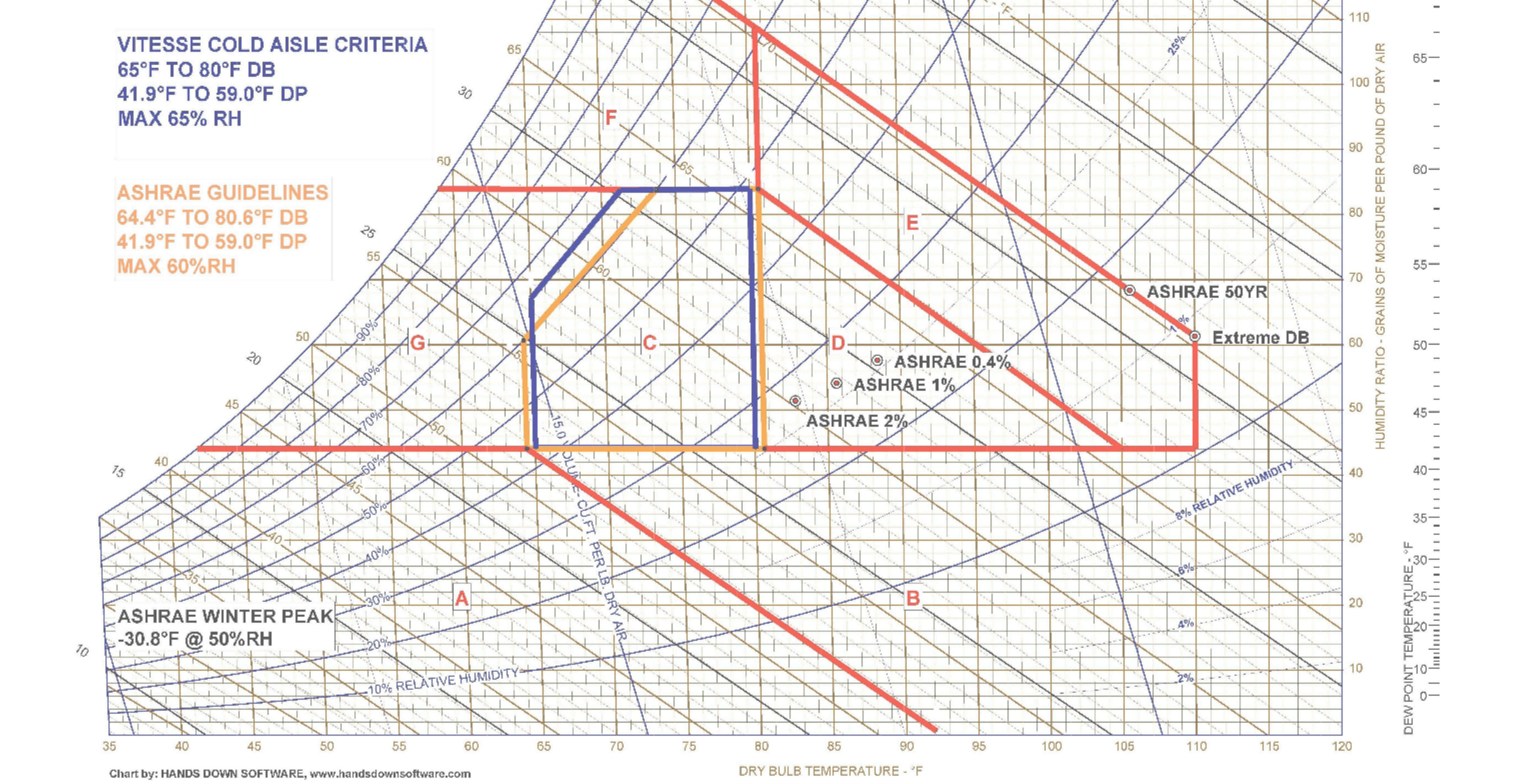
Data centers: network topology

Ankit Singla

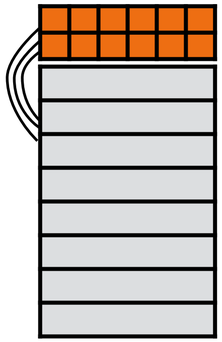
ETH Zürich Spring 2017



[Data Center v1.0, Open Compute Project

[Image: Trower, NASA]

A server rack

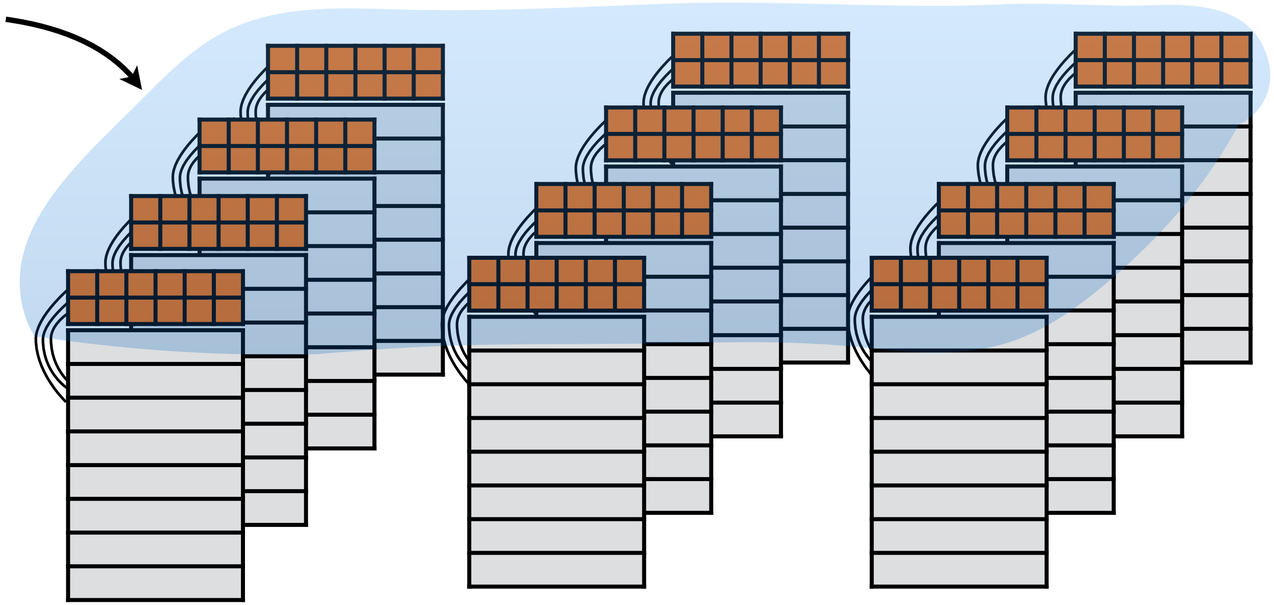


A top-of-rack switch

A rack of servers

Lots of racks

How to network the racks?

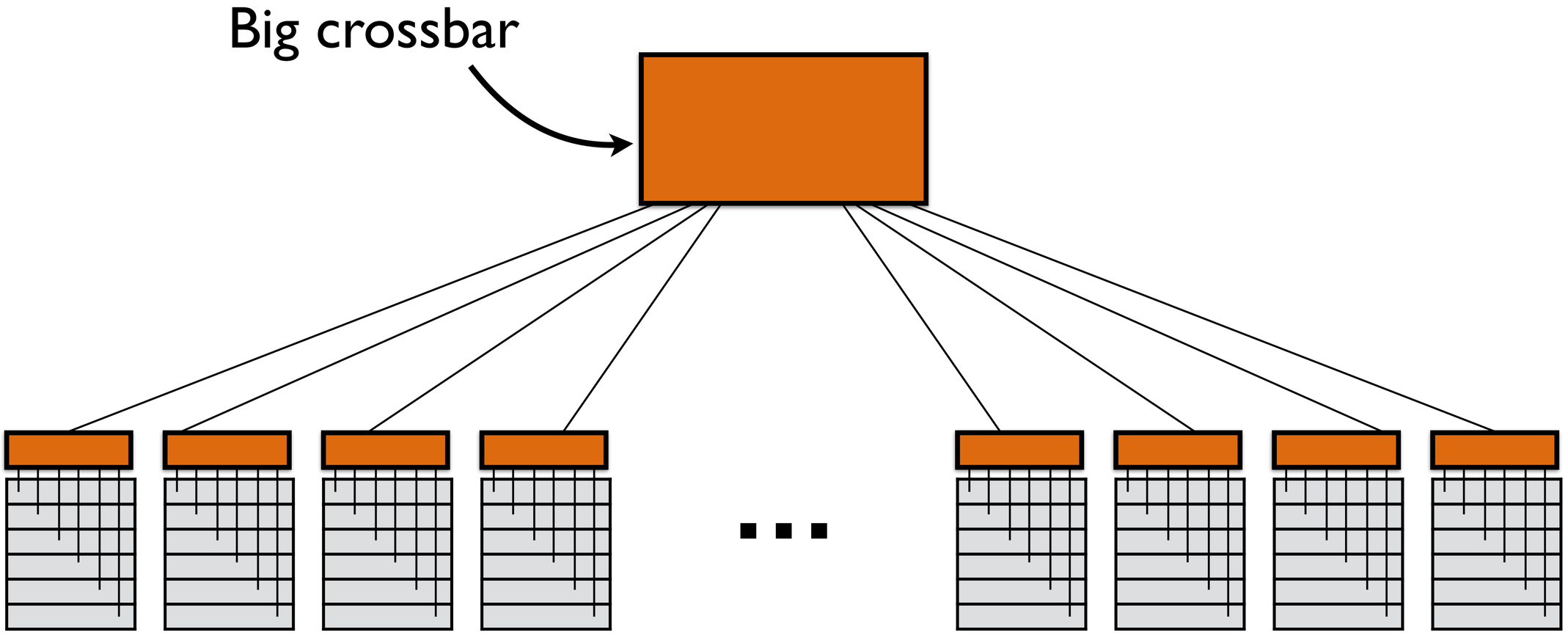


Facebook: machine-machine traffic “doubling at an interval of less than a year”

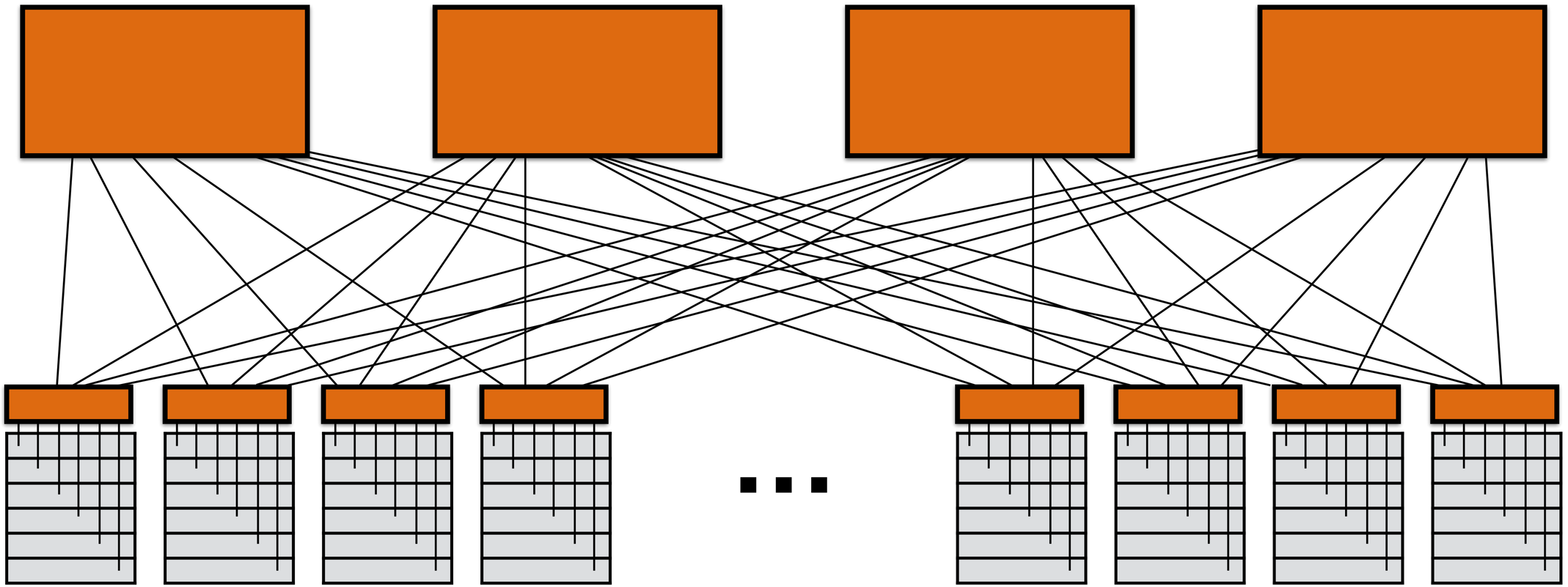
Need high throughput networks to ...

- Support big data analytics
- Ease virtual machine placement

“Big switch” approach

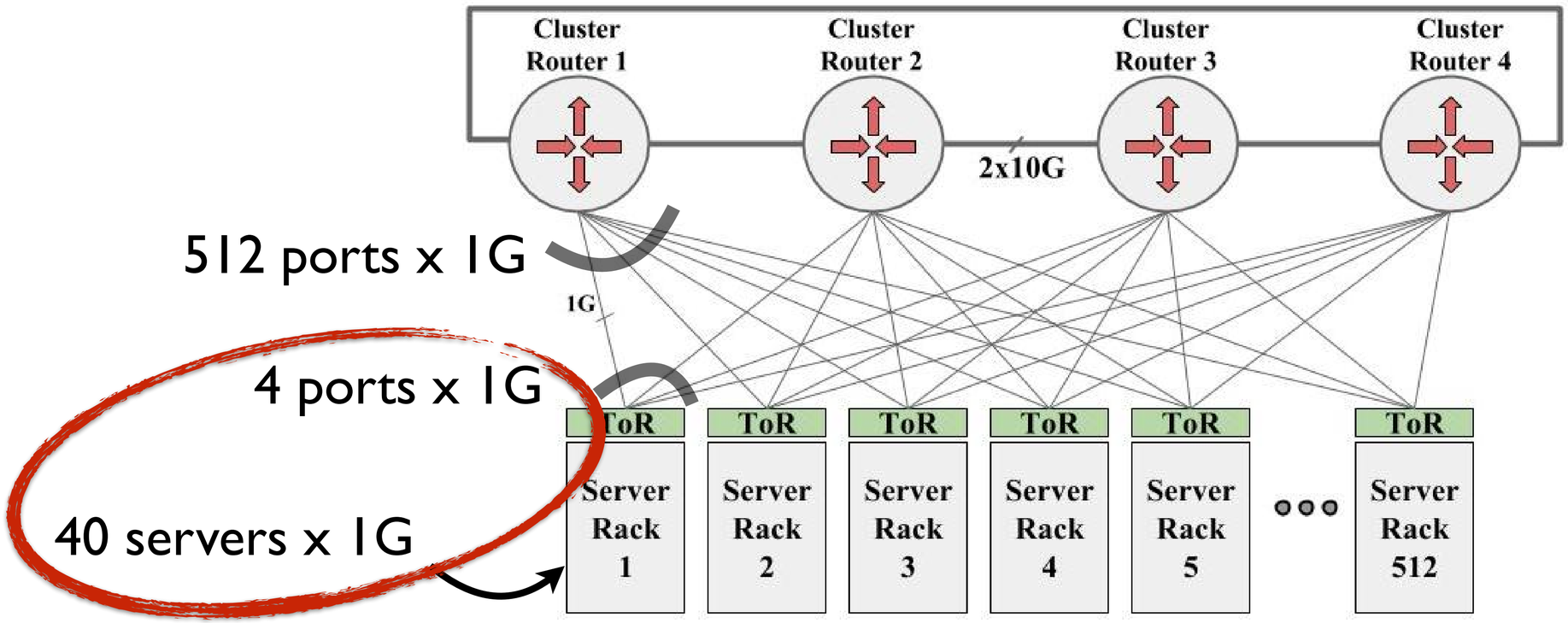


“Big switch” approach

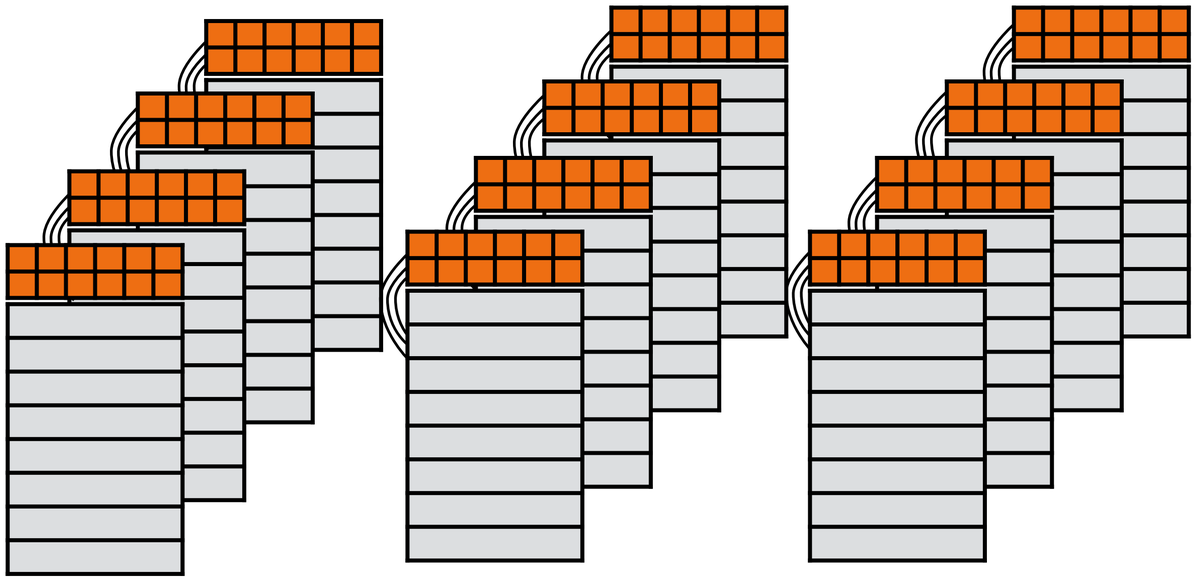


Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network

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.

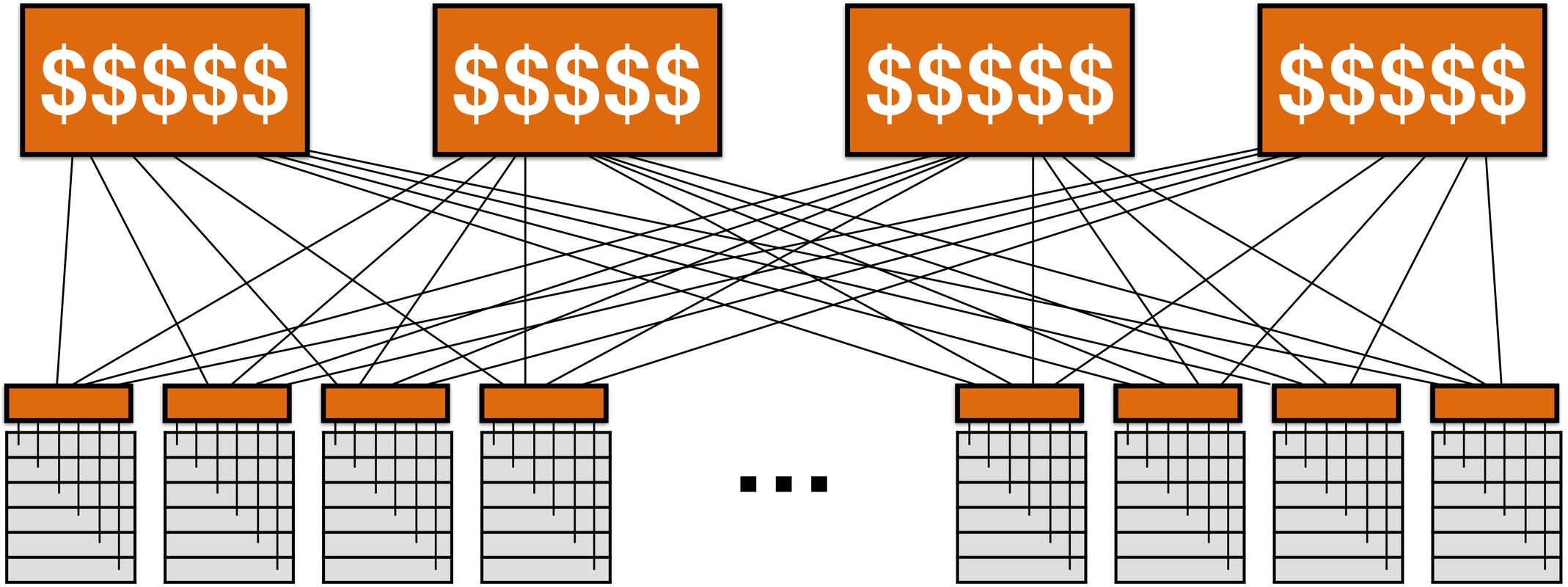


Side-note: workload management

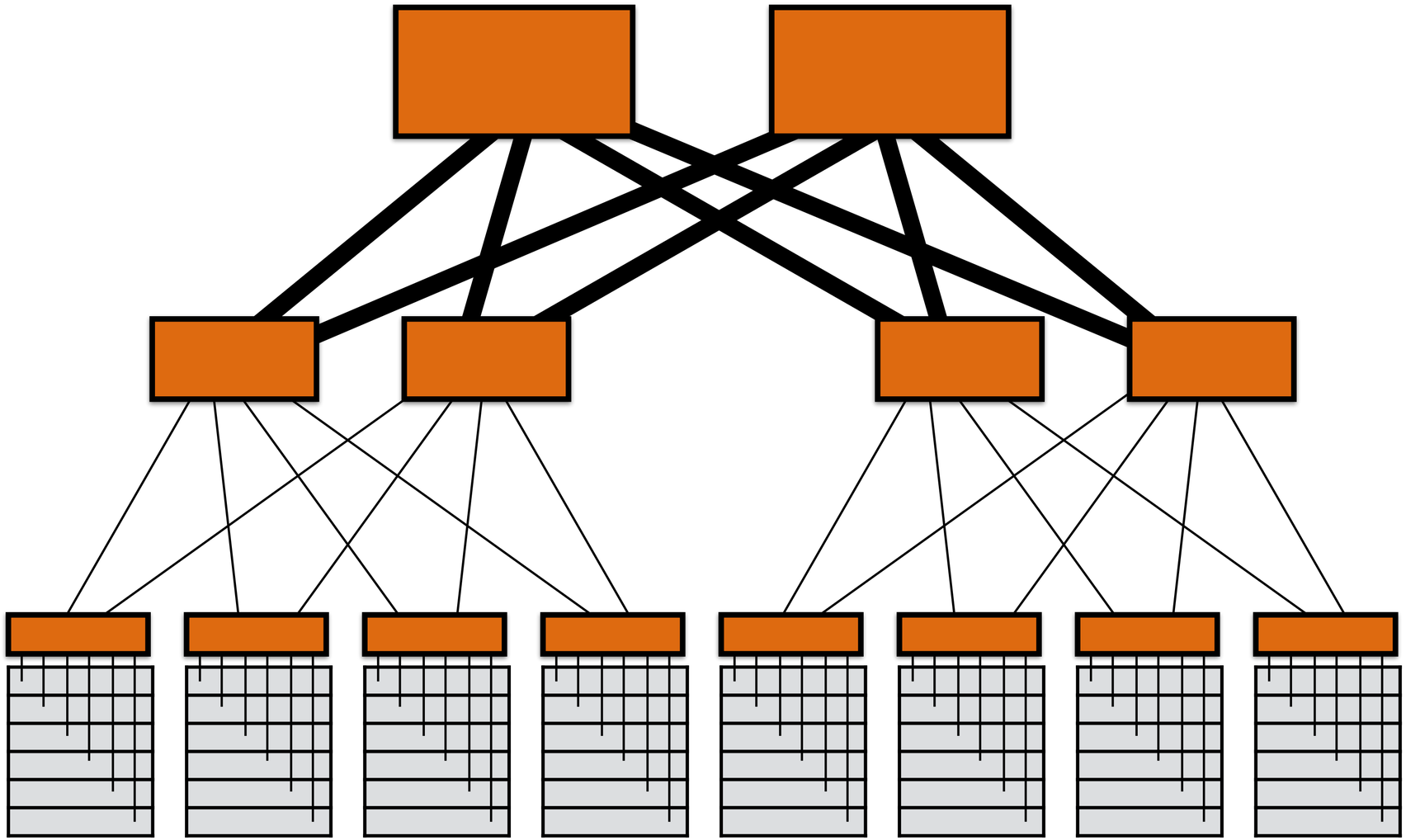


If your Map-Reduce job needs 24 servers, which ones would you take?

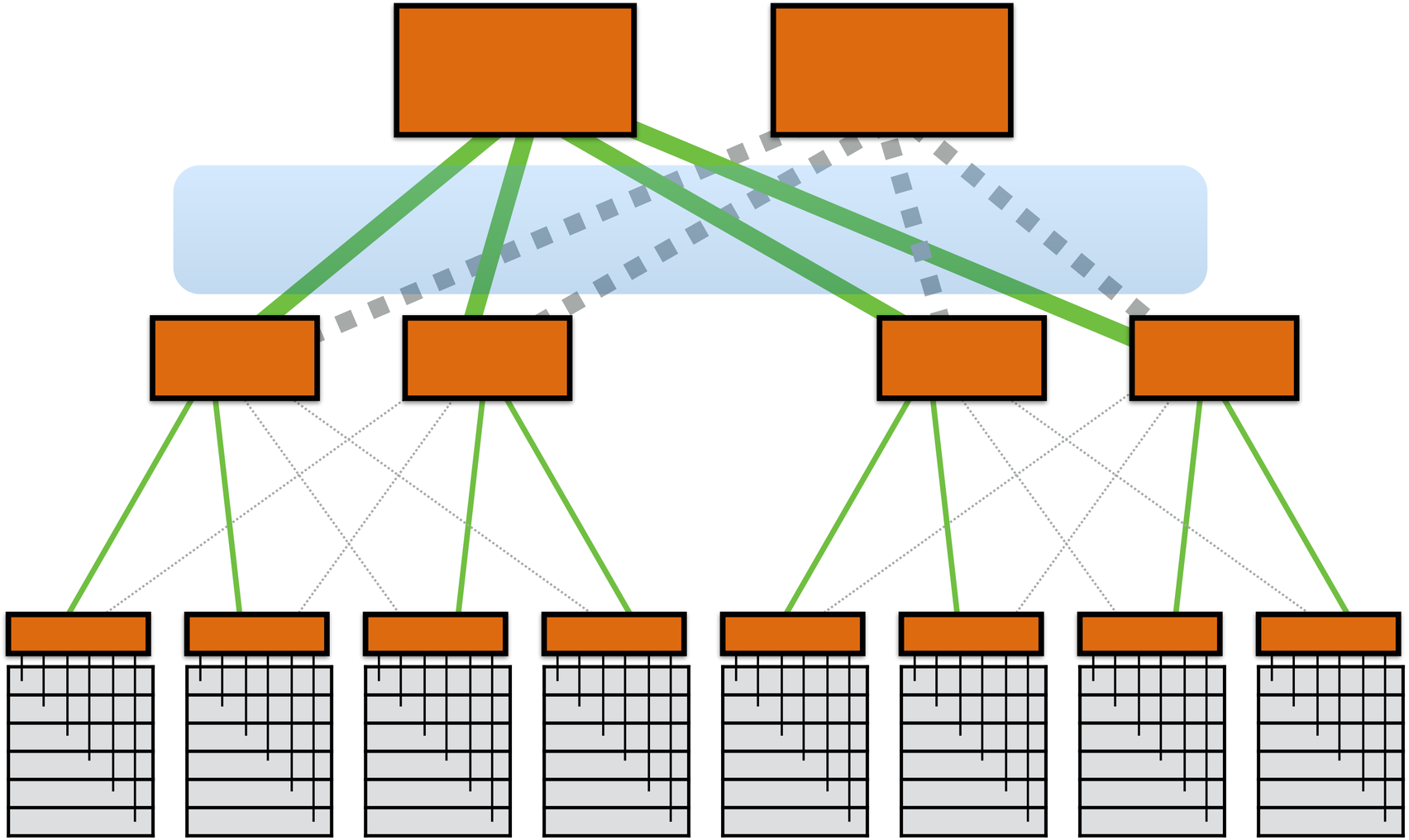
“Big switch” approach



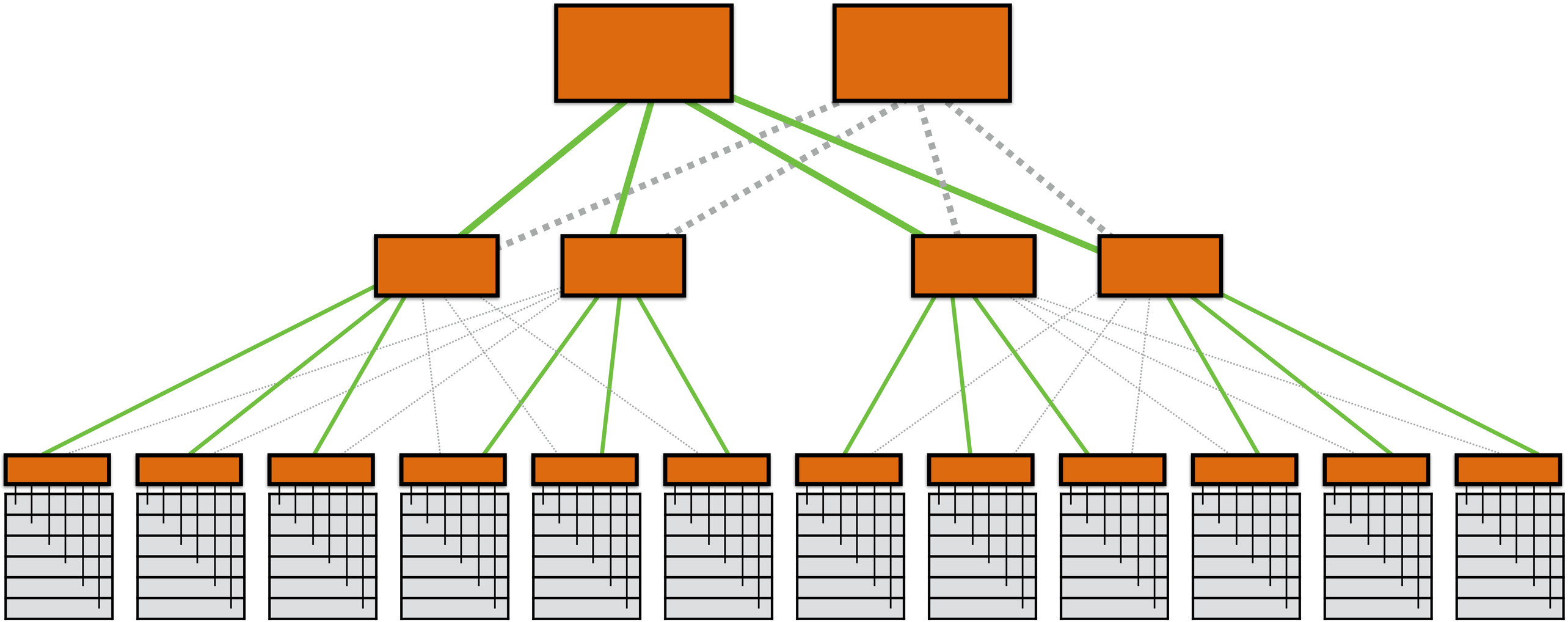
Alternative: tree network



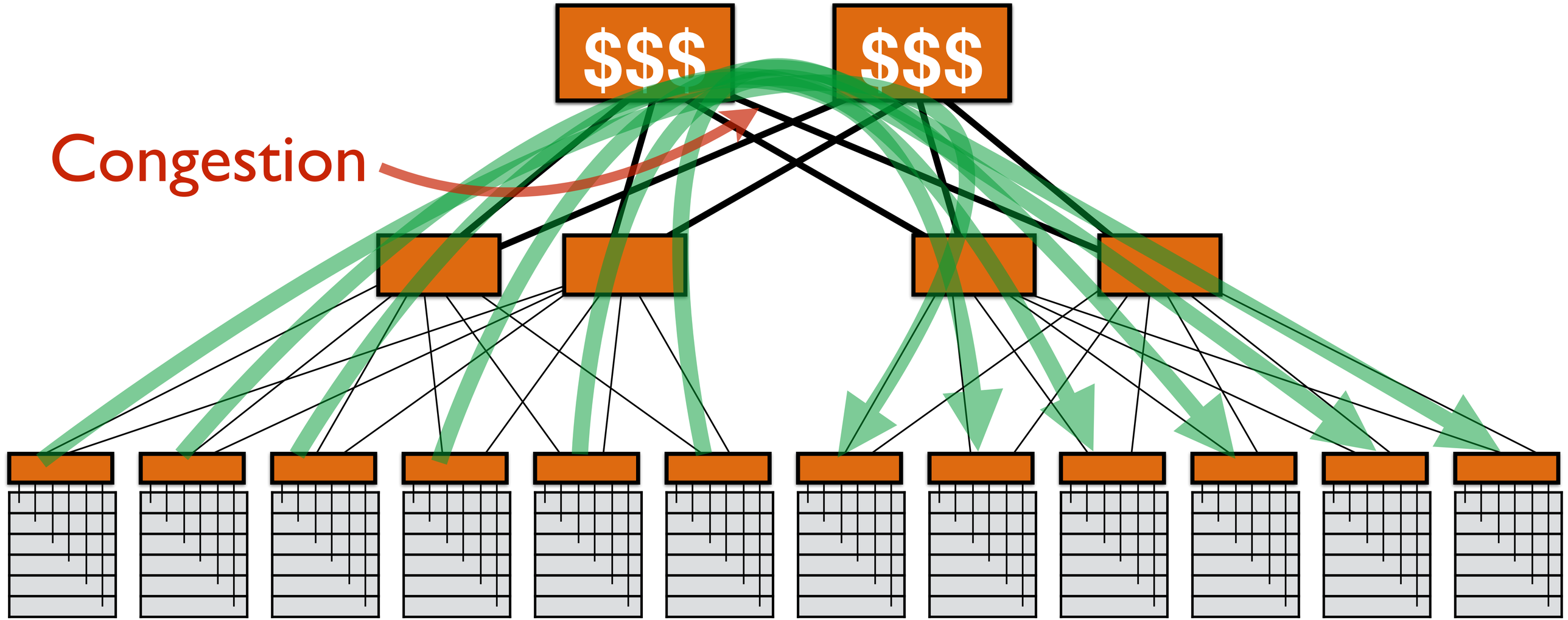
Alternative: tree network



Alternative: tree network



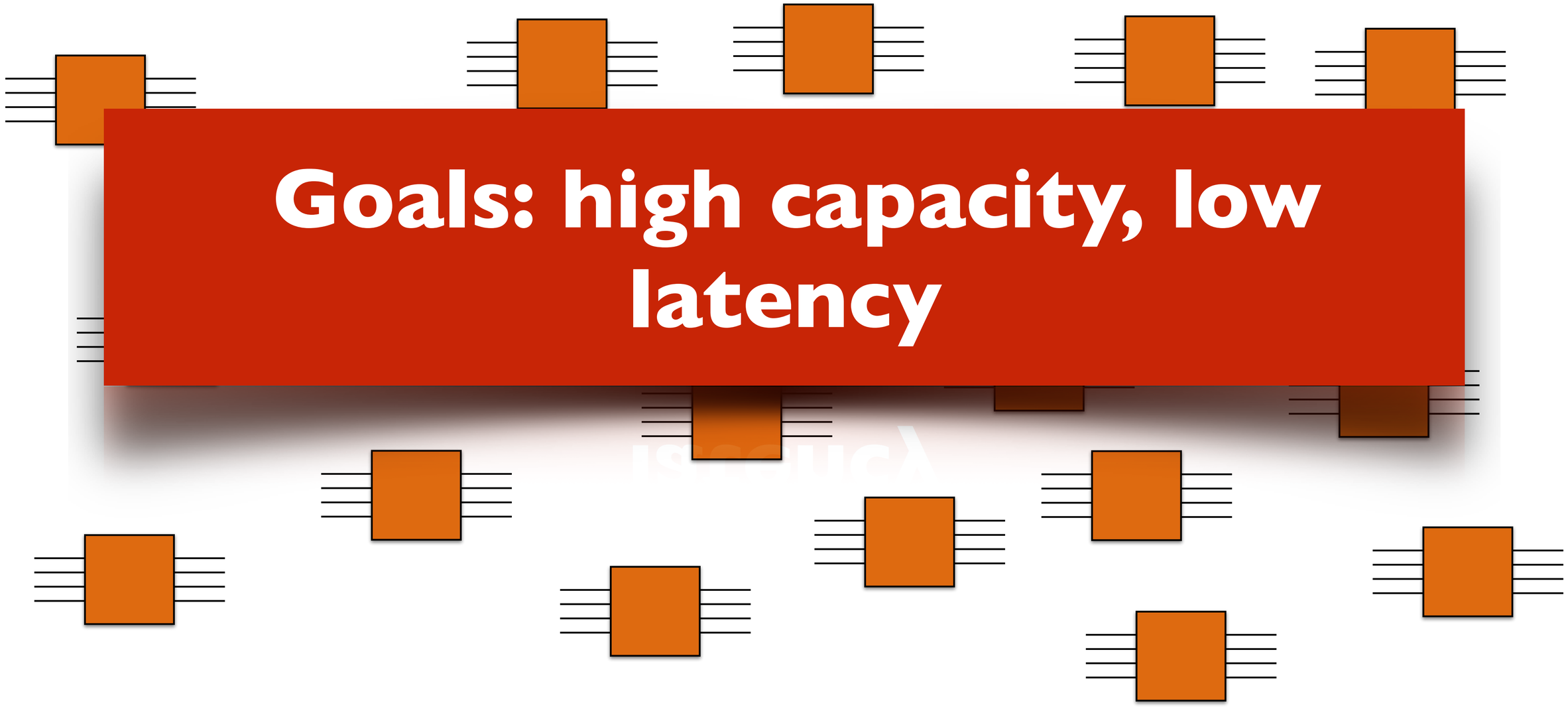
Alternative: tree network



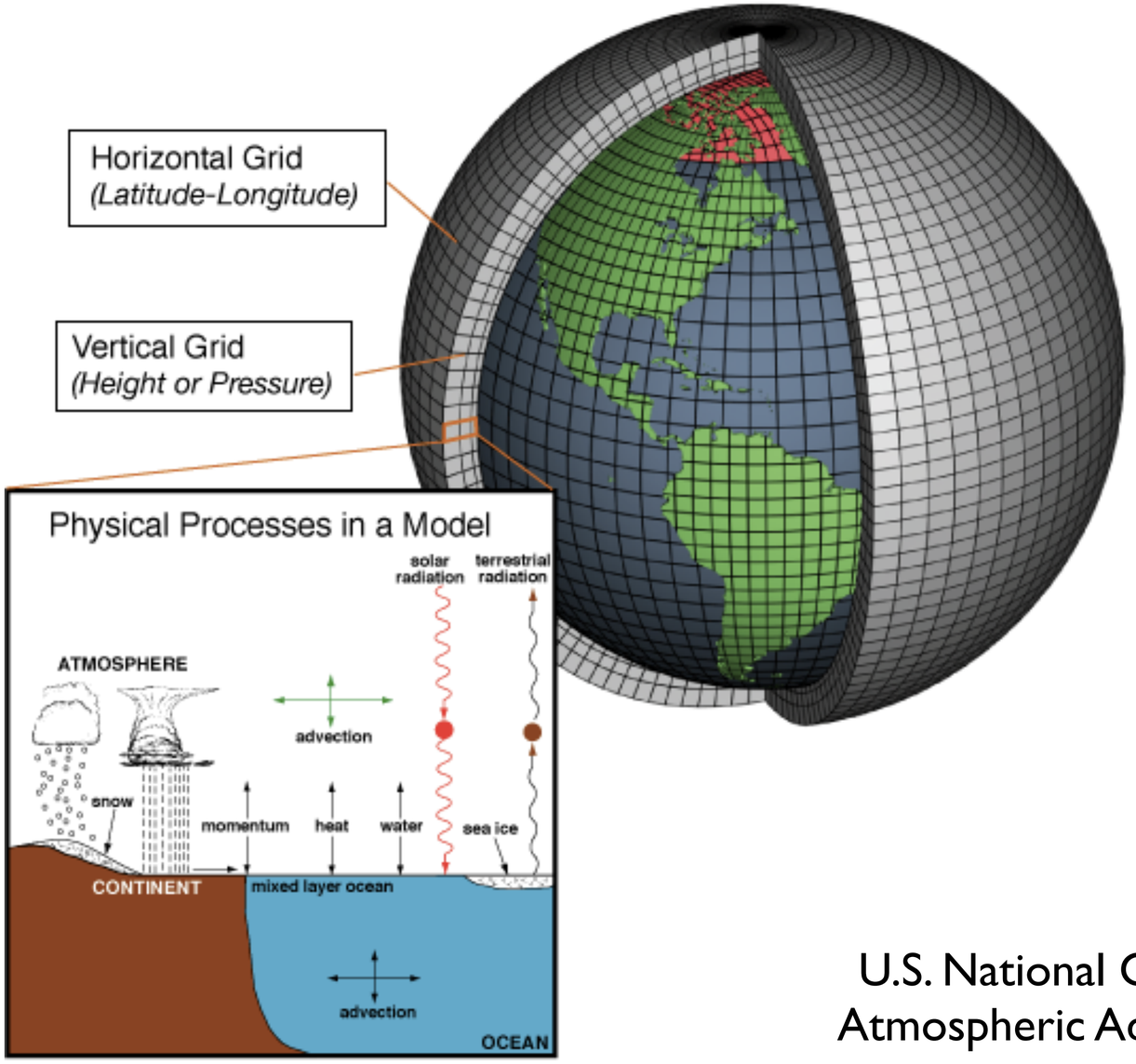
Build with identical switches throughout?

“Scaling out” vs. “scaling up”

Connect many cheap, identical switches?

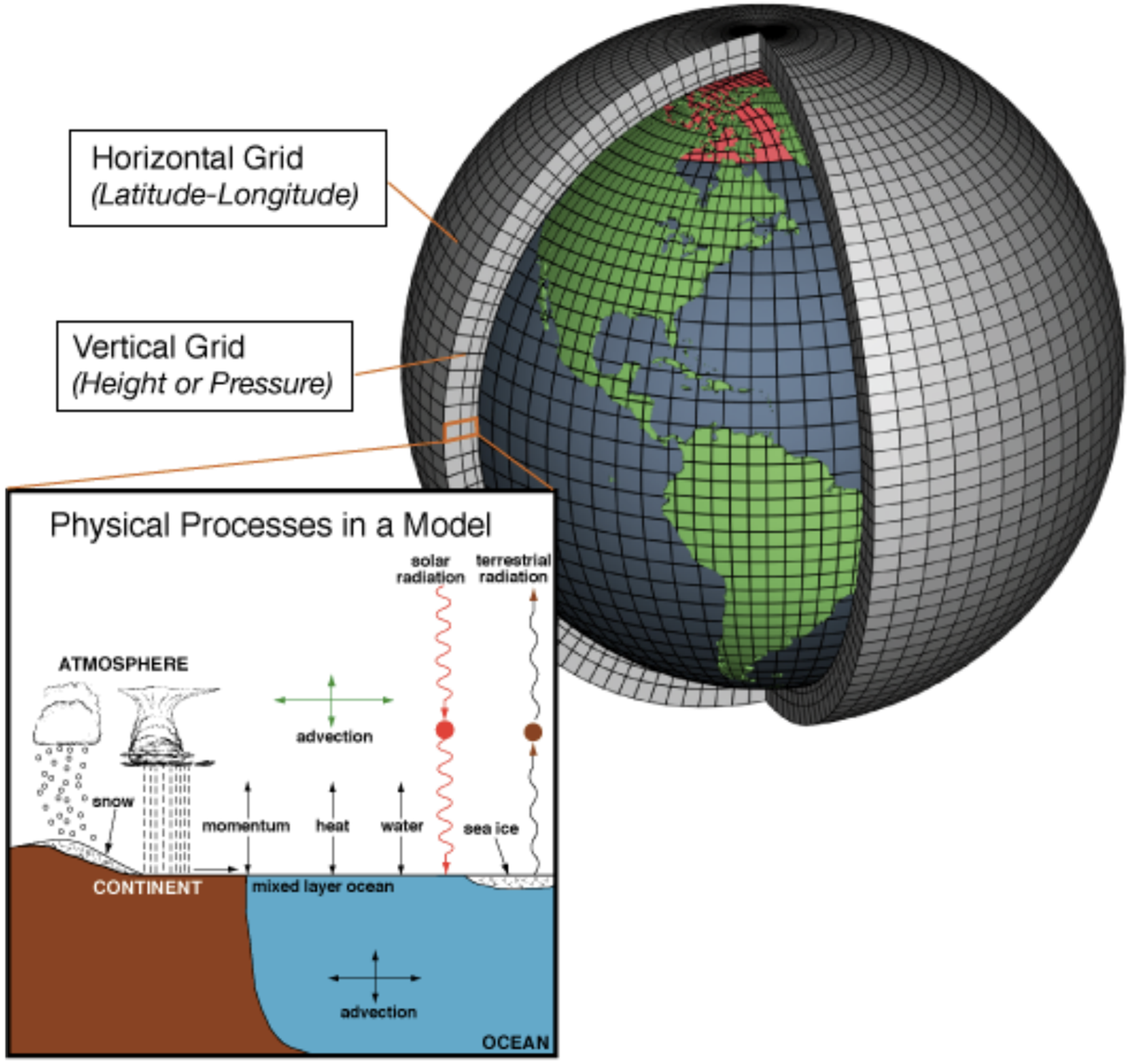


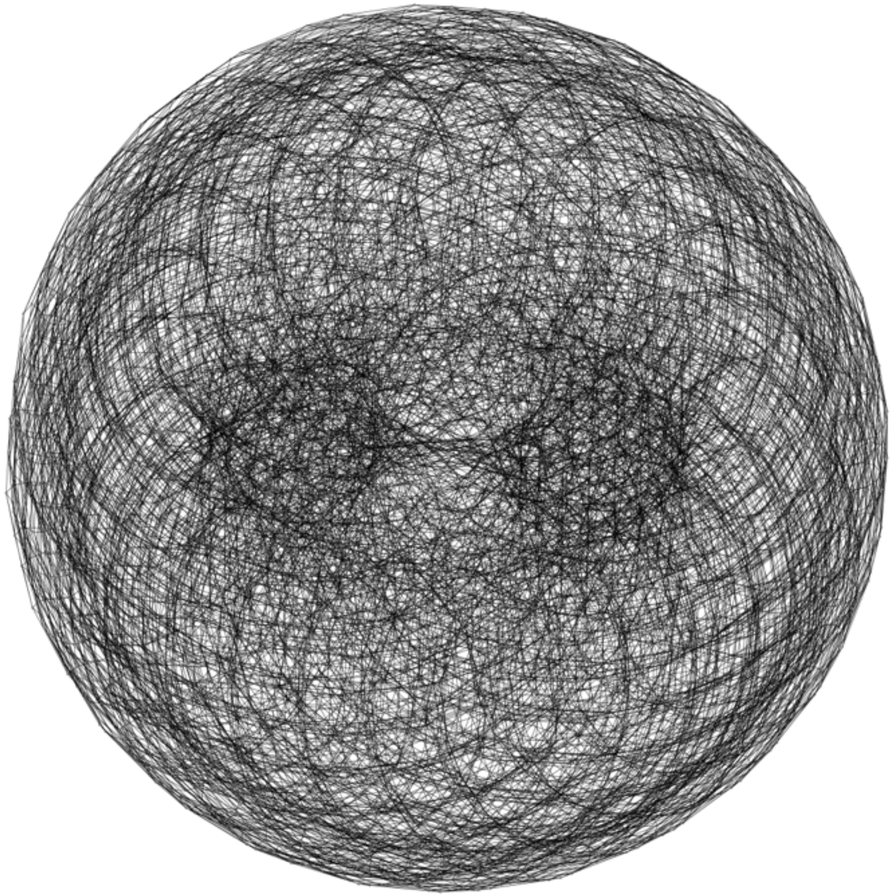
If you know your application …



U.S. National Oceanic and Atmospheric Administration

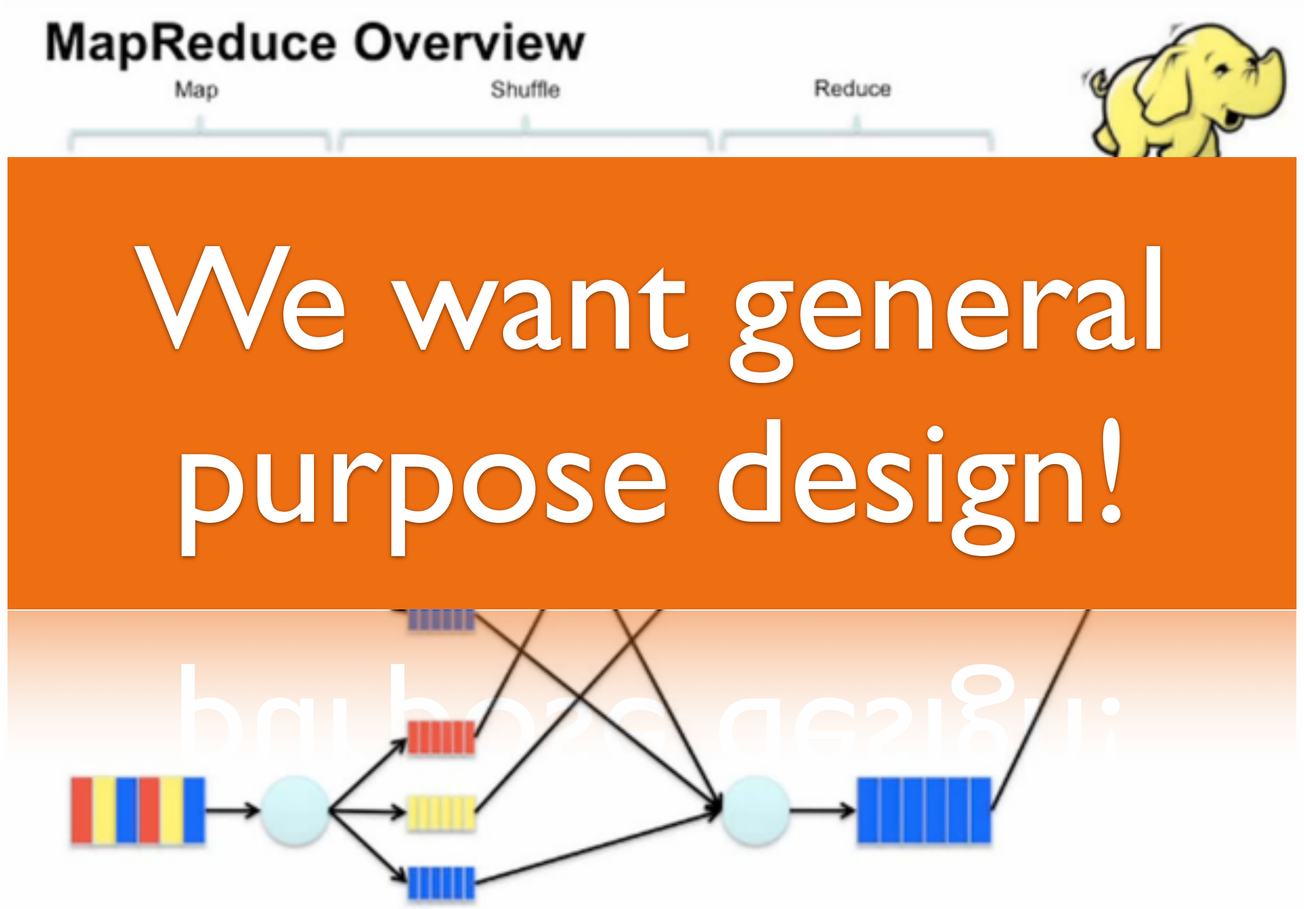
… design for it



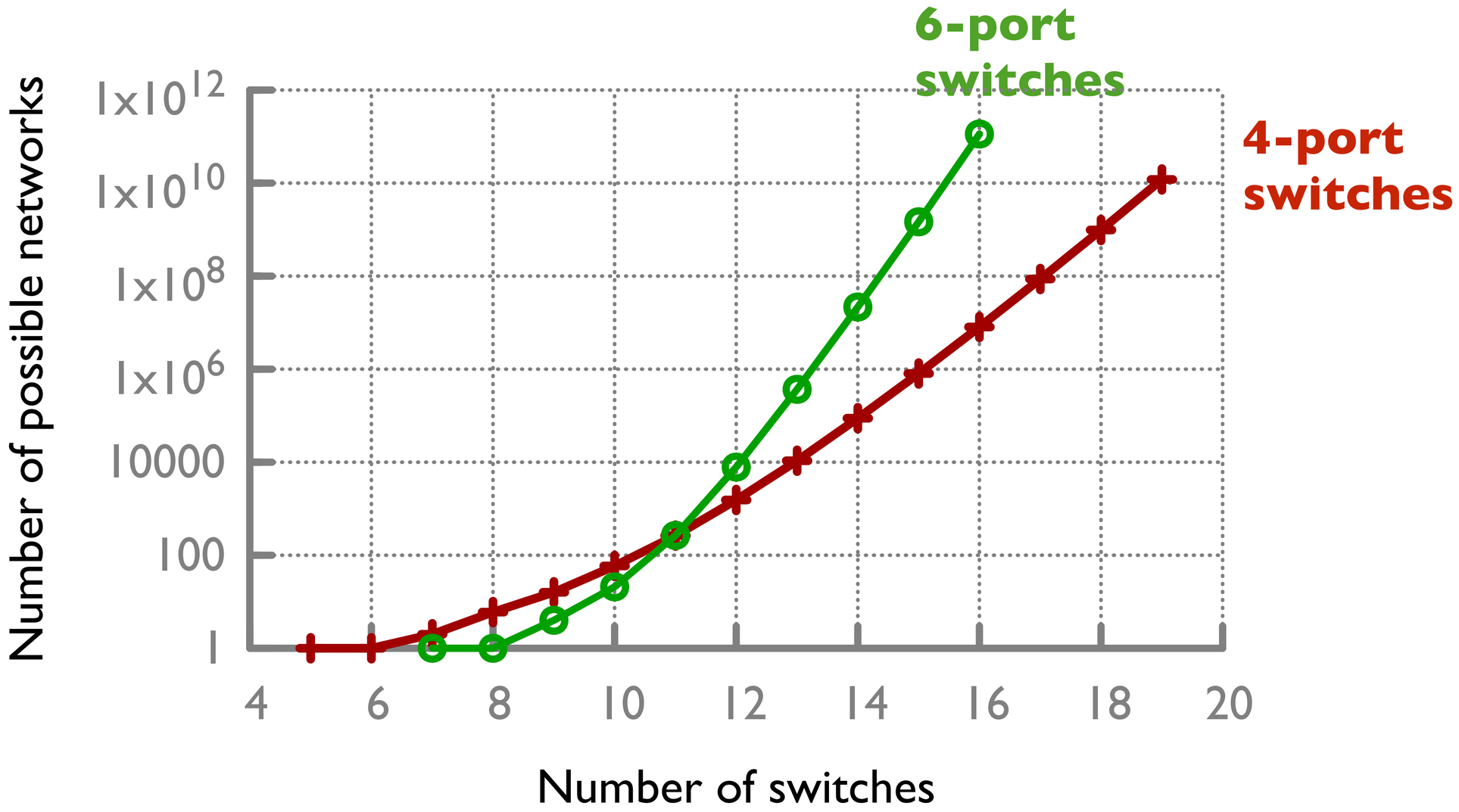


"Hopper", NERSC

But, other apps may not work well …

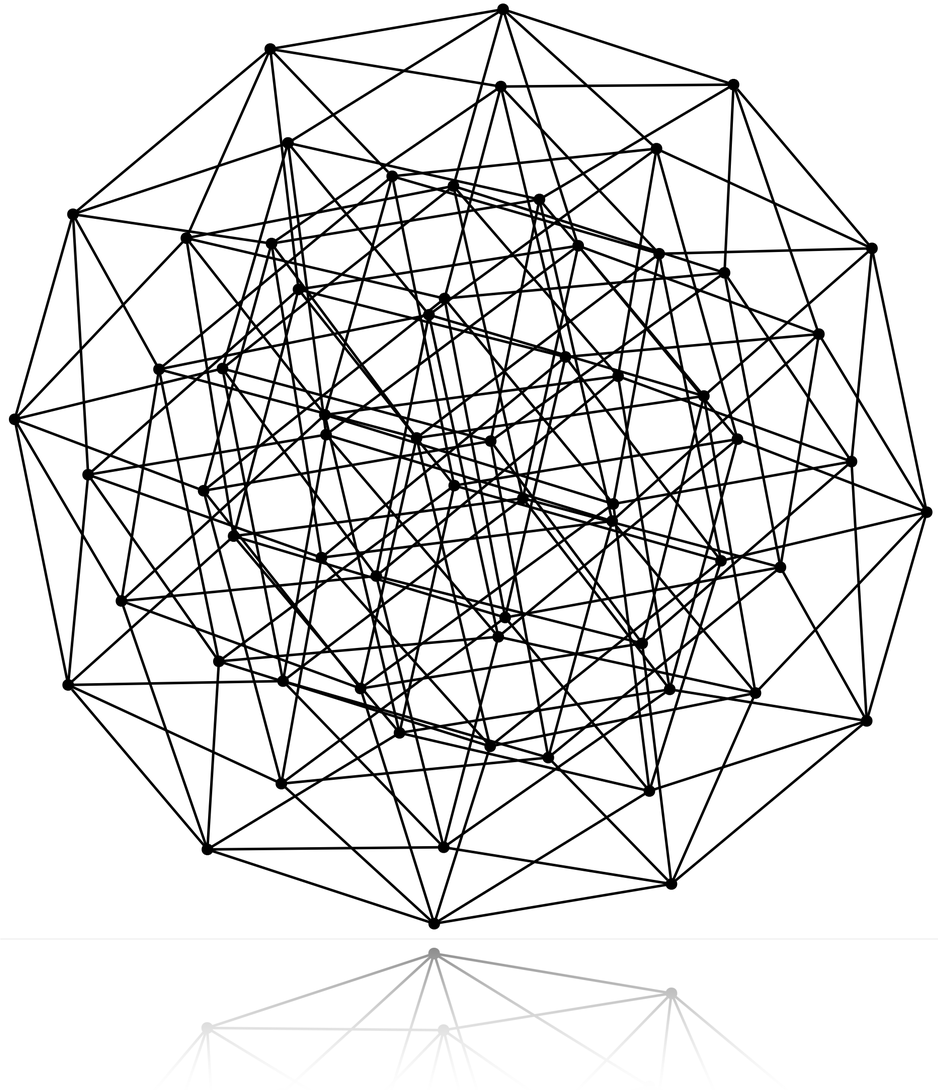


What's so hard about this?

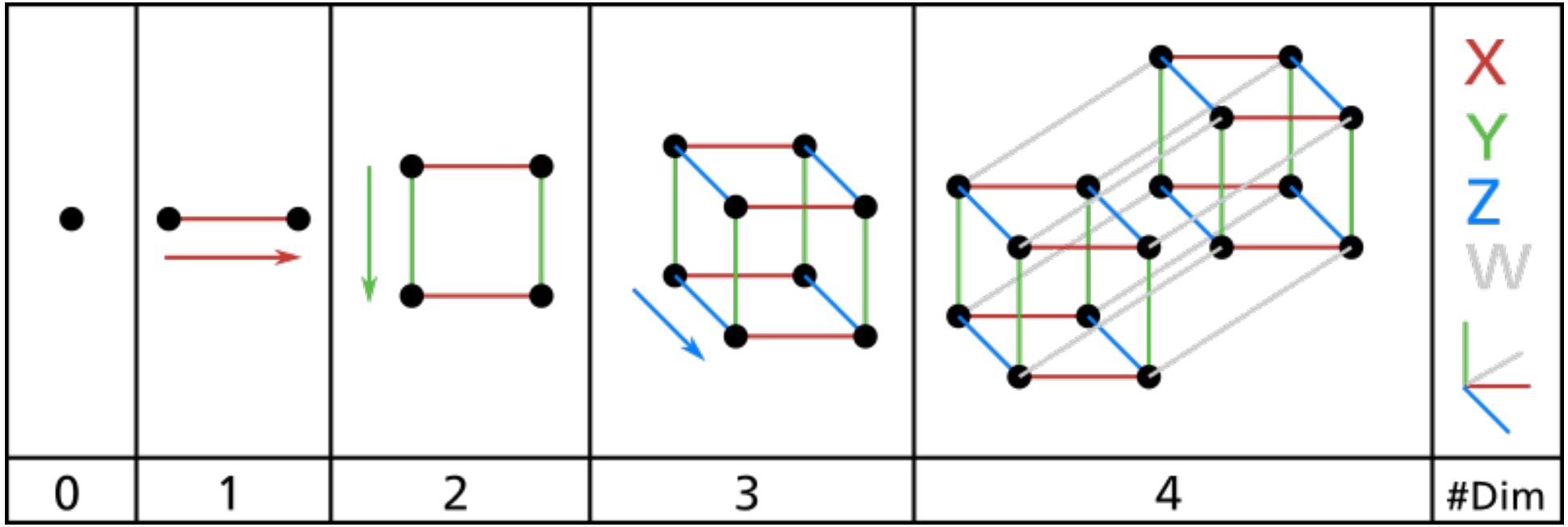


How would YOU think about this problem?!

So people pick known good candidates

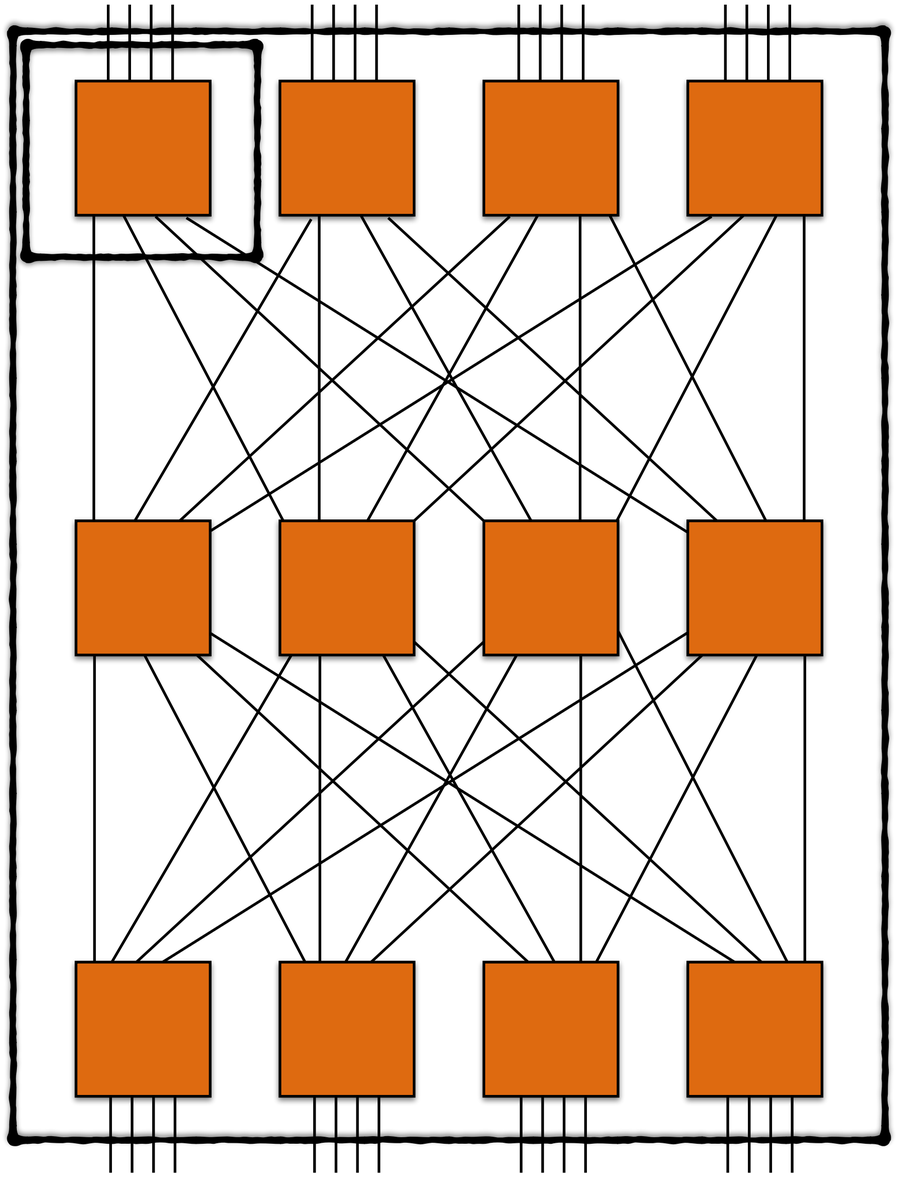


Hypercube



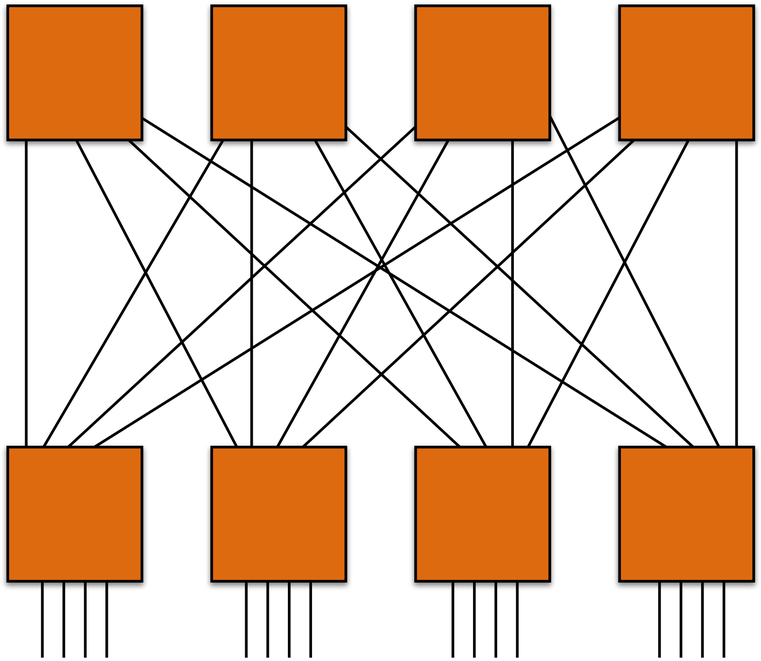
NerdBoy1392 [CC BY-SA 3.0], via Wikimedia

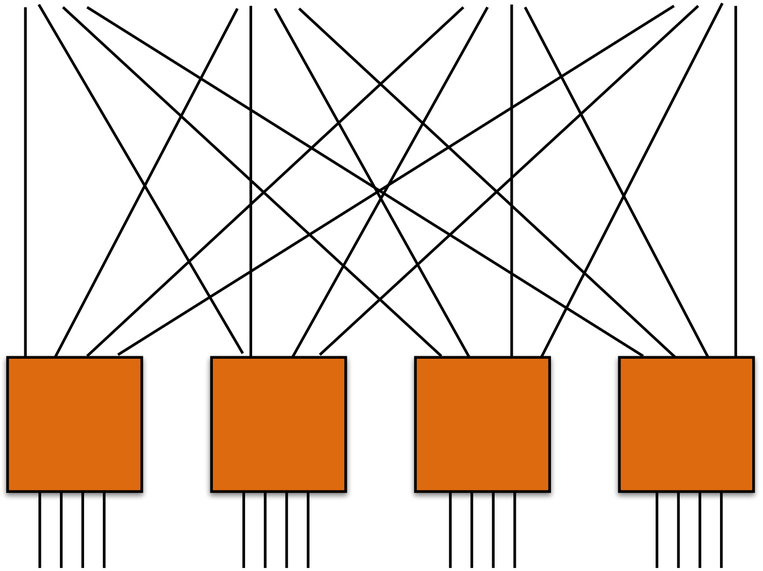
Clos networks



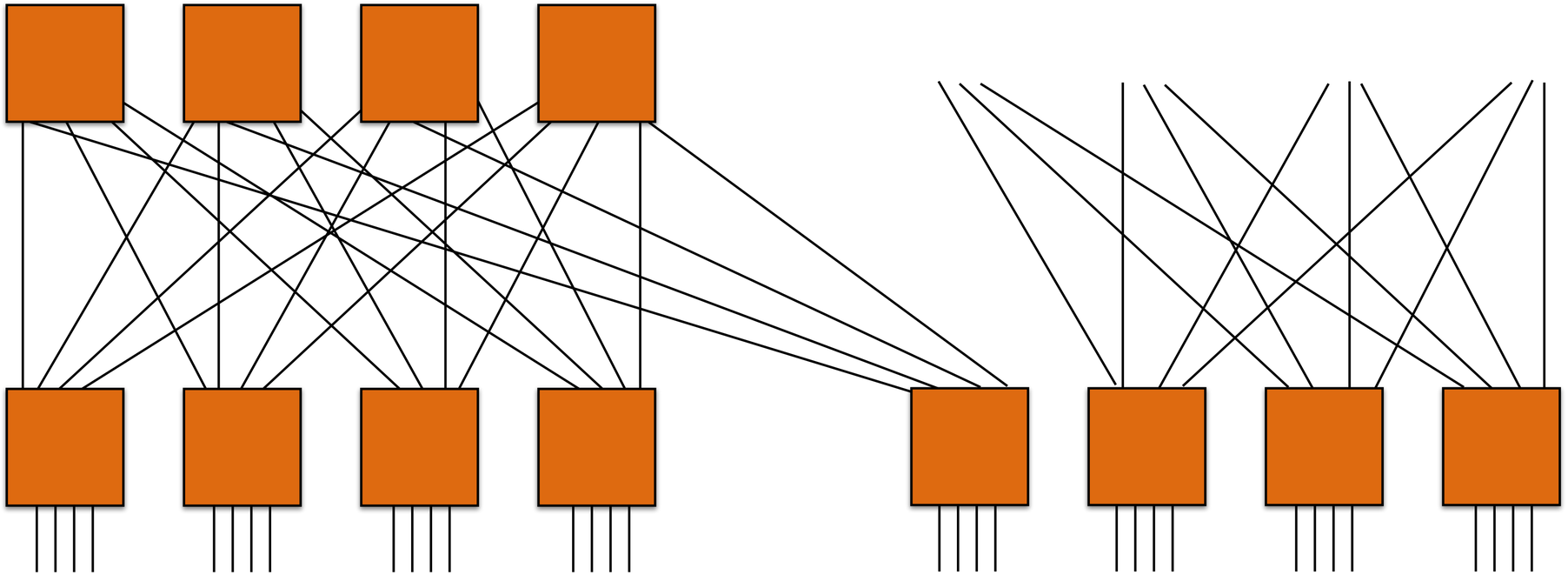
Use small, cheap elemen
to build large networks!!

Folded-Clos?



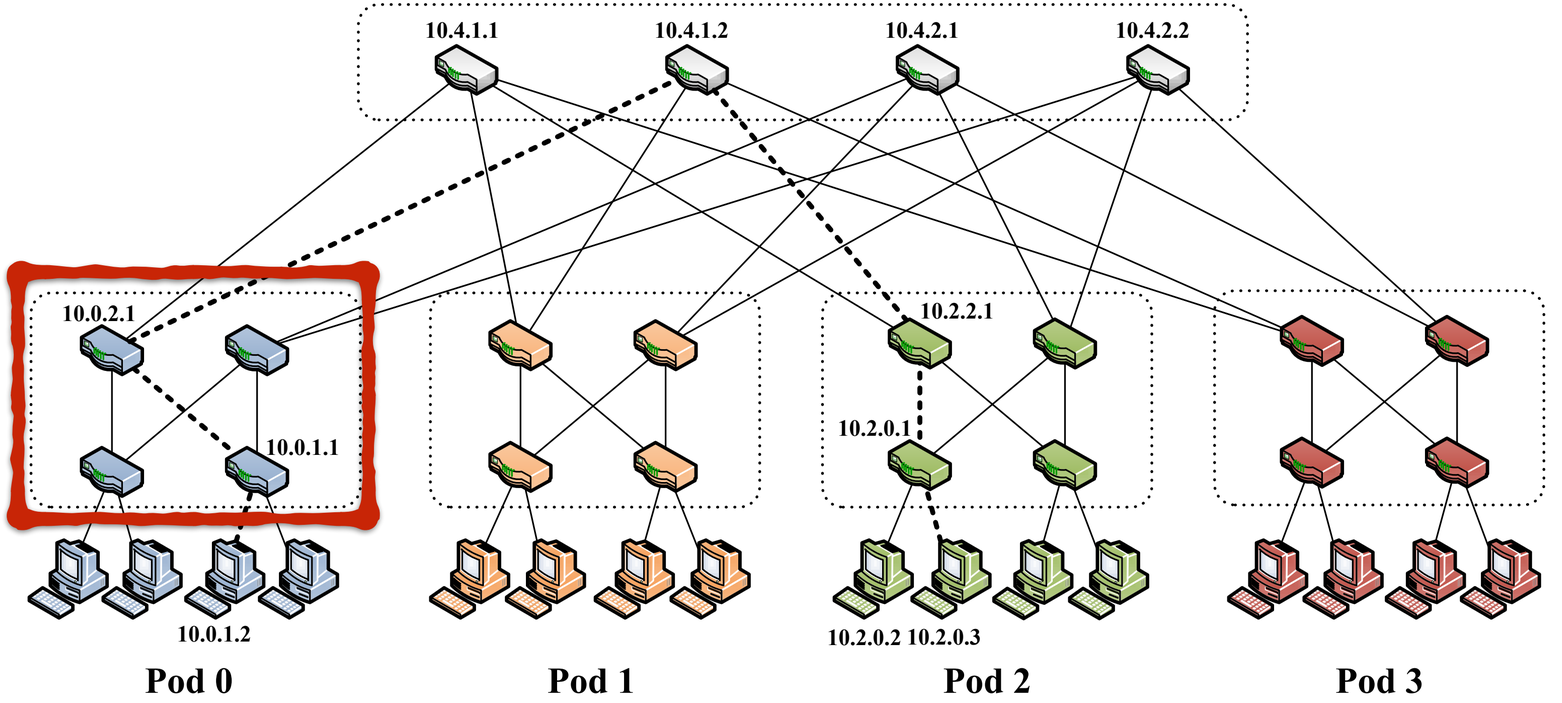


Folded-Clos?



Wikipedia user Nachoman-au

Fat-tree network



ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Fat-tree network

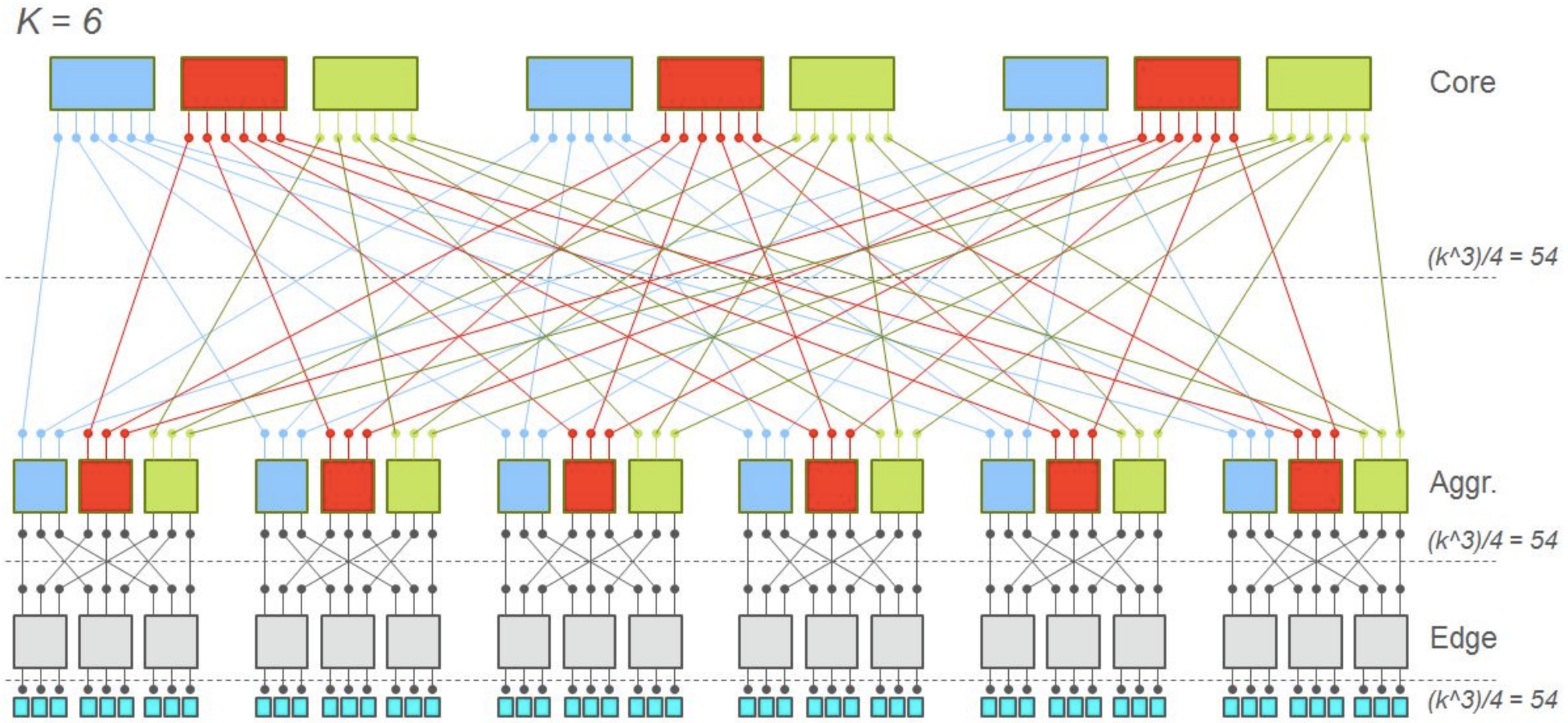
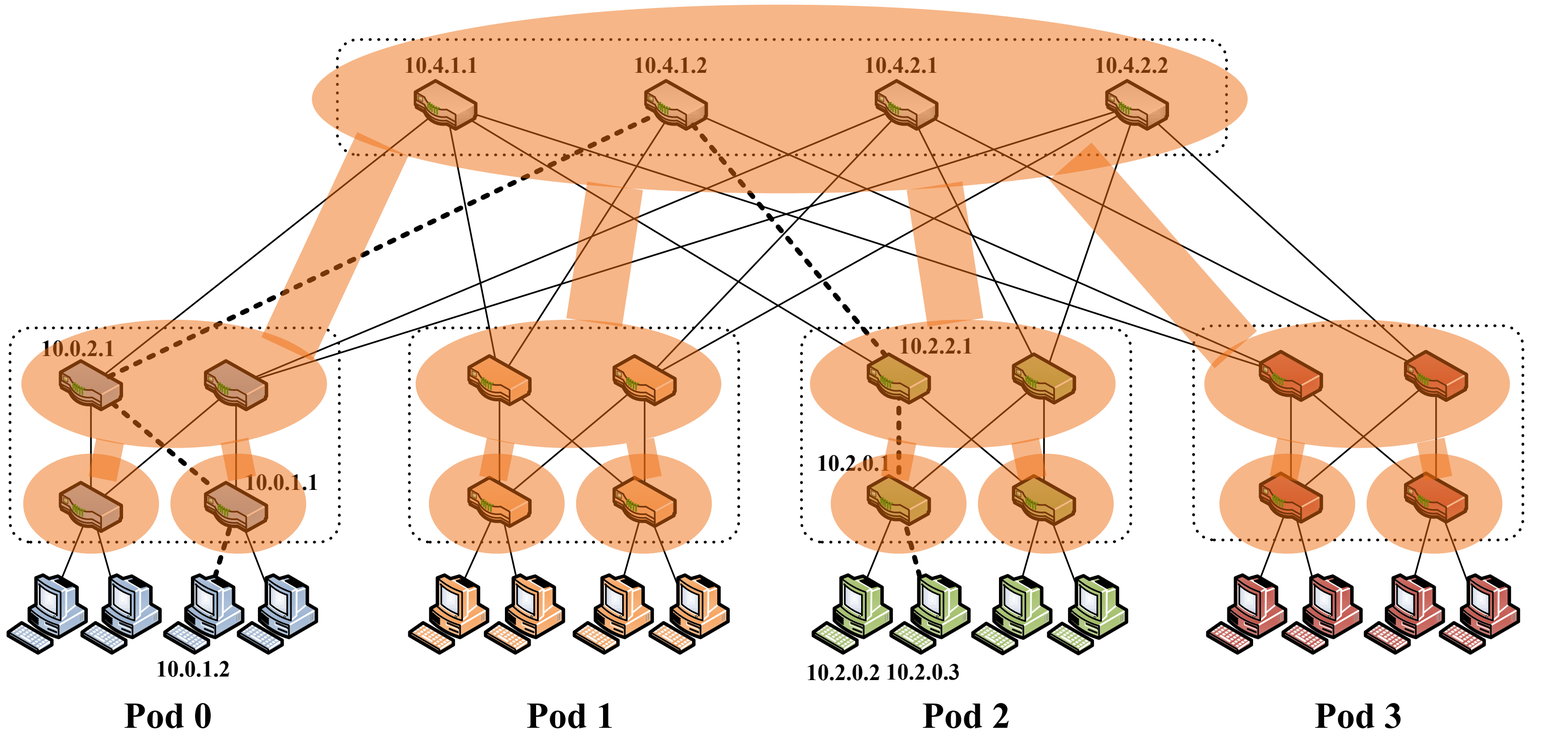


Image source: Francesco Celestino

Fat-tree network



Core

Aggregation

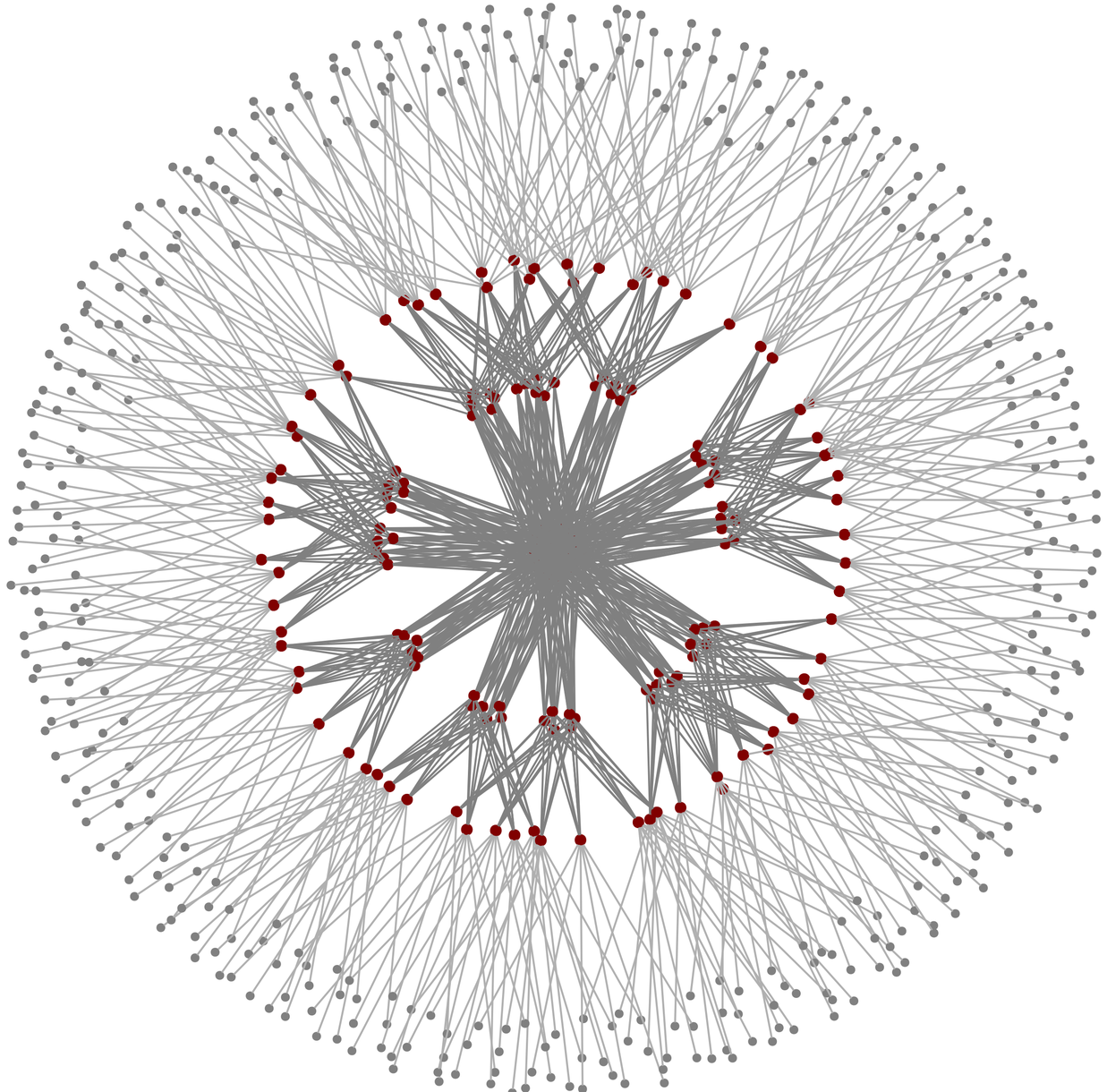
Servers

ACM SIGCOMM, 2008

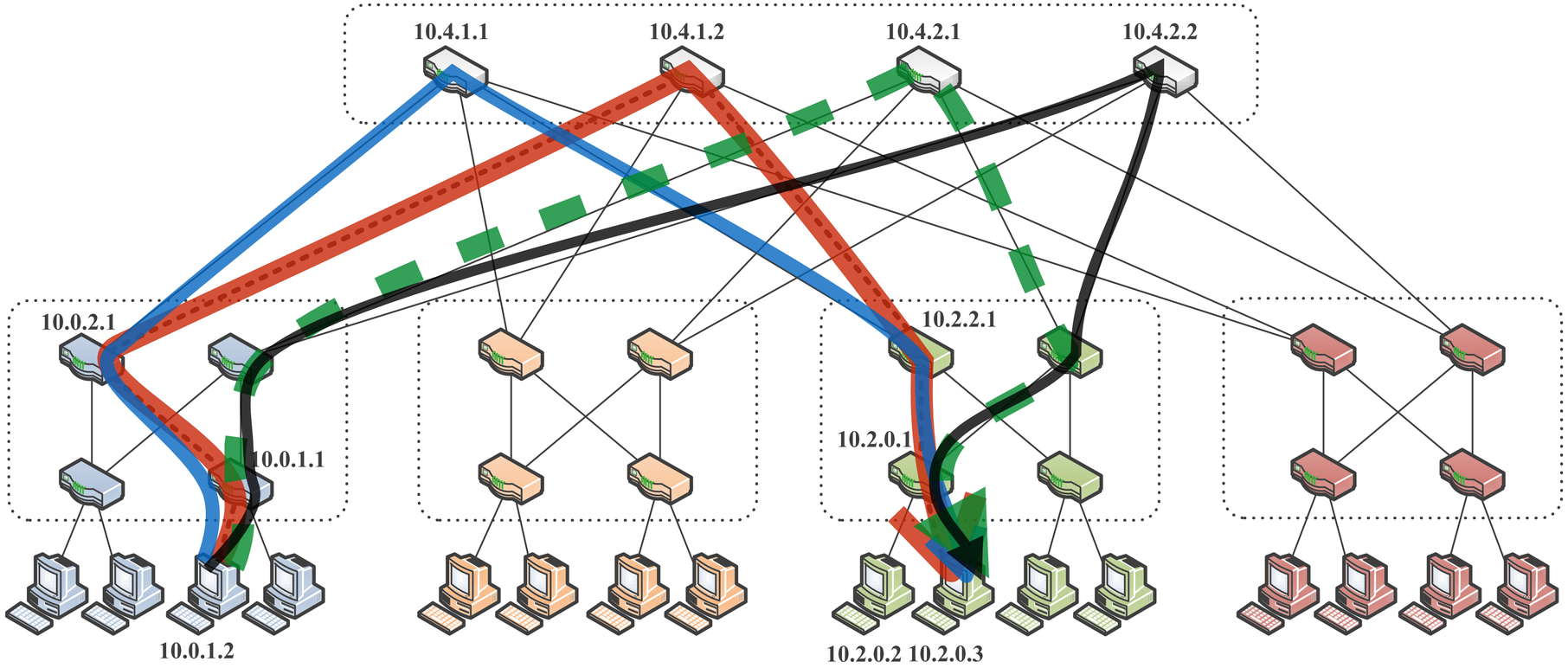
A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Fat-tree network



Fat-tree network

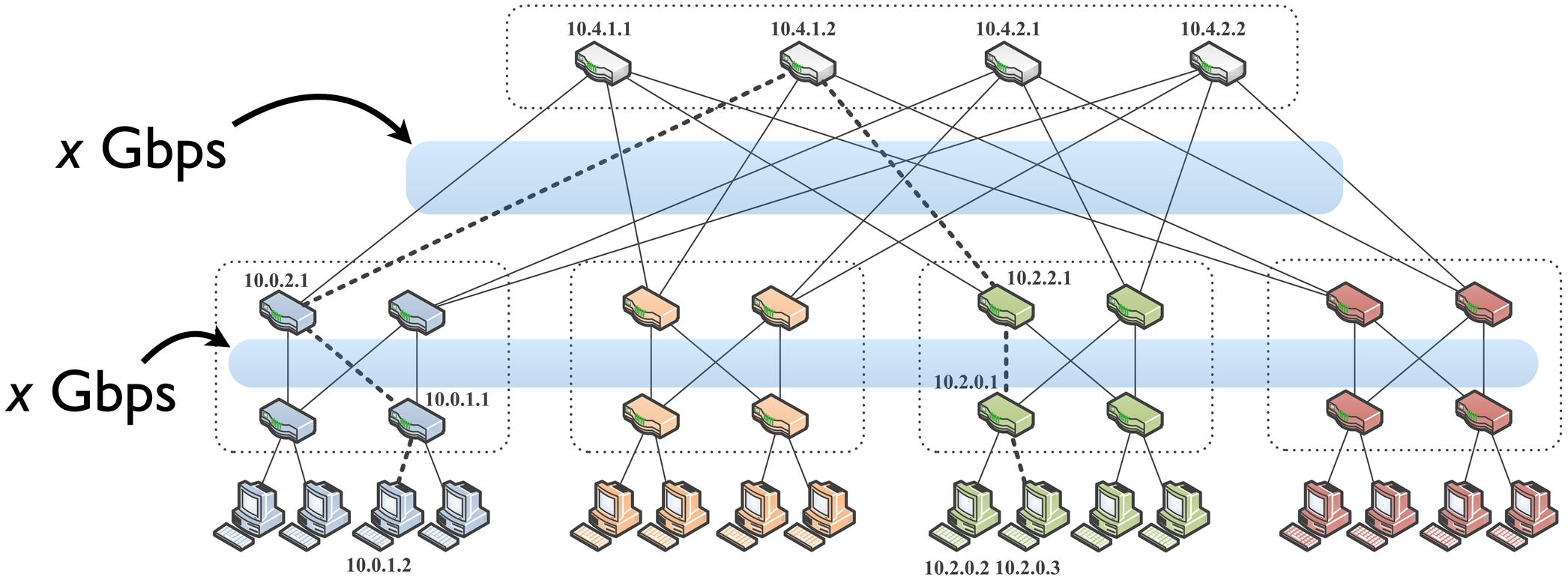


ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Fat-tree network

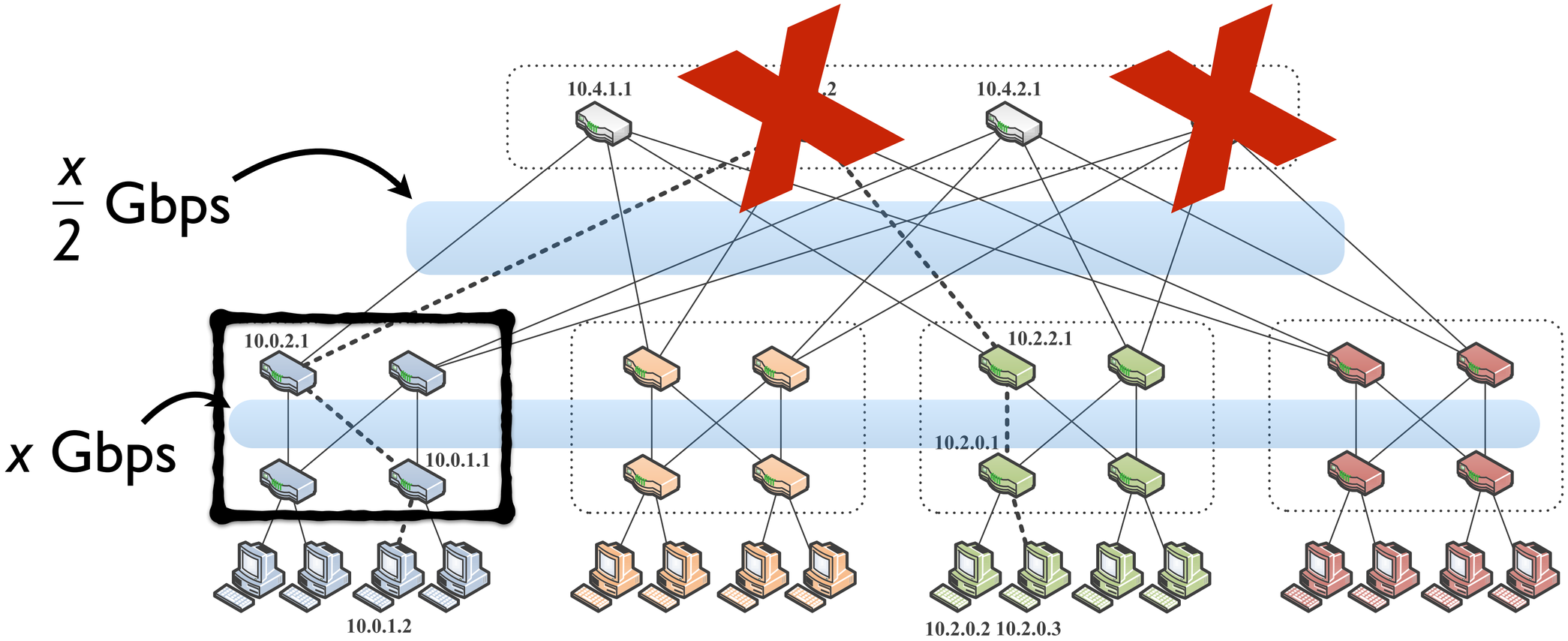


ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Fat-tree network



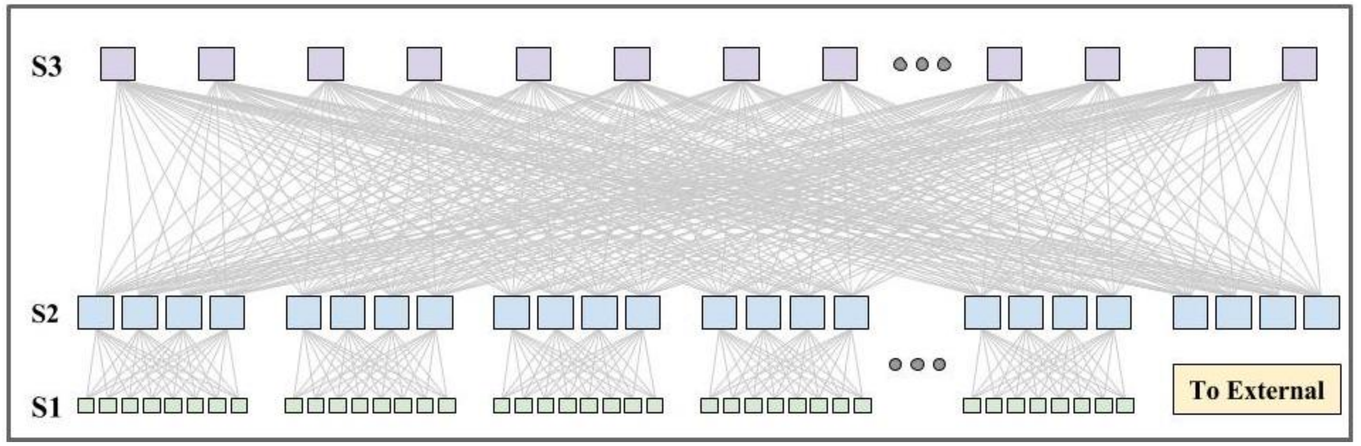
ACM SIGCOMM, 2008

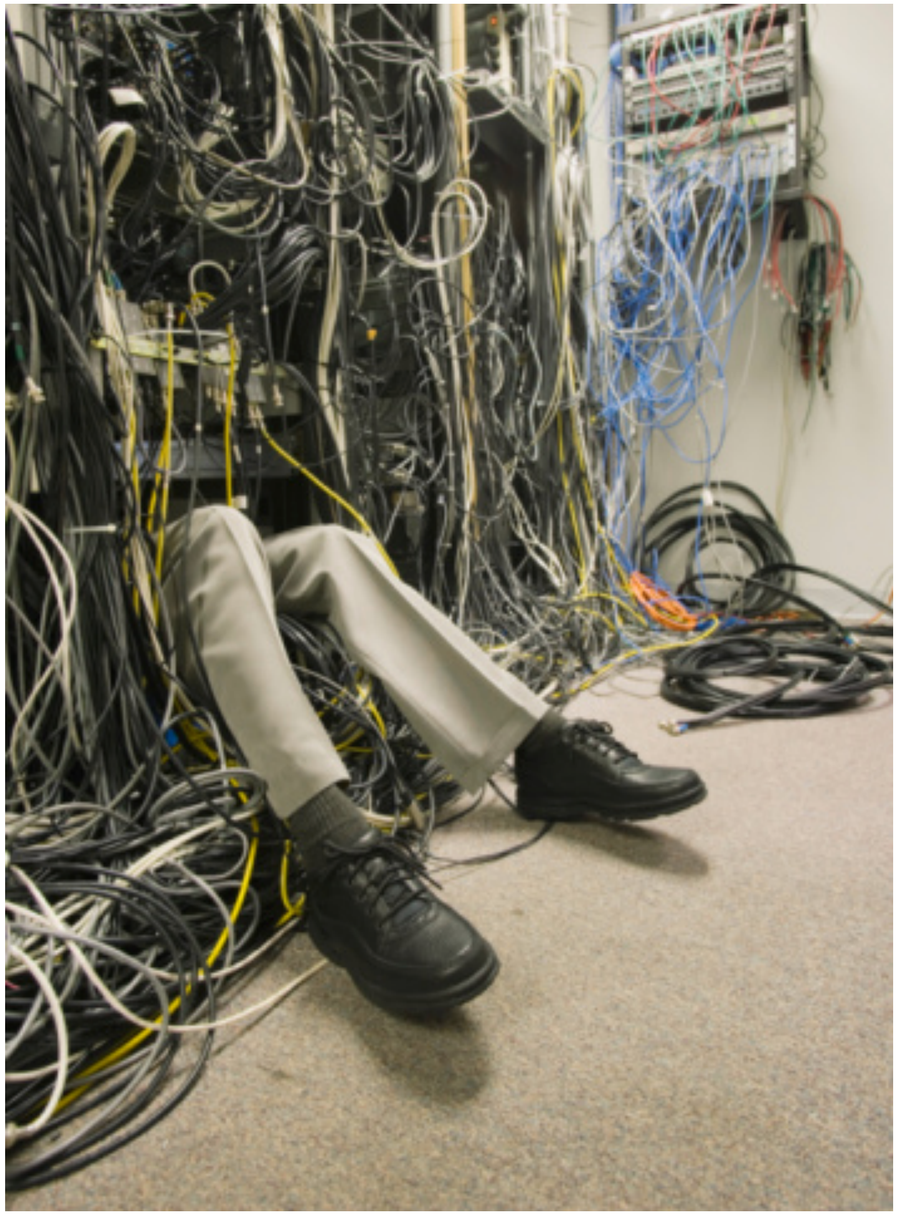
A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

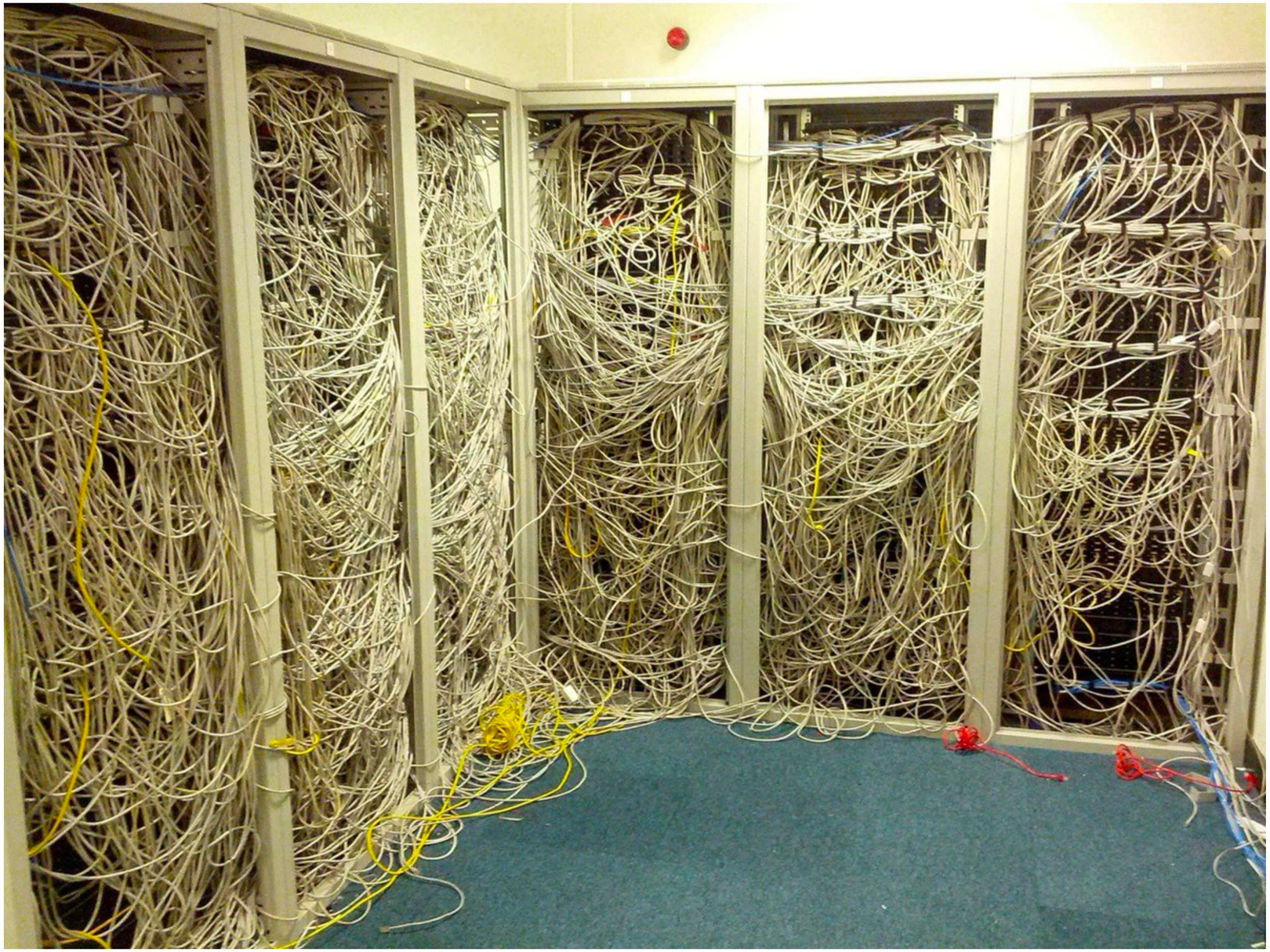
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network

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.





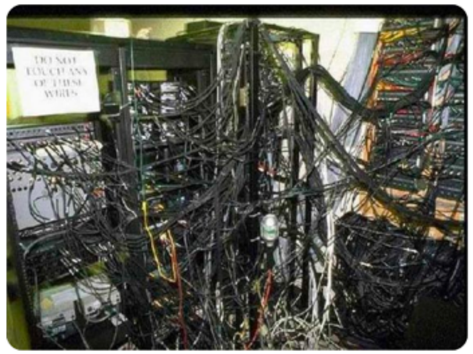
[David Samuel Robbins, gettyimages.ch]



[@AlexCWheeler, Twitter]

Explore Wiring Fail, Wiring Jobs, and more!

Computers | Cable | Thoughts | The spider | Wells | Lord of the rings | Need to | Yellow



Wiring Fail | Wiring Jobs | Safe Wiring

Poor data center cable management. I'm expecting Shelob the spider from Lord of the Rings to emerge any moment.

See More

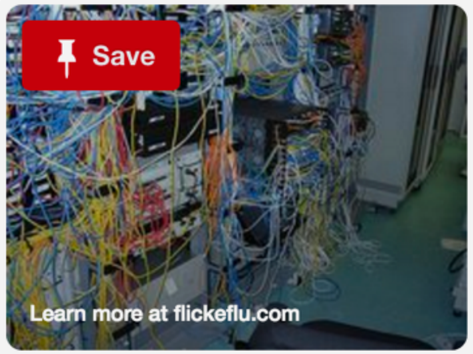
9 ♥ 1

Datacenter Di | Datacenter Google | Ahhh

Un giro nei #datacenter di #Google

See More

12



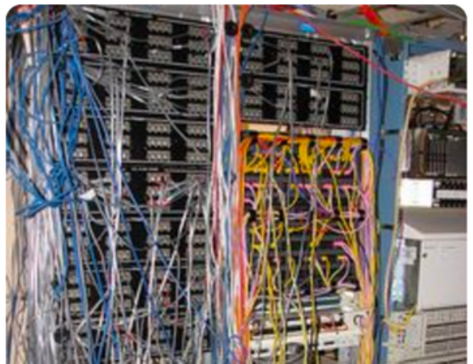
Horrible Data | Wiring Disasters | Center

Aaaah! What a horrible data center disaster.

See More

by Eric Brandwine

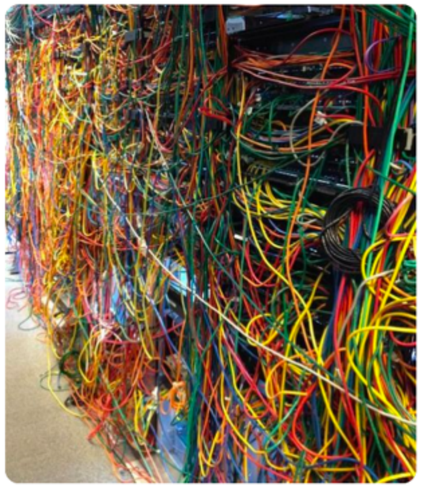
1



Messy Cable Closets & Serv

13 Pins **43** Followers

Cable | To work | Search | Wire | Tech | Try again | Gazebo | Google



Room Disasters | Computer Disasters

Server Room disaster

See More

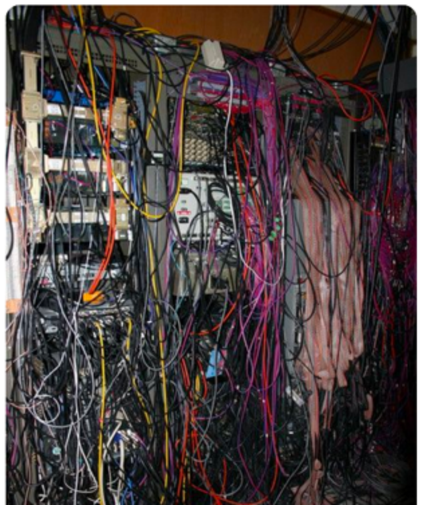
3

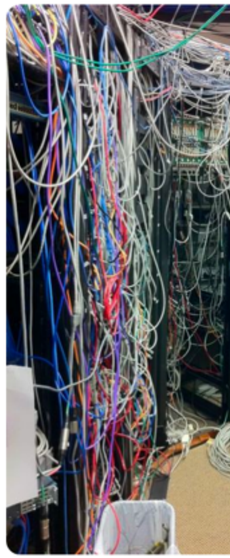
Room Nightmares | Cabling Nightmares

Real-world server room nightmares

See More

2





Messy Cable | Worst Se

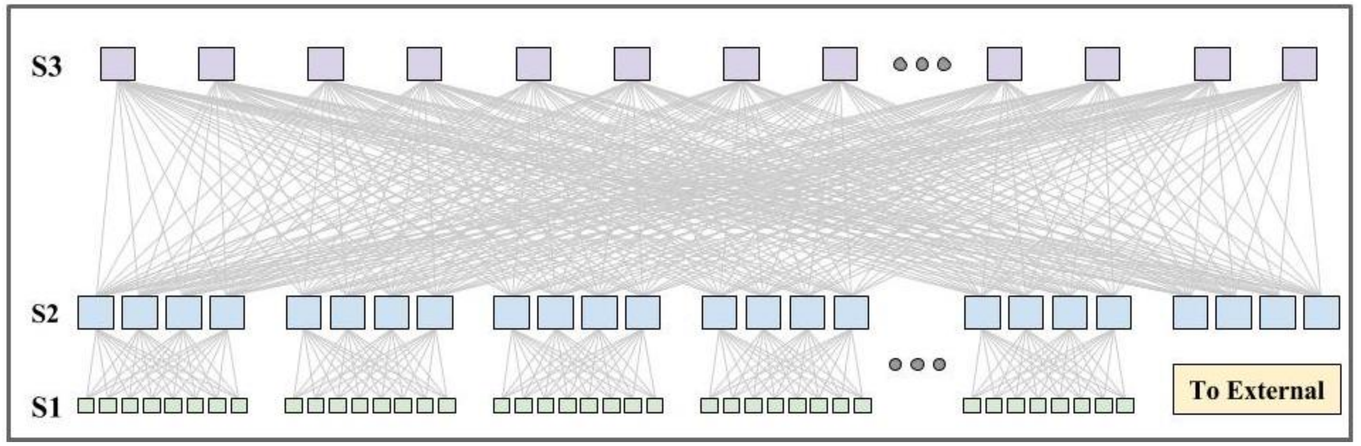
Real-world server roo

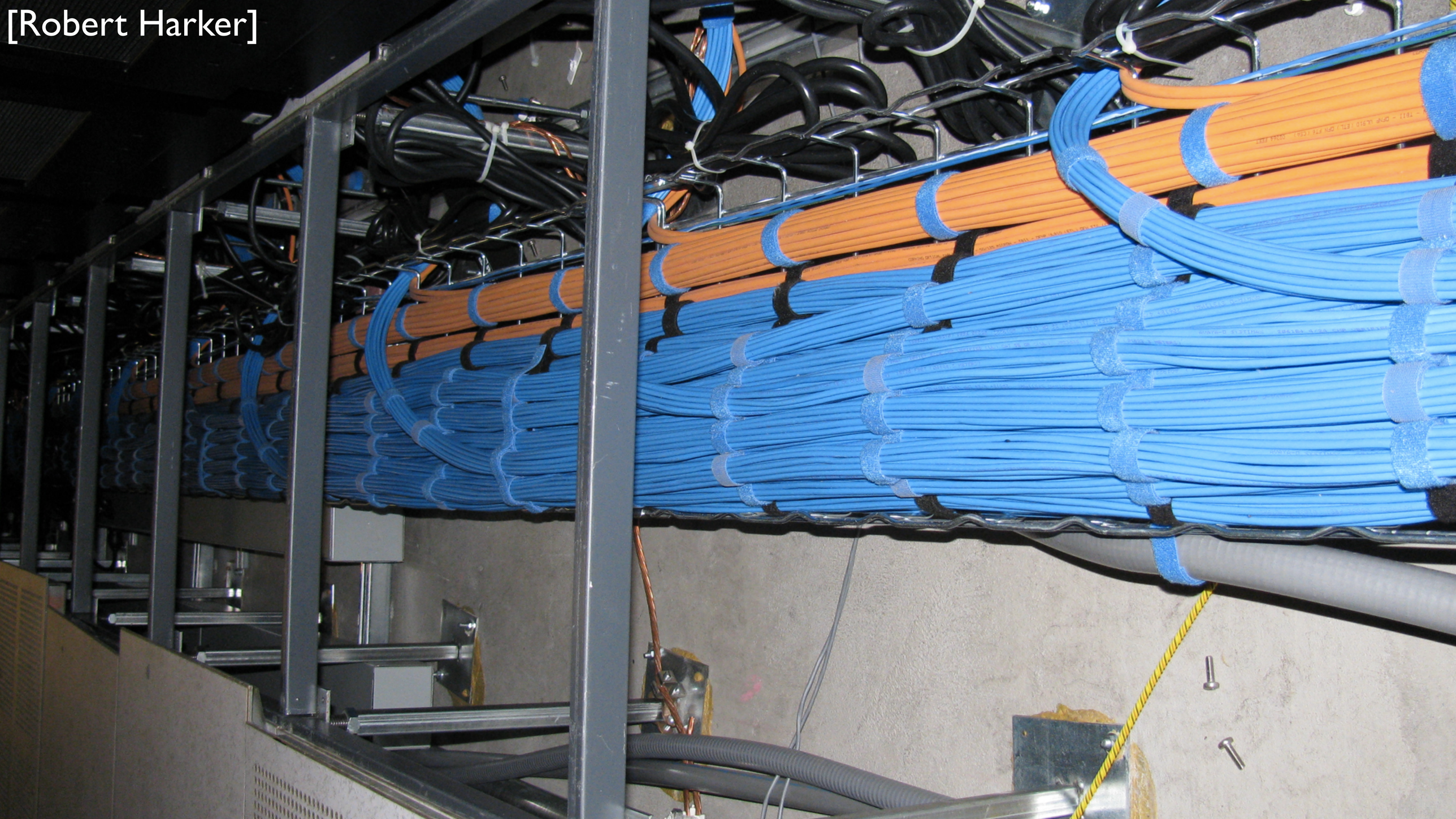
See More

1

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

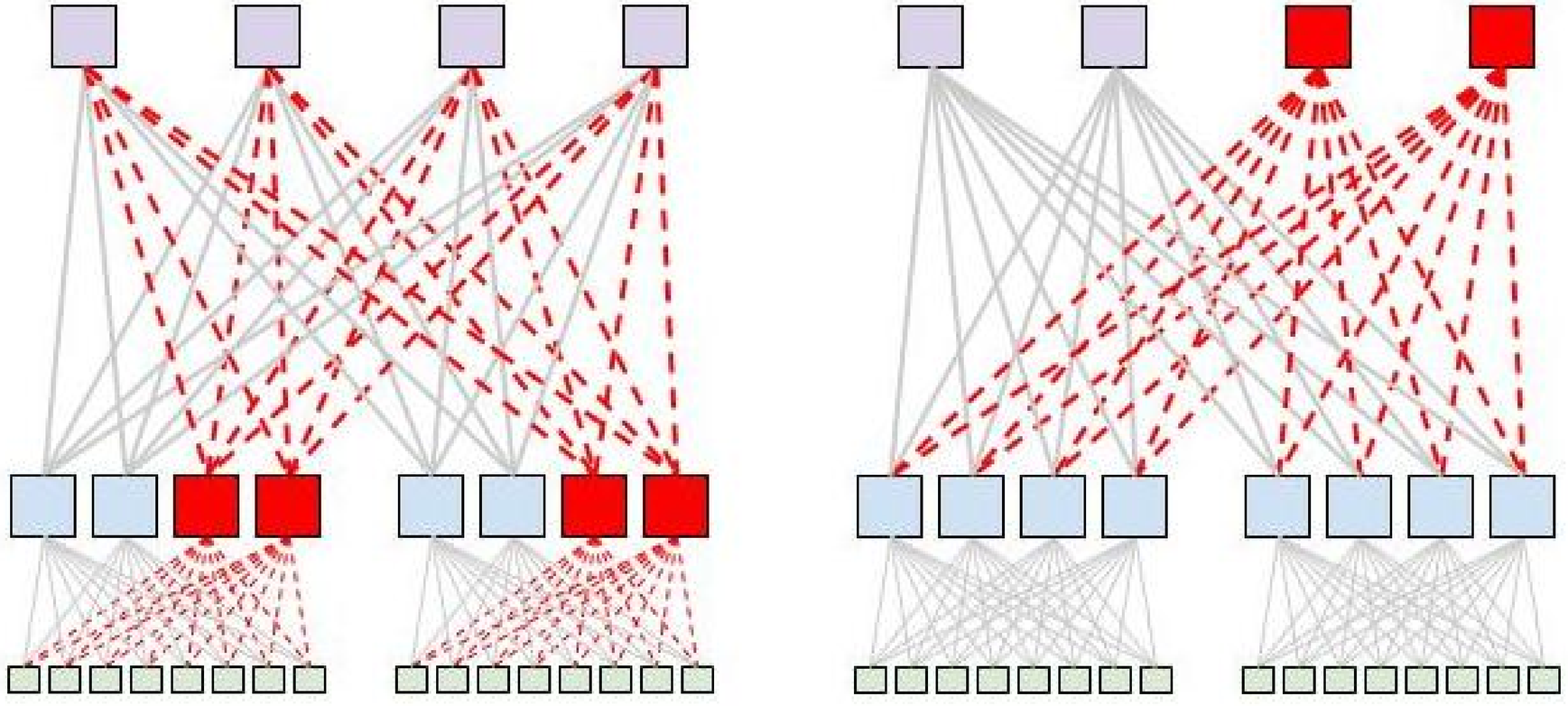
Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, and Amin Vahdat
Google, Inc.



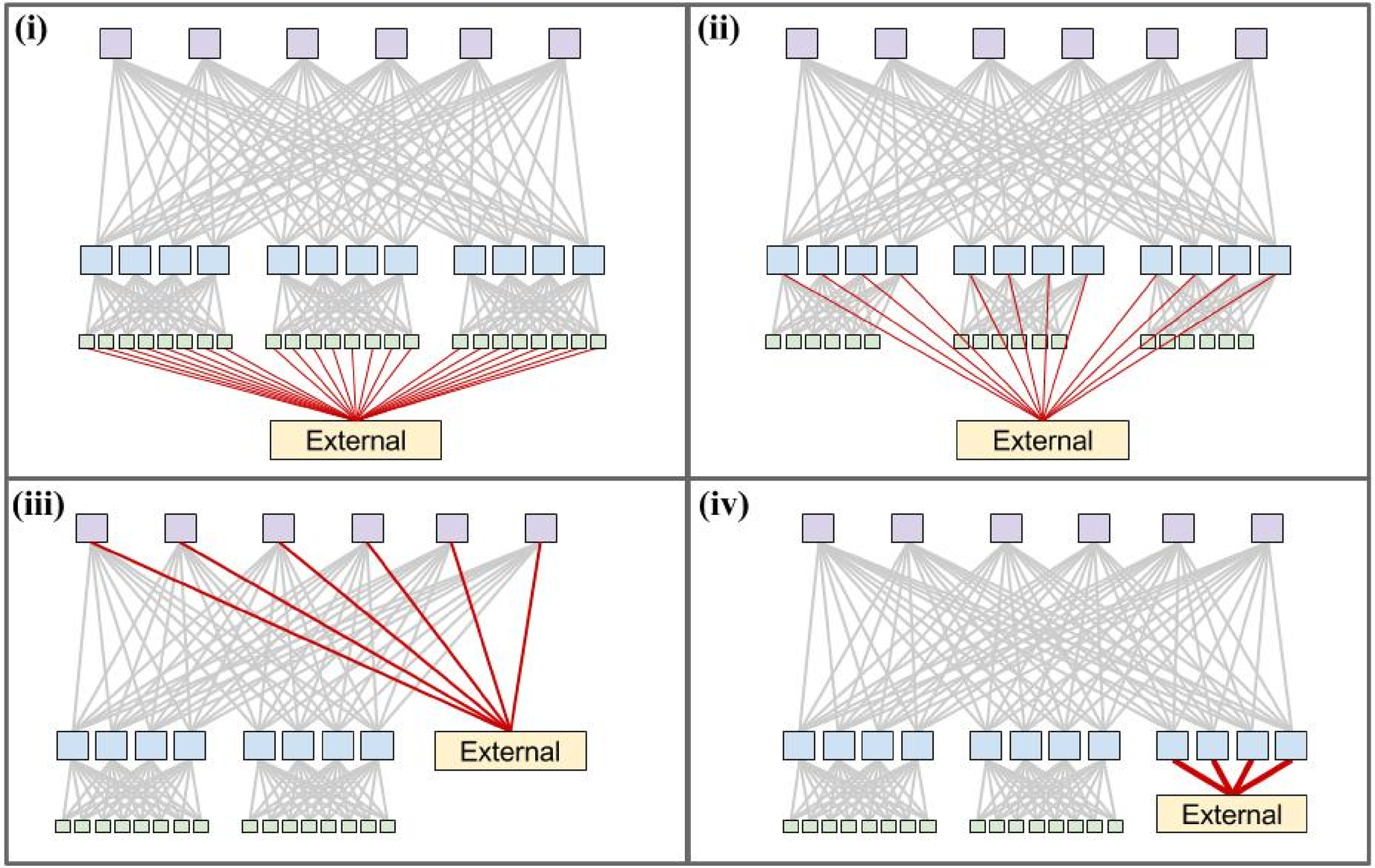

[Robert Harker]

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

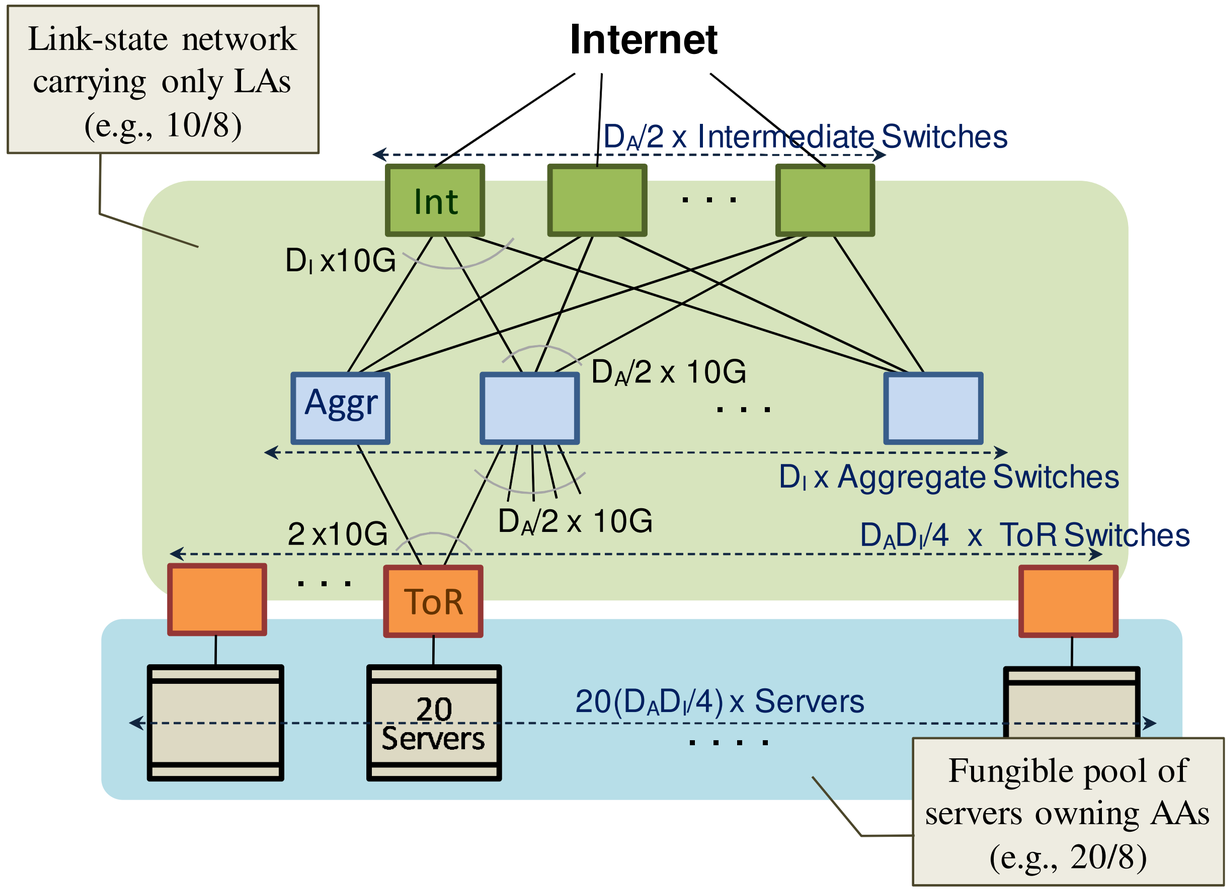
Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, and Amin Vahdat
Google, Inc.



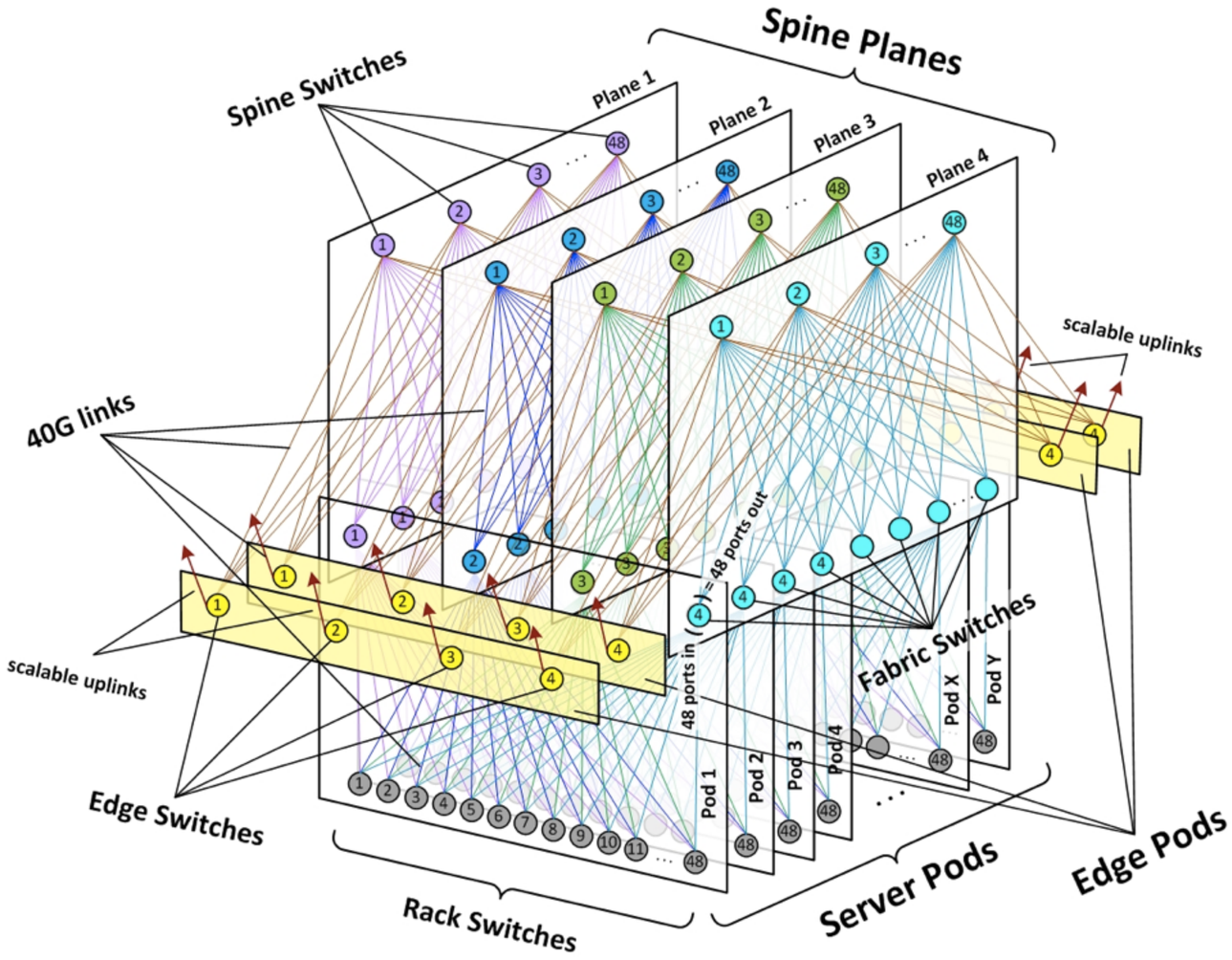
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network



Variants of this design are common



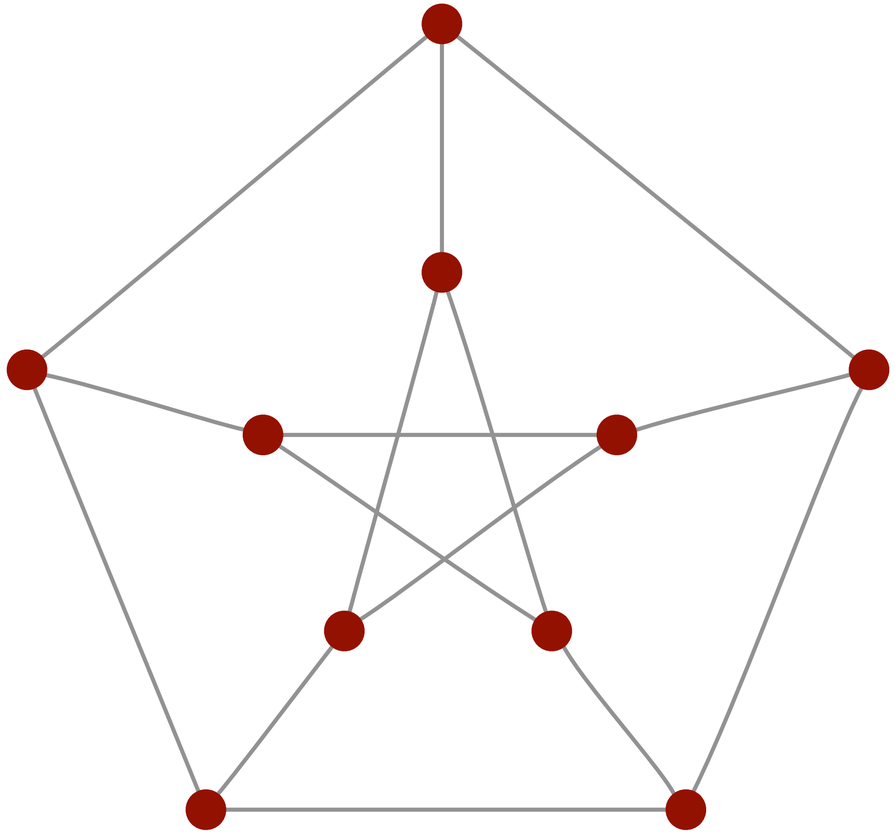
VL2 @ **Microsoft**, ACM SIGCOMM'09
Greenburg, Hamilton, Jain, Kandula, Kim, Lahiri, Maltz, Patel, Sengupta



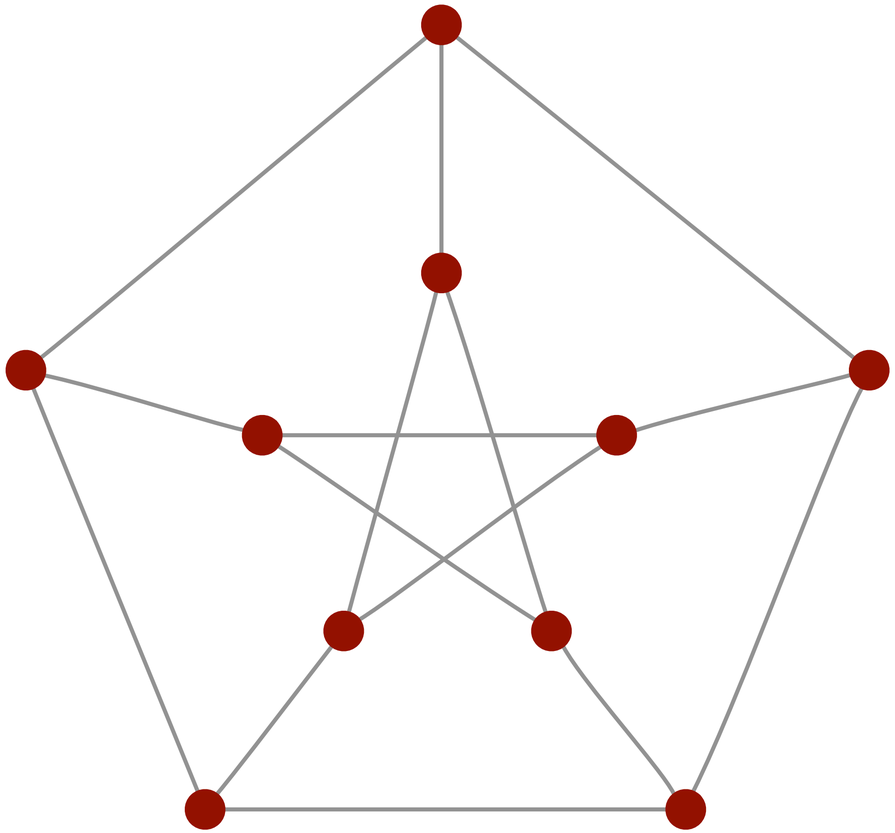
"Introducing data center fabric, the next-generation **Facebook** data center network", Alexey Andreyev, 2015

Connect many cheap, identical switches?

Yes …



Yes …



[Jan Kristian Haugland, neutreeko.net]

Shorter paths ⇒ high capacity

A packet that travels on a short path
consumes a small amount of network capacity

A simple upper bound on throughput

flows • capacity used per flow

≤ total capacity

A simple upper bound on throughput

flows • capacity used per flow

≤ total capacity

A simple upper bound on throughput

flows • throughput per flow • mean path length

≤ total capacity

A simple upper bound on throughput

$$\text{throughput per flow} \leq \frac{\text{total capacity}}{\text{\# flows} \bullet \text{mean path length}}$$

A simple upper bound on throughput

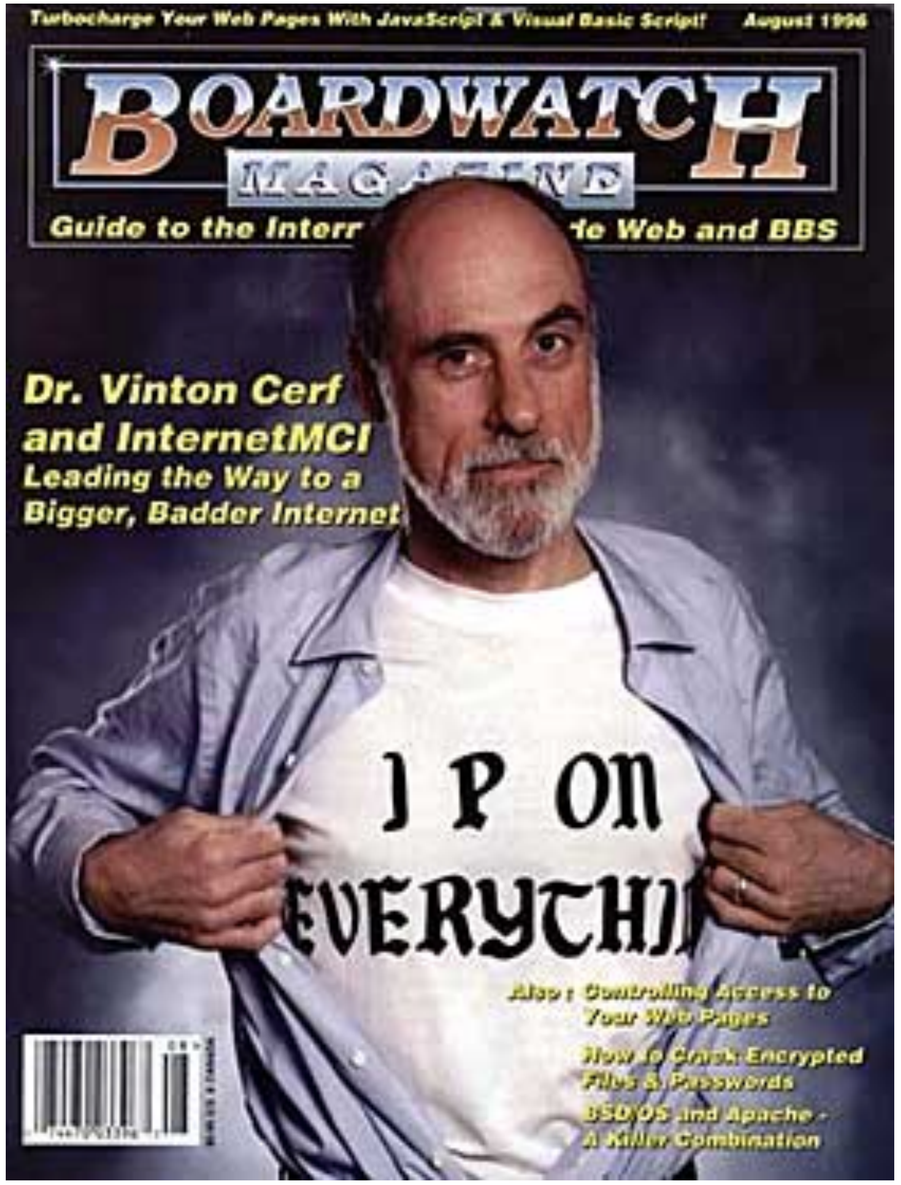
$$\text{throughput per flow} \leq \frac{\sum_{\text{links}} \text{capacity}(\textit{link})}{\#\text{ flows} \cdot \text{mean path length}}$$

Lower bound this!

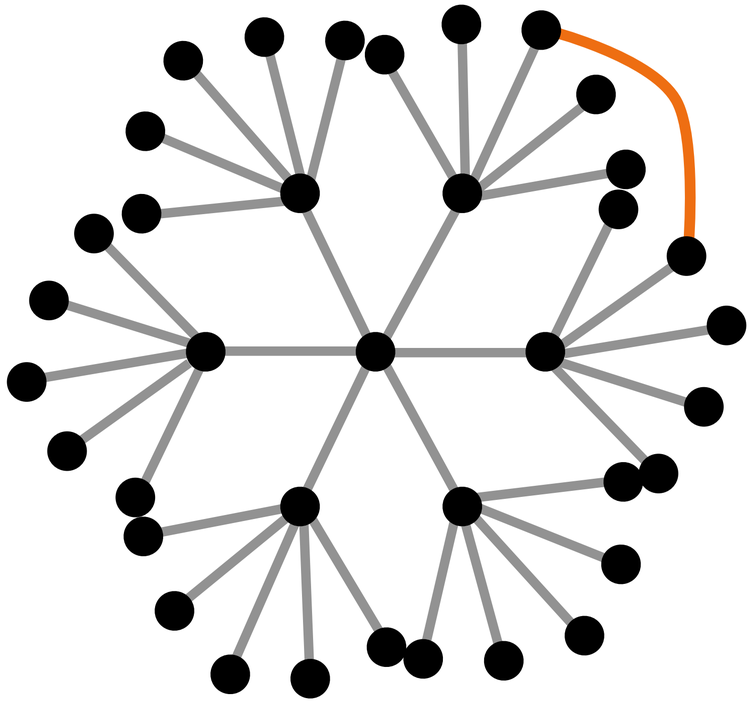
[Arxiv 2013, NSDI 2014] High Throughput Data Center Topology Design

Ankit Singla, P. Brighten Godfrey, Alexandra Kolla

Lower bound on mean path length



Lower bound on mean path length

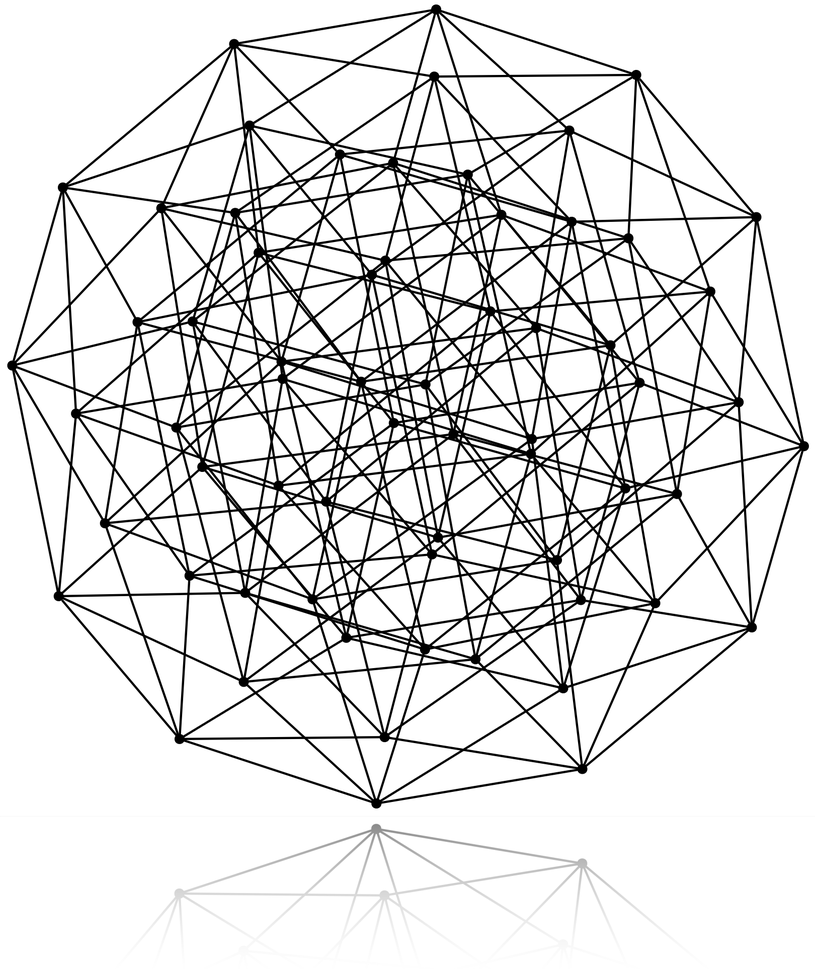


Distance	# Nodes
1	6
2	6^2 - 6

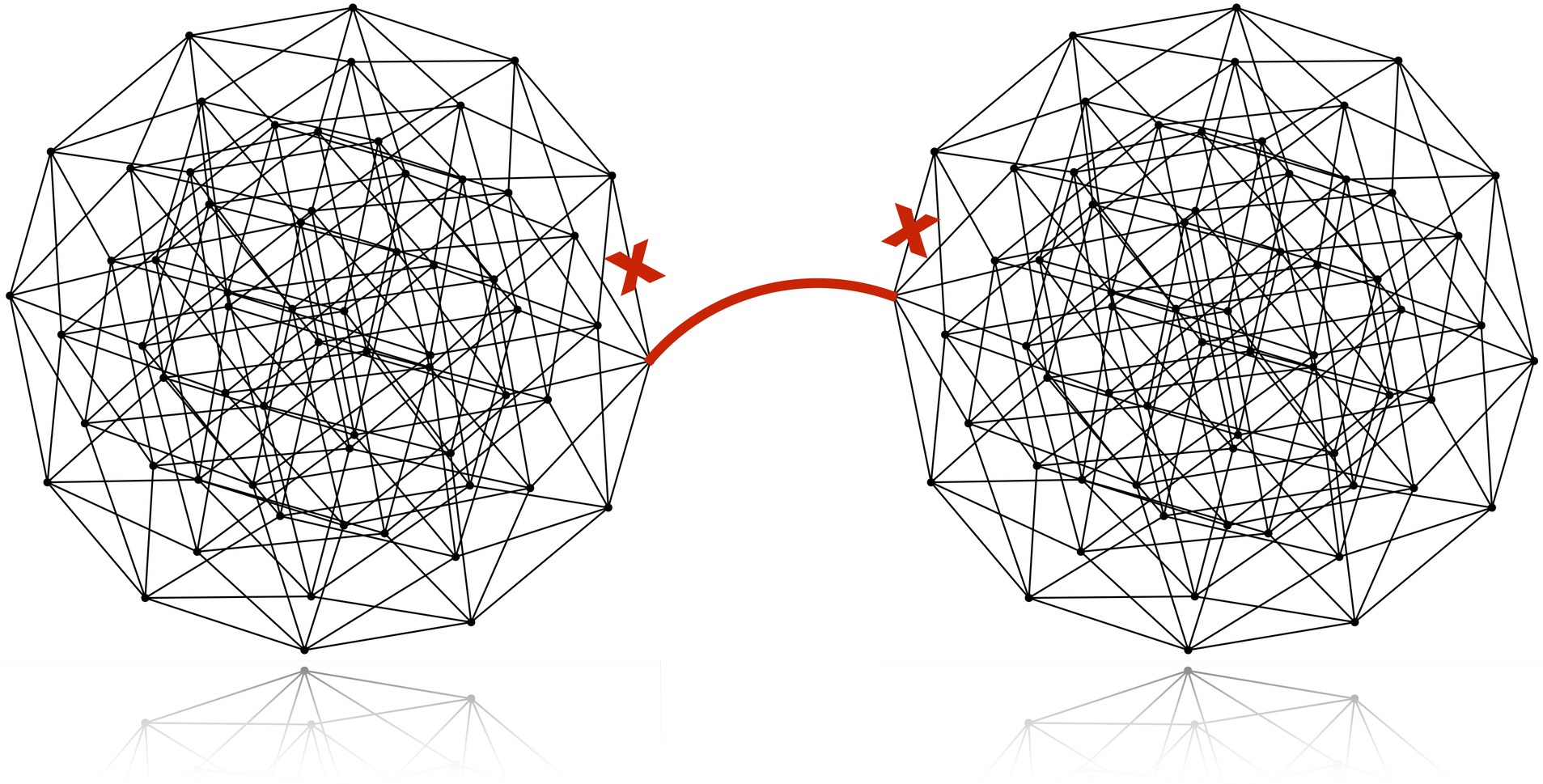
(Ugliness omitted)

[Cerf et al., "A lower bound on the average shortest path length in regular graphs", 1974]

… but cuts matter too



… but cuts matter too

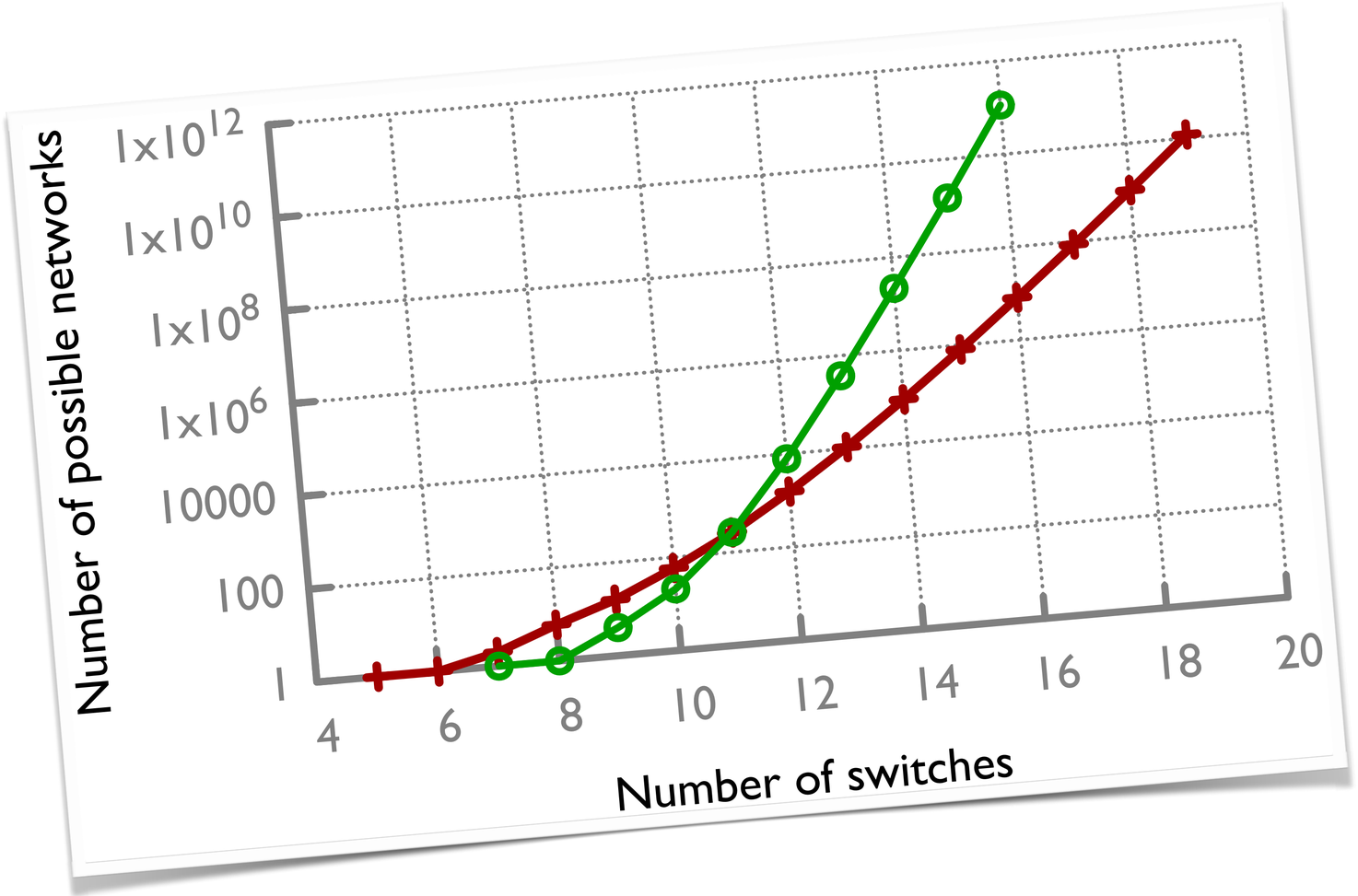


Shorter paths ⇒ high capacity

A packet that travels on a short path
consumes a small amount of network capacity

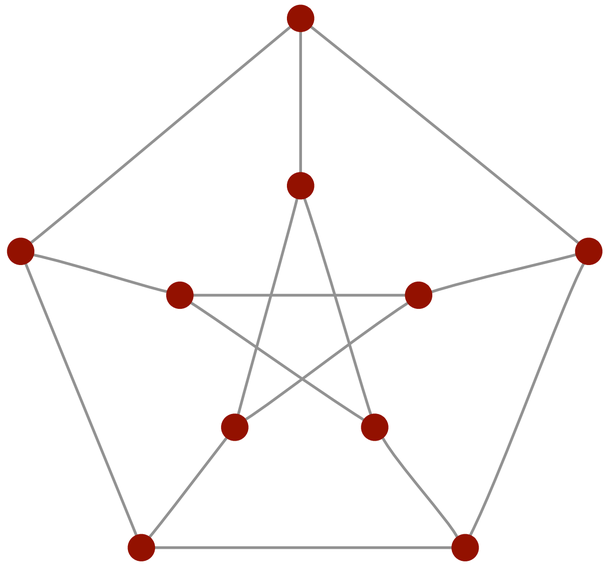
Degree-diameter problem

"Out of all these, give me one with lowest diameter …"



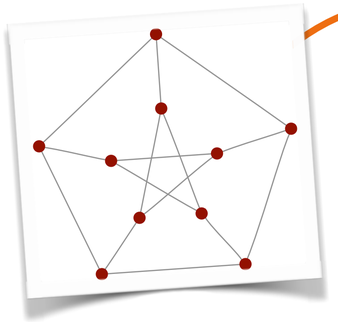
Degree-diameter problem

What is the maximum number of nodes
in any graph with degree ∂ and diameter d?



Petersen graph

Degree-diameter problem



Diameter

Degree

d \ k	2	3	4	5	6	7	8	9	10
3	10	20	38	70	132	196	336	600	1250
4	15	41	96	364	740	1 320	3 243	7 575	17 703
5	24	72	210	624	2 772	5 516	17 030	57 840	187 056
6	32	110	390	1404	7 917	19 383	76 461	307 845	1 253 615
7	50	168	672	2 756	11 988	52 768	249 660	1 223 050	6 007 230
8	57	253	1 100	5 060	39 672	131 137	734 820	4 243 100	24 897 161
9	74	585	1 550	8 200	75 893	279 616	1 686 600	12 123 288	65 866 350
10	91	650	2 286	13 140	134 690	583 083	4 293 452	27 997 191	201 038 922
11	104	715	3 200	19 500	156 864	1 001 268	7 442 328	72 933 102	600 380 000
12	133	786	4 680	29 470	359 772	1 999 500	15 924 326	158 158 875	1 506 252 500
13	162	851	6 560	40 260	531 440	3 322 080	29 927 790	249 155 760	3 077 200 700
14	183	916	8 200	57 837	816 294	6 200 460	55 913 932	600 123 780	7 041 746 081
15	186	1 215	11 712	76 518	1 417 248	8 599 986	90 001 236	1 171 998 164	10 012 349 898
16	198	1 600	14 640	132 496	1 771 560	14 882 658	140 559 416	2 025 125 476	12 951 451 931

[Wikipedia: https://en.wikipedia.org/wiki/Table_of_the_largest_known_graphs_of_a_given_diameter_and_maximal_degree]

Degree-diameter problem

Just use the best-known degree-diameter graphs?!

[NSDI 2012] Jellyfish: Networking Data Centers Randomly
Ankit Singla, Chi-Yao Hong, Lucian Popa, P. Brighten Godfrey

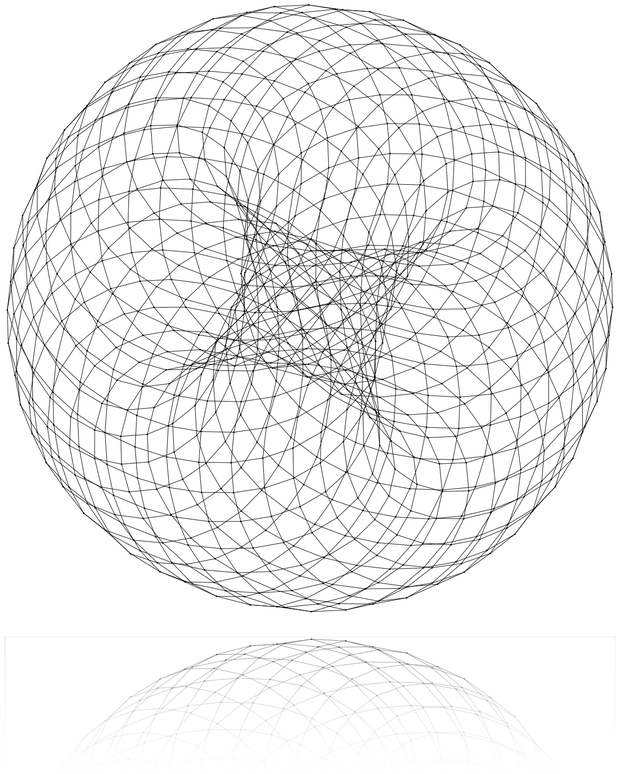
[SC 2014] Slim Fly: A Cost Effective Low-Diameter Network Topology
Maciej Besta, Torsten Hoefler

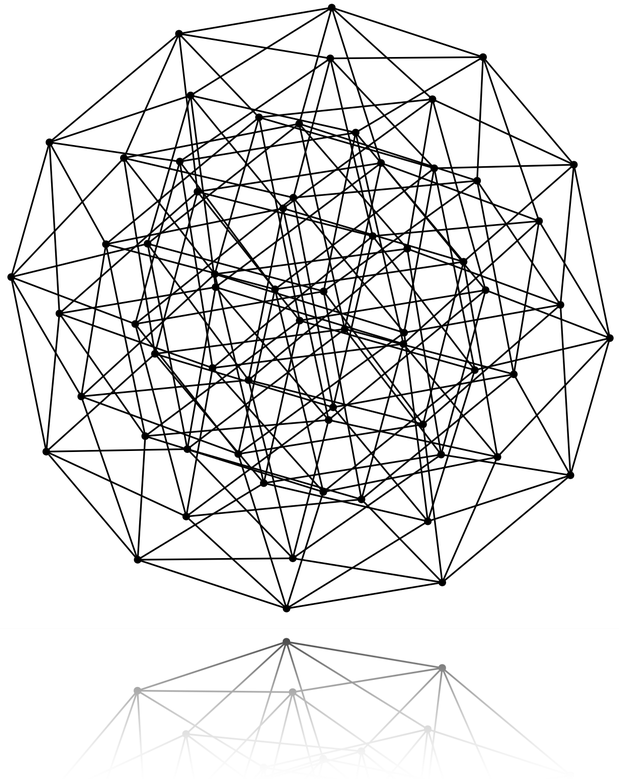
Degree-diameter problem

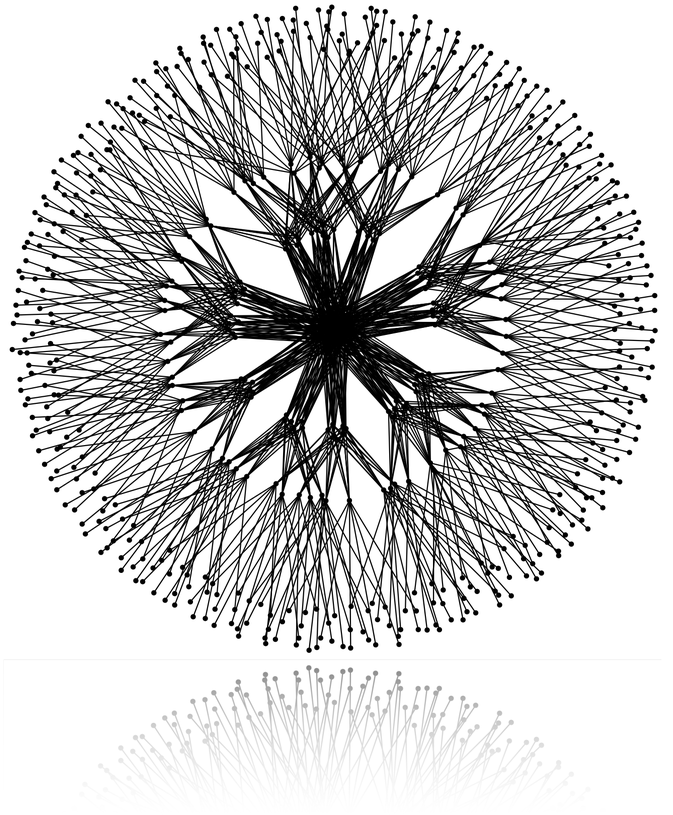
Just use the best-known degree-diameter graphs?!

Problem: Lack of flexibility

Today’s structured networks



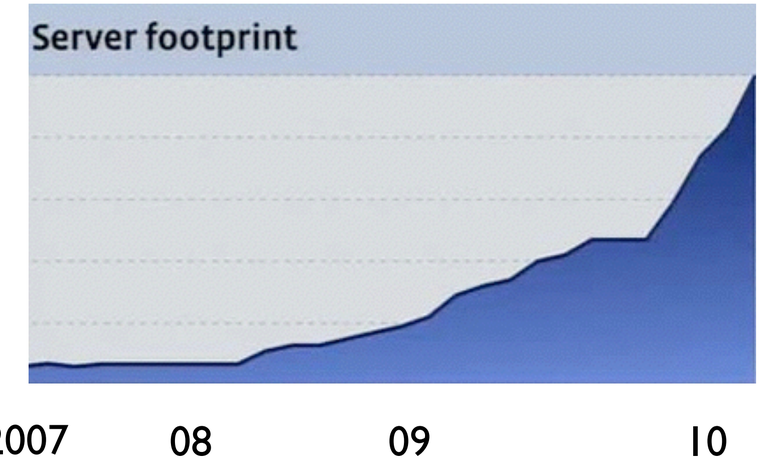




Lack of flexibility …

Coarse design points

- Hypercube: 2^k switches
- de Bruijn-like: 3^k switches
- 3-level fat tree: $5k^2/4$ switches



Facebook "adding capacity on a daily basis"

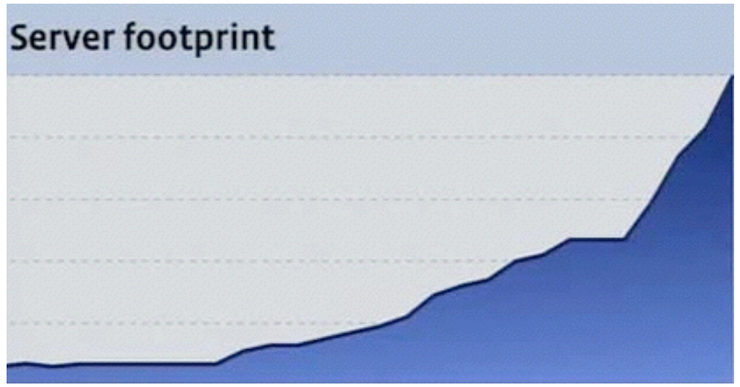
Unclear how to maintain structure incrementally

- Over-utilize switches? Uneven / constrained bandwidth
- Leave ports free for later? Wasted investment

Lack of flexibility …

Coarse design points

- Hypercube: 2^k identical switches
- de Bruijn-like: 3^k identical switches
- 3-level fat tree: $5k^2/4$ identical switches



Facebook "adding capacity on a daily basis"

Unclear how to maintain structure incrementally

- Over-utilize switches? Uneven / constrained bandwidth
- Leave ports free for later? Wasted investment

How do we handle heterogeneity?

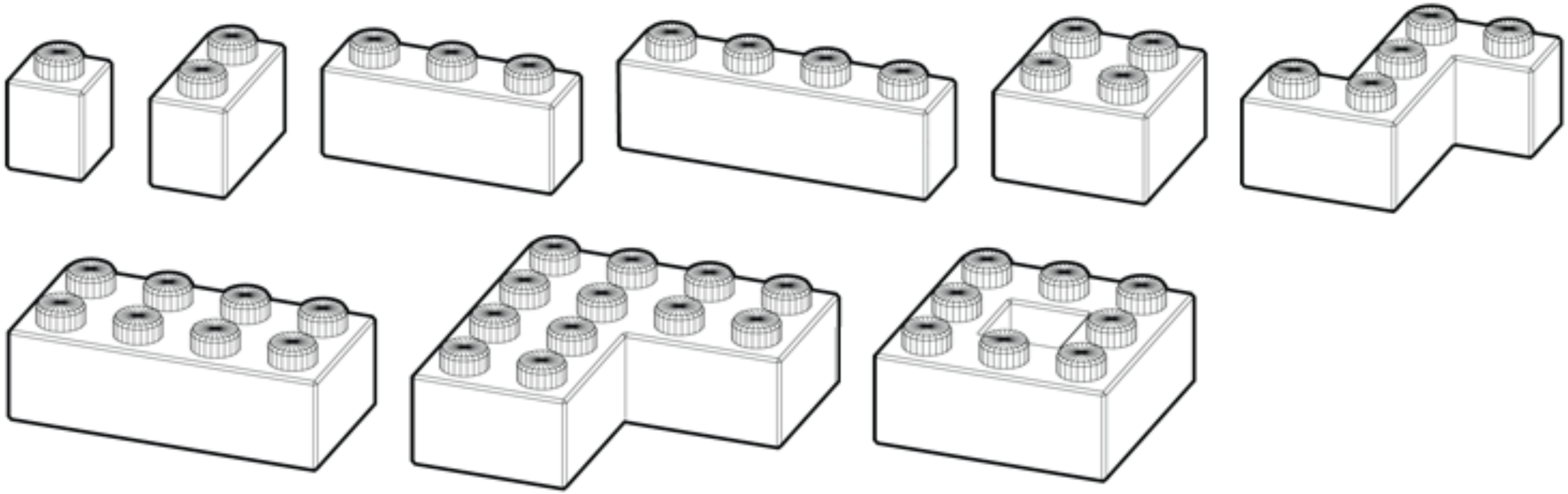
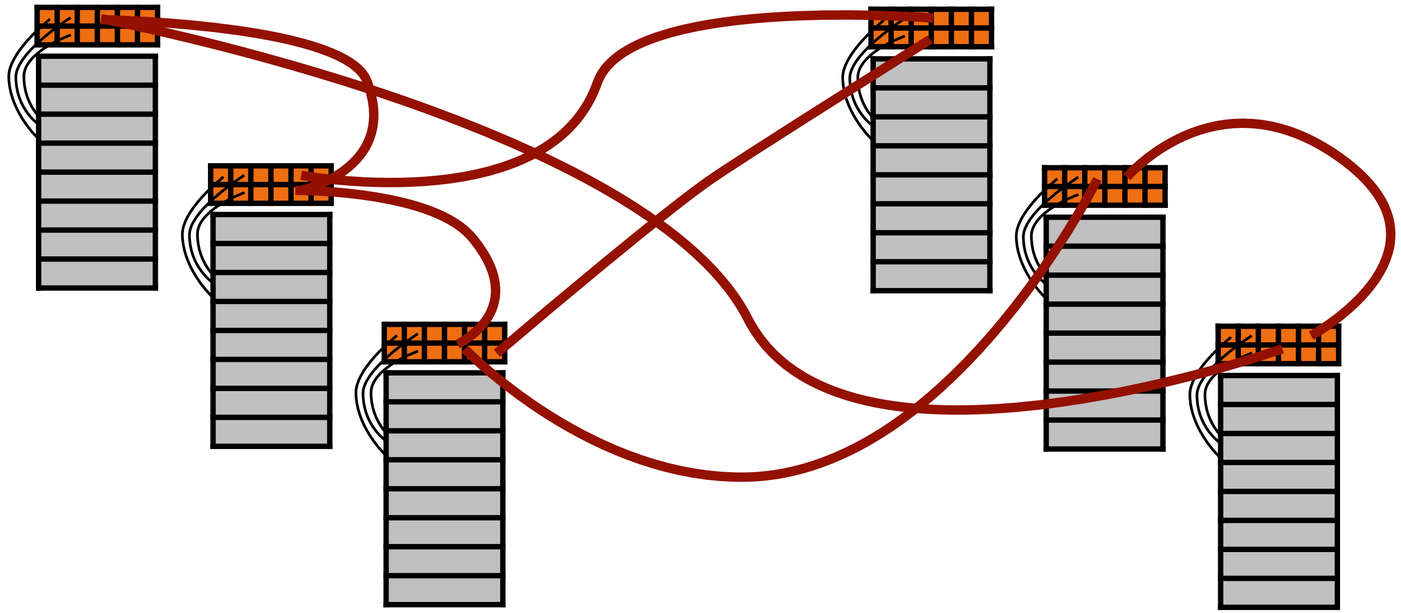


Image credit: Legolizer (www.drububu.com)

Forget about structure –
let's have no structure at all!

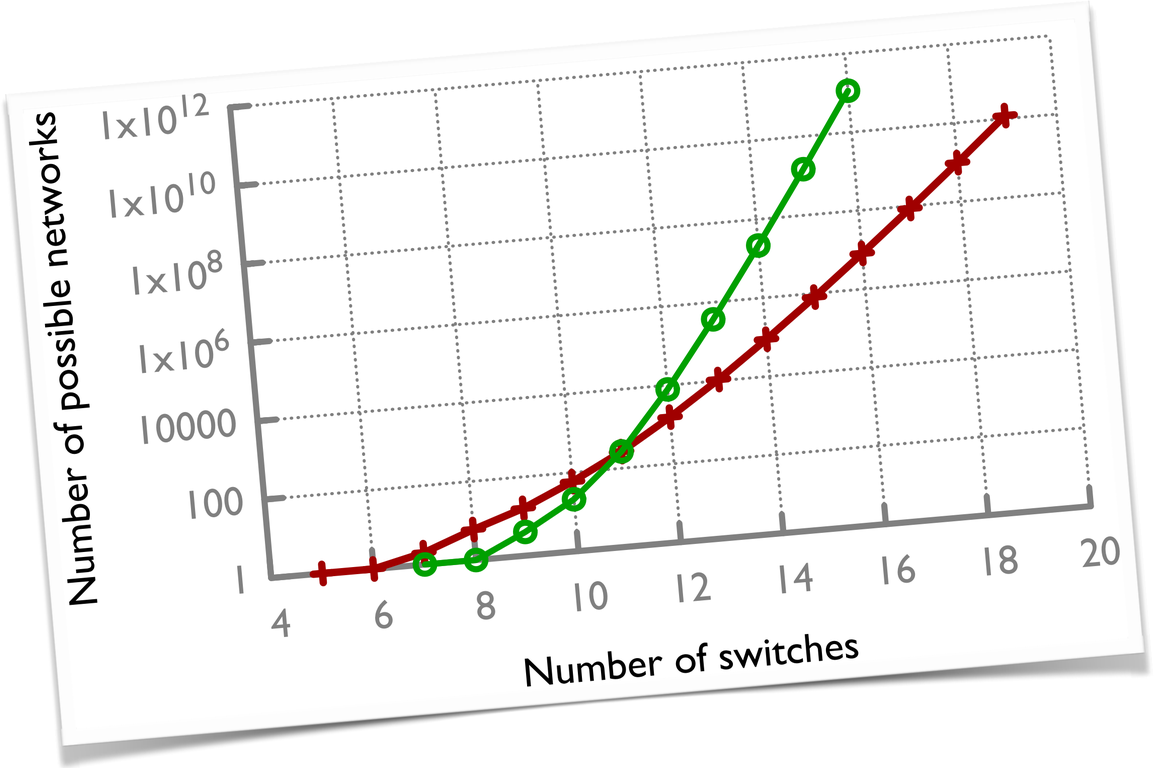
Jellyfish



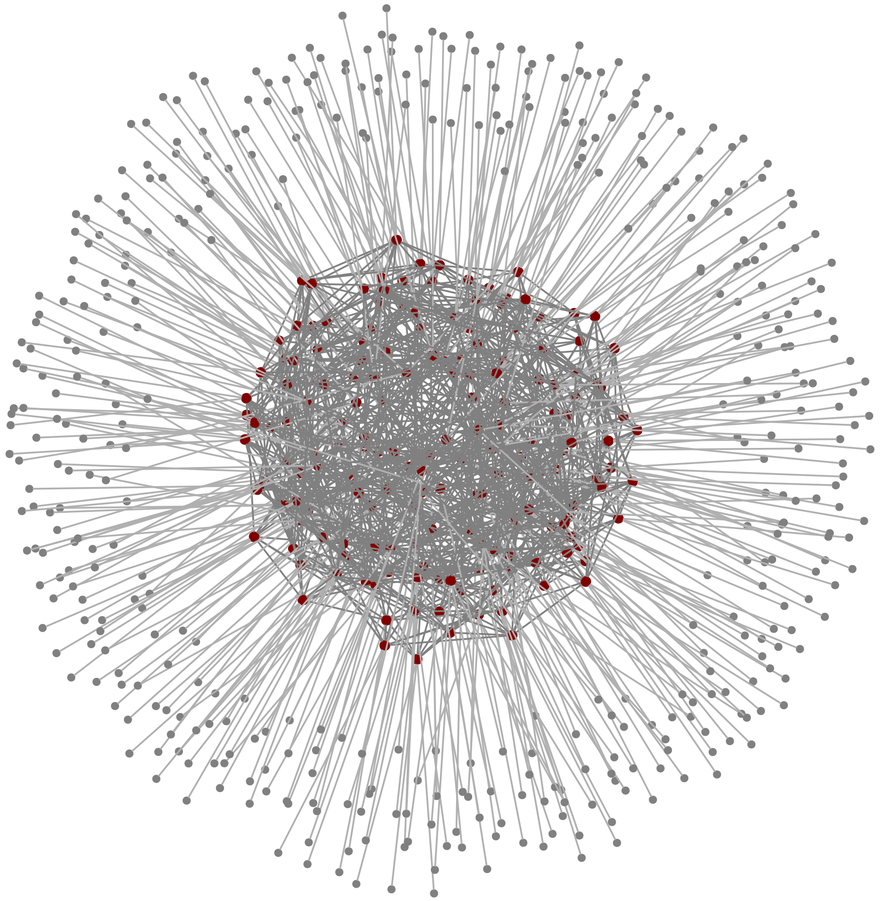
Servers connected to top-of-rack switch

Switches form uniform-random interconnections

Out of this large space, pick uniformly at random!



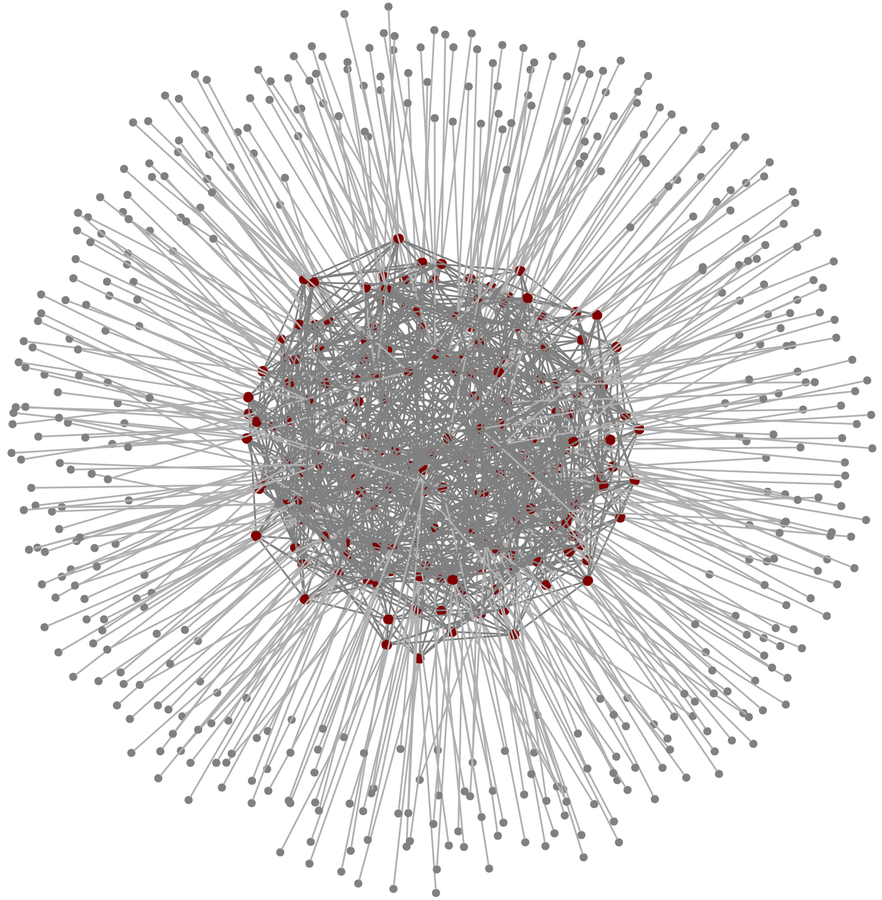
Capacity as a fluid



Jellyfish random graph

432 servers, 180 switches, degree 12

Capacity as a fluid



Jellyfish random graph

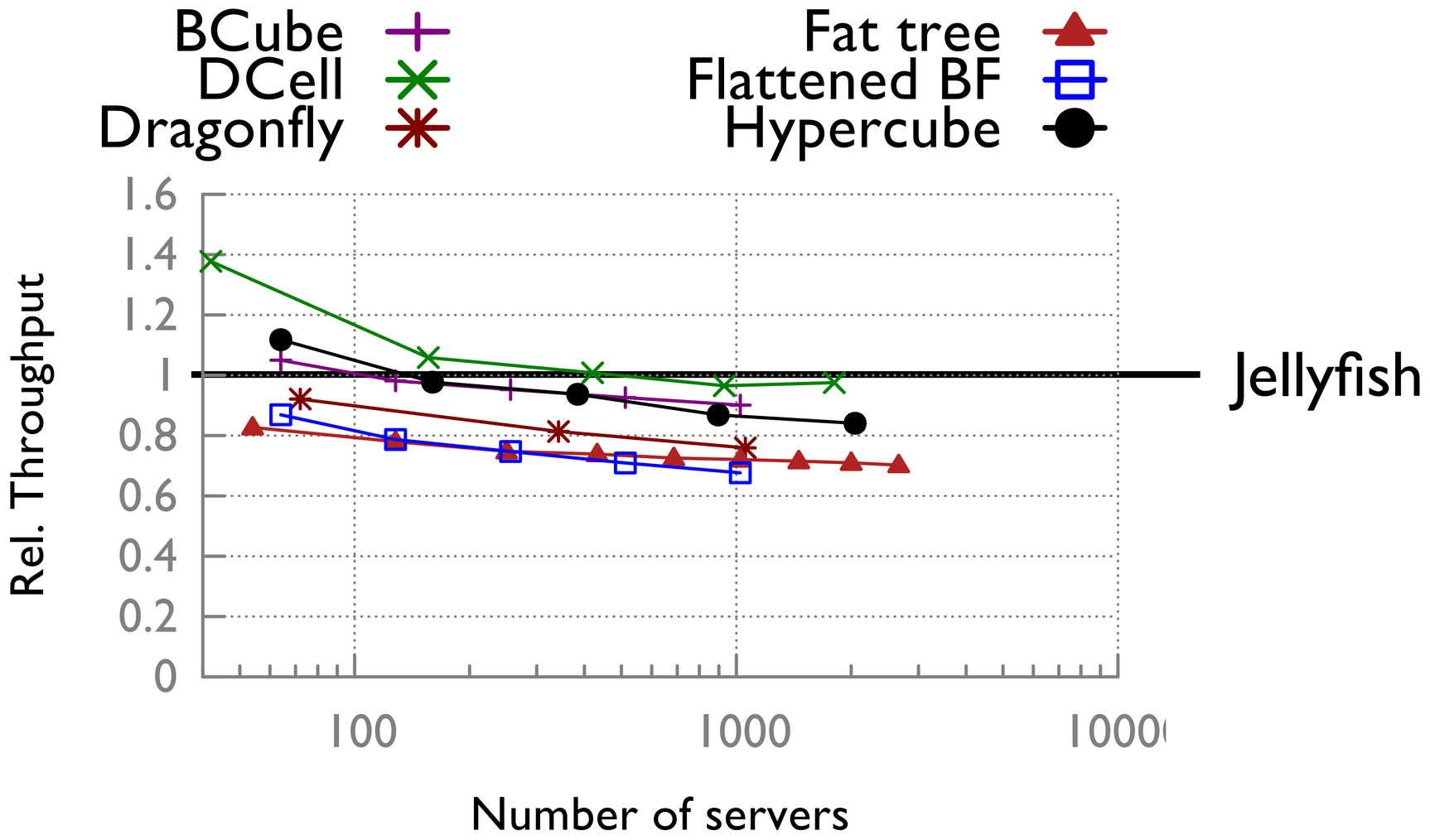
432 servers, 180 switches, degree 12

Jellyfish

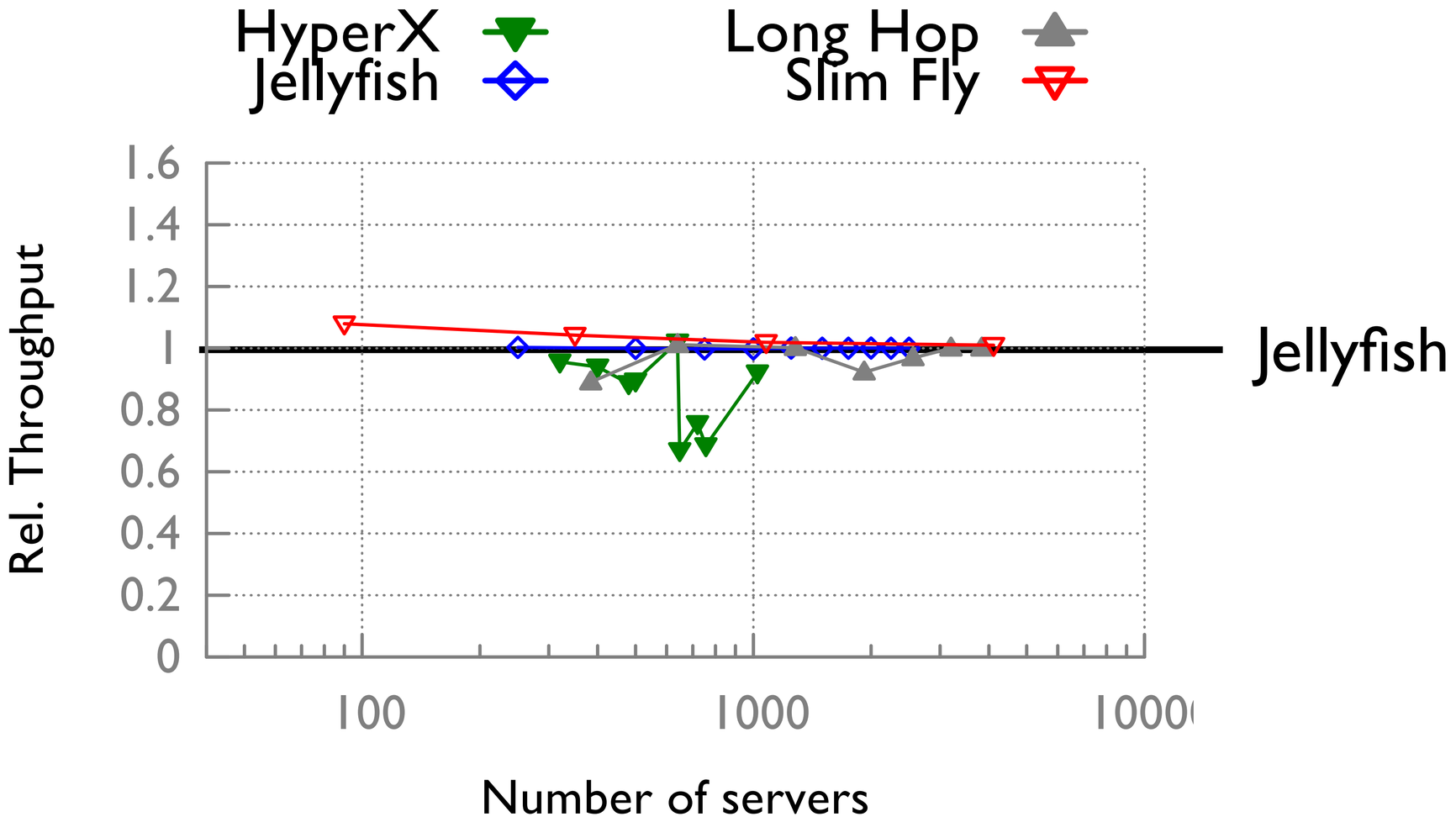
Crossota norvegica

Photo: Kevin Raskoff

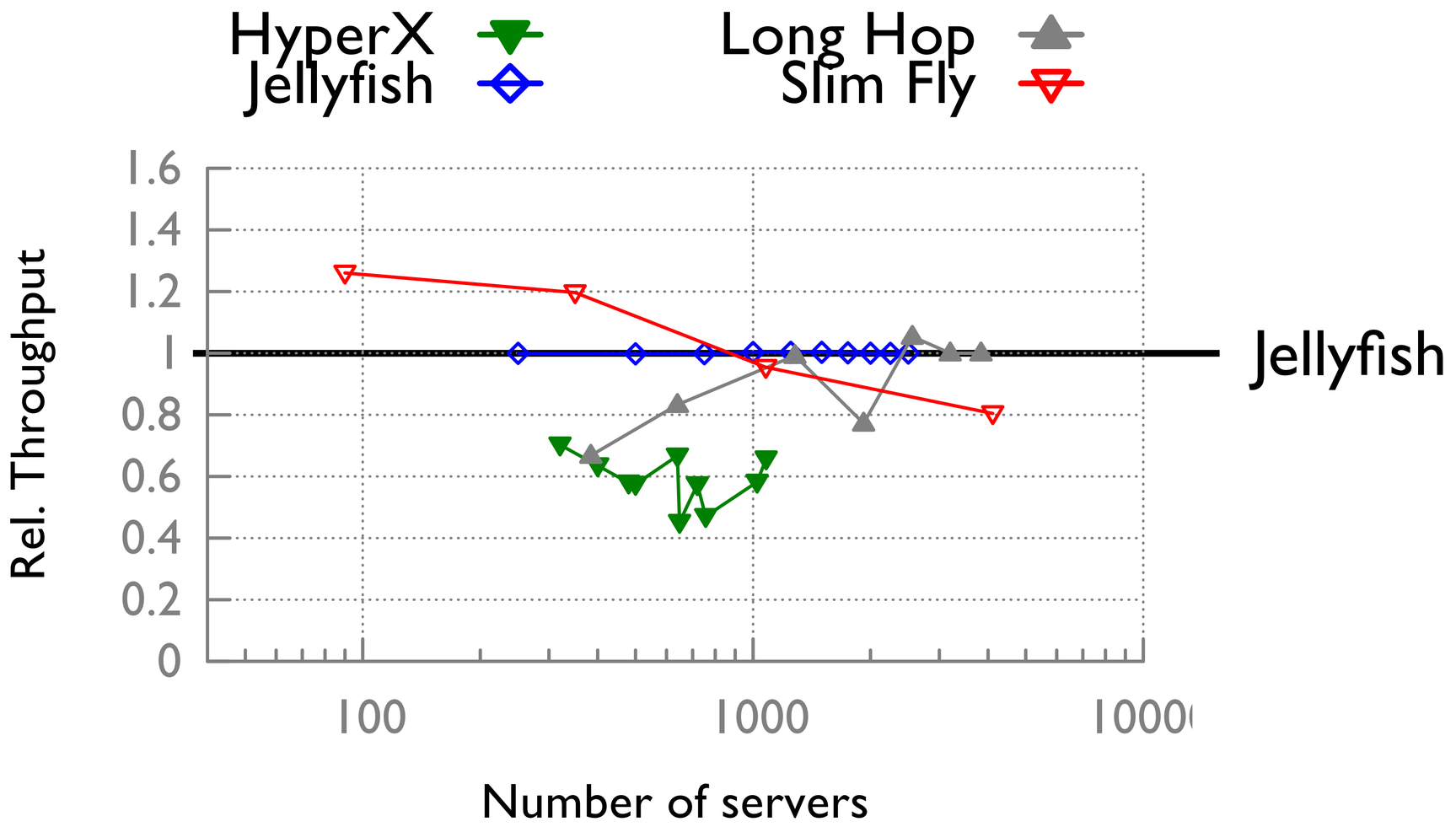
Really? Random could work?



Really? Random could work?



Really? Random could work?



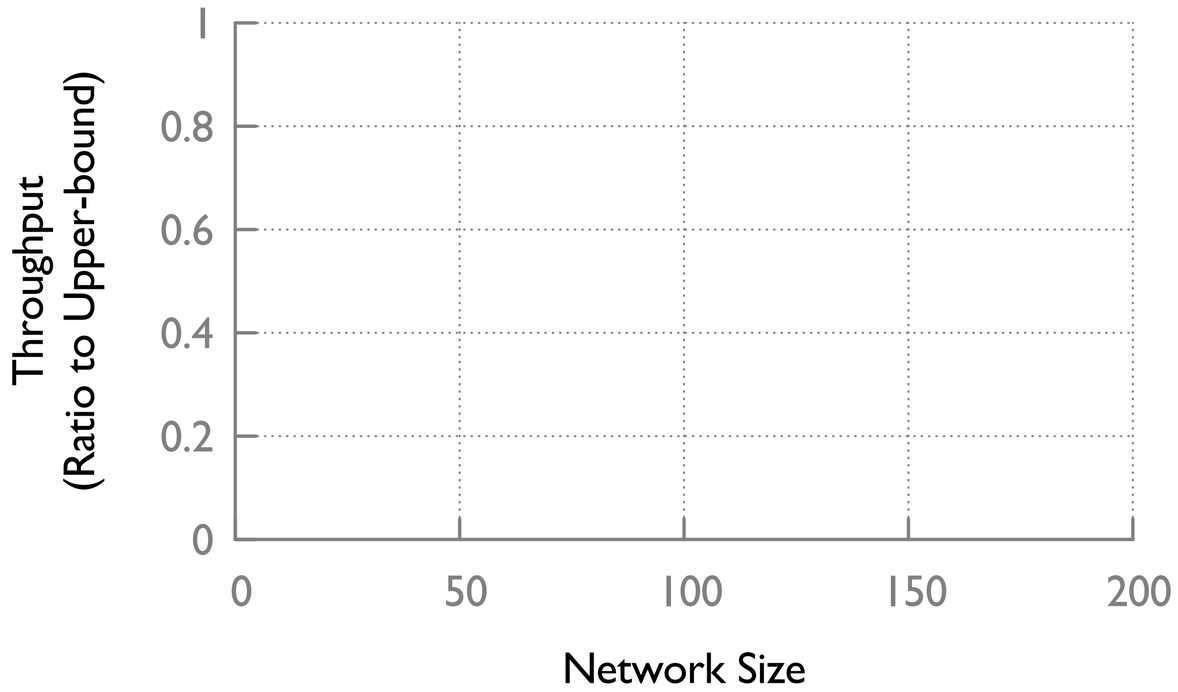
In fact, nothing will do much better!

A simple upper bound on throughput

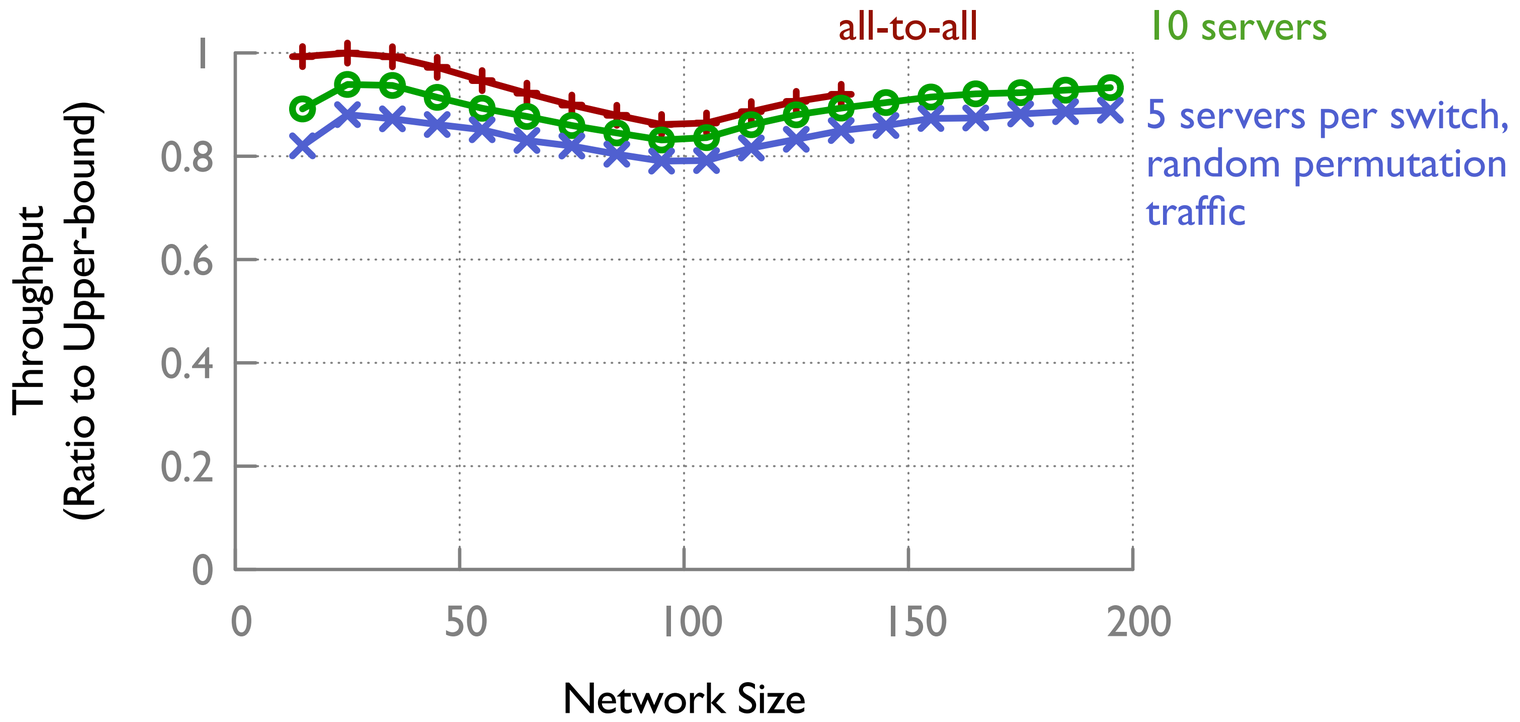
$$\text{throughput per flow} \leq \frac{\sum_{\text{links}} \text{capacity}(\textit{link})}{\#\text{ flows} \bullet \text{mean path length}}$$

Lower bound this!

Random graphs vs. bound

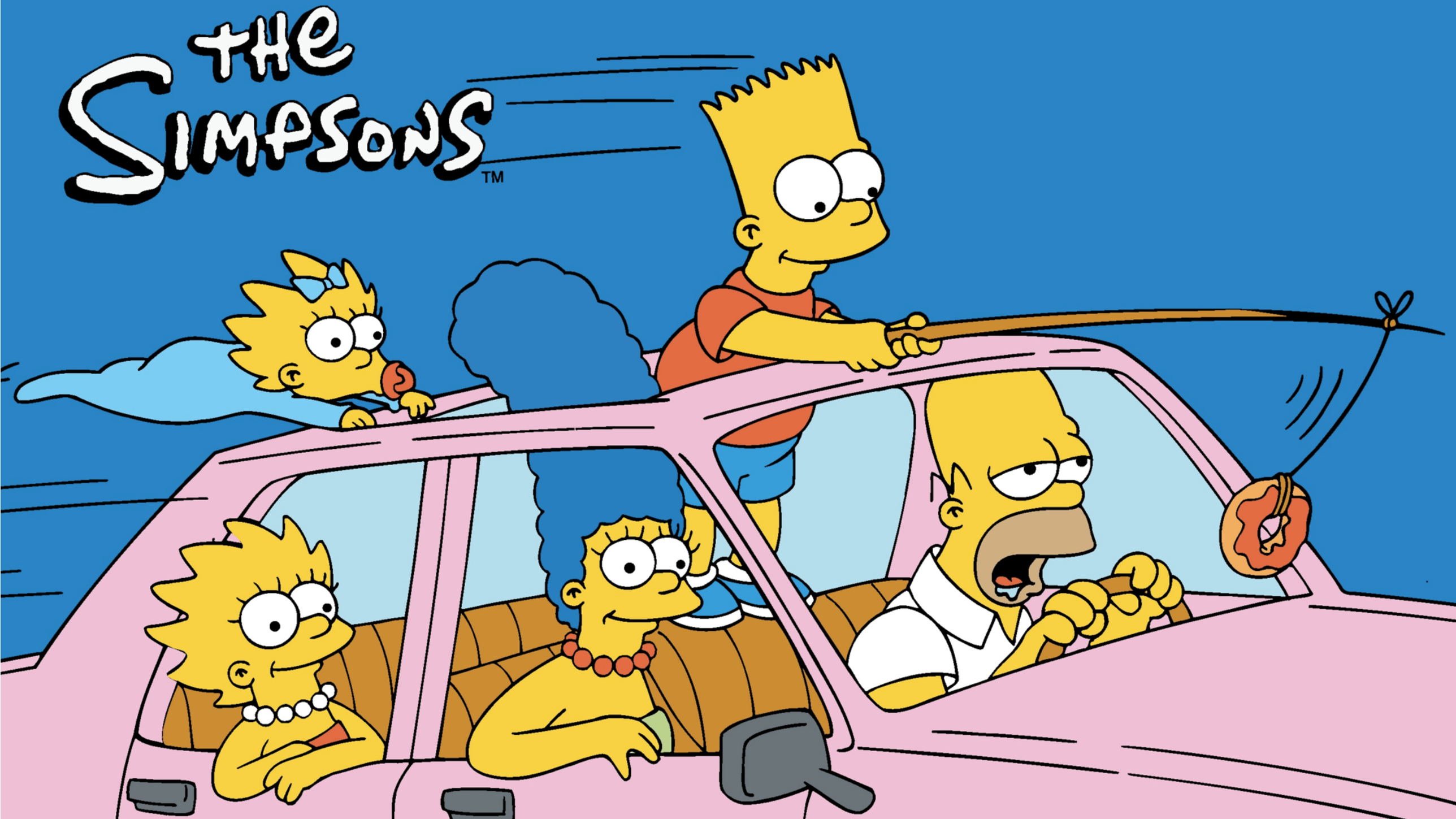


Random graphs vs. bound



Random graphs within a few percent of optimal!

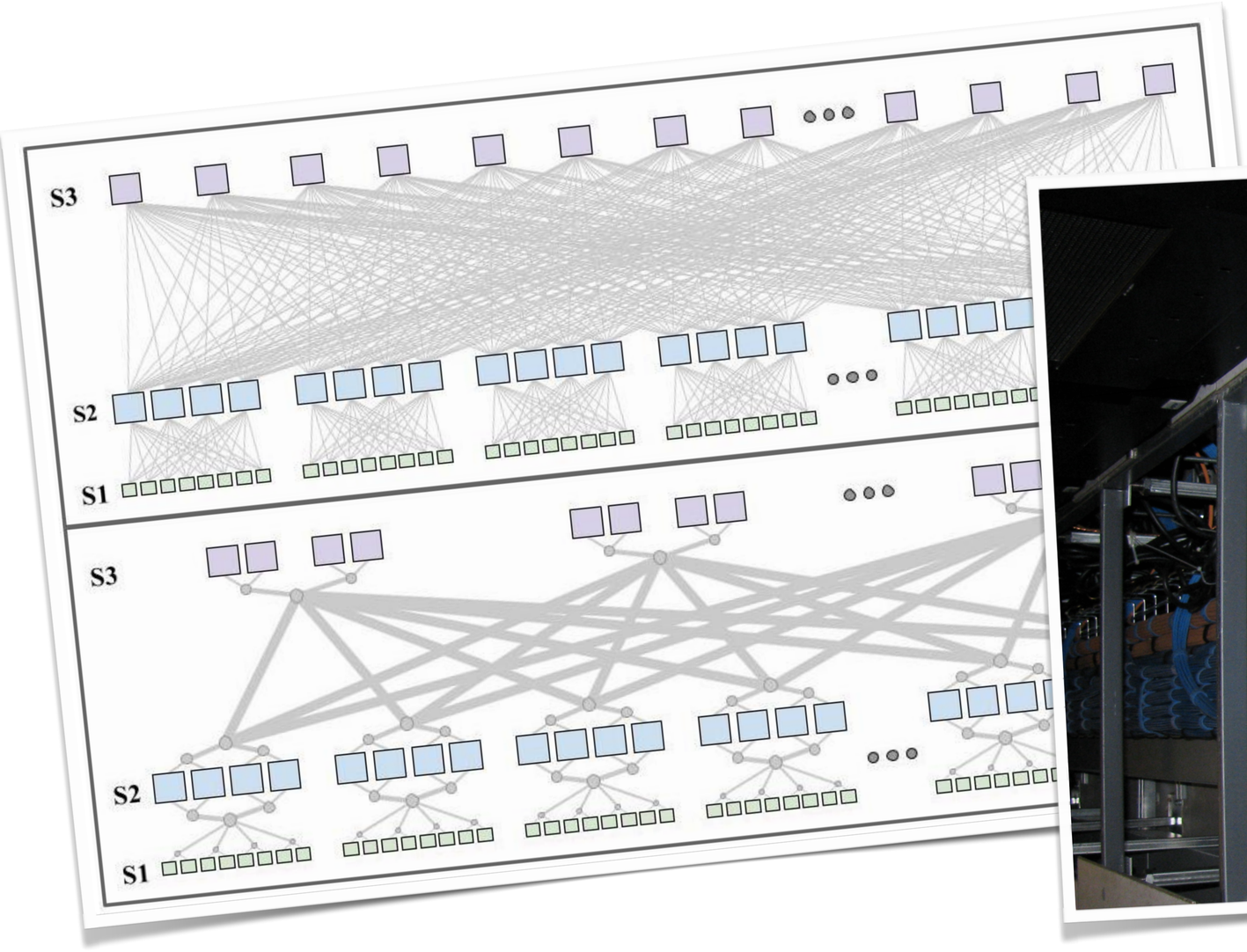

These stunts are performed by trained professionals ...

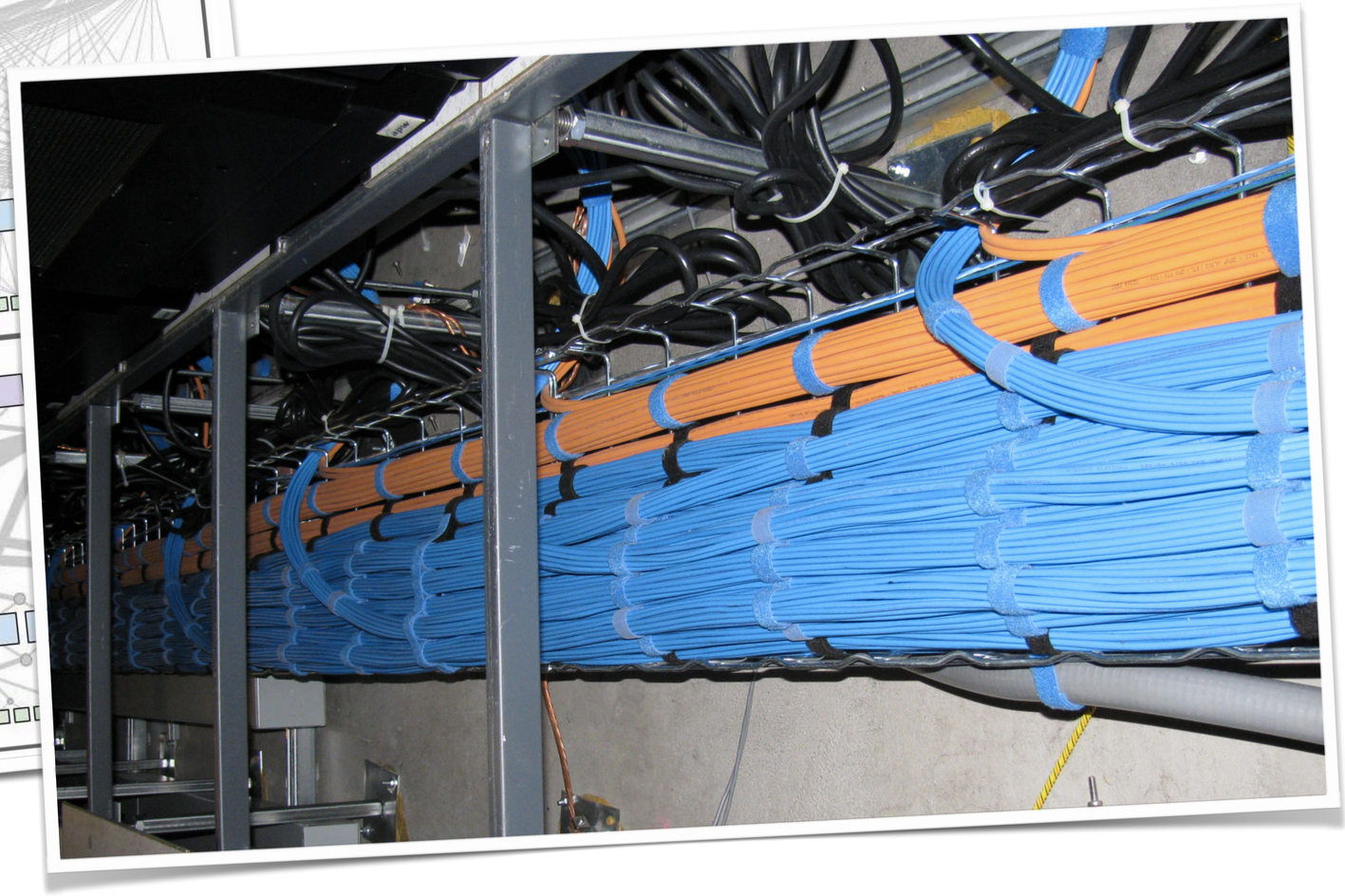

THE SIMPSONS
TM

Impact so far …

Ankit Singla

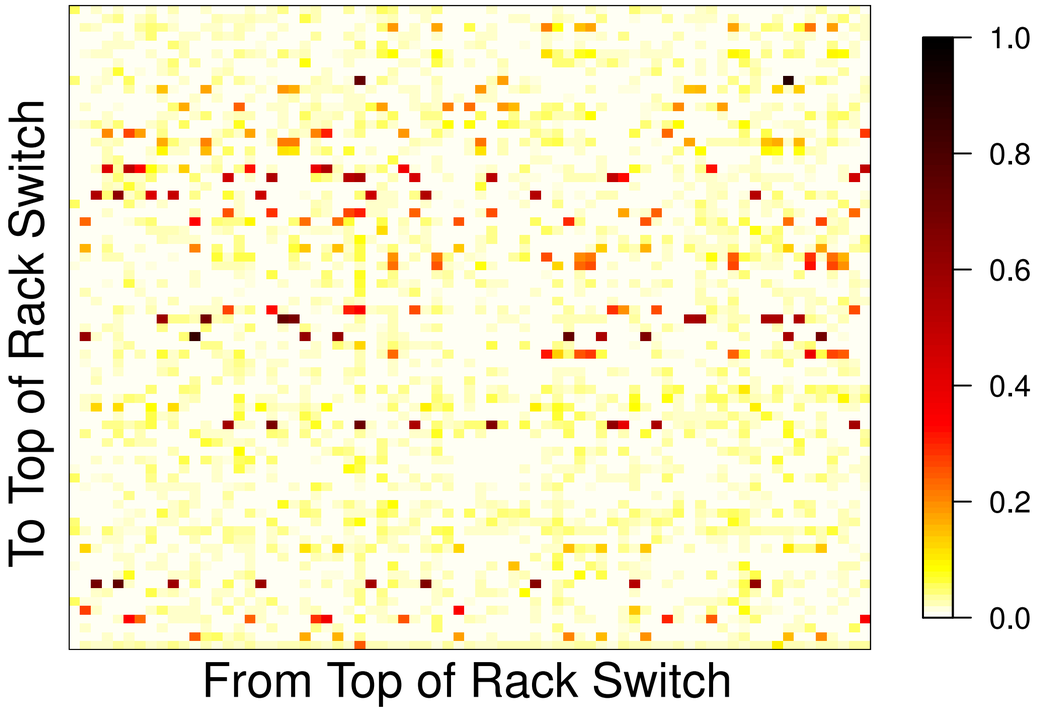
Assistant Professor, CS / Systems, ETH
PhD, Computer Science, UIUC, 2015

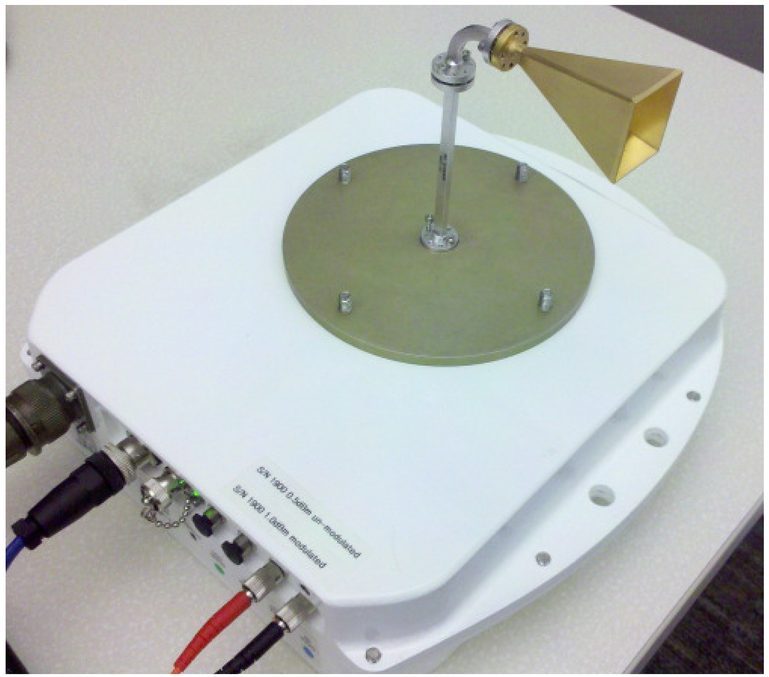

S3
S2
S1
S3
S2
S1



Augmenting Data Center Networks with Multi-Gigabit Wireless Links

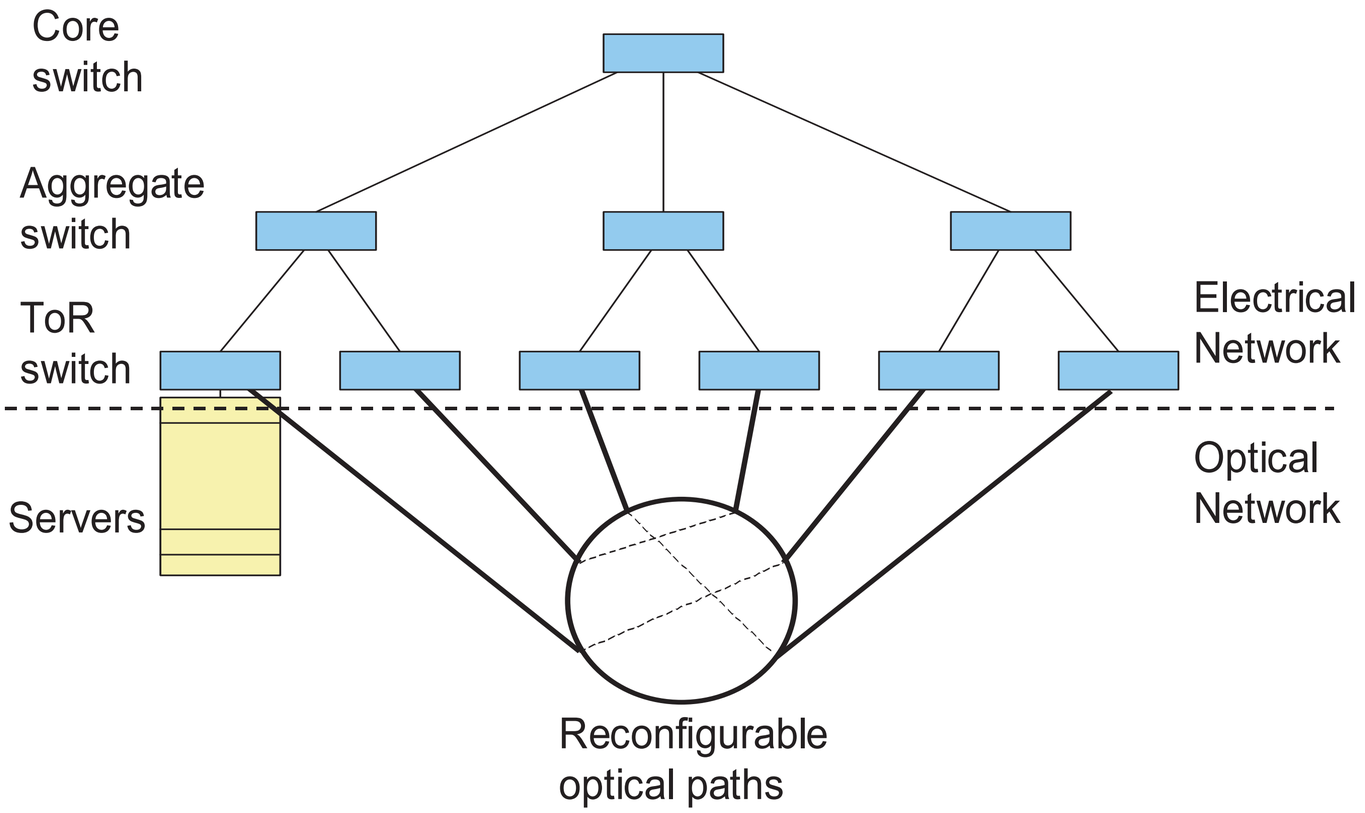
Daniel Halperin*†, Srikanth Kandula†, Jitendra Padhye†, Paramvir Bahl†, and David Wetherall*
Microsoft Research† and University of Washington*





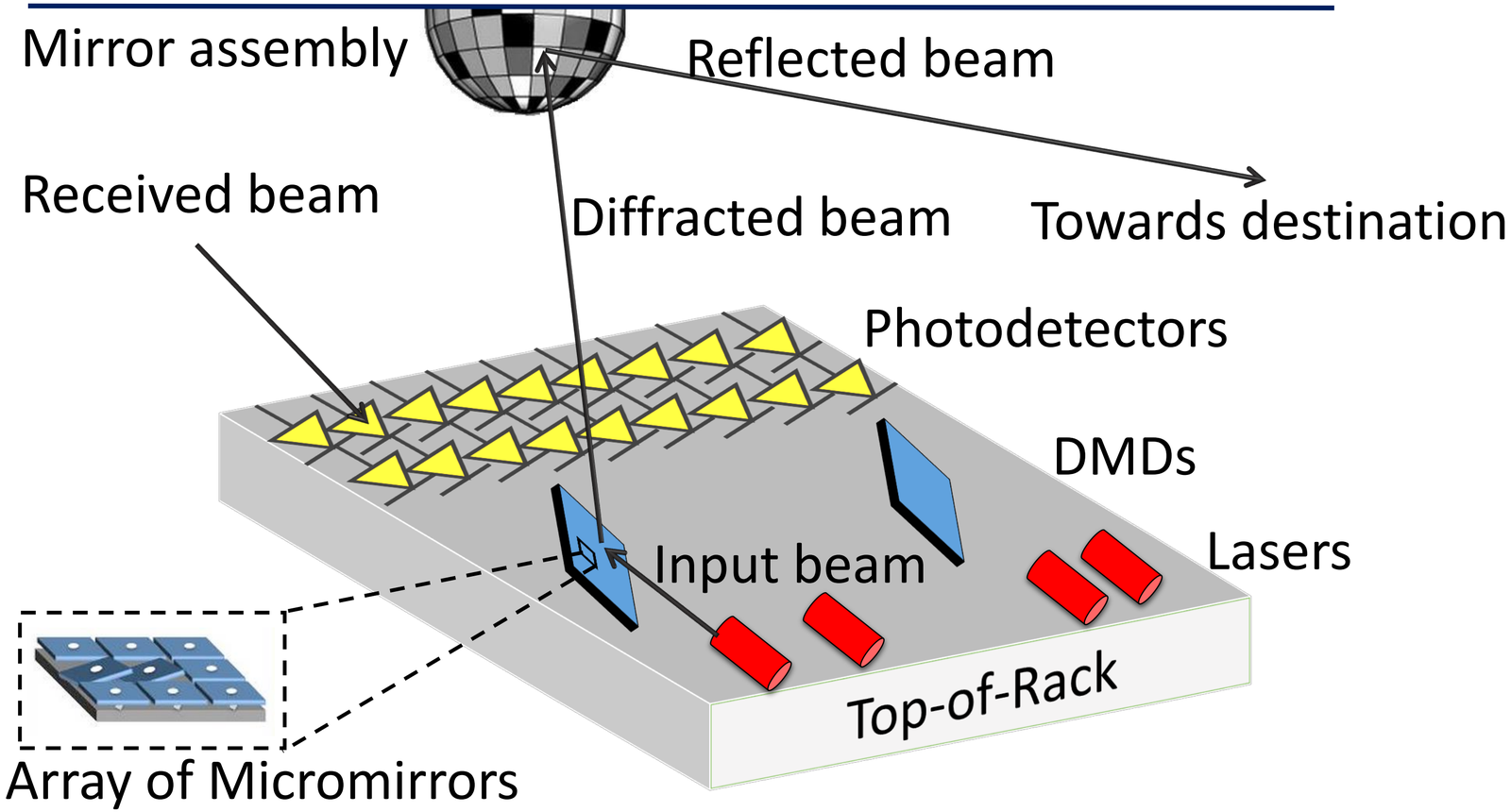
c-Through: Part-time Optics in Data Centers

Guohui Wang*, David G. Andersen†, Michael Kaminsky‡, Konstantina Papagiannaki‡,
T. S. Eugene Ng*, Michael Kozuch‡, Michael Ryan‡
*Rice University, †Carnegie Mellon University, ‡Intel Labs Pittsburgh



ProjecToR: Agile Reconfigurable Data Center Interconnect

Monia Ghobadi Ratul Mahajan Amar Phanishayee
Nikhil Devanur Janardhan Kulkarni Gireeja Ranade
Pierre-Alexandre Blanche† Houman Rastegarfar† Madeleine Glick† Daniel Kilper†
Microsoft Research †University of Arizona

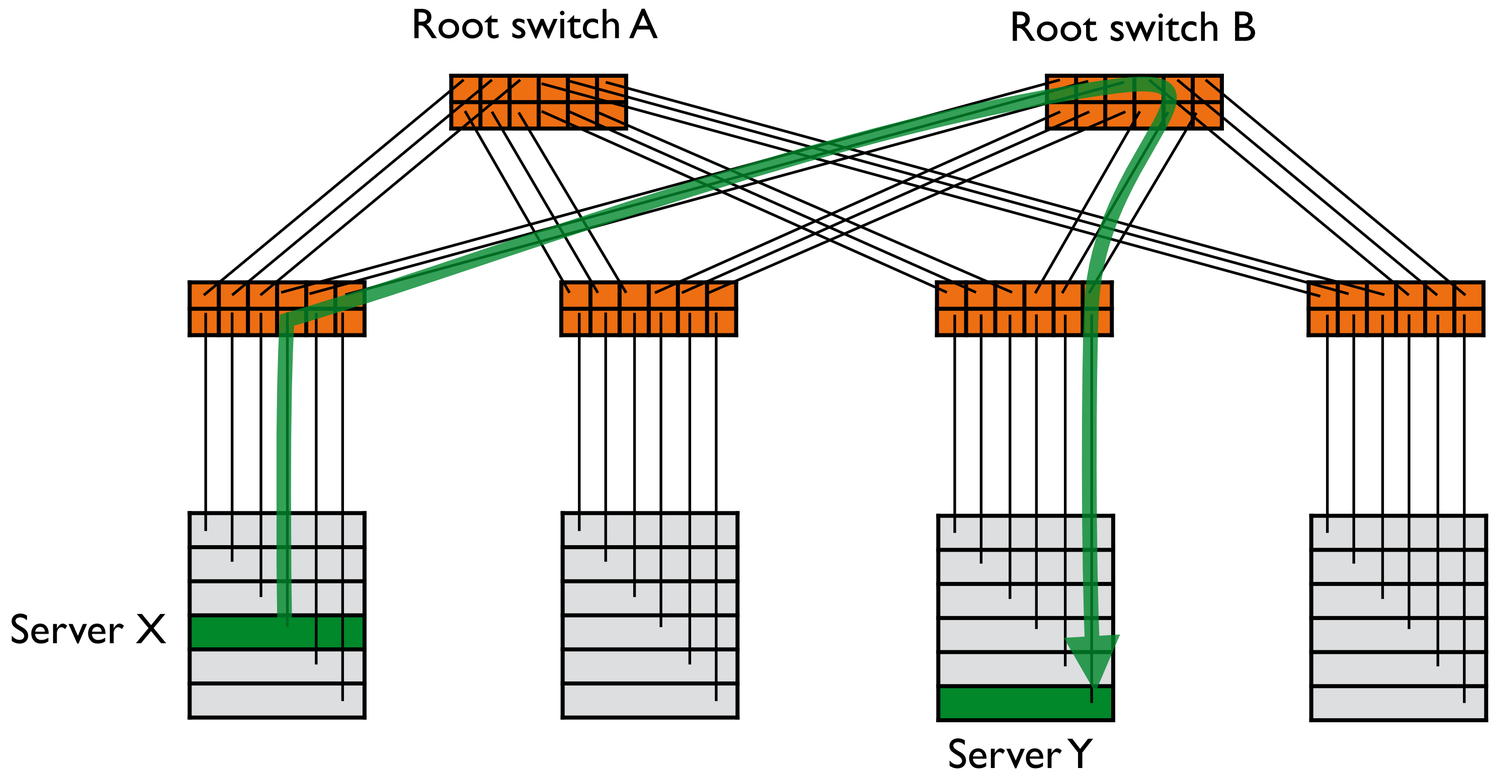


Next lecture …

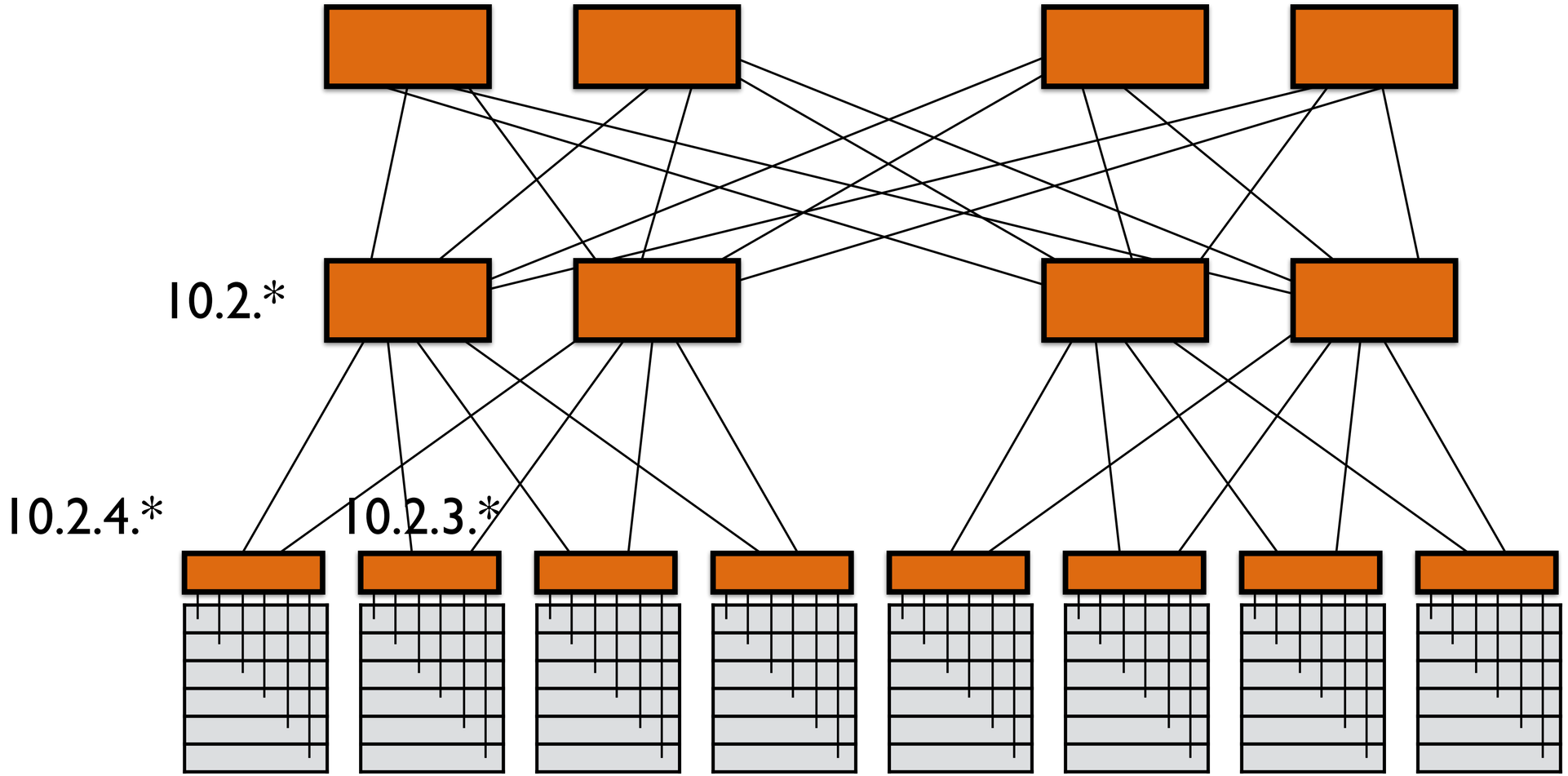
How do we do routing?

Useful tools for routing

1 Exploiting structure in the network

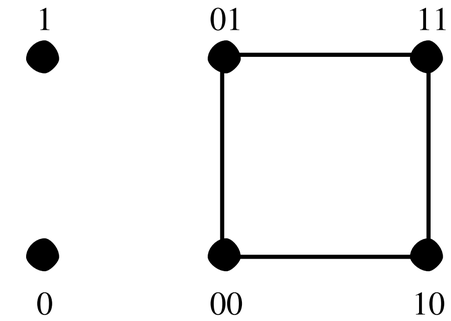


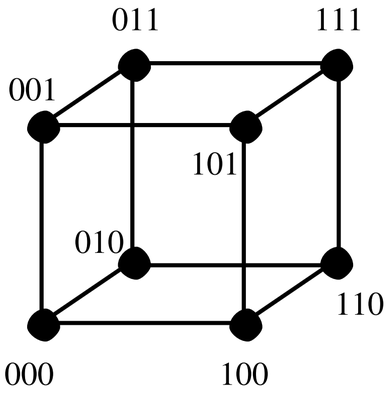
Useful tools for routing

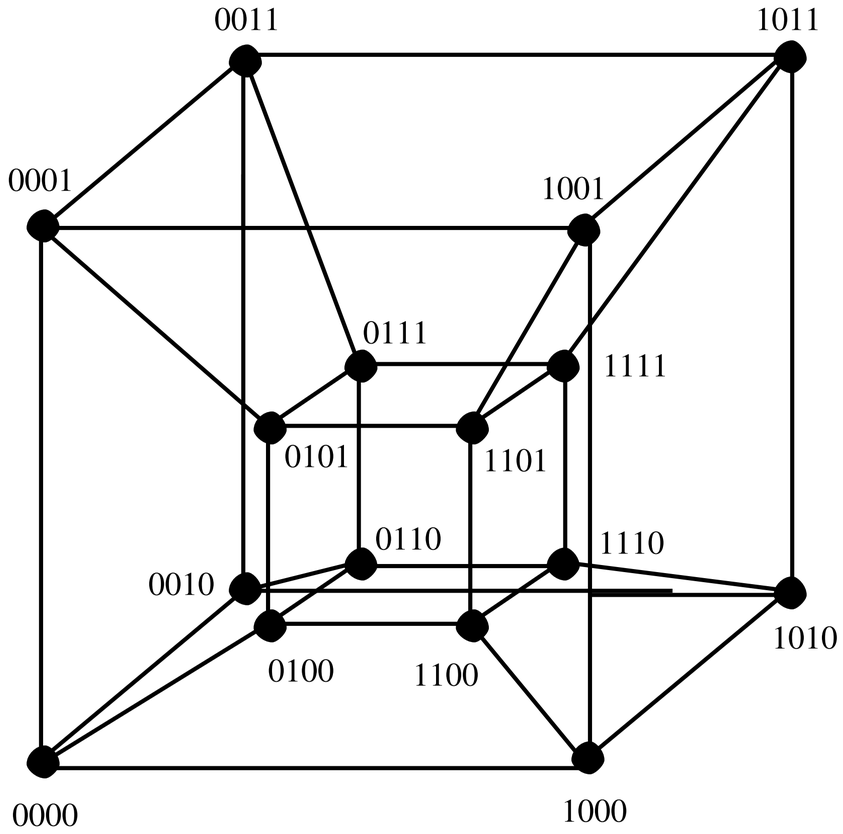


Useful tools for routing

2 Incorporate routing information in node names





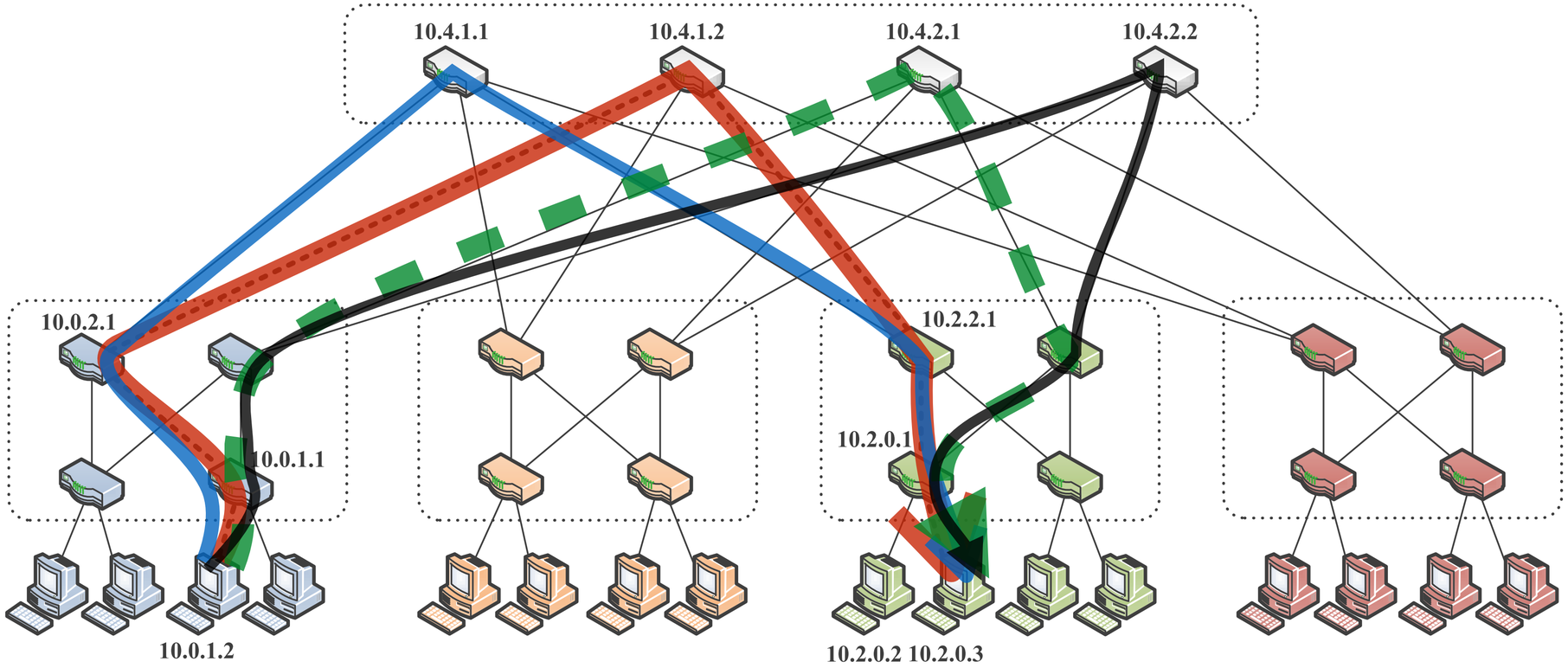


Nikolopoulos, "Optimal Gray-code labeling and recognition algorithms for hypercubes", Elsevier 2000.

Routing: an information encoding problem

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

Forwarding: *using encoded routing information*



ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat

Bleeding-edge of DC research!

Under submission, 2017

Beyond fat-trees without antennae, mirrors, and disco-balls

Simon Kassing

Weekly reading guide

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