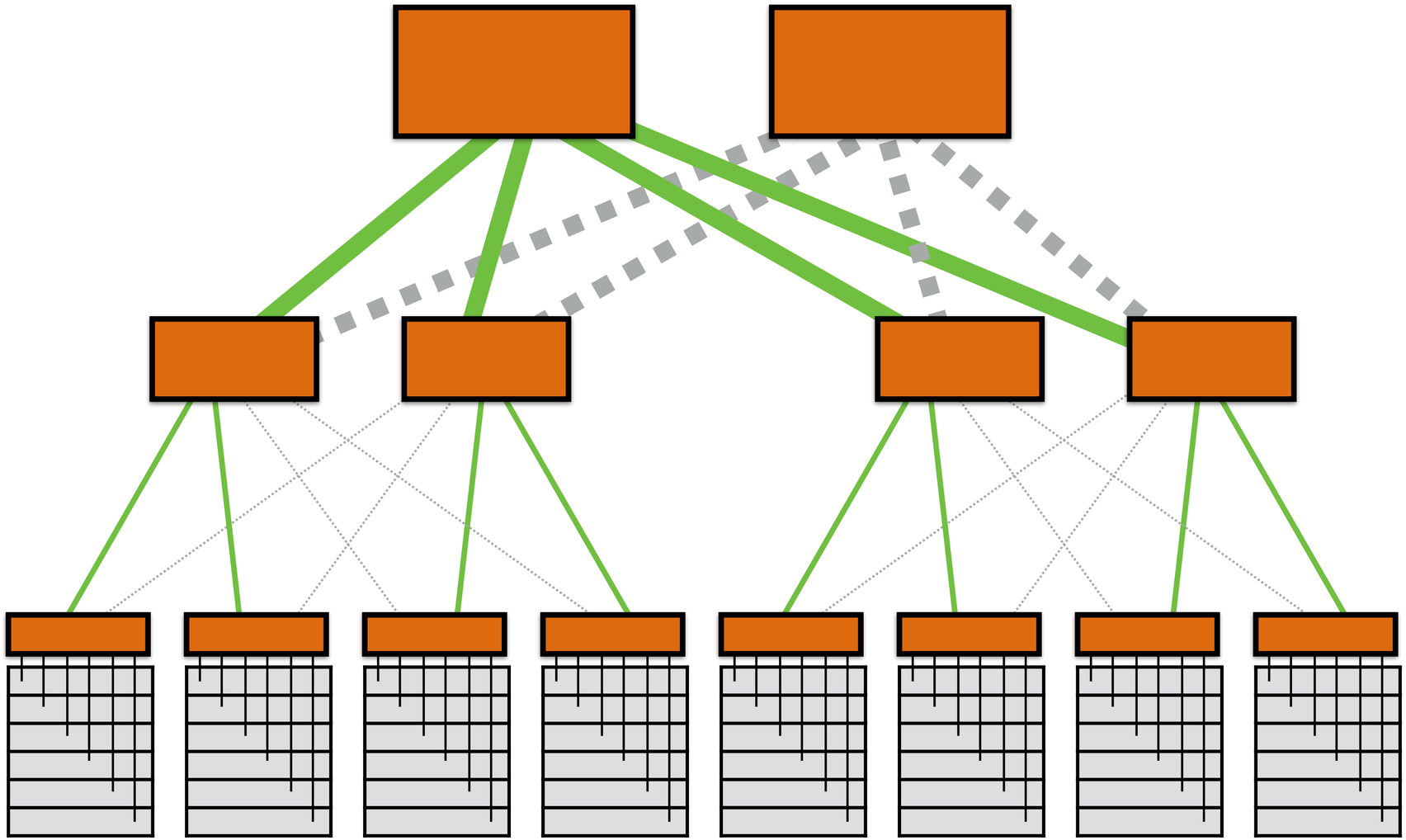
Radia Perlman: Don't Call Me the Mother of the Internet

A very important contribution!

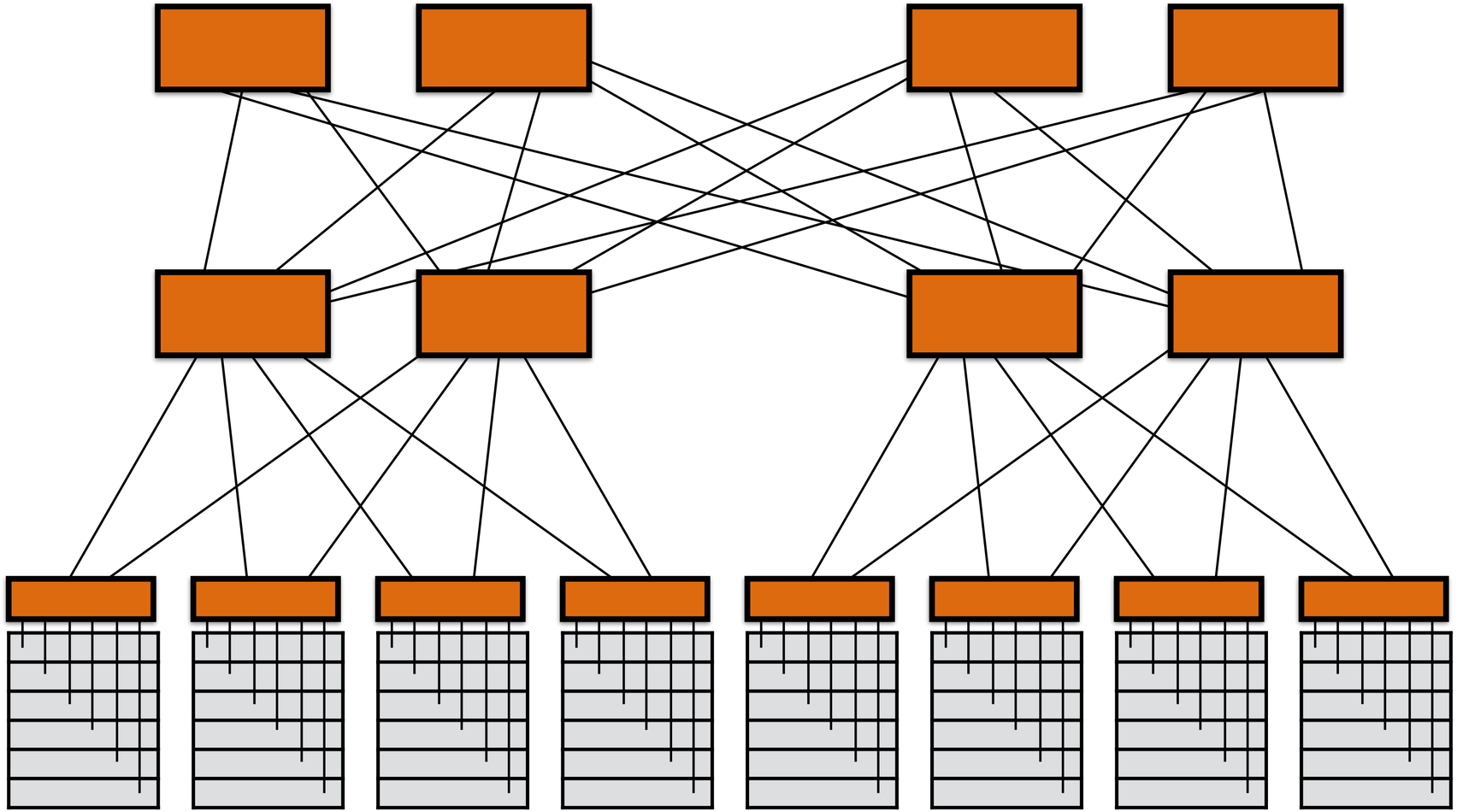
STP works for tree-like networks



STP works for tree-like networks



STP works for tree-like networks



FabricPath

M-LAG

TRILL

QFabric

SPB

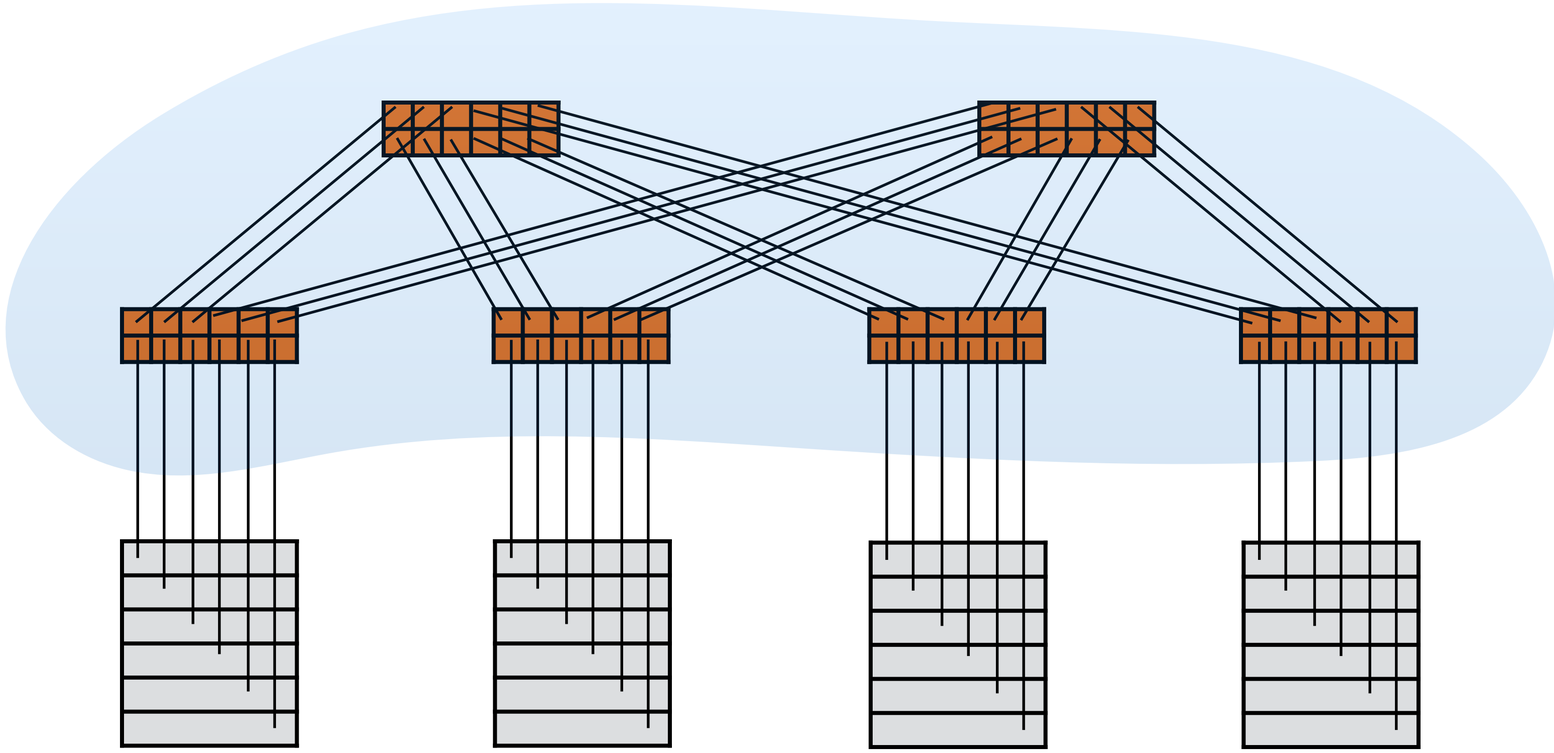
VCS

Image adapted from *"Buchstabensuppe"* by Till Westermayer

Recall: several routing options

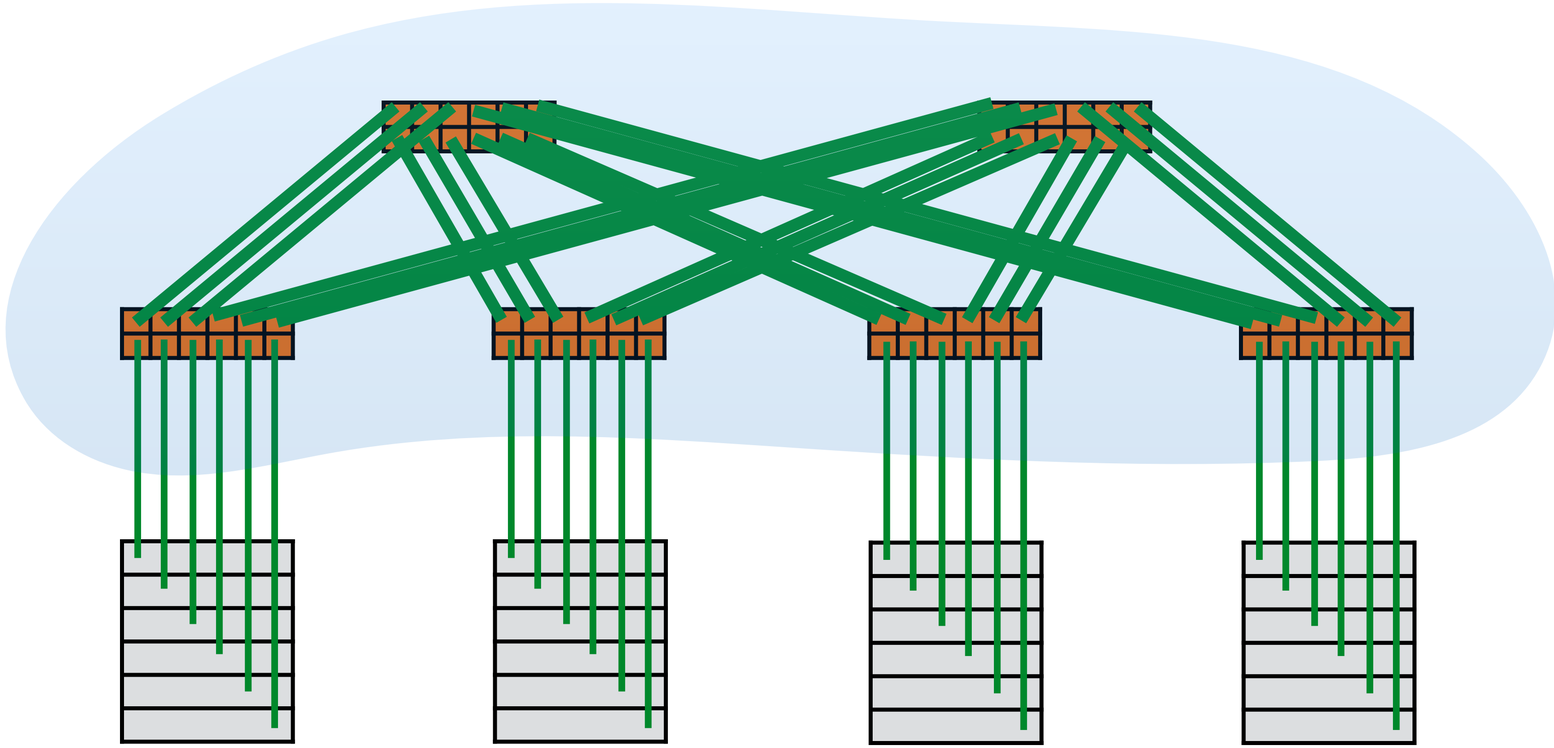
1. Structure the network itself!
2. Incorporate routing information in node names
3. Store information in routing tables
4. Store maps of the network at some or all devices

Transparent Interconnection of Lots of Links



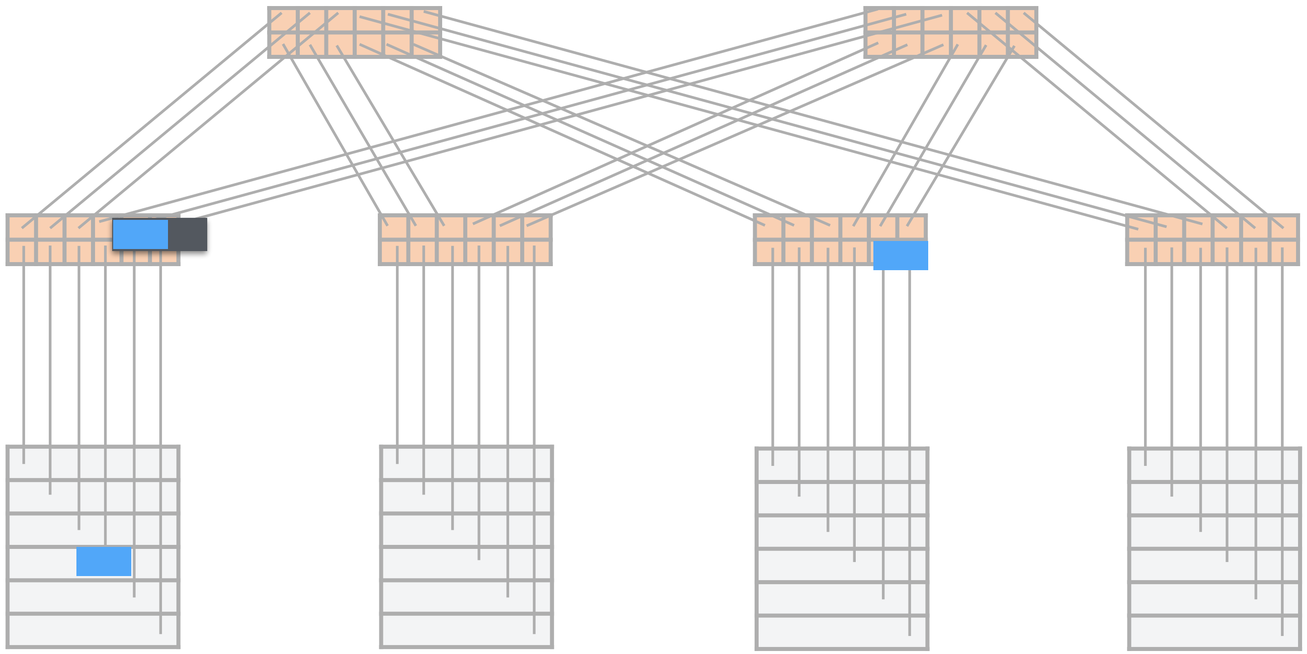
Link-state protocol between switches

TRILL



All links usable!

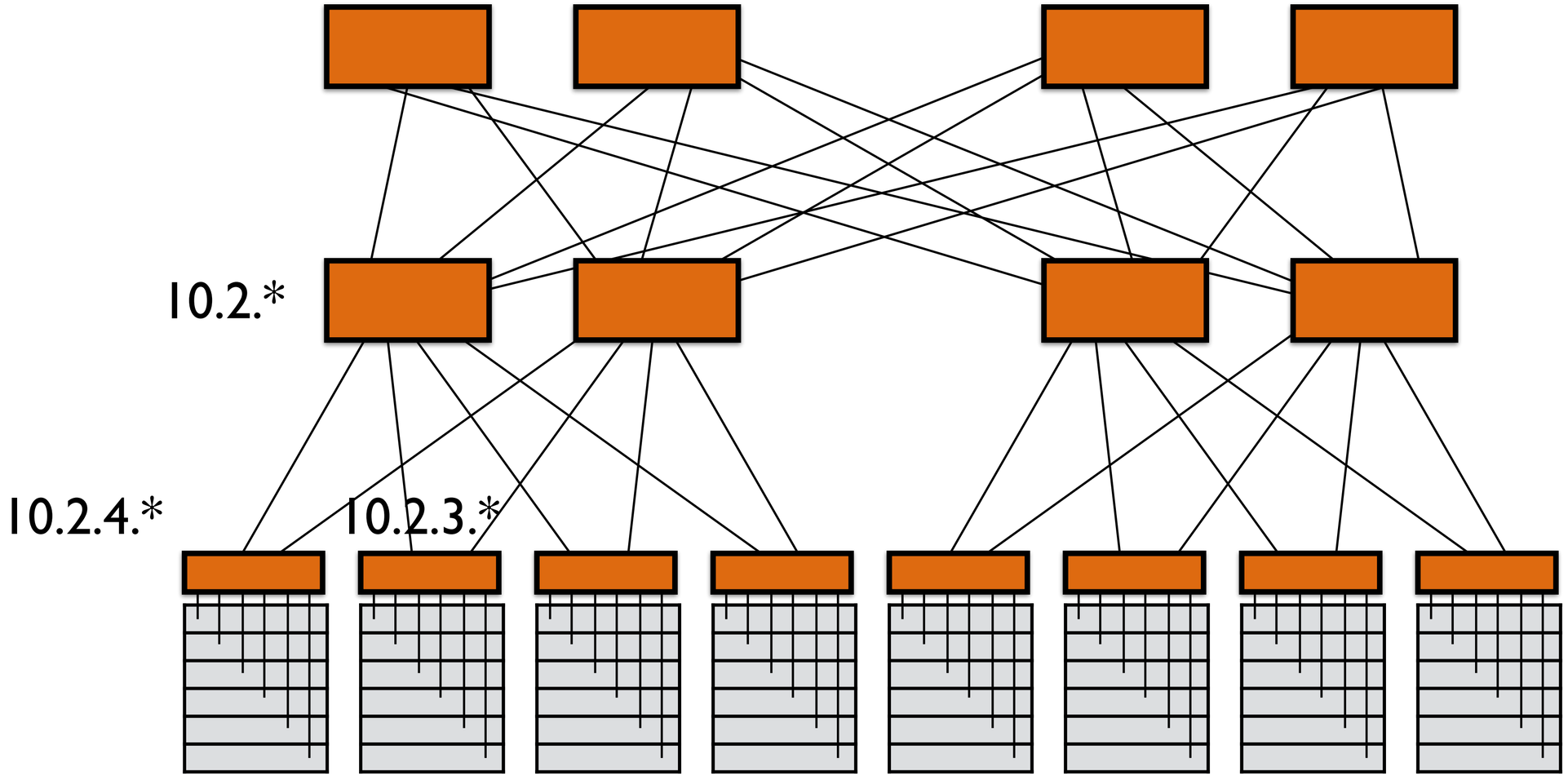
TRILL



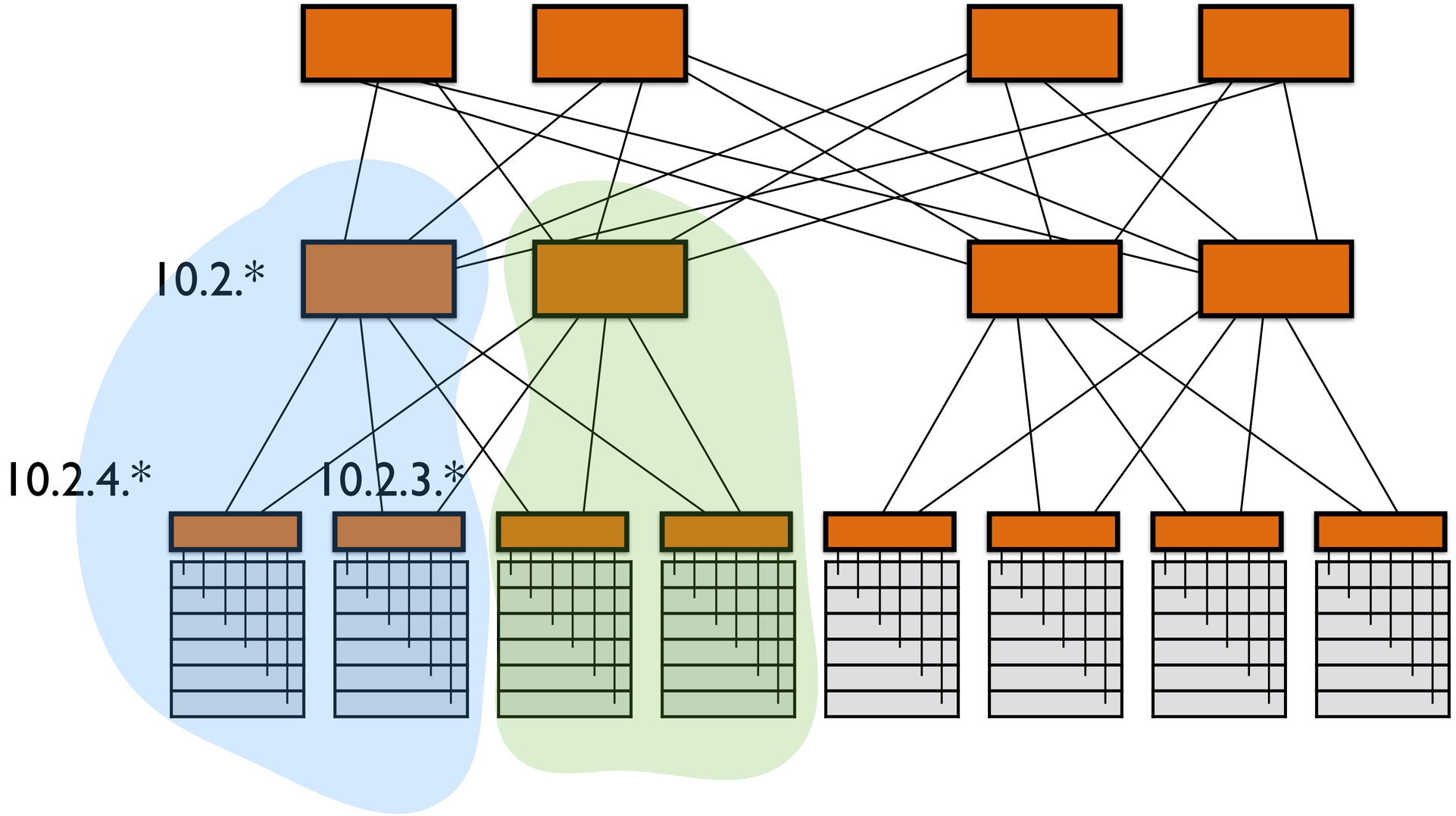
How is host connectivity determined?

Scale
Scale
Scale
Scale
Scale
Scale
Scale

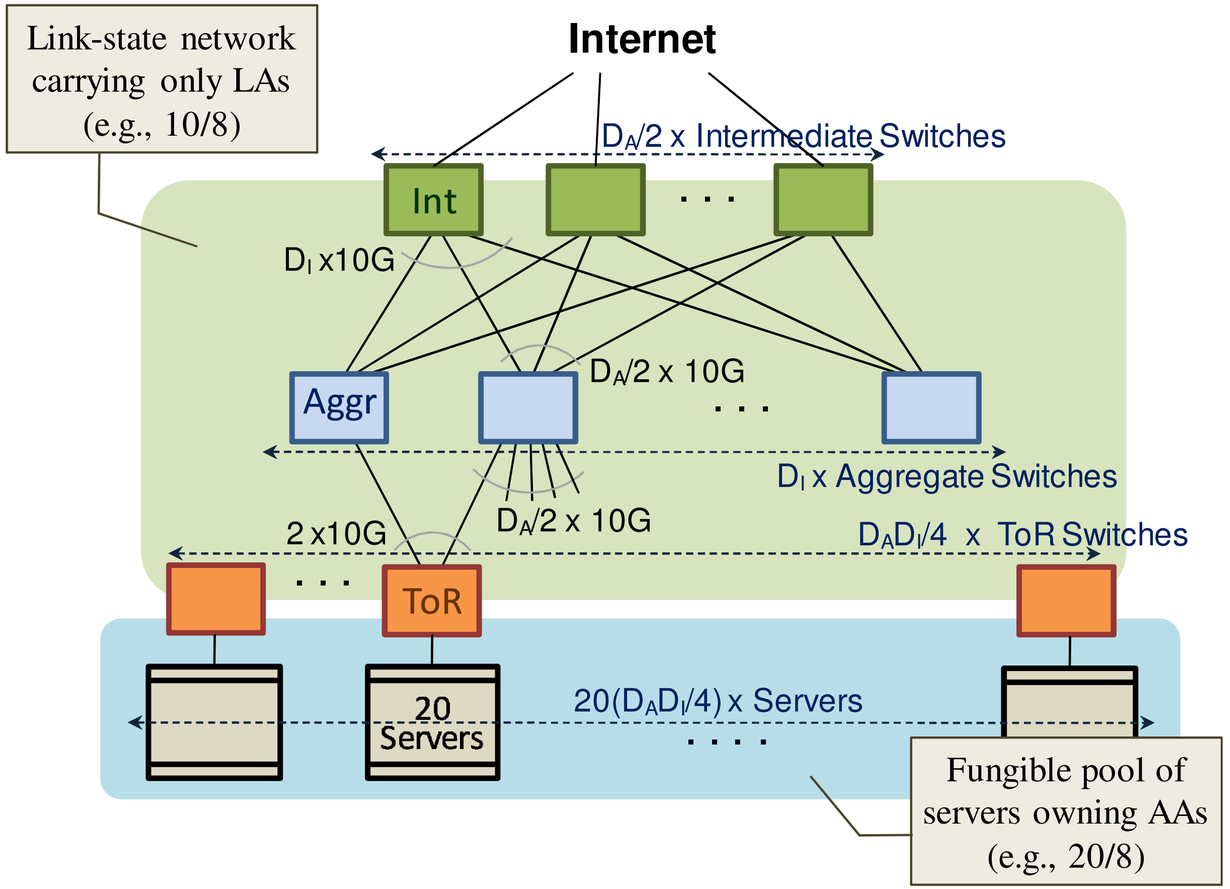
Traditional solution: subnets



Traditional solution: VLANs



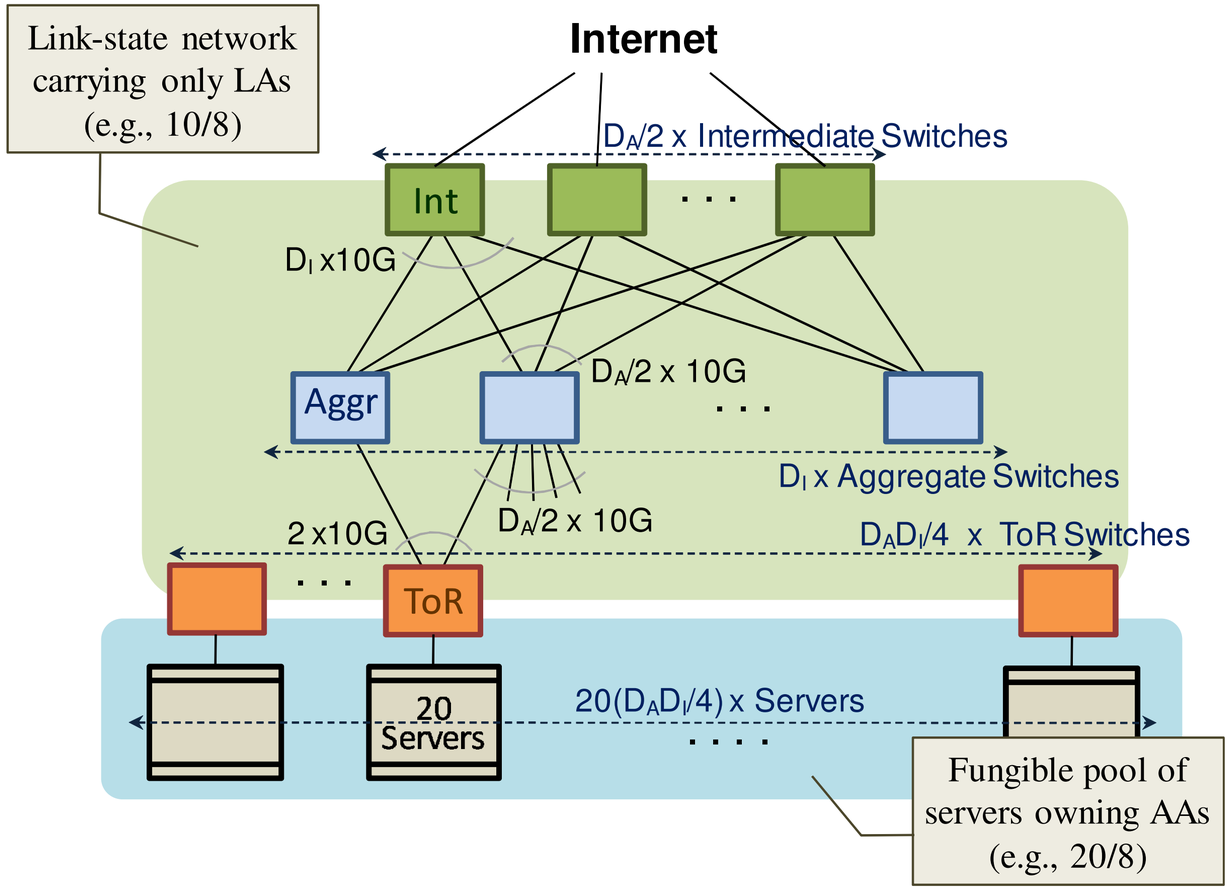
VL2: "Virtual Layer 2"



VL2 @ **Microsoft**, ACM SIGCOMM'09

Greenburg, Hamilton, Jain, Kandula, Kim, Lahiri, Maltz, Patel, Sengupta

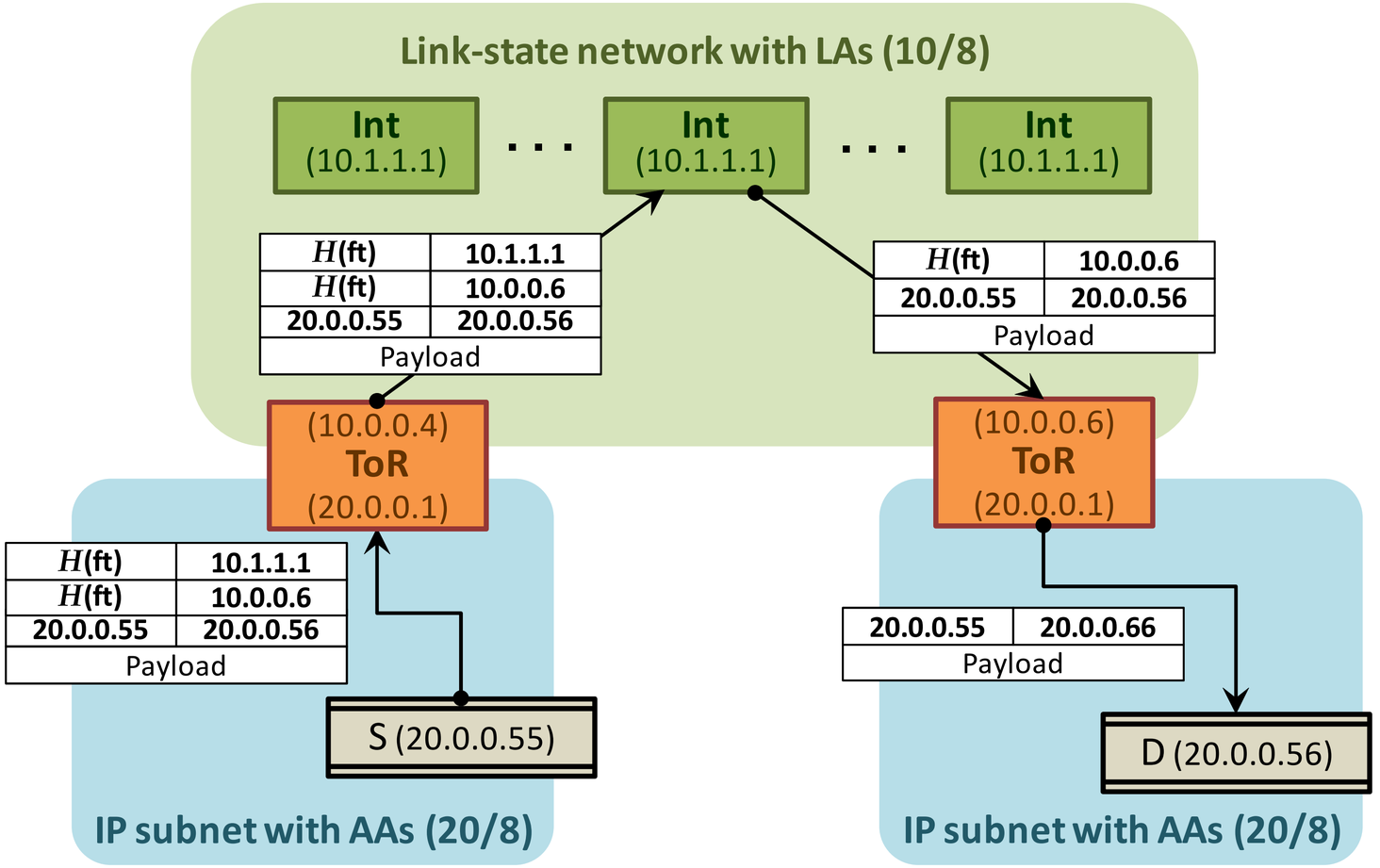
VL2: "Virtual Layer 2"



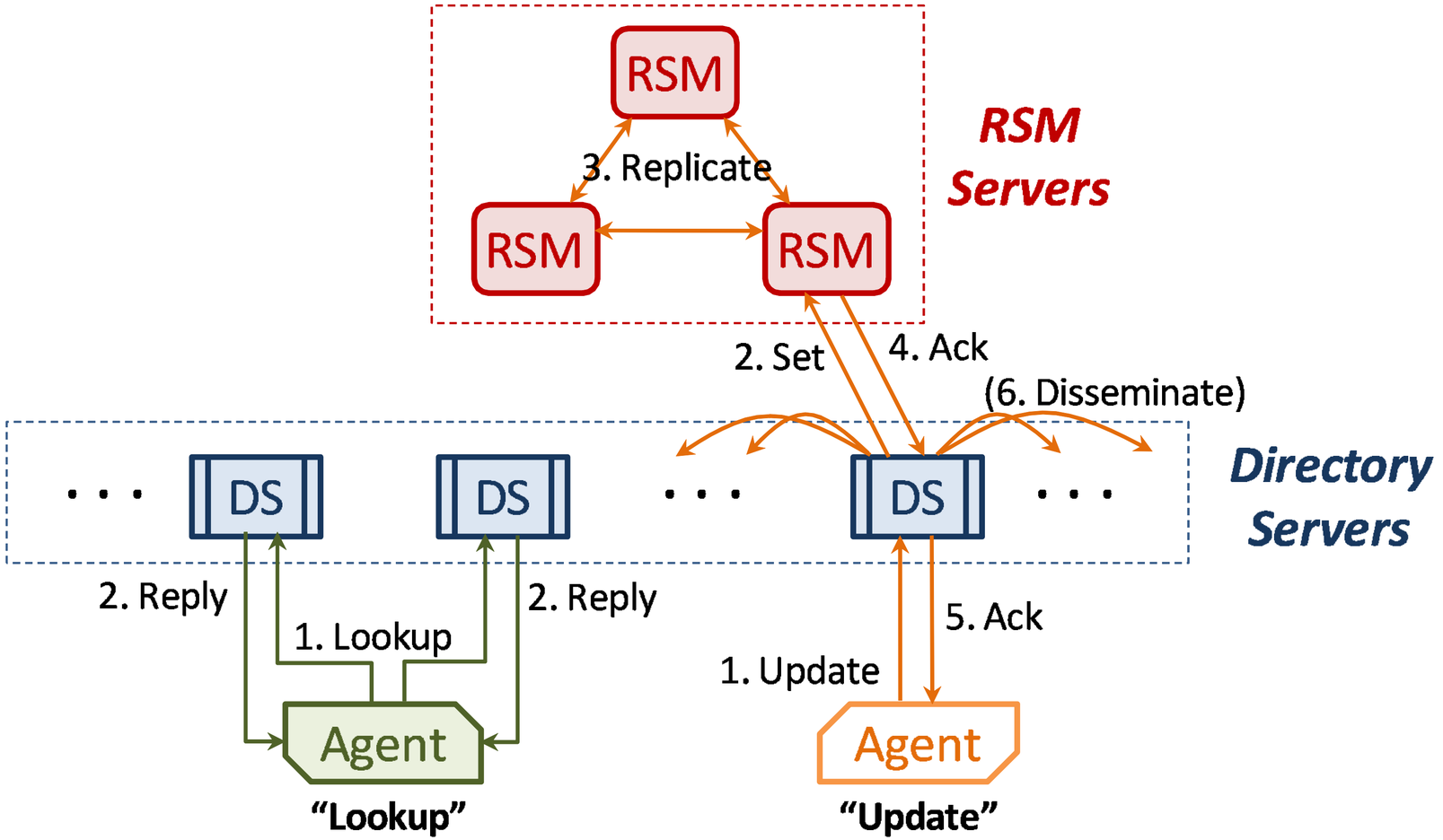
- Clos network
- Name ≠ Location
- Randomized load balancing

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

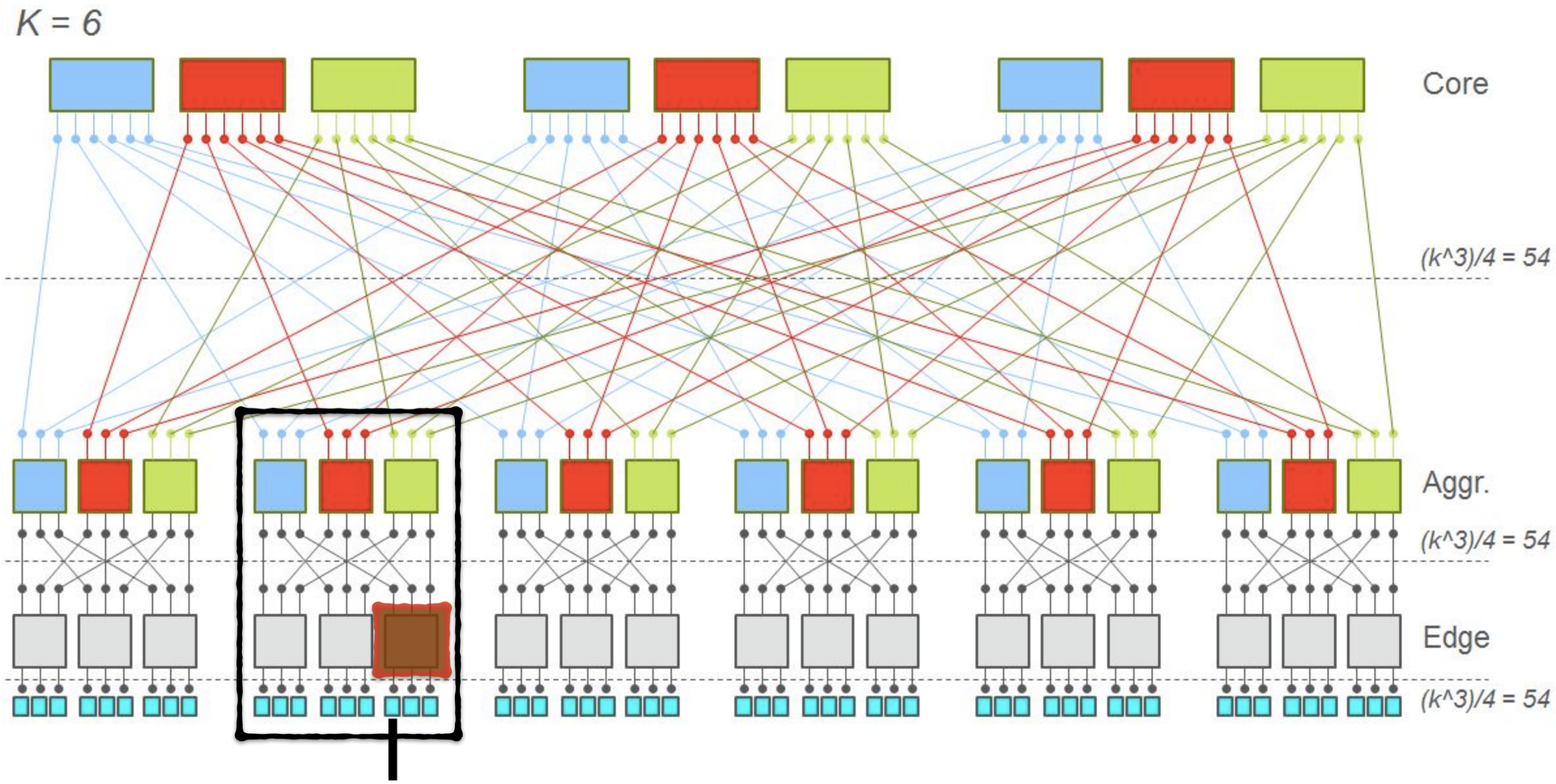
VL2: Location vs. App addresses



VL2: Address translation



Alternative solution: PortLand

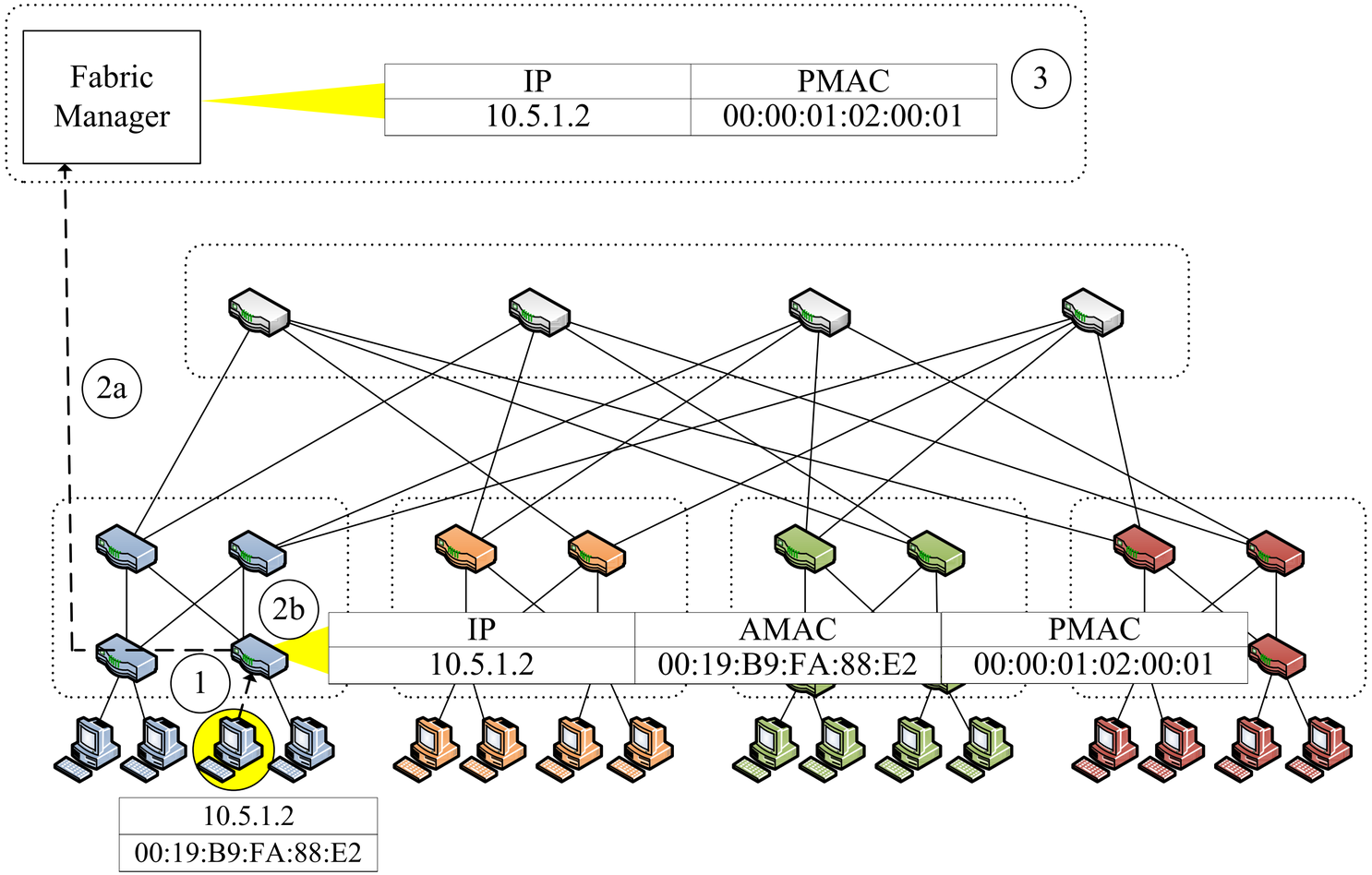


PMAC = pod.position.port.vmid

2 3 1 38

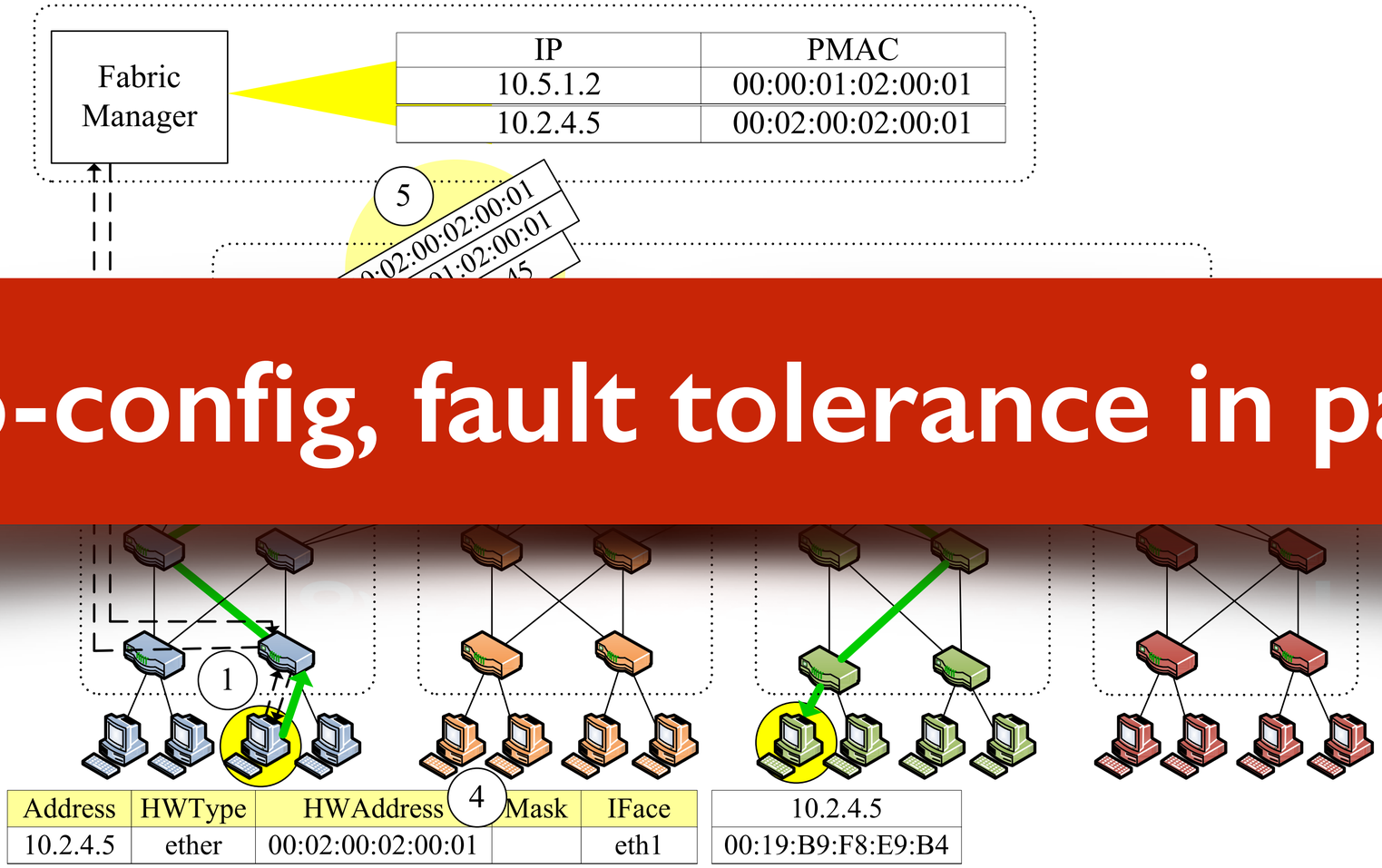
Image: Francesco Celestino

Alternative solution: PortLand



PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric
Mysore et al., ACM SIGCOMM'09

Alternative solution: PortLand



Auto-config, fault tolerance in paper

PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric
Mysore et al., ACM SIGCOMM'09

Routing in data centers

Internet Engineering Task Force (IETF)
Request for Comments: 7938
Category: Informational
ISSN: 2070-1721

P. Lapukhov
Facebook
A. Premji
Arista Networks
J. Mitchell, Ed.
August 2016

Use of BGP for Routing in Large-Scale Data Centers

ACM CCR, 2014

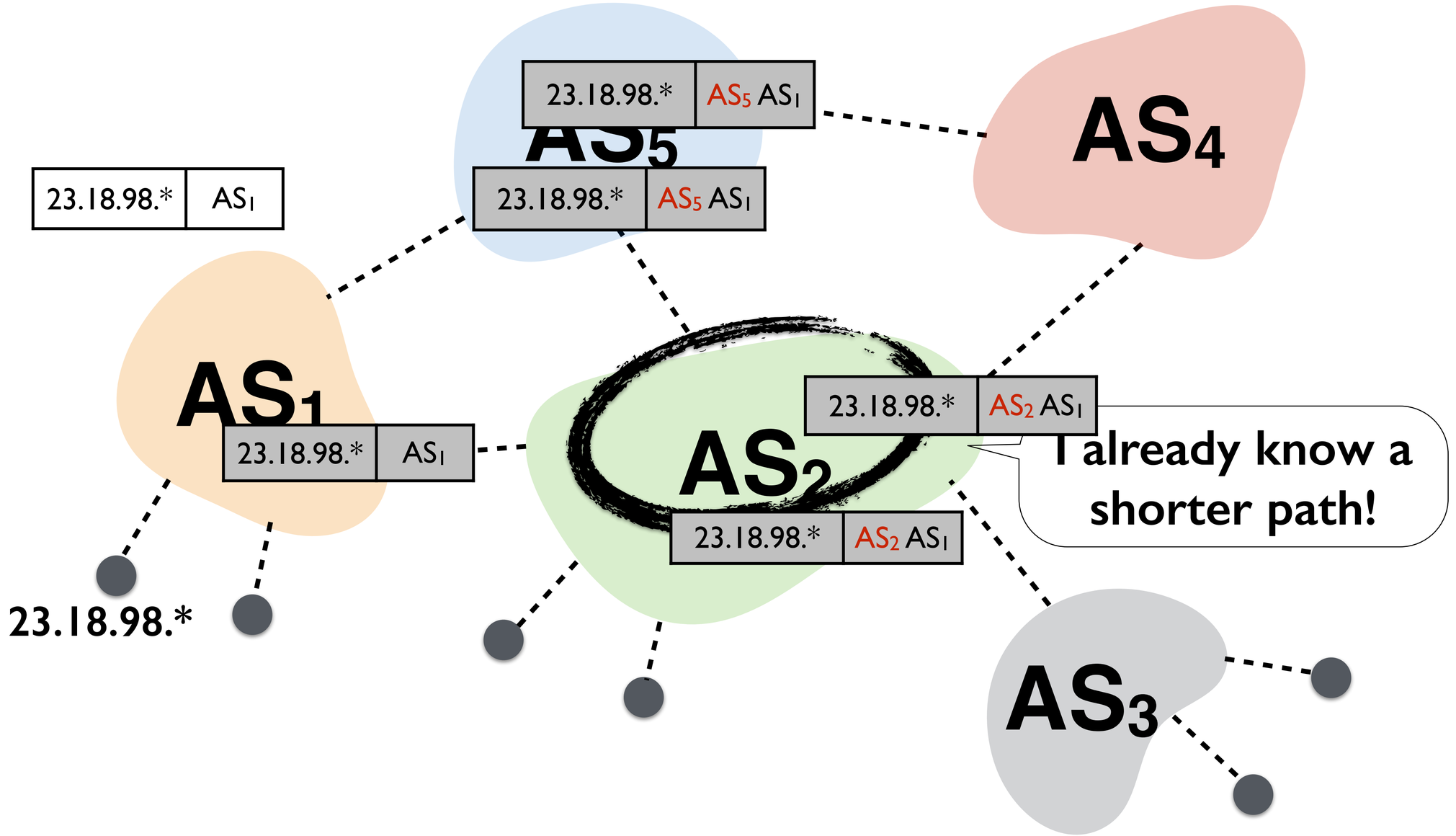
The Road to SDN: An Intellectual History of Programmable Networks

Nick Feamster
Georgia Tech
feamster@cc.gatech.edu

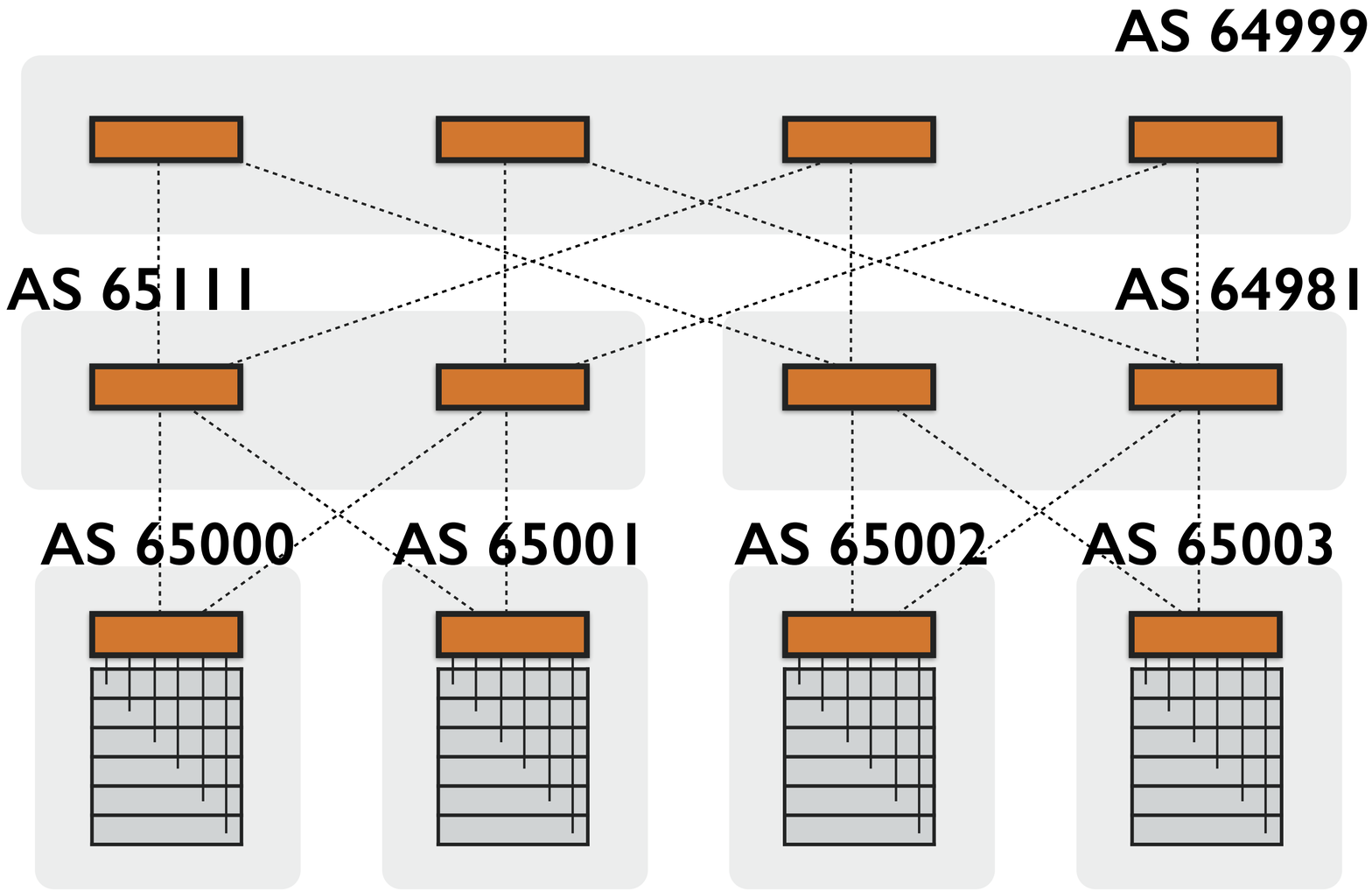
Jennifer Rexford
Princeton University
jrex@cs.princeton.edu

Ellen Zegura
Georgia Tech
ewz@cc.gatech.edu

BGP: Border Gateway Protocol

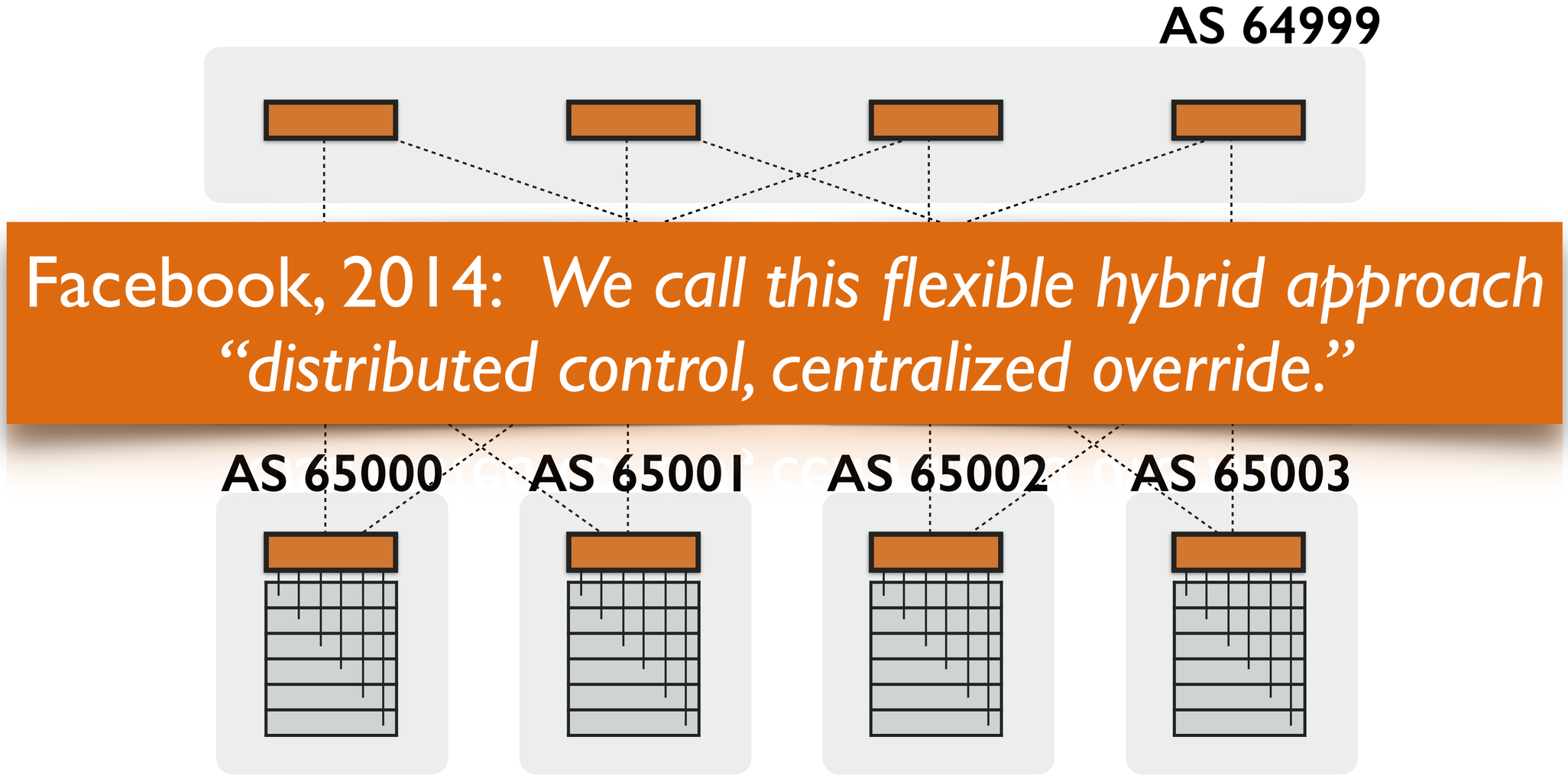


BGP in the data center



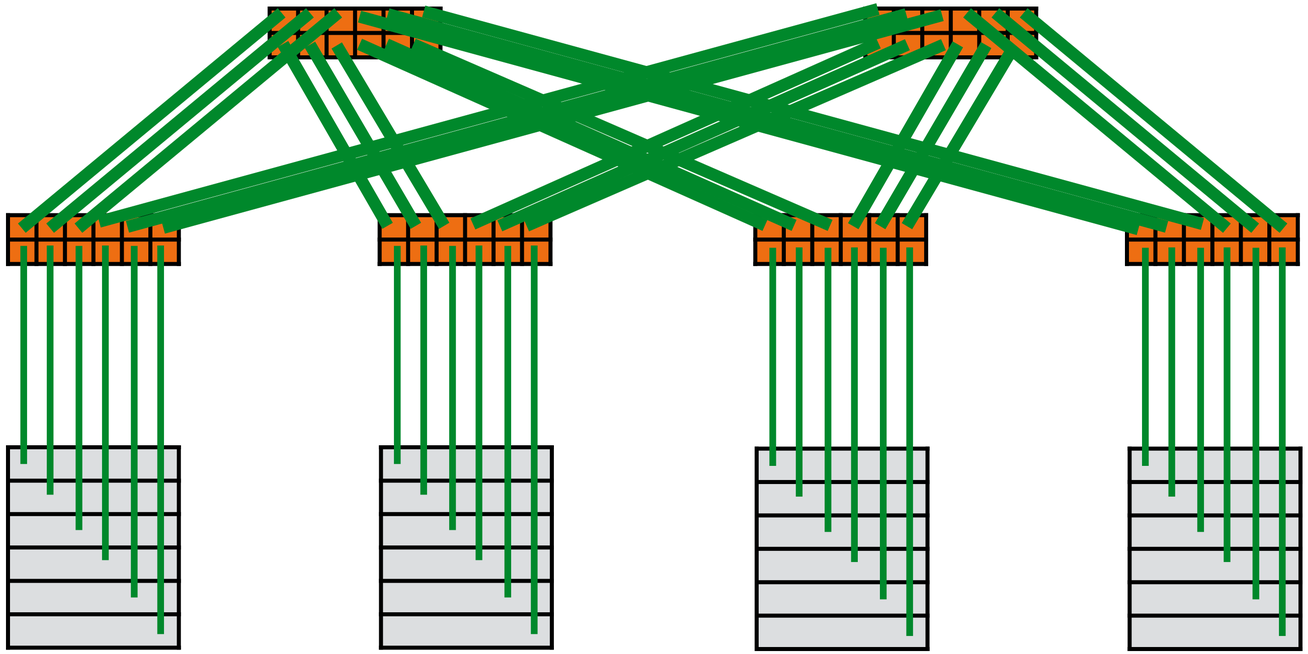
Divide groups of switches and servers into ASes; use BGP!

BGP in the data center

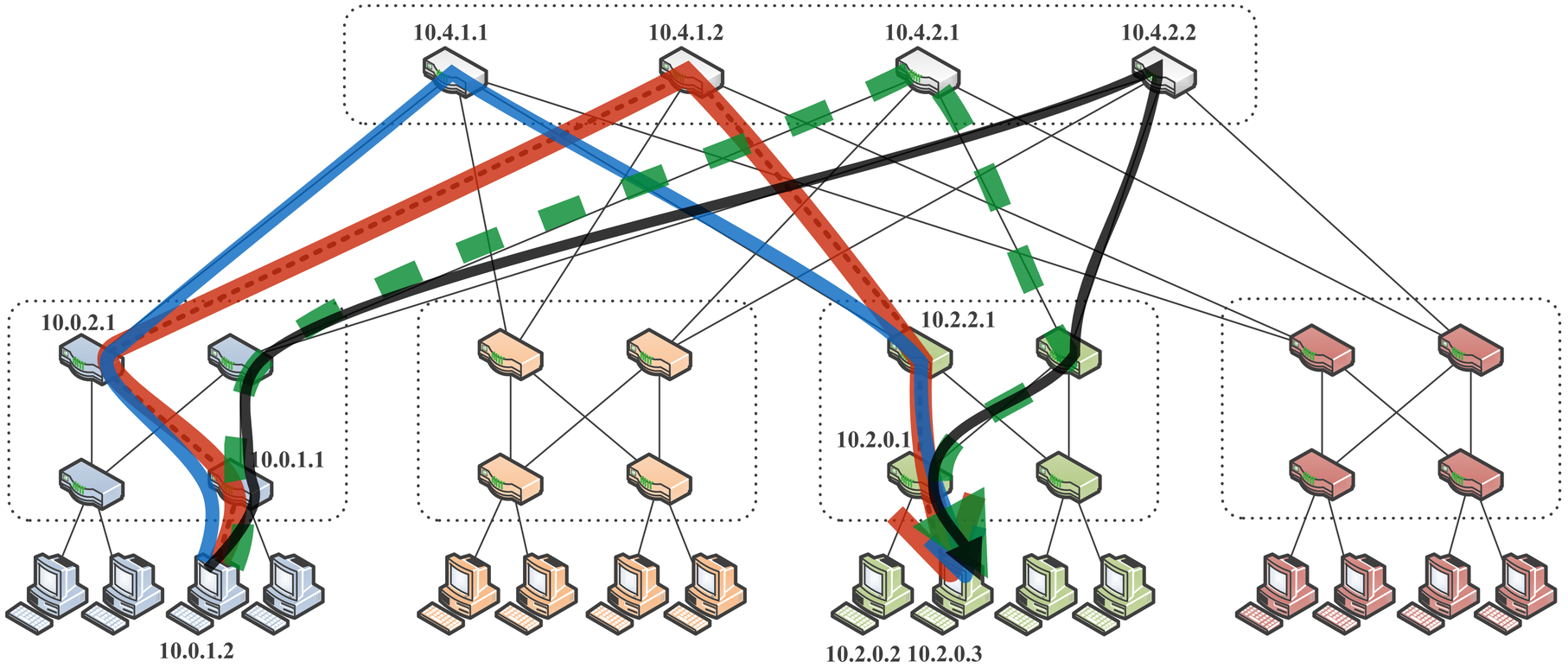


Divide groups of switches and servers into ASes; use BGP!

Routing



Multipath routing



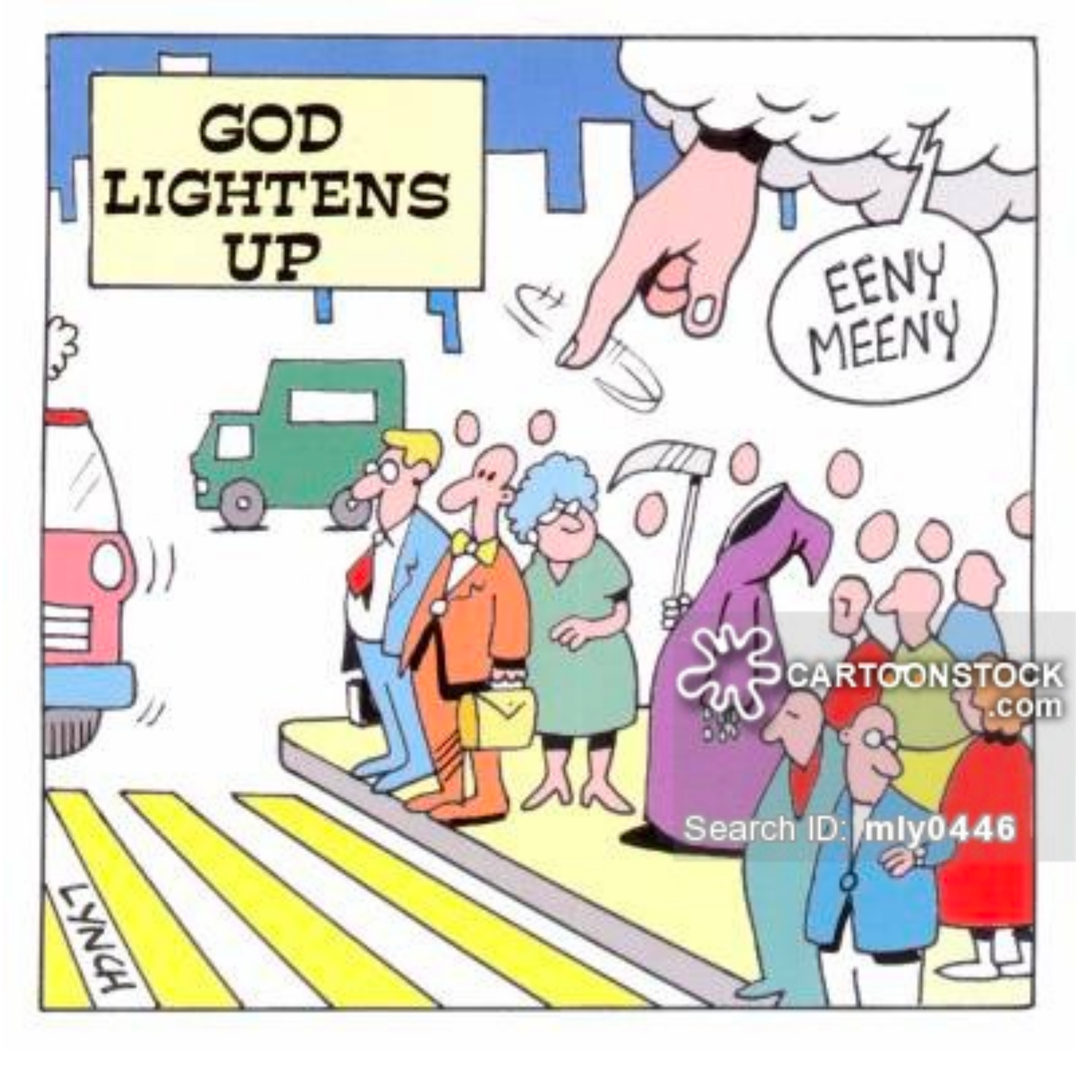
A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares
malfares@cs.ucsd.edu

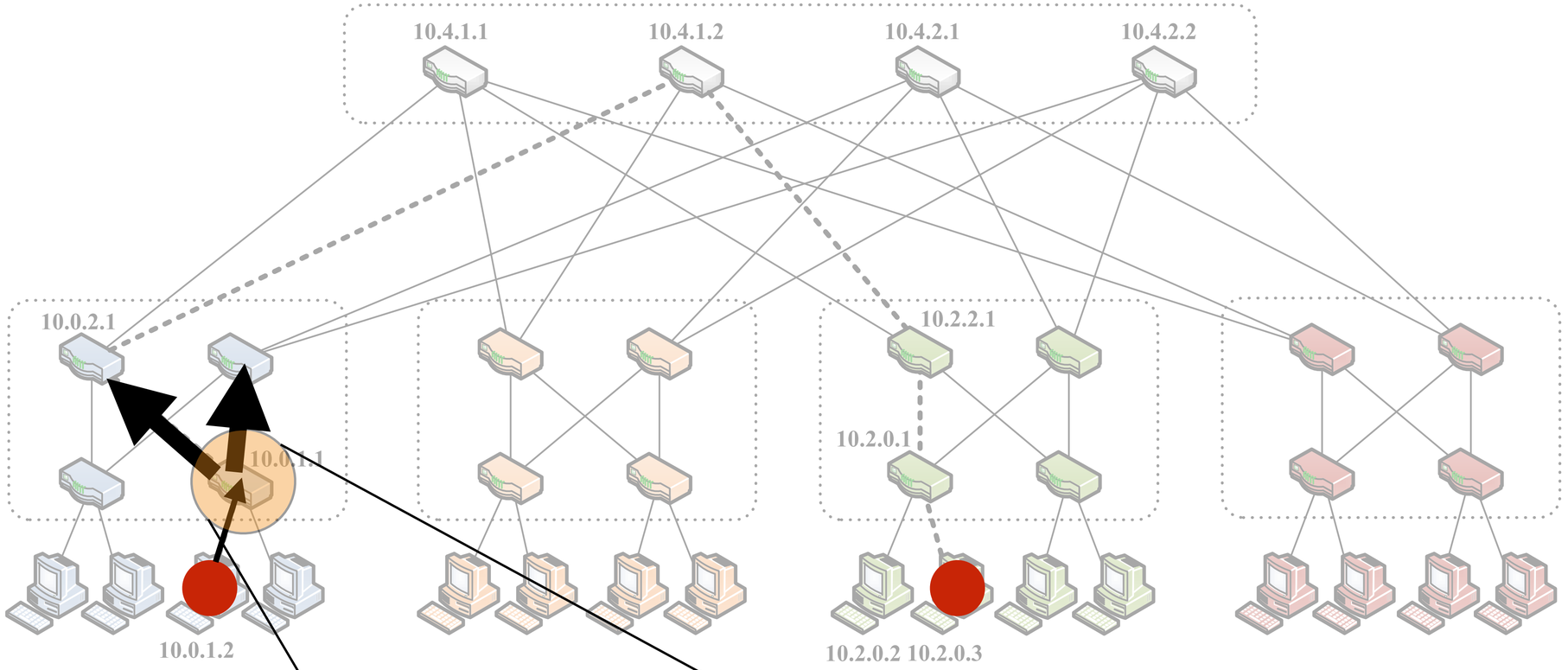
Alexander Loukissas
aloukiss@cs.ucsd.edu

Amin Vahdat
vahdat@cs.ucsd.edu

Randomization strikes again!



Equal cost multi-path (ECMP)



10.2.0.3	port 1
10.2.0.3	port 2

Choose uniformly at random!

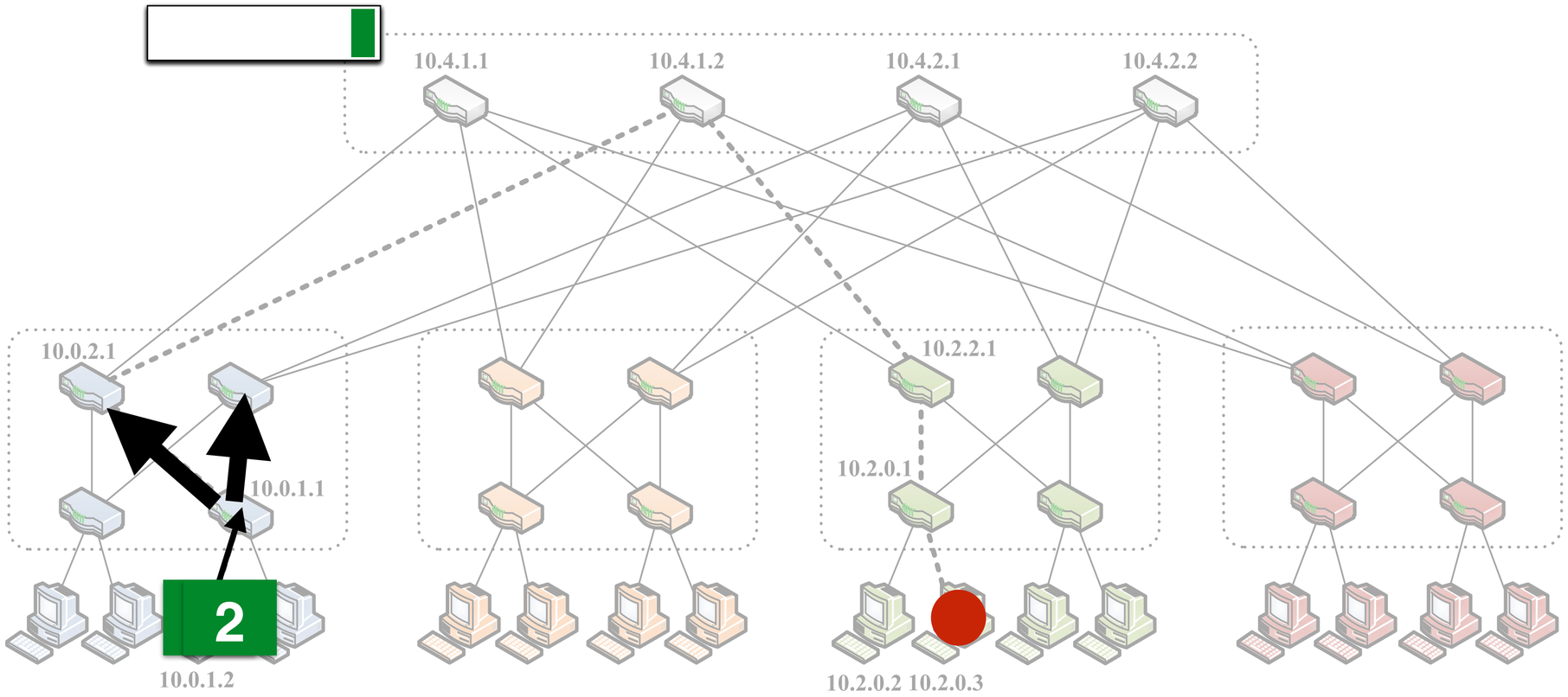
Equal cost multi-path (ECMP)

What problems might you run into?

10.2.0.3	port 1
10.2.0.3	port 2

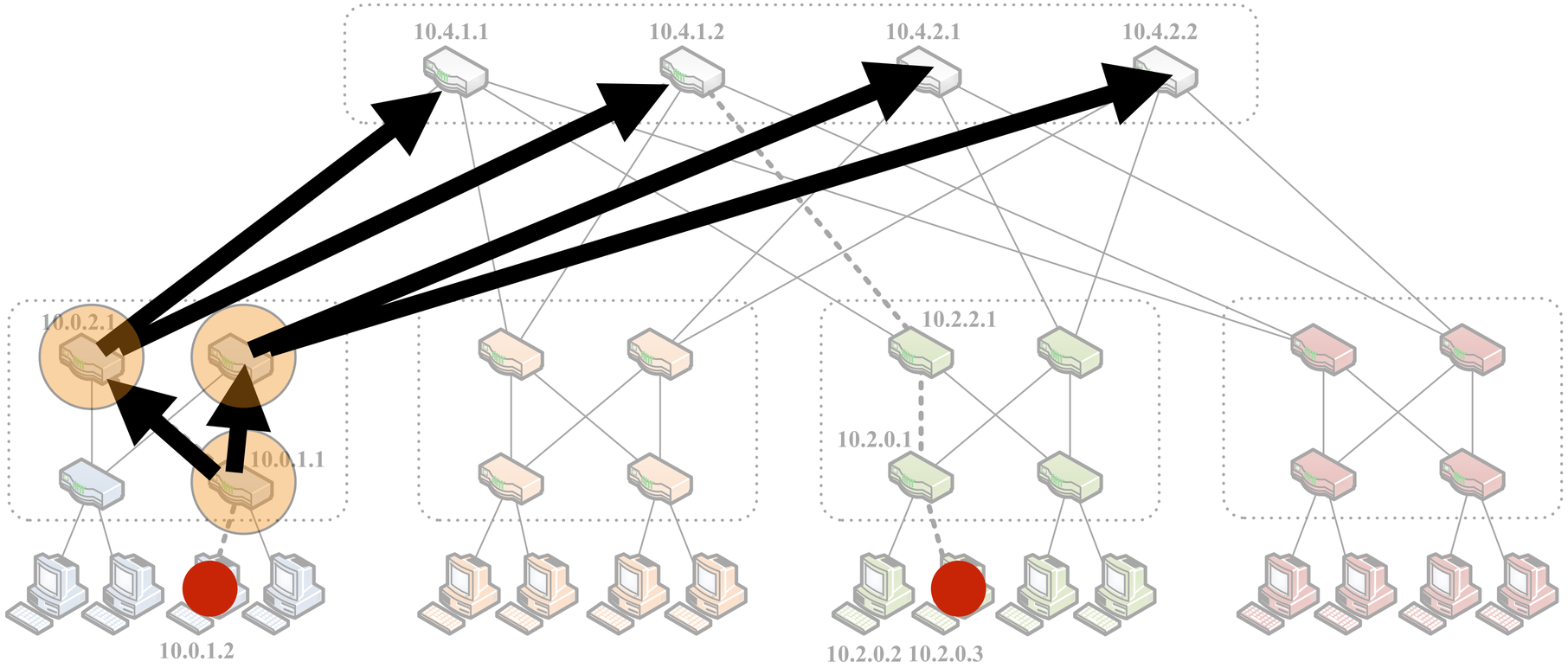
Choose uniformly at random!

Equal cost multi-path (ECMP)

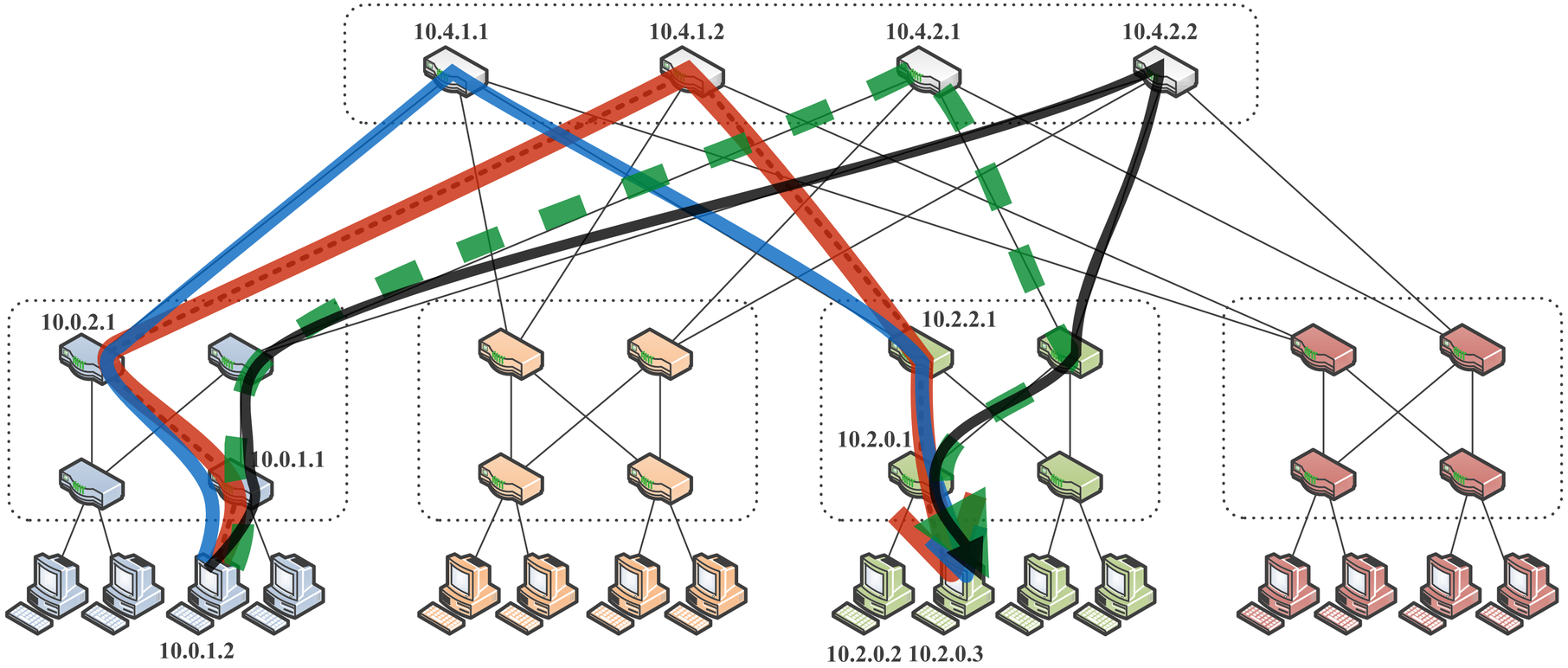


Output port = **hash** (packet header)

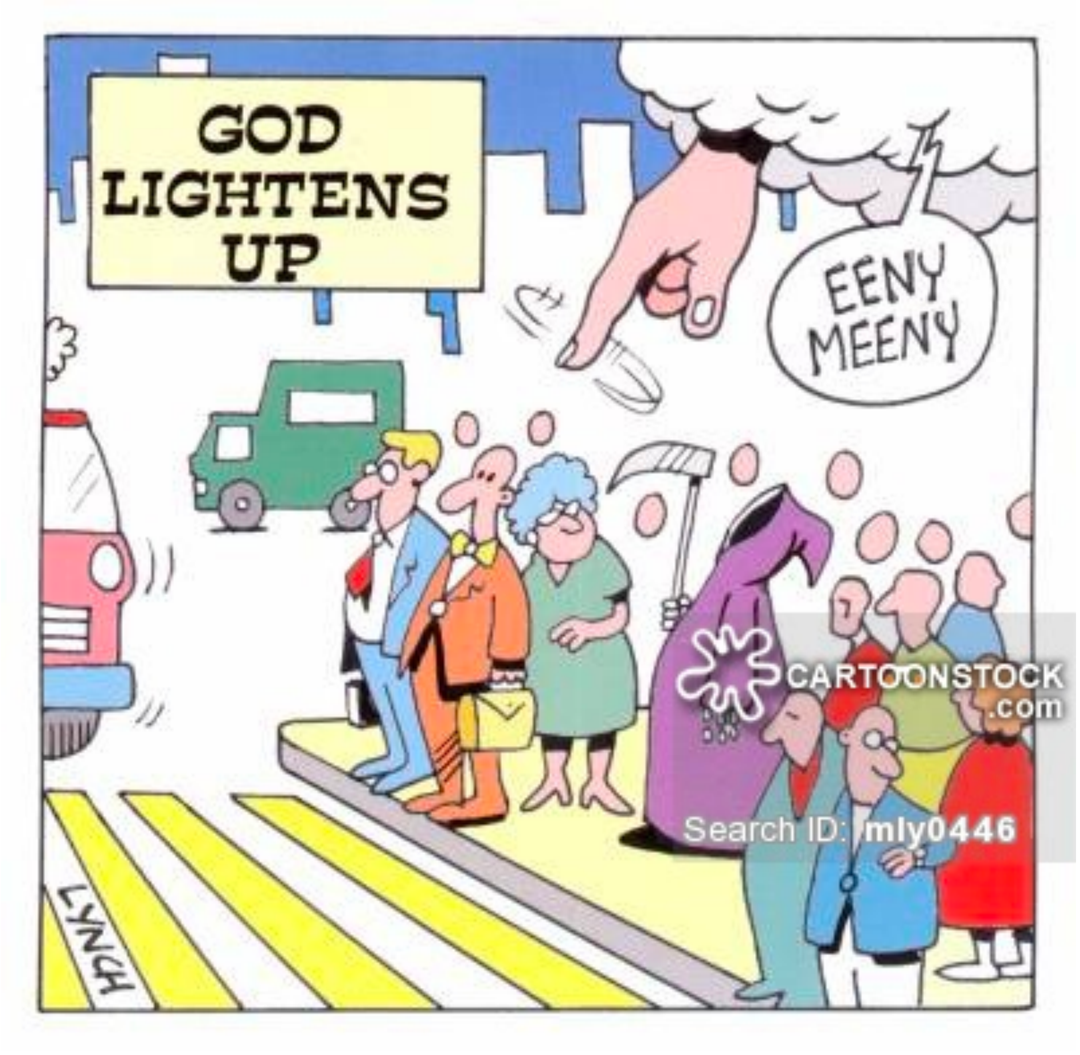
Equal cost multi-path (ECMP)



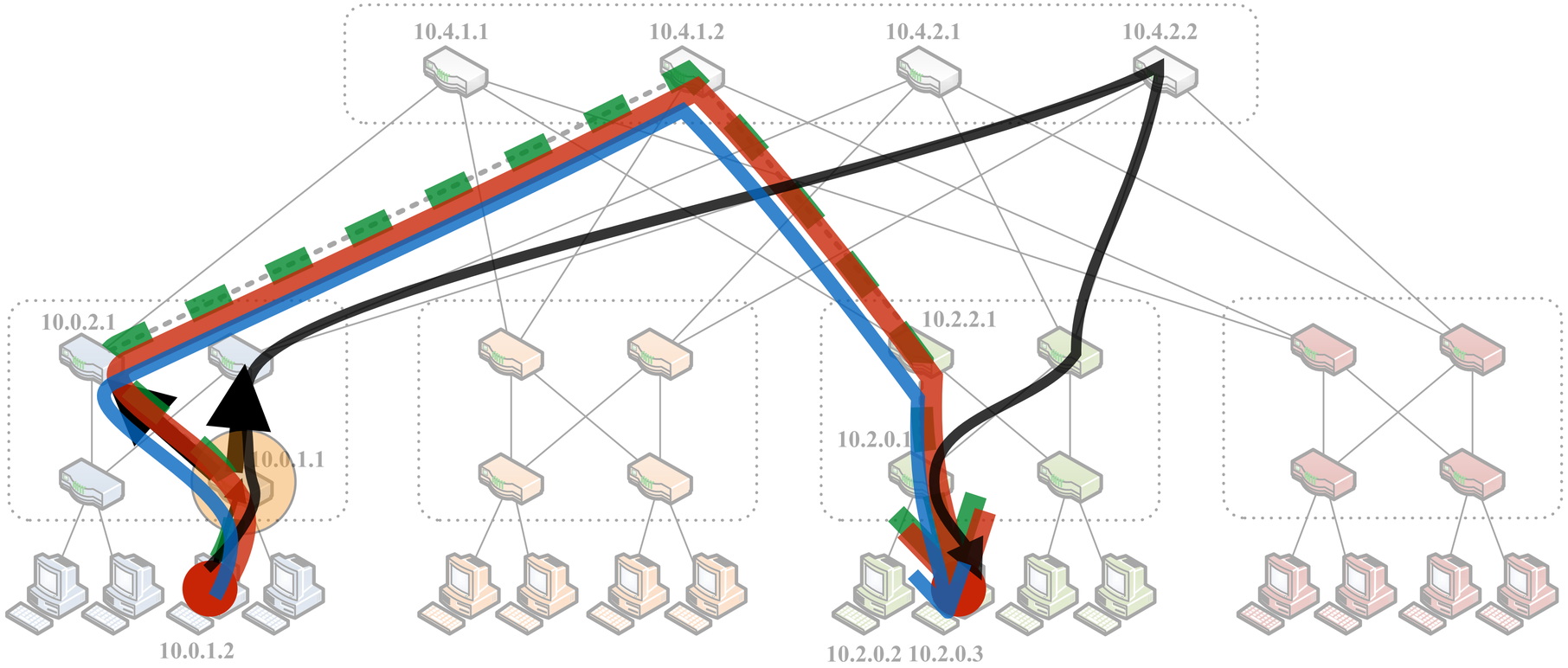
Equal cost multi-path (ECMP)



Randomization strikes again?!

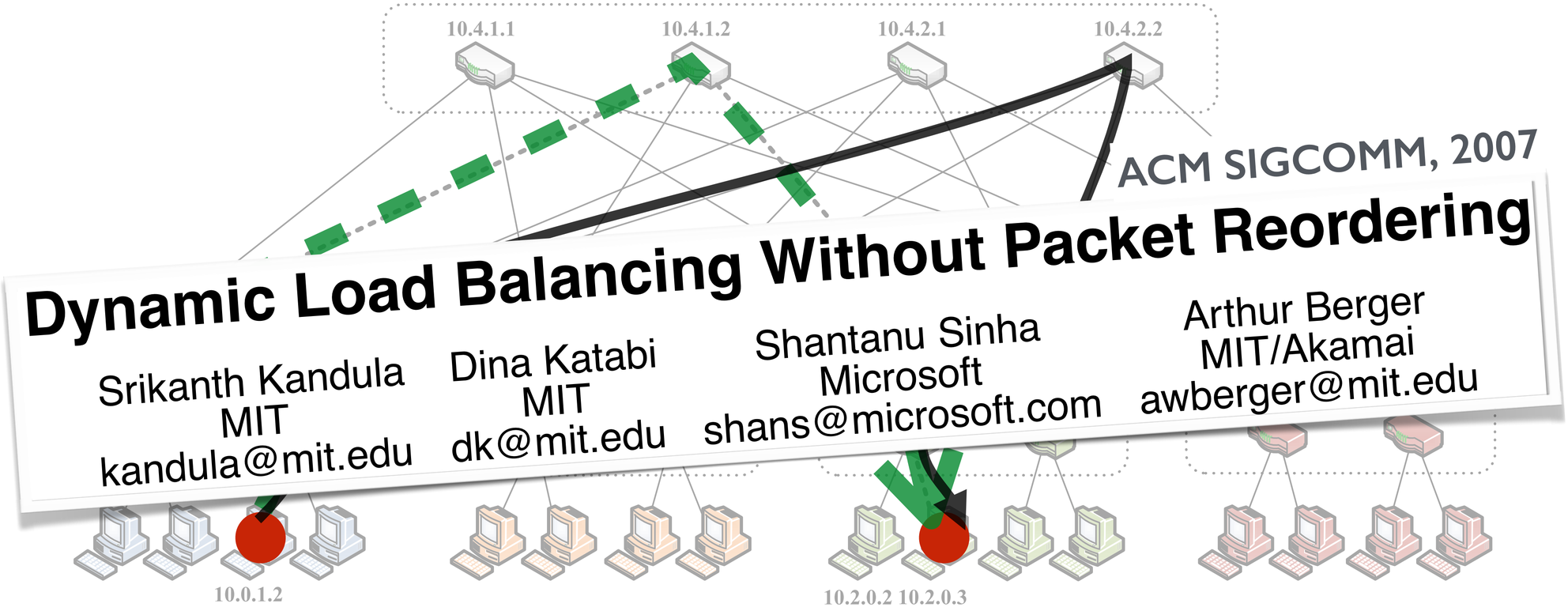


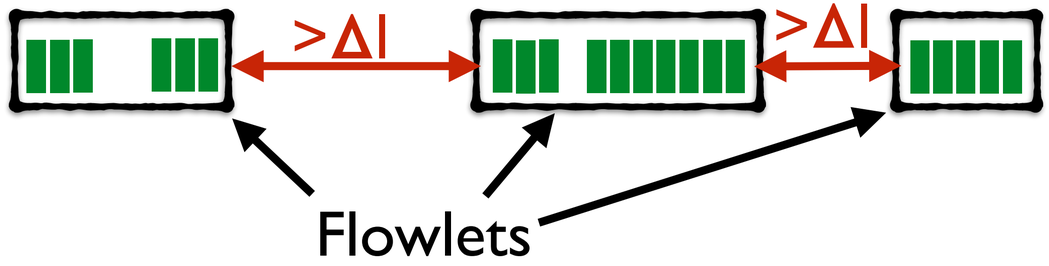
ECMP: traffic imbalance



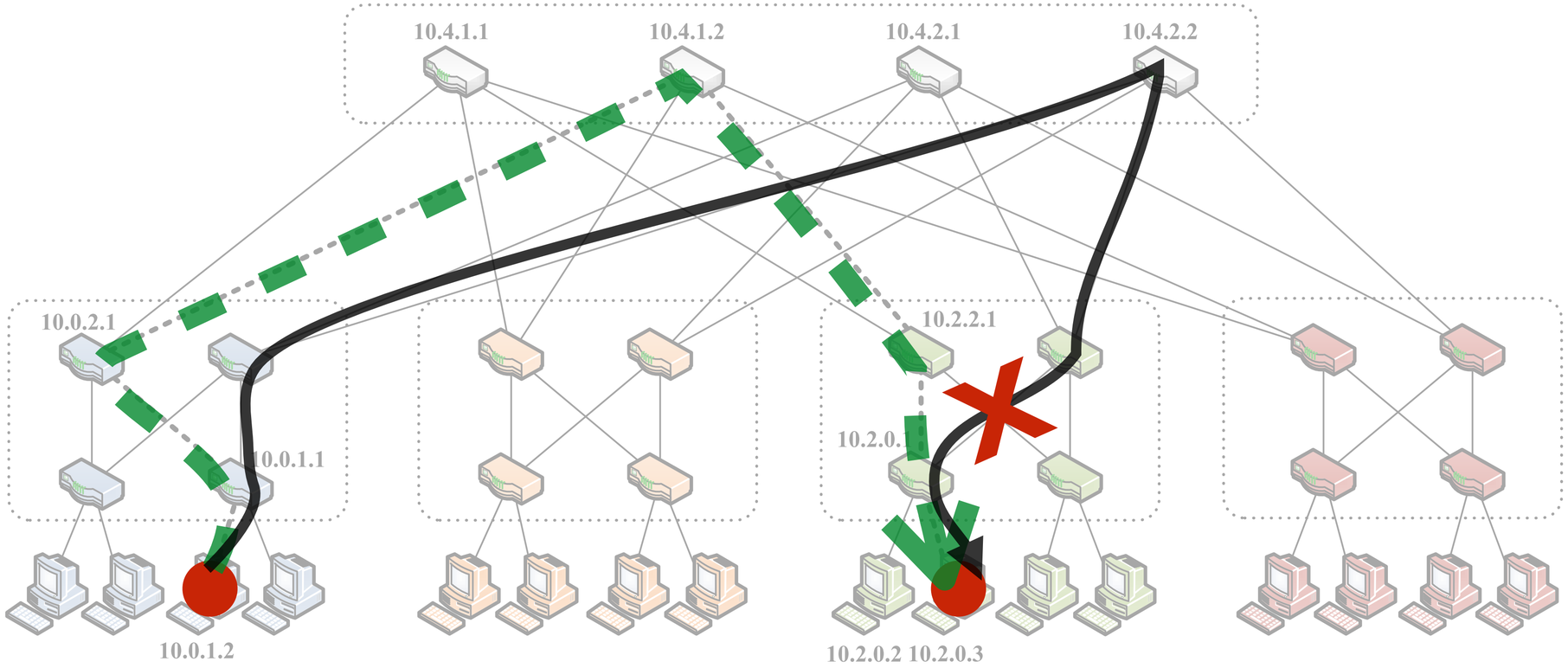
Output port = **hash** (packet header)

Flowlets


10.4.1.1
10.4.1.2
10.4.2.1
10.4.2.2
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
10.0.1.2
10.2.0.2 10.2.0.3

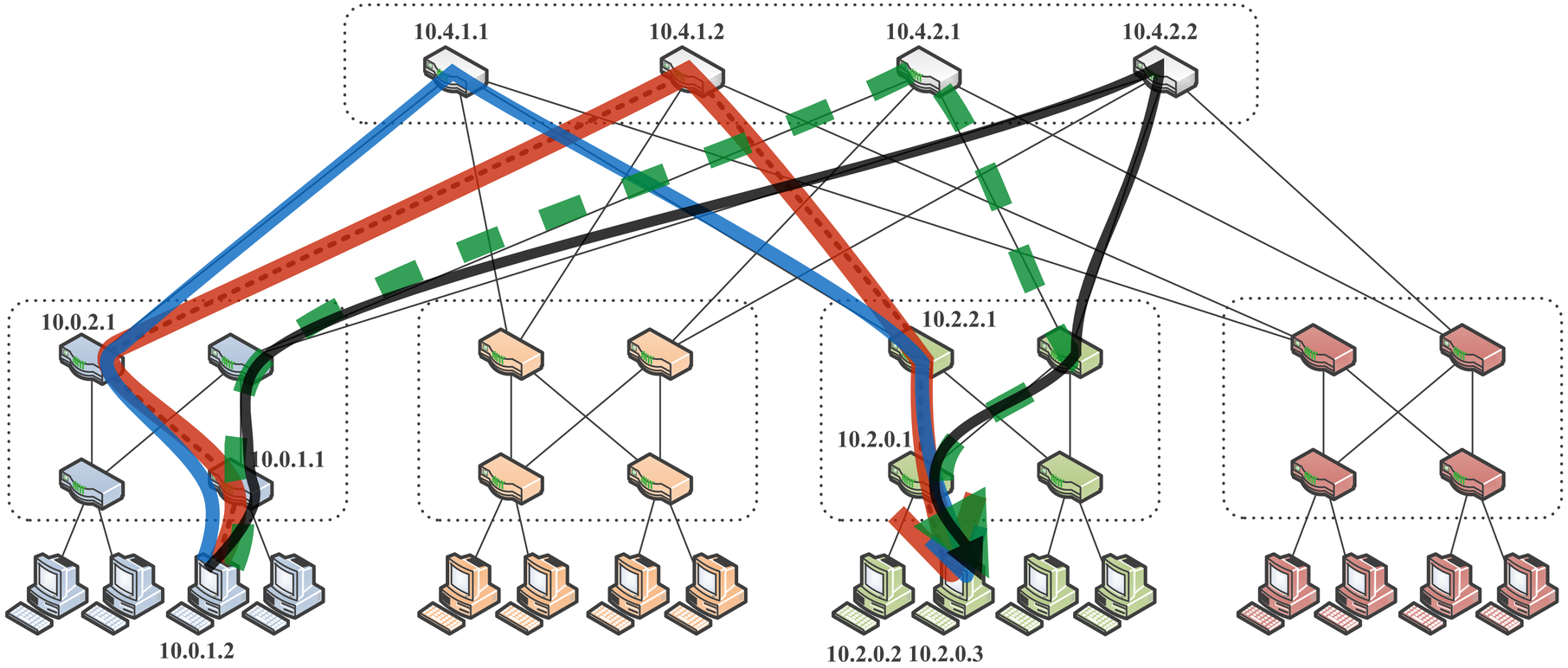

>Δ|
>Δ|
Flowlets

ECMP: local, oblivious choices



Cannot avoid distant failures!

Equal cost multi-path (ECMP)



Bonus: what if all switches use the same hash() ?

CONGA: edge based monitoring

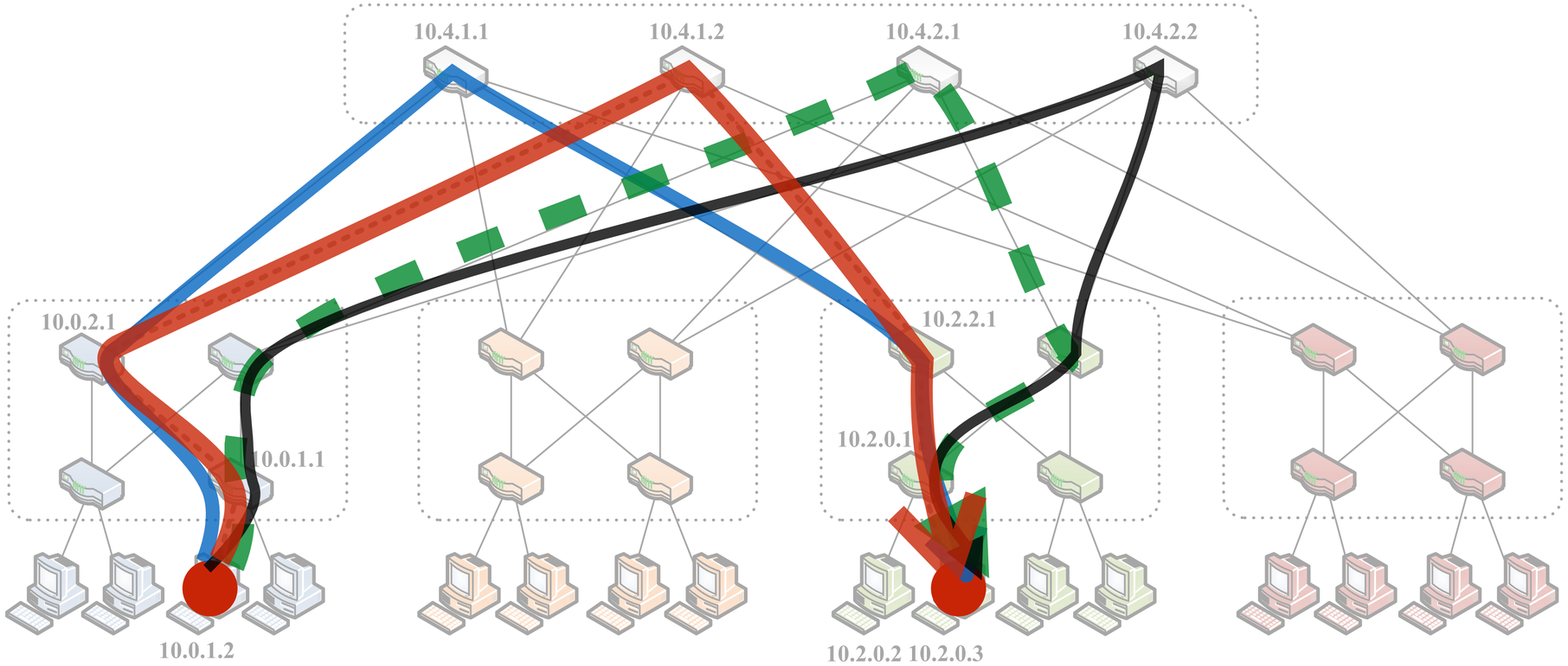
ACM SIGCOMM, 2014

CONGA: Distributed Congestion-Aware Load Balancing for Datacenters

Mohammad Alizadeh, Tom Edsall, Sarang Dharmapurikar, Ramanan Vaidyanathan, Kevin Chu, Andy Fingerhut, Vinh The Lam (Google), Francis Matus, Rong Pan, Navindra Yadav, George Varghese (Microsoft)

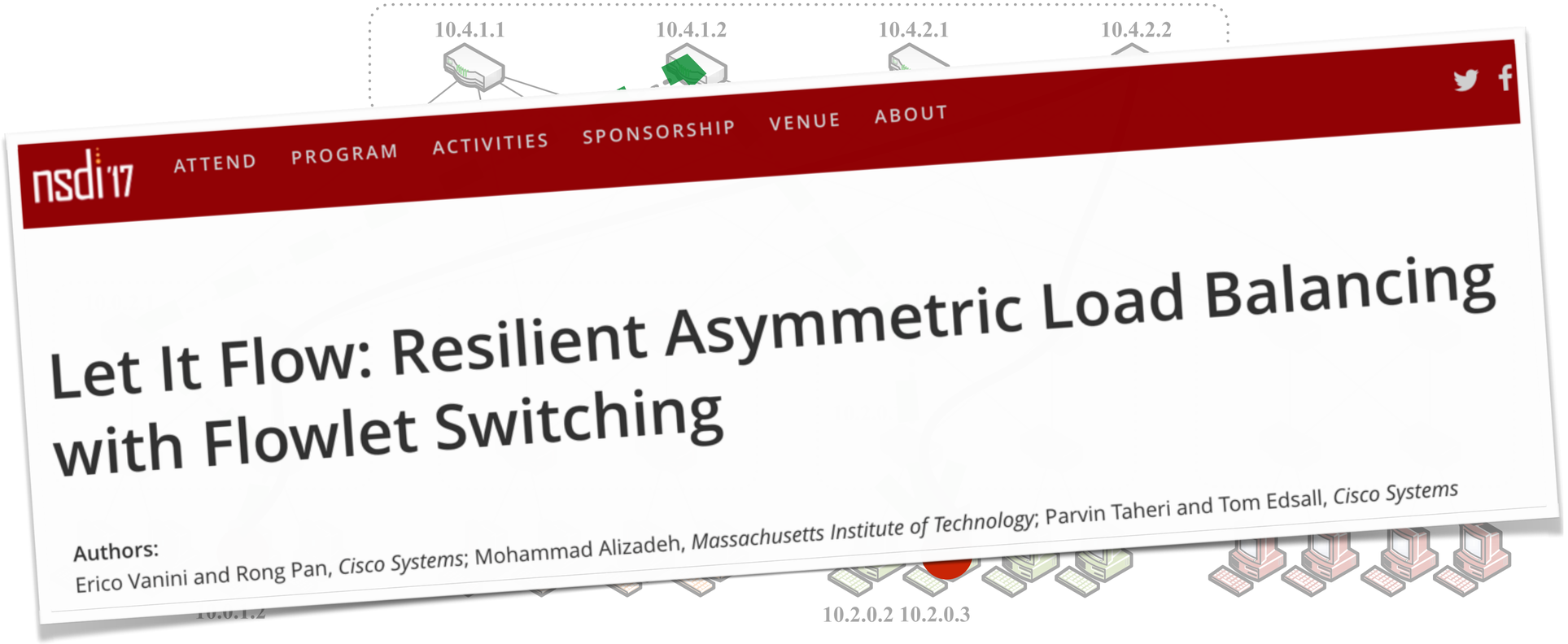
Cisco Systems

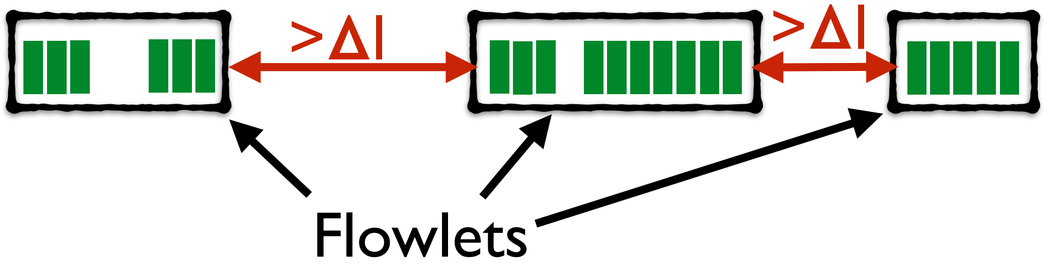
CONGA: edge based monitoring



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

Flowlets


10.4.1.1
10.4.1.2
10.4.2.1
10.4.2.2
nsdi'17
ATTEND
PROGRAM
ACTIVITIES
SPONSORSHIP
VENUE
ABOUT
Let It Flow: Resilient Asymmetric Load Balancing with Flowlet Switching
Authors:
Erico Vanini and Rong Pan, Cisco Systems; Mohammad Alizadeh, Massachusetts Institute of Technology; Parvin Taheri and Tom Edsall, Cisco Systems
10.2.0.2 10.2.0.3


>Δl
>Δl
Flowlets

Other possibilities?

Great freedom and resources!

Next time: SDN

The Problem

Networks are complicated

- Just like any computer system
- Worse: it's distributed
- Even worse: no clean programming APIs, only "knobs and dials"

Network equipment is proprietary

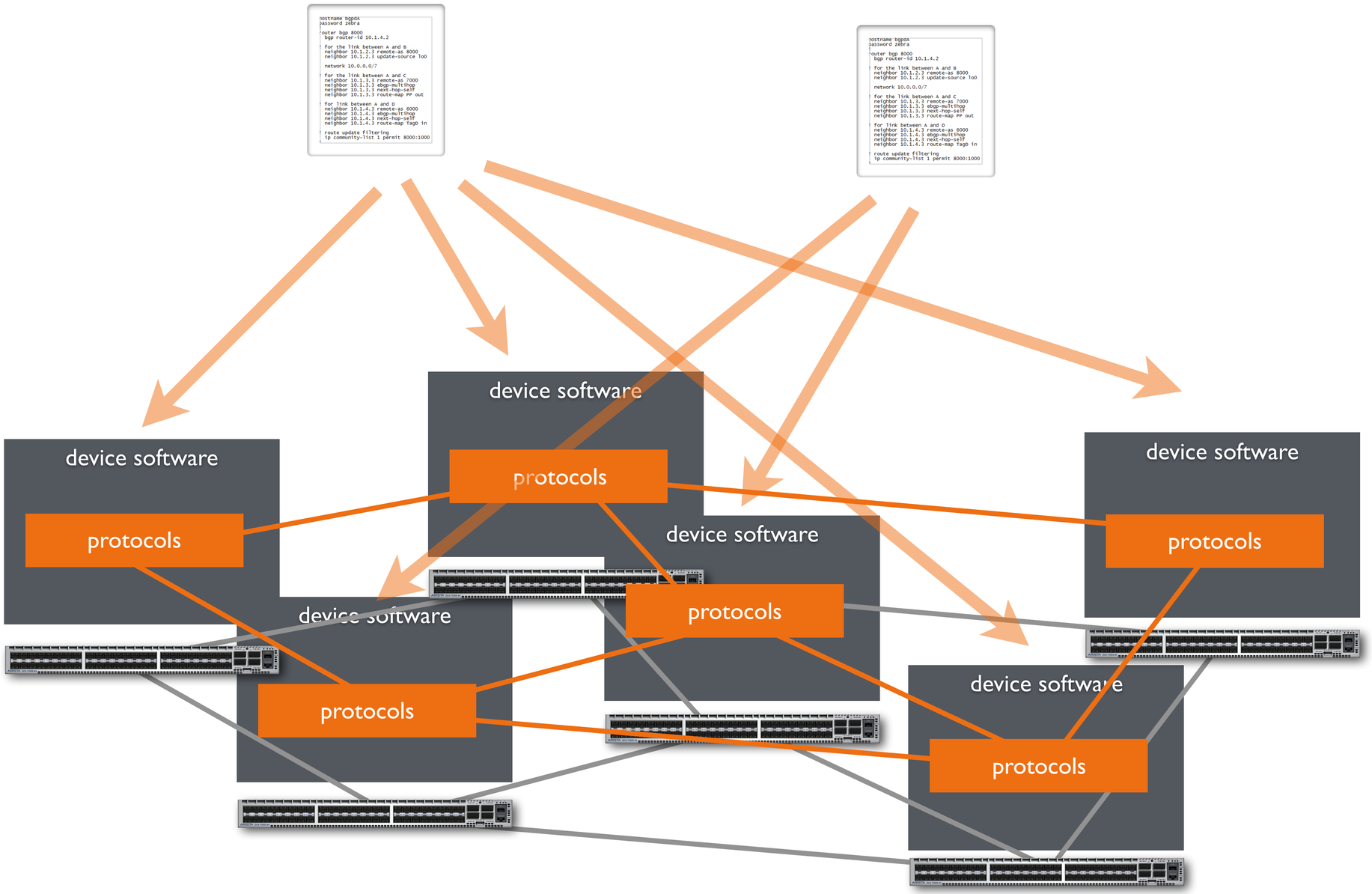
- Integrated solutions (software, configuration, protocol implementations, hardware) from major vendors

Result: Hard to innovate and modify networks

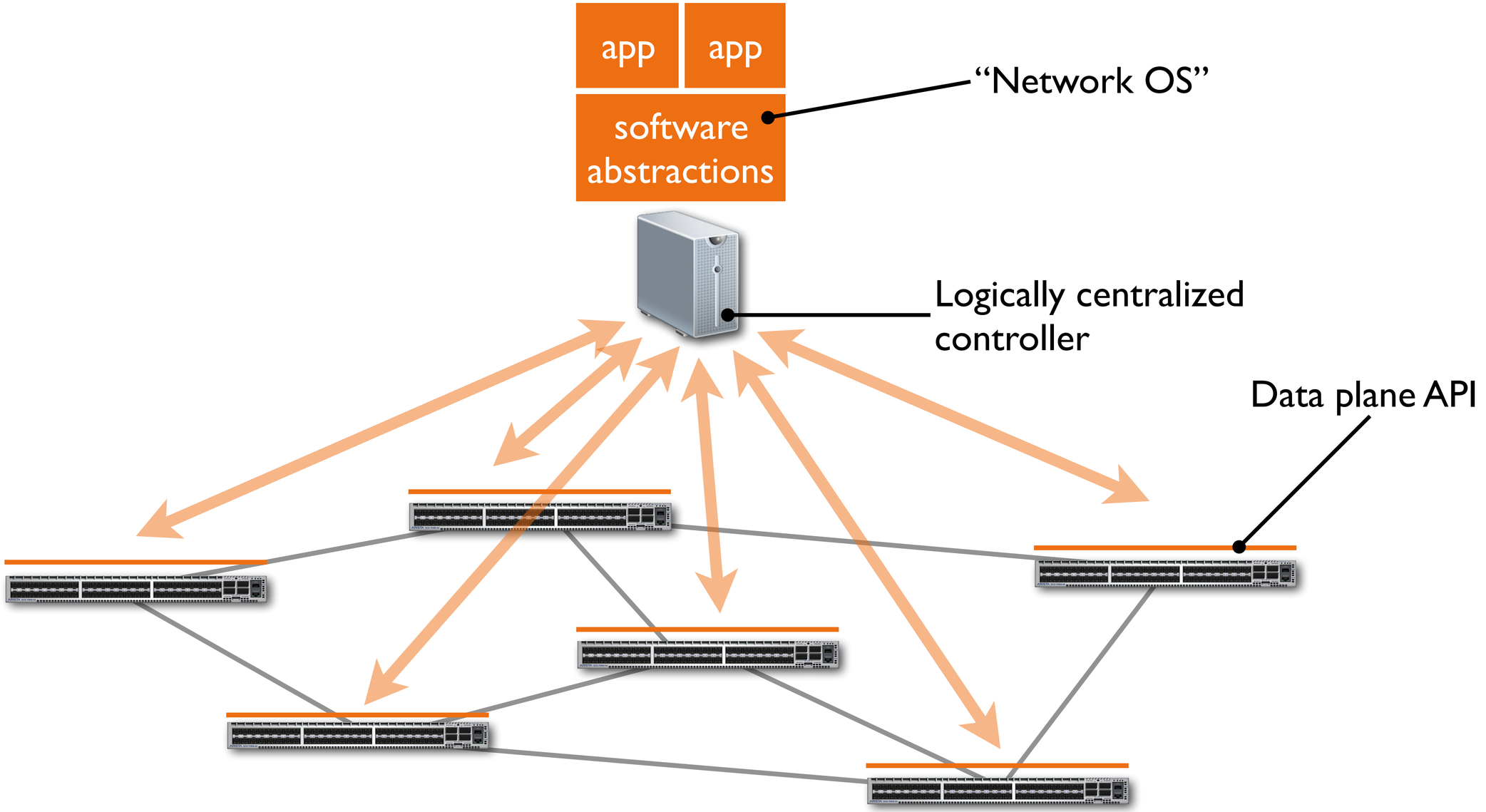
Traditional network

```
hostname bgpdA
password zebra
!
router bgp 8000
  bgp router-id 10.1.4.2

! for the link between A and B
  neighbor 10.1.2.3 remote-as 8000
  neighbor 10.1.2.3 update-source lo0

  network 10.0.0.0/7

! for the link between A and C
  neighbor 10.1.3.3 remote-as 7000
  neighbor 10.1.3.3 ebgp-multihop
  neighbor 10.1.3.3 next-hop-self
  neighbor 10.1.3.3 route-map PP out

! for link between A and D
  neighbor 10.1.4.3 remote-as 6000
  neighbor 10.1.4.3 ebgp-multihop
  neighbor 10.1.4.3 next-hop-self
  neighbor 10.1.4.3 route-map TagD in

! route update filtering
  ip community-list 1 permit 8000:1000
!
```

Traditional network



Software-defined network



Weekly reading guide

Network programmability

ACM CCR, 2014

P4: Programming Protocol-Independent Packet Processors

Pat Bosshart[†], Dan Daly[*], Glen Gibb[†], Martin Izzard[†], Nick McKeown[‡], Jennifer Rexford[**], Cole Schlesinger[**], Dan Talayco[†], Amin Vahdat[¶], George Varghese[§], David Walker[**]
[†]Barefoot Networks [*]Intel [‡]Stanford University [**]Princeton University [¶]Google [§]Microsoft Research

- A new stage of the SDN trend
- Programmable switches!
- Hot topic now: papers, startups, industrial involvement