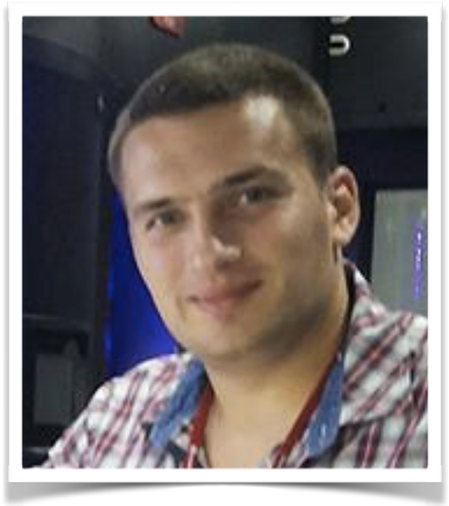
Data centers: apps & traffic

Ankit Singla

ETH Zürich Spring 2017

Email Vojislav to join us on Slack



Vojislav Đukić (vdukic)

- Why or why not source routing in DCs?
- Google's new congestion control design

About the assignments …

- Programming assignments (graded)
- Written assignments (not graded)
  - Hand-in Tue or mail-in Wed
  - Group feedback, not individual (ML E12)
  - Ask questions


Me


You

More seriously …

- Pains of scale …
- Goals of exercises
  - Solidify / practice lecture concepts
  - "What will the exam look like?"
- We encourage you to ask questions
  - Yes, even silly ones!

Hey, where do baby cats come from?!

Data centers, obviously!!

This lecture …

- Introduction to data center networking
- What do DC applications look like?
- What does "typical" DC traffic look like?

- Intro: "Data center topologies & routing"
  - Reading: A Scalable, Commodity Data Center Network Architecture




Dropbox

coursera

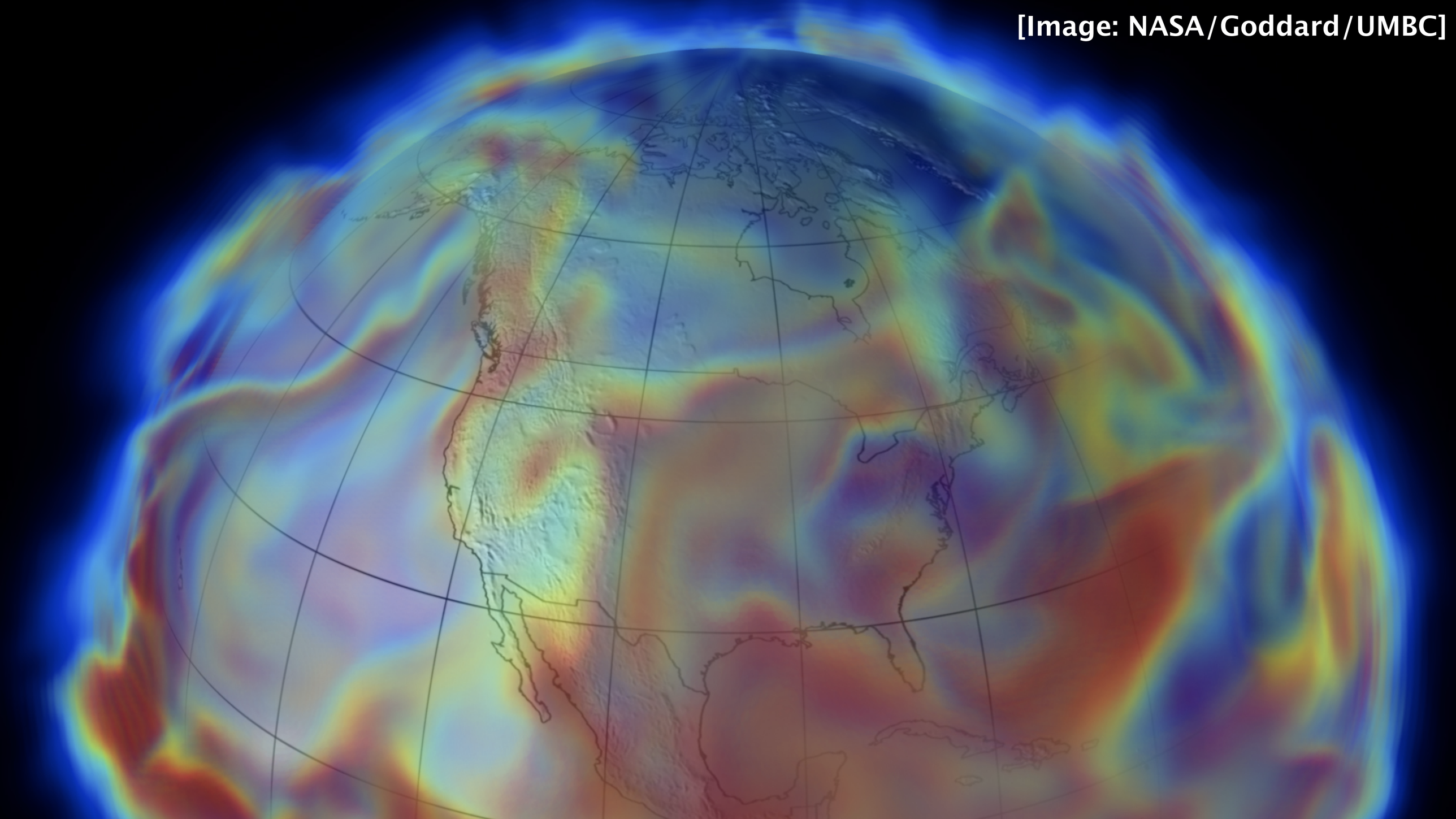


amazon

Google

bing

NETFLIX


[Image: NASA/Goddard/UMBC]

How a Web search works

How a Web search works

How a Web search works

Scatter-gather traffic pattern

Extremely short response deadlines for each server — 10ms

"Up to 150 stages, degree of 40, path lengths of 10 or more"

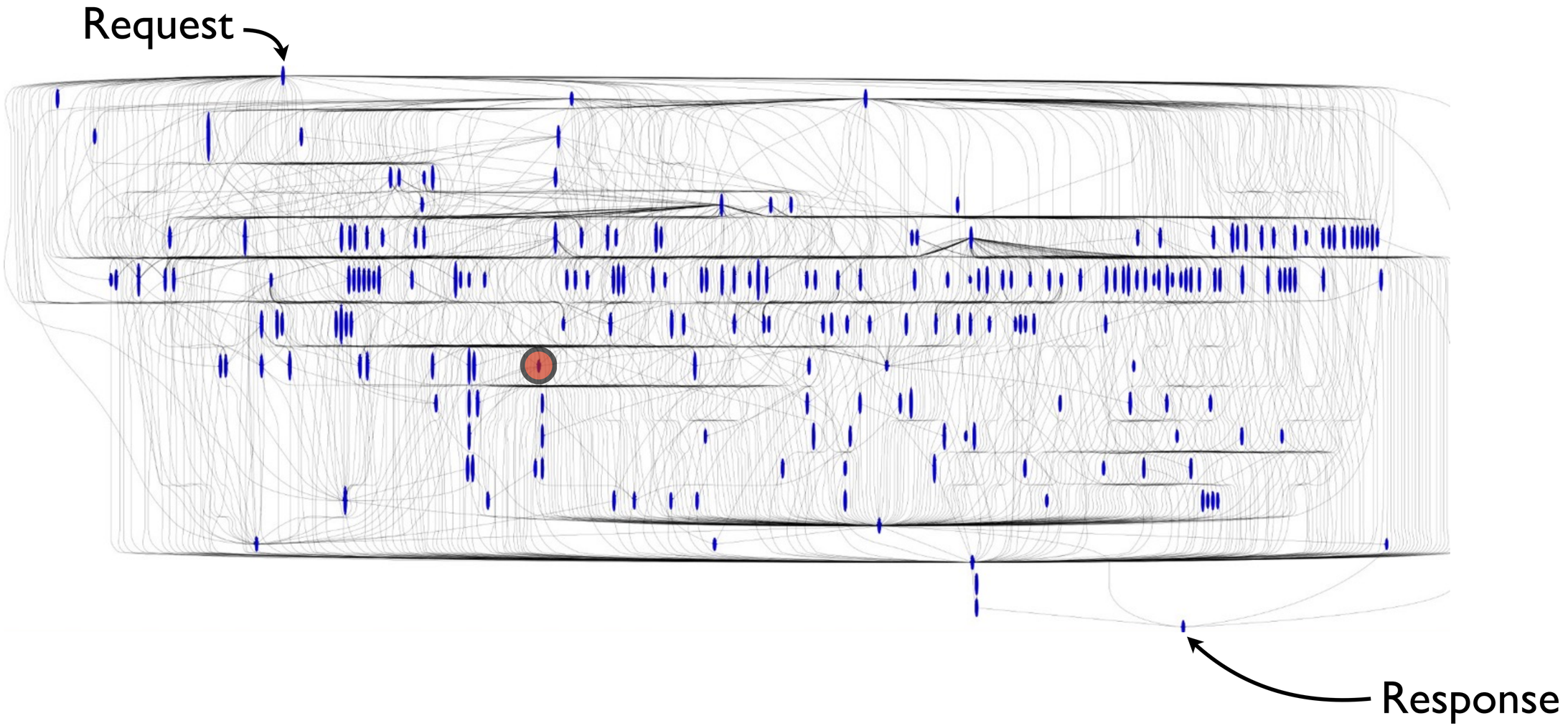


Image source: Talk on "Speeding up Distributed Request-Response Workflows"
by Virajith Jalaparti at ACM SIGCOMM'13

Benson Kua [CC BY-SA 2.0] via Wikimedia

Other Web application traffic

One popular page loaded ⇒ average of **521** distinct `memcache` fetches

95th percentile: **1740** distinct `memcache` fetches

Big data analytics

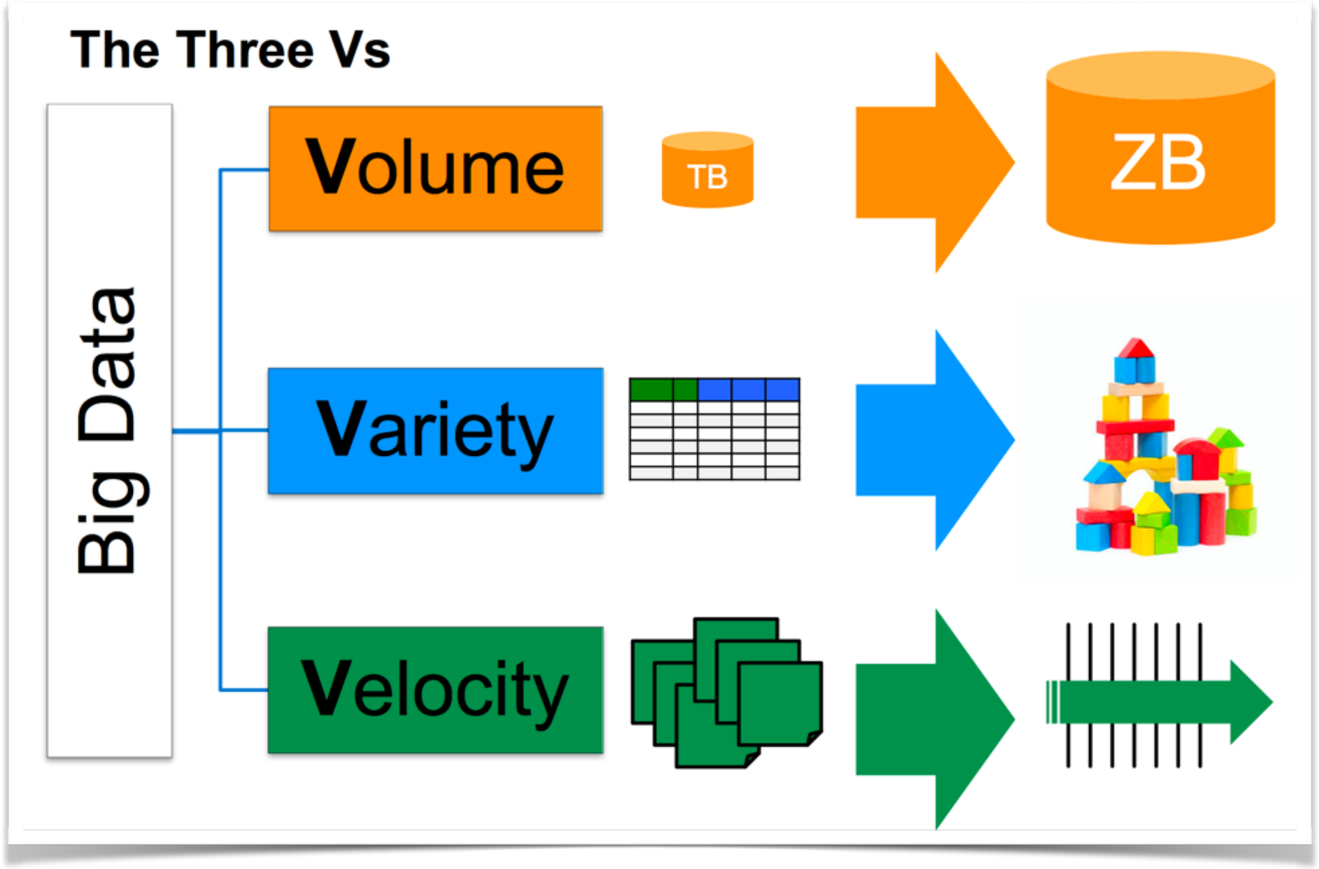
Hadoop

Spark

Database *joins*

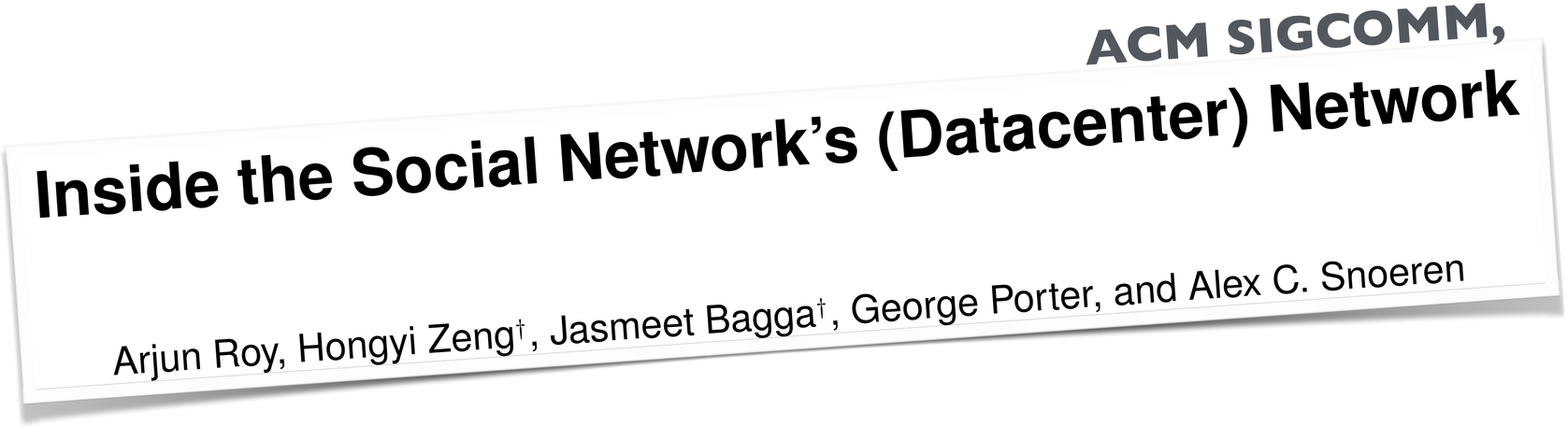
⋮

Big Data class …



What does data center traffic look like?

Data center applications and traffic



- Only one data point
- How did they collect measurements?
- Workload + management = observations?
- Data is available to play with; link on course-page

How do we measure DCs?

- Log everything all the time?!
  Expensive / infeasible (?)
- Log with sampling?!
  Samples might miss problems
- Re-play with logging ON?!
  Useless for “Heisenbugs”

Maybe you could do this at the servers?

Sidenote: packet processing in software

Flexible **slow, CPU-expensive**

10 Gbps, 84 Byte packets ⇒ 67 ns time budget

Context: CPU-memory takes tens of ns

Sidenote: packet processing in software

Flexible **slow, CPU-expensive**

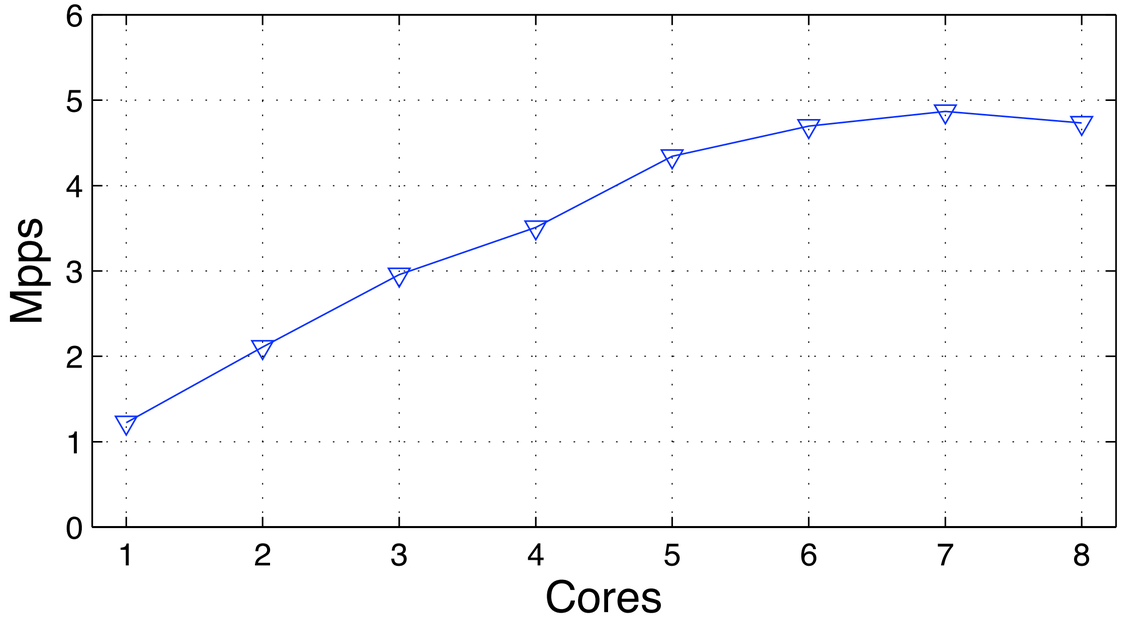
- Packet I/O
- Context-switching overheads
- Packet classification

Sidenote: packet processing in software

Understanding the Packet Processing Capability of Multi-Core Servers

Norbert Egi[‡], Mihai Dobrescu[†], Jianqing Du[†], Katerina Argyraki[†], Byung-Gon Chun[§], Kevin Fall[§], Gianluca Iannaccone[§], Allan Knies[§], Maziar Manesh[§], Laurent Mathy[‡], Sylvia Ratnasamy[§]

[§] Intel Research, [†] EPFL, [‡] Lancaster University



2.5 Gbps

Figure 14: *Per-flow monitoring performance*

How do we measure DCs?

- **IF** you were able to monitor and log every single packet, what would you do?

- What if you knew (a significant chunk of) the traffic in advance?

A LOT of measurement infrastructure, data

PlanetLab

Seattle P2P testbed

RouteViews

DIMES

CAIDA

M-Lab

iPlane

My MOOC!

…

A challenge for you!

The Science Behind Foldit

Foldit is a revolutionary crowdsourcing computer game enabling *you* to contribute to important scientific research. This page describes the science behind Foldit and how your playing can help.

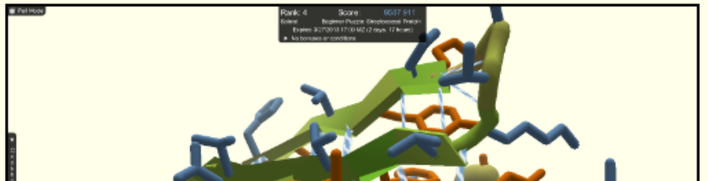
Page Contents:

Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks

The online game poses protein-folding puzzles, and participants provided insights recently that solved the structure of an enzyme involved in reproduction of HIV

What is protein folding?

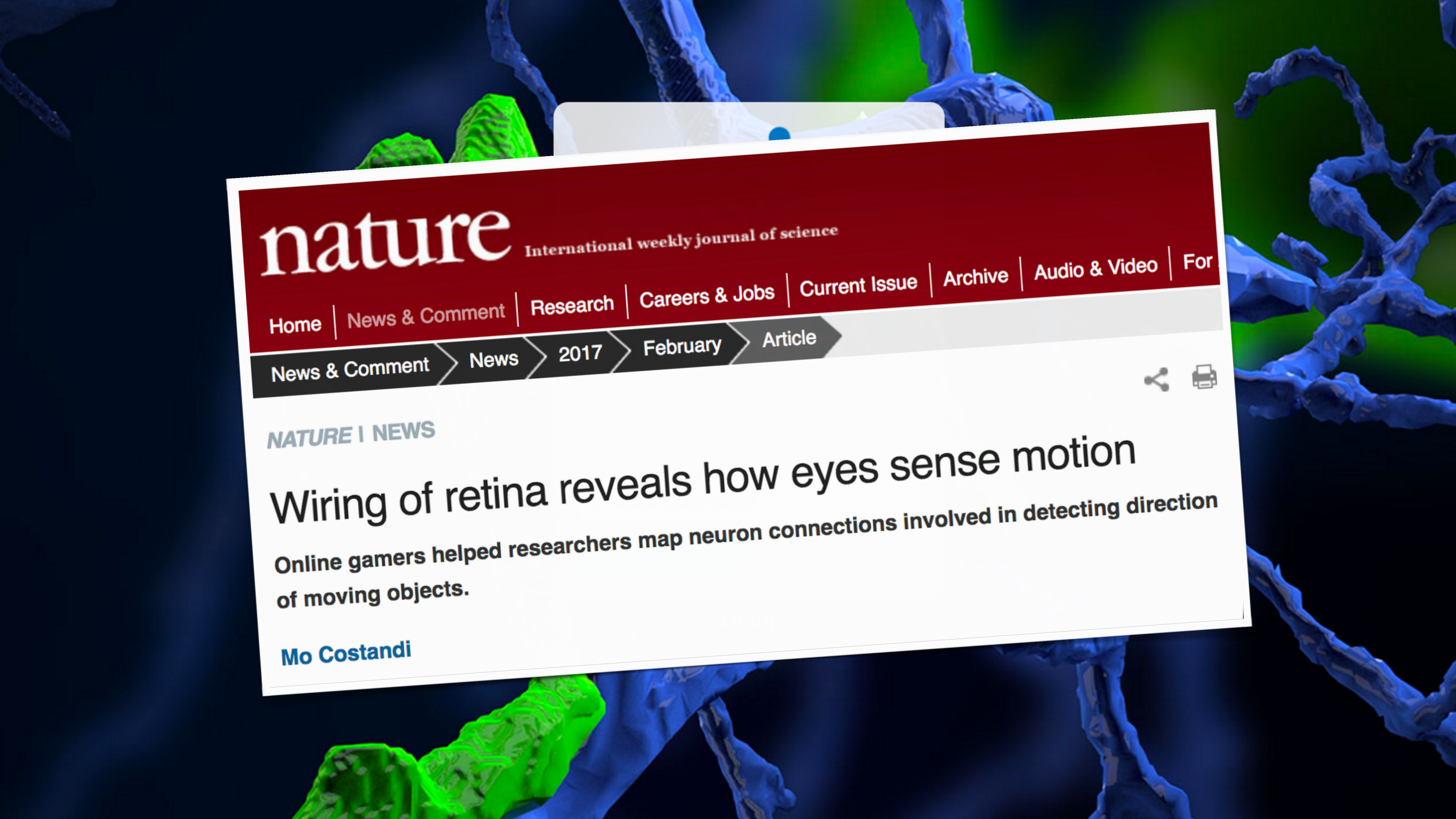
What is a protein? Proteins are the workhorses in every cell of every living thing. Your body is made up of trillions of cells, of all different kinds: muscle cells, brain cells, blood cells, and more.



Ready to discover new worlds?

Citizen Scientists Help Find Alien Planets

By Mike Wall, Space.com Senior Writer | April 15, 2011 05:55pm ET



nature
International weekly journal of science

NATURE | NEWS

Wiring of retina reveals how eyes sense motion

Online gamers helped researchers map neuron connections involved in detecting direction of moving objects.

Mo Costandi

Unfun the Headline!

Hide instructions

- Below, you are **given** the **headline** of a **satirical** news article. (Click on it to see the full article.)
- Your task: **Turn it into** a headline that could have been published by a **real news source** (e.g., *CNN*, *Fox News*, *The New York Times*) or **magazine** (e.g., *Cosmopolitan*, *Rolling Stone*, *Scientific American*), **changing as few words as possible** (it's case-insensitive).
- Your version doesn't need to refer to an actual event, but it must sound realistic enough such that it *could* refer to an actual event.
- Reward: **50 points** right away **+ up to 1,000 points** when other players think your headline is real (more points the more real they think it is, and more points the fewer words you change).

STUDY FINDS OWNING COOL LEATHER JACKET MORE REWARDING THAN RAISING CHILDREN

Type unfunned version here!

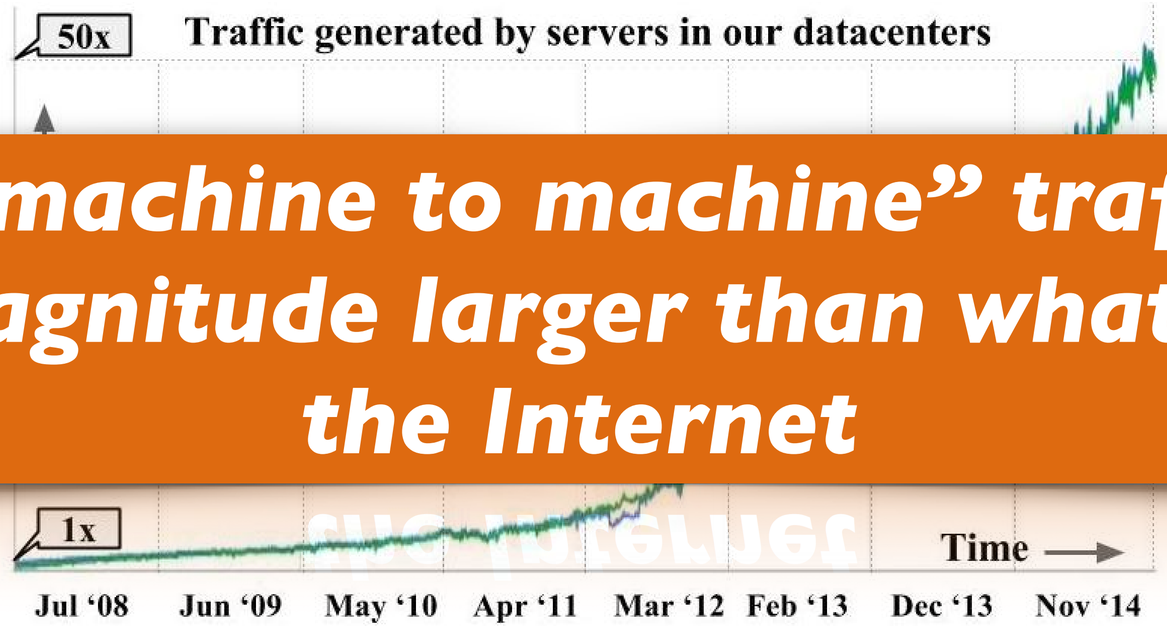
Copy original | Words changed: **11**

Done! Skip »

What can we do for networking?!

What does data center traffic look like?

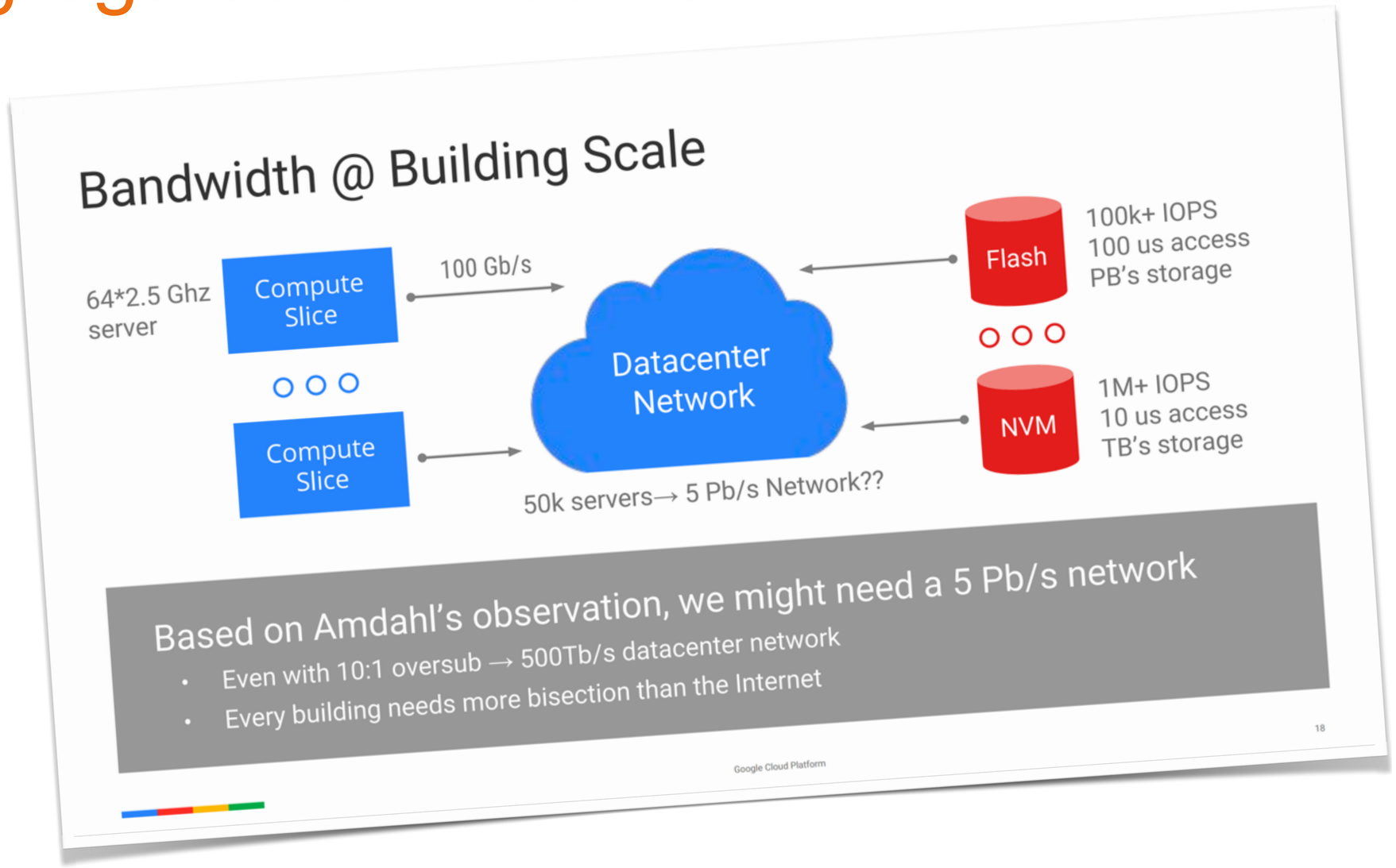
Traffic characteristics: growing volume



Facebook: *"machine to machine" traffic is several orders of magnitude larger than what goes out to the Internet*

"Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", Arjun Singh et al. @ **Google**, ACM SIGCOMM'15

“Disaggregated data centers”



“Cloud 3.0 and Software Defined Networking”
Talk by Amin Vahdat, Oct. 2016

Traffic characteristics: rack locality

Facebook

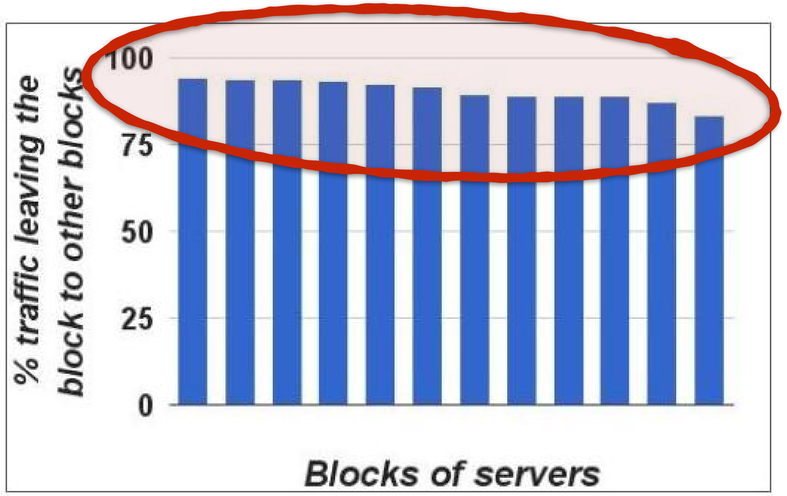
"Inside the Social Network's (Datacenter) Network"
Arjun Roy et al., ACM SIGCOMM'15

Locality	All	Hadoop	FE	Svc.	Cache	DB
Rack	12.9	13.3	2.7	12.1	0.2	0
Cluster	57.5	80.9	81.3	56.3	13.0	30.7
DC	11.9	3.3	7.3	15.7	40.7	34.5
Inter-DC	17.7	2.5	8.6	15.9	16.1	34.8
Percentage		23.7	21.5	18.0	10.2	5.2

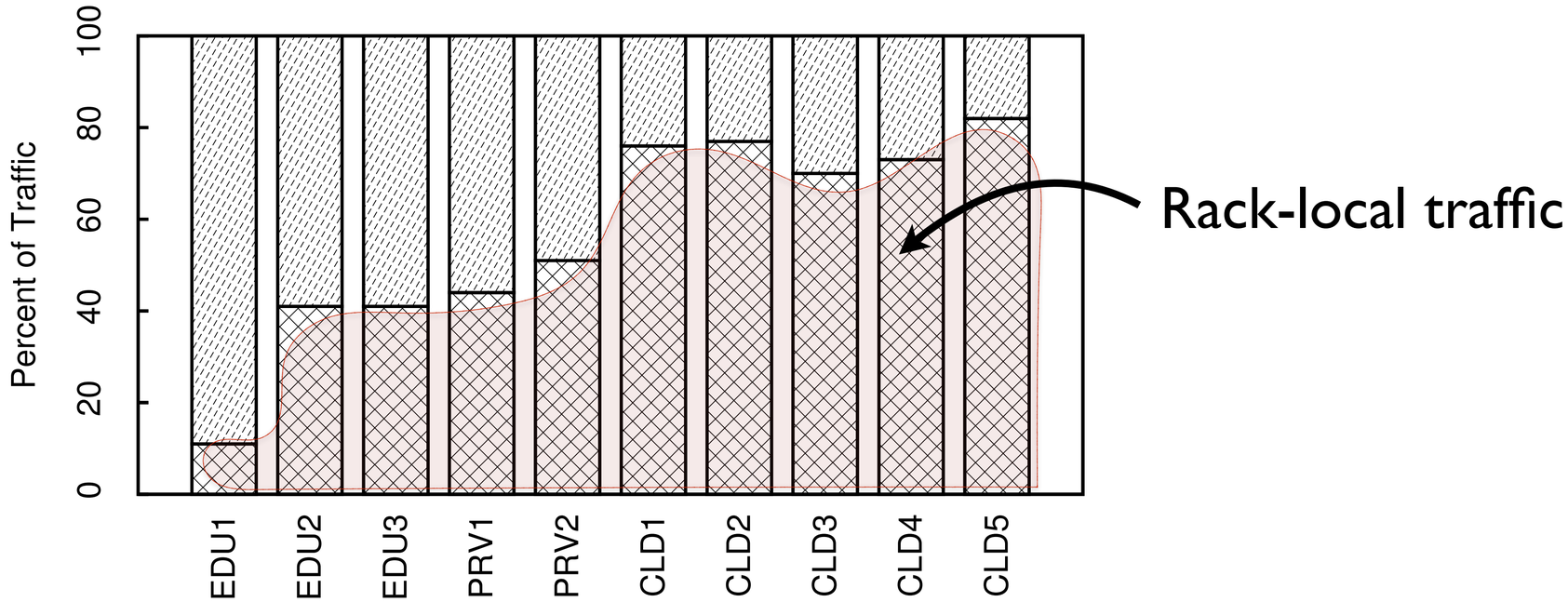
Google

"Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network"
Arjun Singh et al., ACM SIGCOMM'15

Job Category	B/w (%)
Storage	49.3
Search Serving	26.2
Mail	7.4
Ad Stats	3.8
Rest of traffic	13.3



Traffic characteristics: rack locality



"Network Traffic Characteristics of Data Centers in the Wild"
Theophilus Benson et al., ACM IMC'10

Traffic characteristics: concurrent flows

Facebook

"Inside the Social Network's (Datacenter) Network"
Arjun Roy et al., ACM SIGCOMM'15

*"Web servers and cache hosts have **100s to 1000s** of concurrent connections"*
*"Hadoop nodes have approximately **25** concurrent connections on average."*

1500 server cluster @ ??

"The Nature of Datacenter Traffic: Measurements & Analysis"
Srikanth Kandula et al. (Microsoft Research), ACM IMC'09

"median numbers of correspondents for a server are two (other) servers within its rack and four servers outside the rack"

Traffic characteristics: flow arrival rate

Facebook

"Inside the Social Network's (Datacenter) Network"
Arjun Roy et al., ACM SIGCOMM'15

"median inter-arrival times of approximately 2ms"

1500 server cluster @ ??

"The Nature of Datacenter Traffic: Measurements & Analysis"
Srikanth Kandula et al. (Microsoft Research), ACM IMC'09

< 0.1x Facebook's rate

Traffic characteristics: flow sizes

Facebook

"Inside the Social Network's (Datacenter) Network"
Arjun Roy et al., ACM SIGCOMM'15

Hadoop: median flow <1KB
<5% exceed 1MB or 100sec

Caching: most flows are long-lived
… but bursty internally

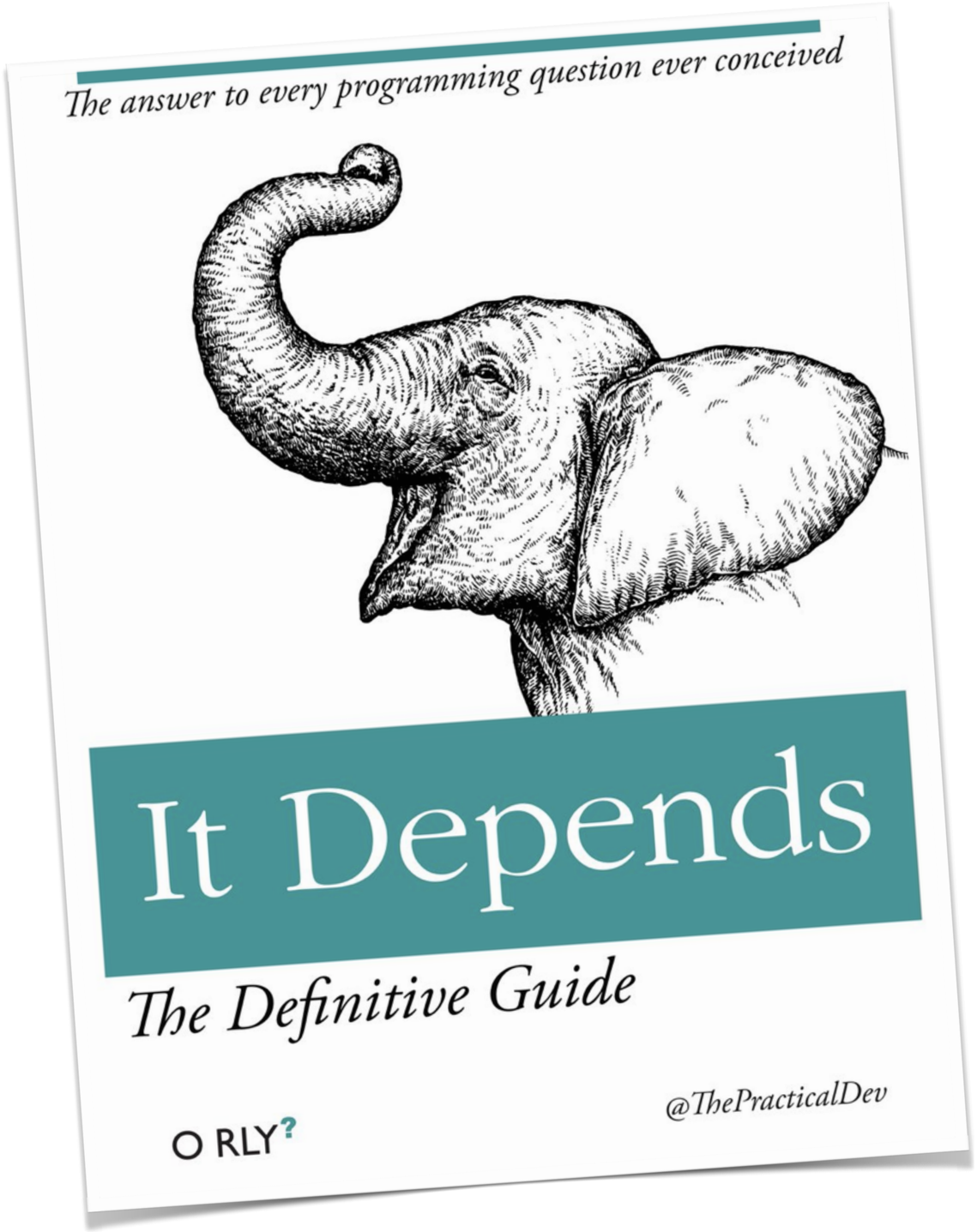
Heavy-hitters ≈ median flow, not persistent

1500 server cluster @ ??

"The Nature of Datacenter Traffic: Measurements & Analysis"
Srikanth Kandula et al. (Microsoft Research), ACM IMC'09

> 80% of the flows last <10sec
> 50% bytes are in flows lasting less <25sec

So … what does data center traffic look like?



on …

- applications
- scale
- network design
- …

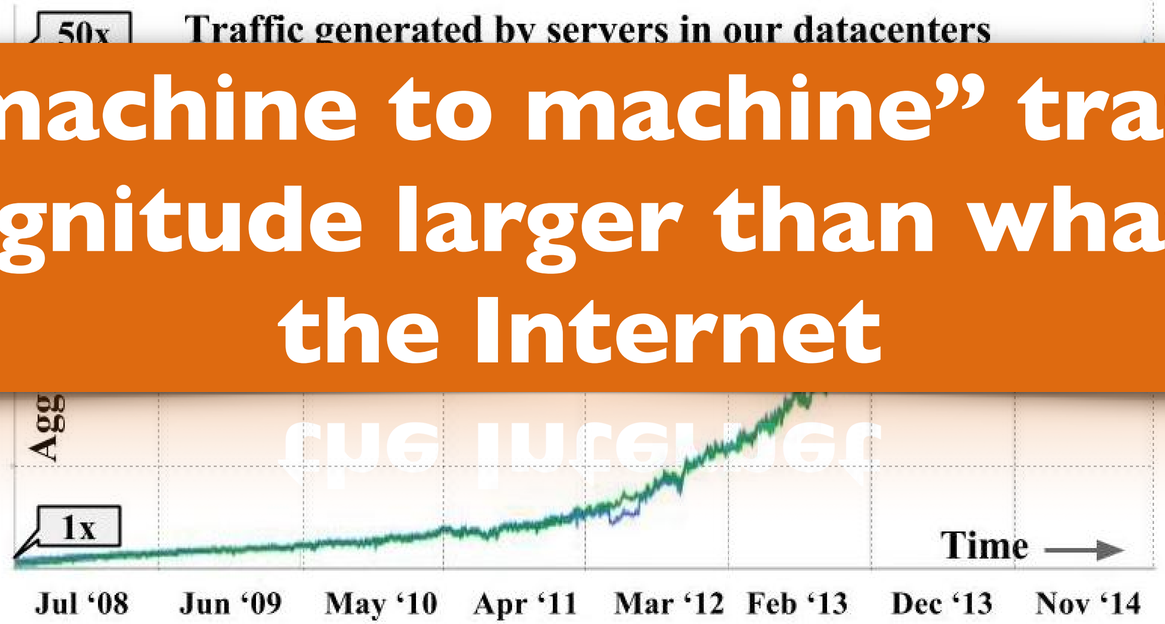
… and very little data is available

Implications for networking

1. Data center internal traffic is BIG
2. Tight deadlines for network I/O
3. Congestion and TCP incast
4. Complex network shared by applications
5. Centralized control at the flow level may be difficult

Implications for networking

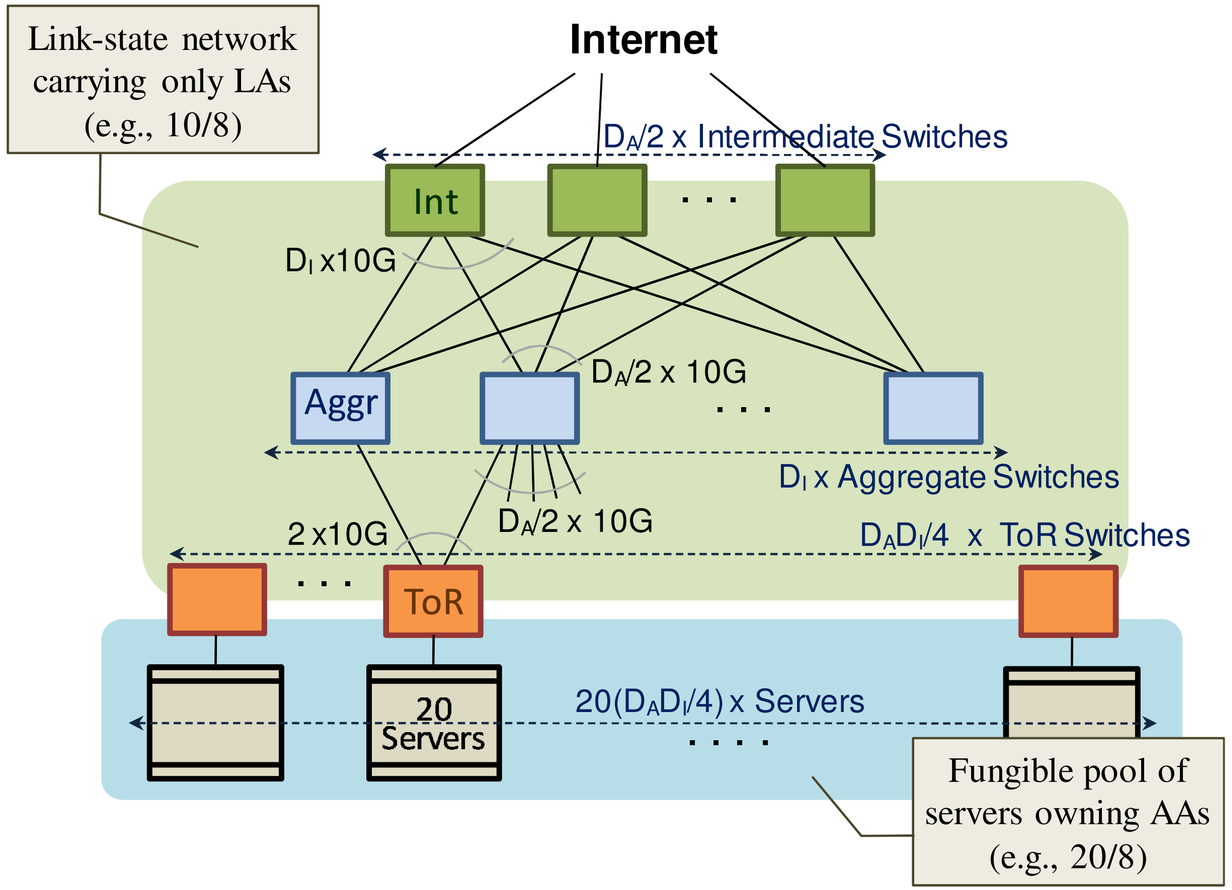
1. Data center internal traffic is BIG

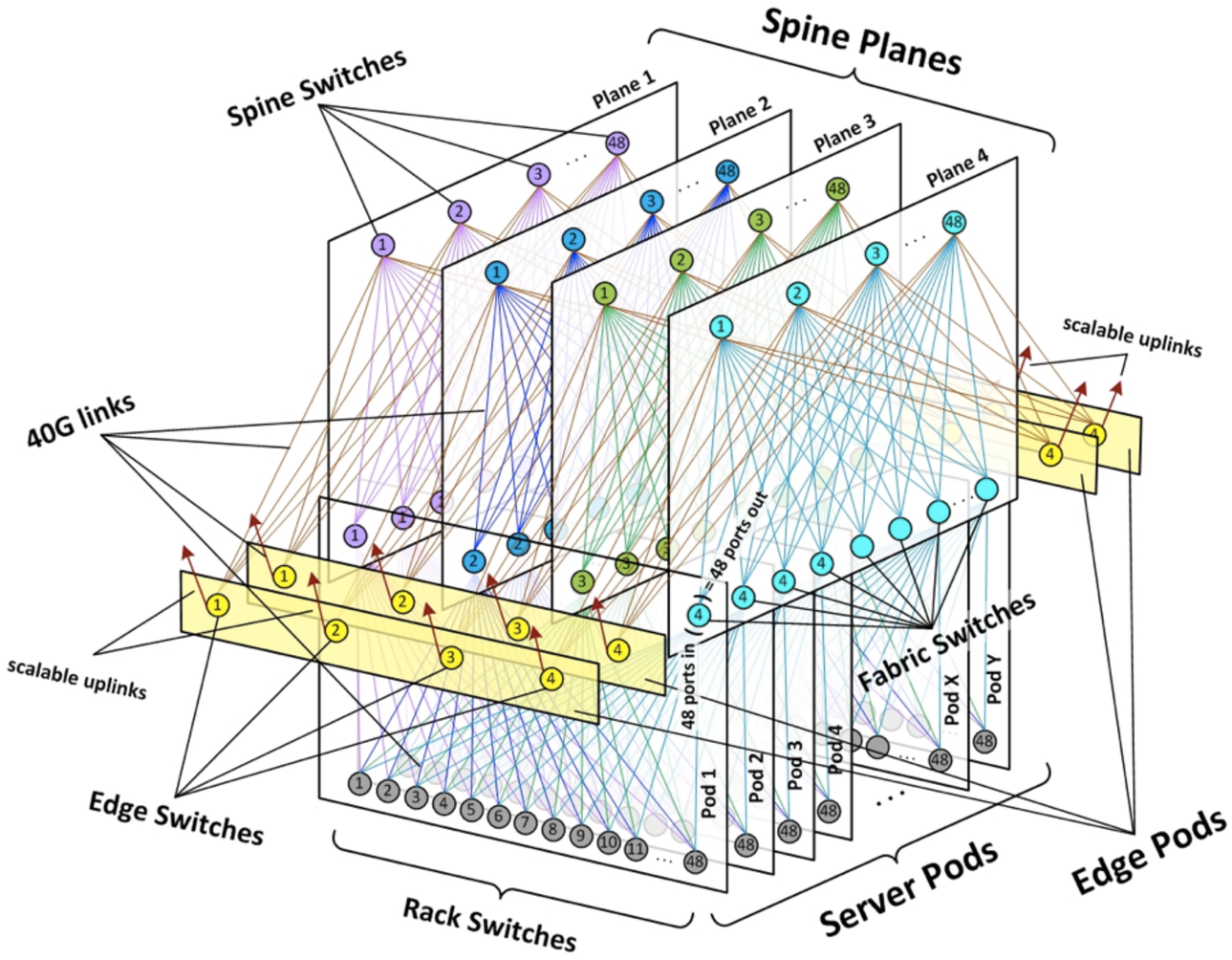


Facebook: "machine to machine" traffic is several orders of magnitude larger than what goes out to the Internet

"Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", Arjun Singh et al. @ **Google**, ACM SIGCOMM'15

How to build networks best?



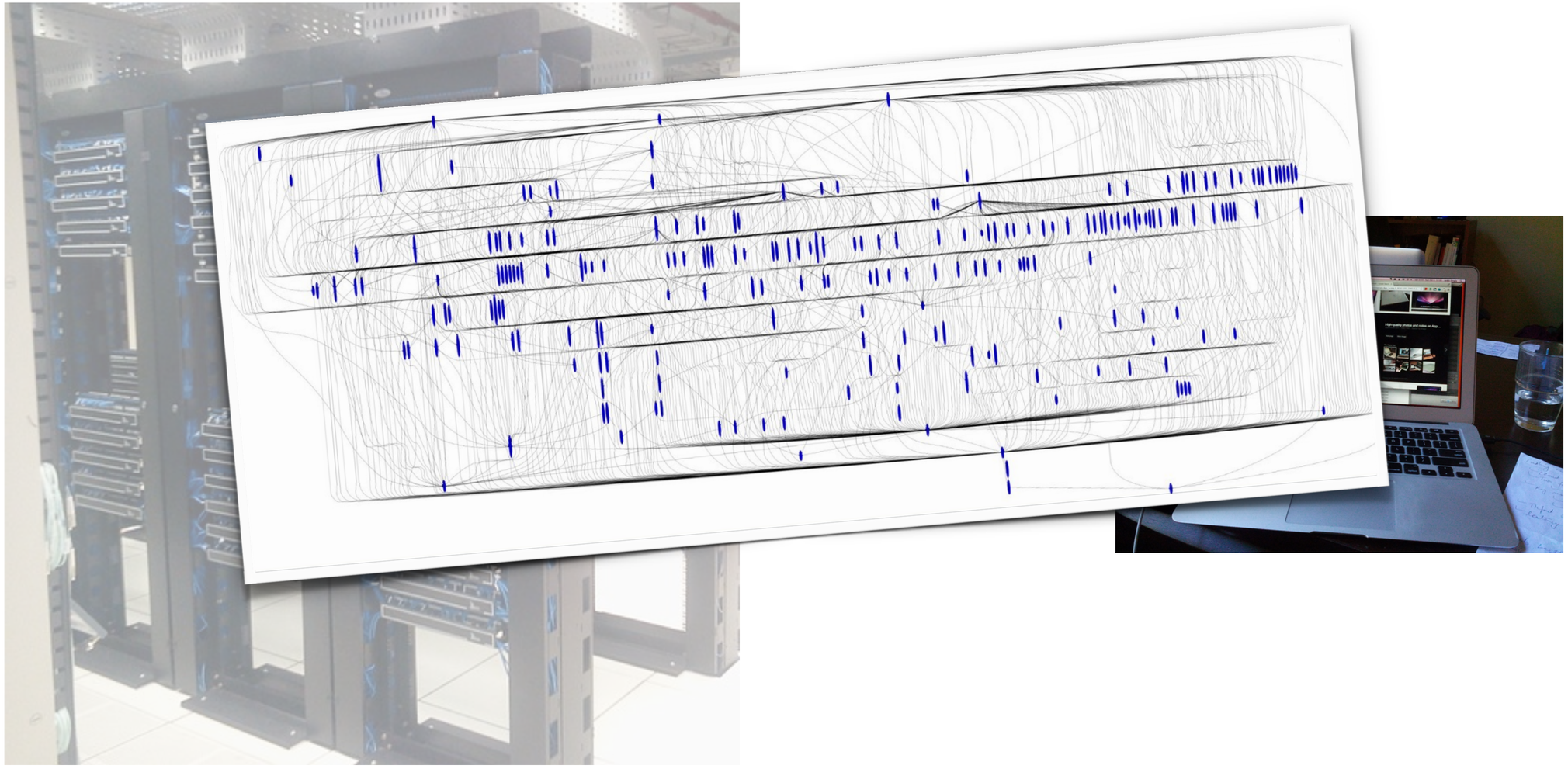


VL2 @ **Microsoft**, ACM SIGCOMM'09
Greenburg, Hamilton, Jain, Kandula, Kim, Lahiri, Maltz, Patel, Sengupta

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

Implications for networking

2. Tight deadlines for network I/O



Implications for networking

2 Tight deadlines for network I/O

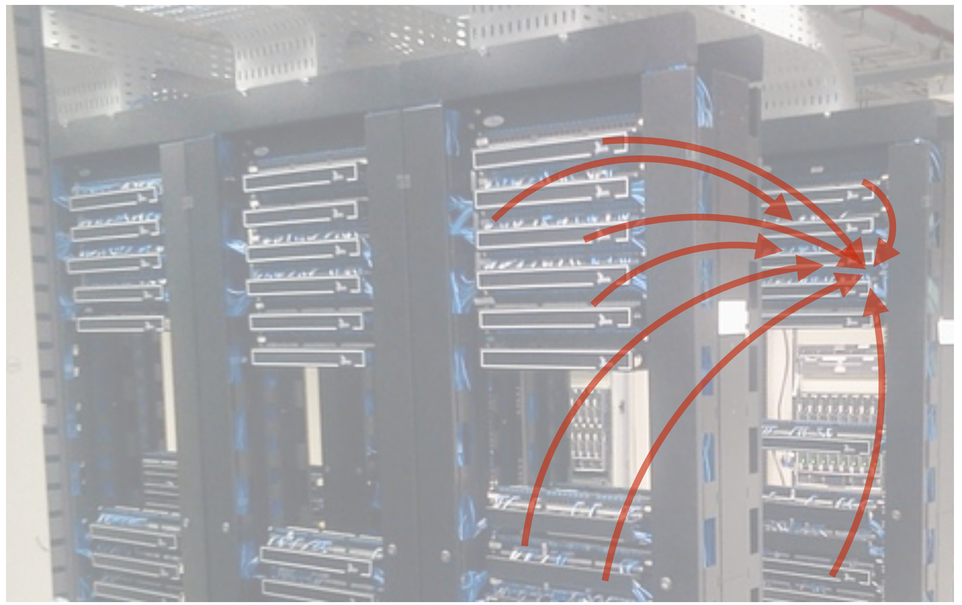
Suppose: server response-time is 10ms for 99% of requests; 1s for 1%

#Servers	Probability job takes >1s
1	1%
100	63%

Need to reduce variability and tolerate *some* variation

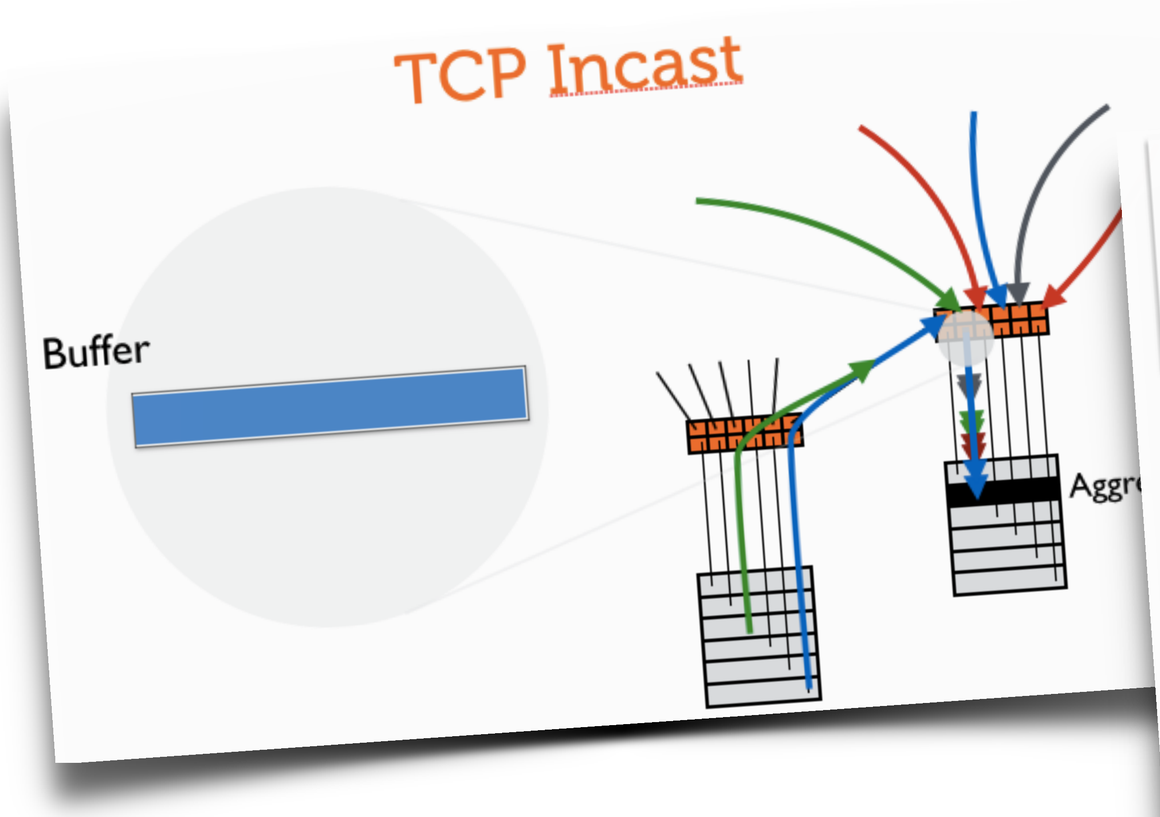
Implications for networking

3 Congestion and TCP incast



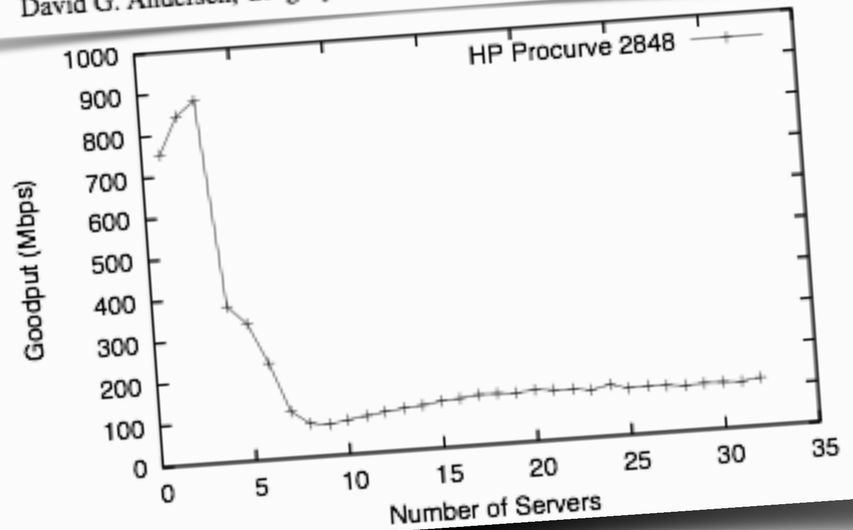
TCP does not work very well

How do we use networks best?



Measurement and Analysis of TCP Throughput Collapse in Cluster-based Storage Systems

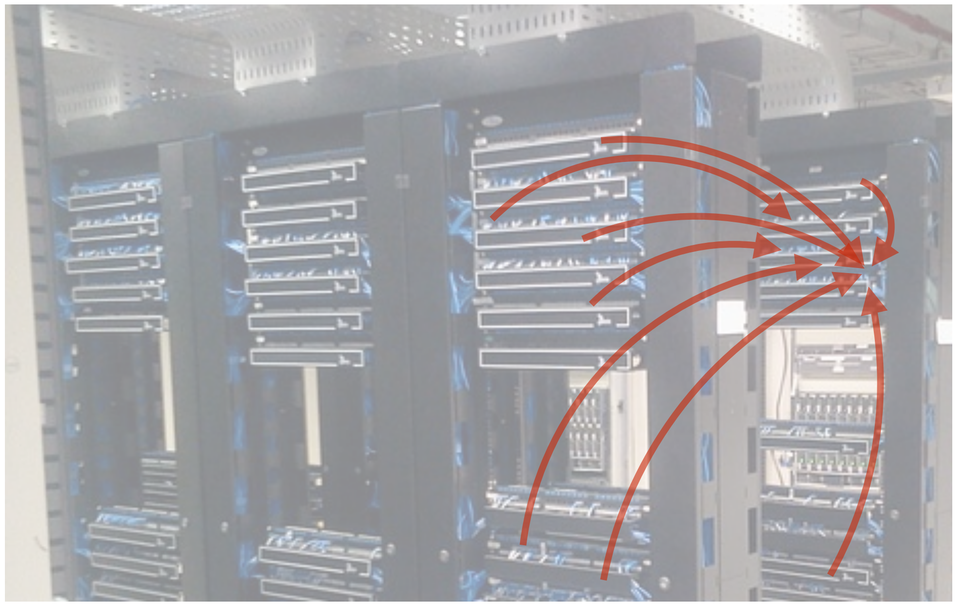
Amar Phanishayee, Elie Krevat, Vijay Vasudevan,
David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Srinivasan Seshan



Congestion control with cwnd <<1?

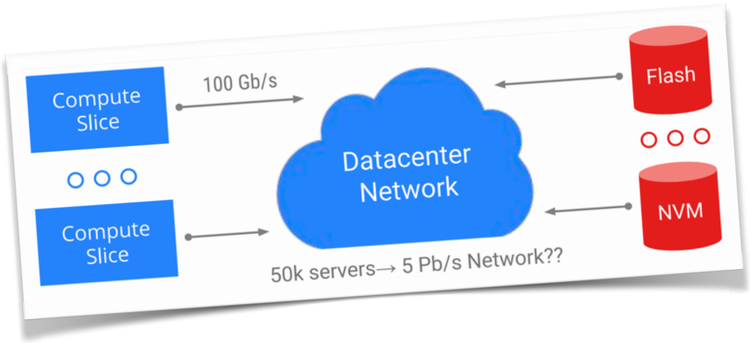
Implications for networking

4 Complex network shared by applications



Applications with different objectives sharing the network

How do we manage sharing?



"Network OS"

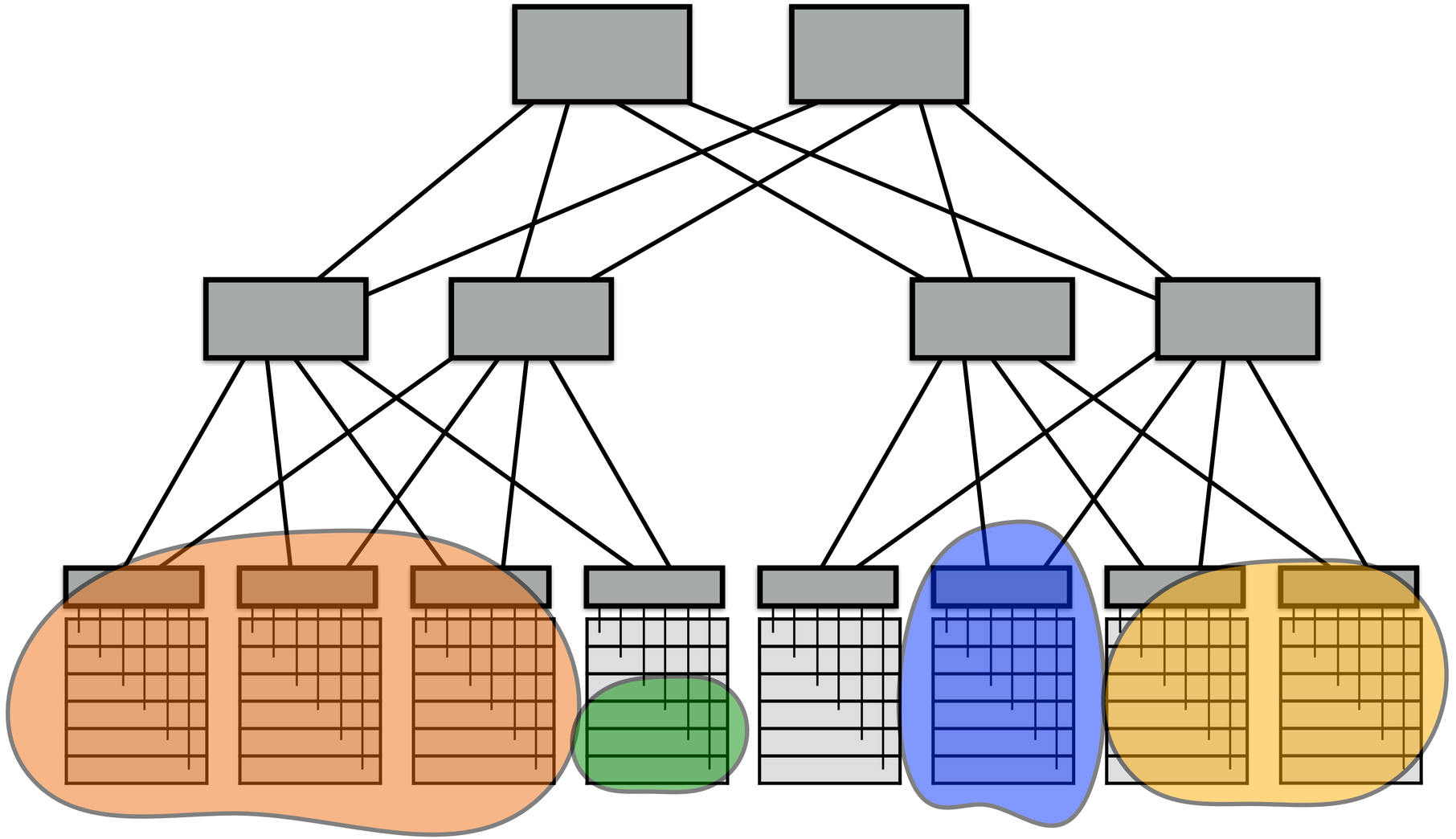
app | app

software abstractions

Logically centralized controller

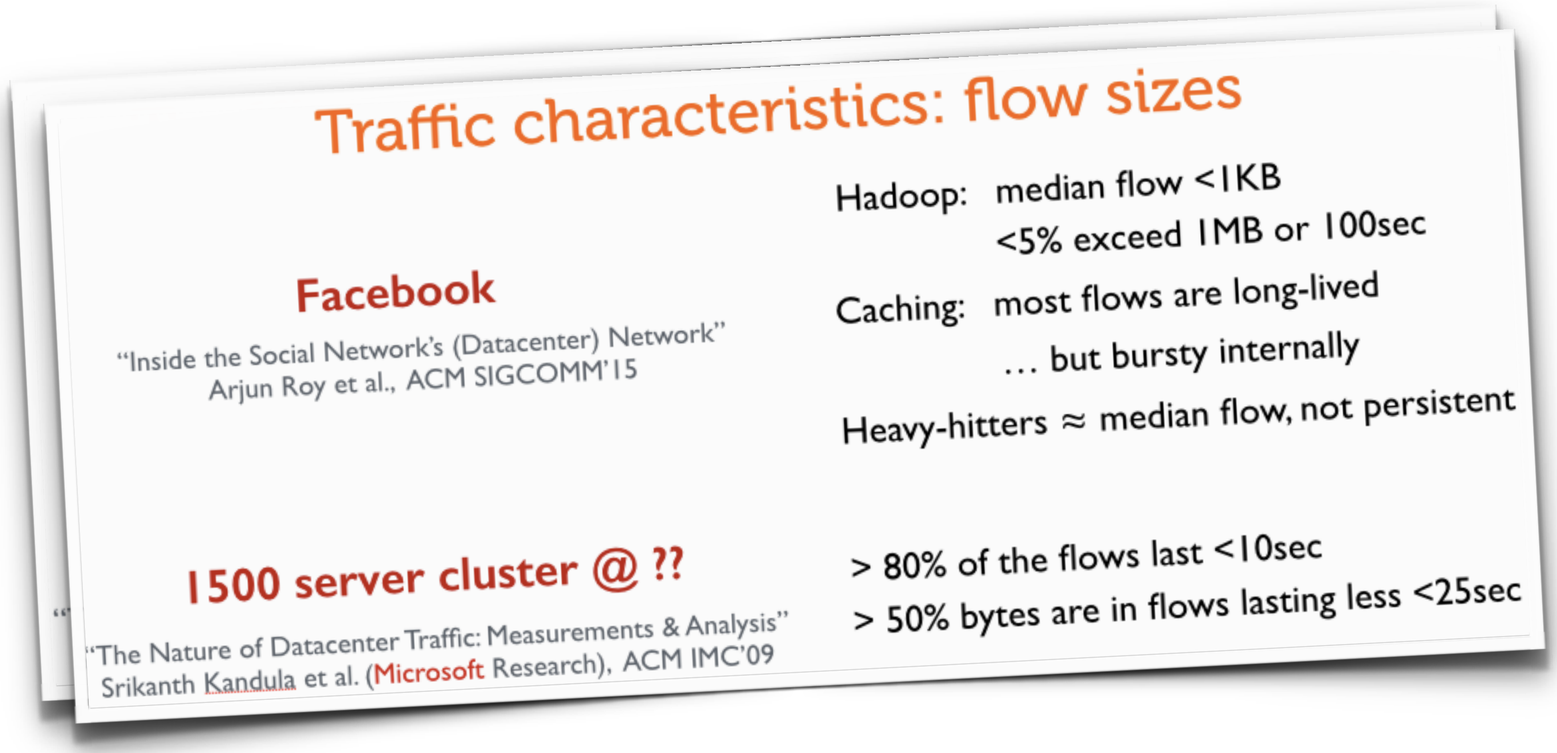
Data plane API

How do we manage sharing?



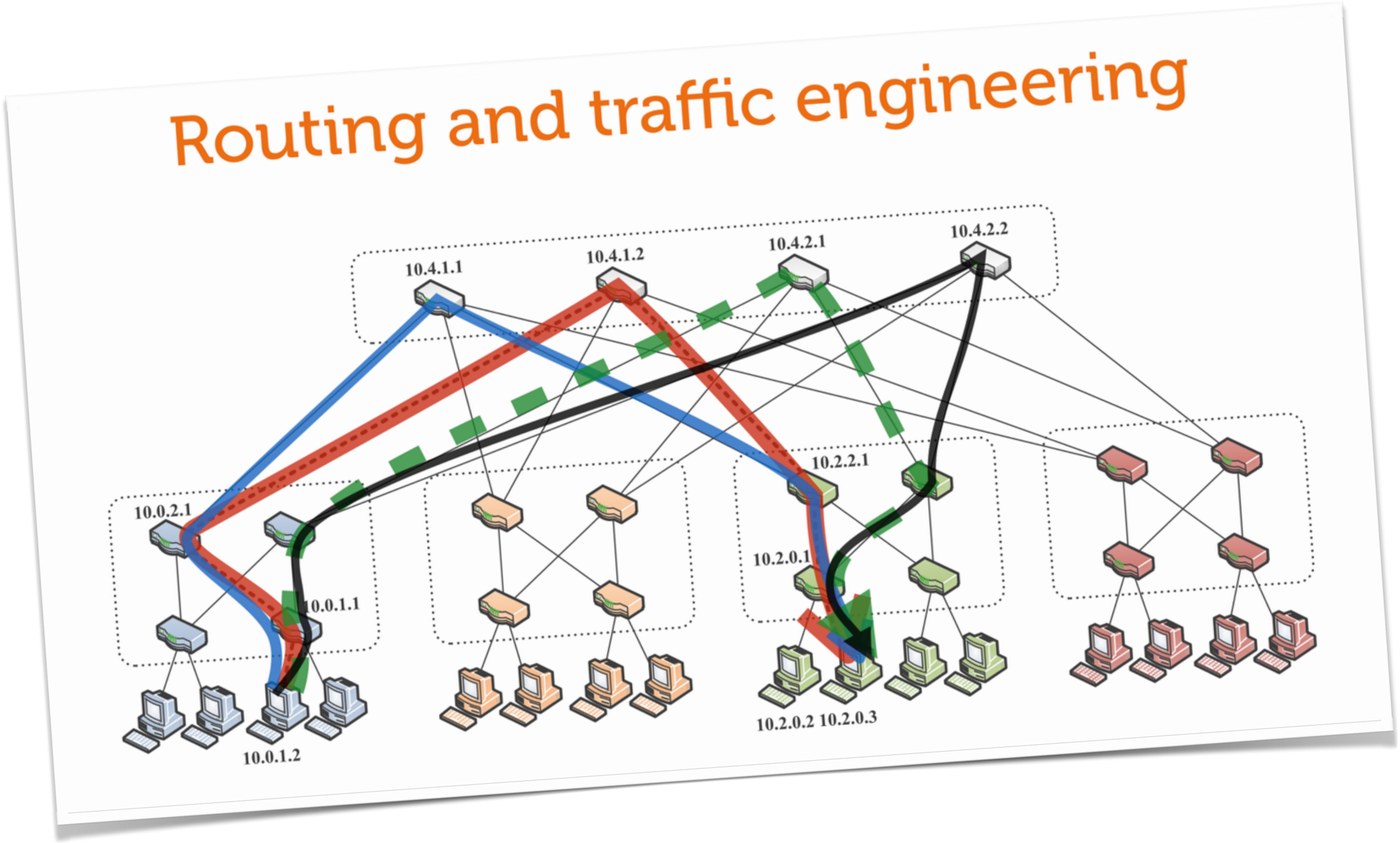
Implications for networking

5 Centralized control at the flow level *may* be difficult



Distributed control, perhaps with some centralized tinkering?

How do we use networks best?



Food for thought …

Are DCs even the right way?

[Mick Stevens for New Yorker]

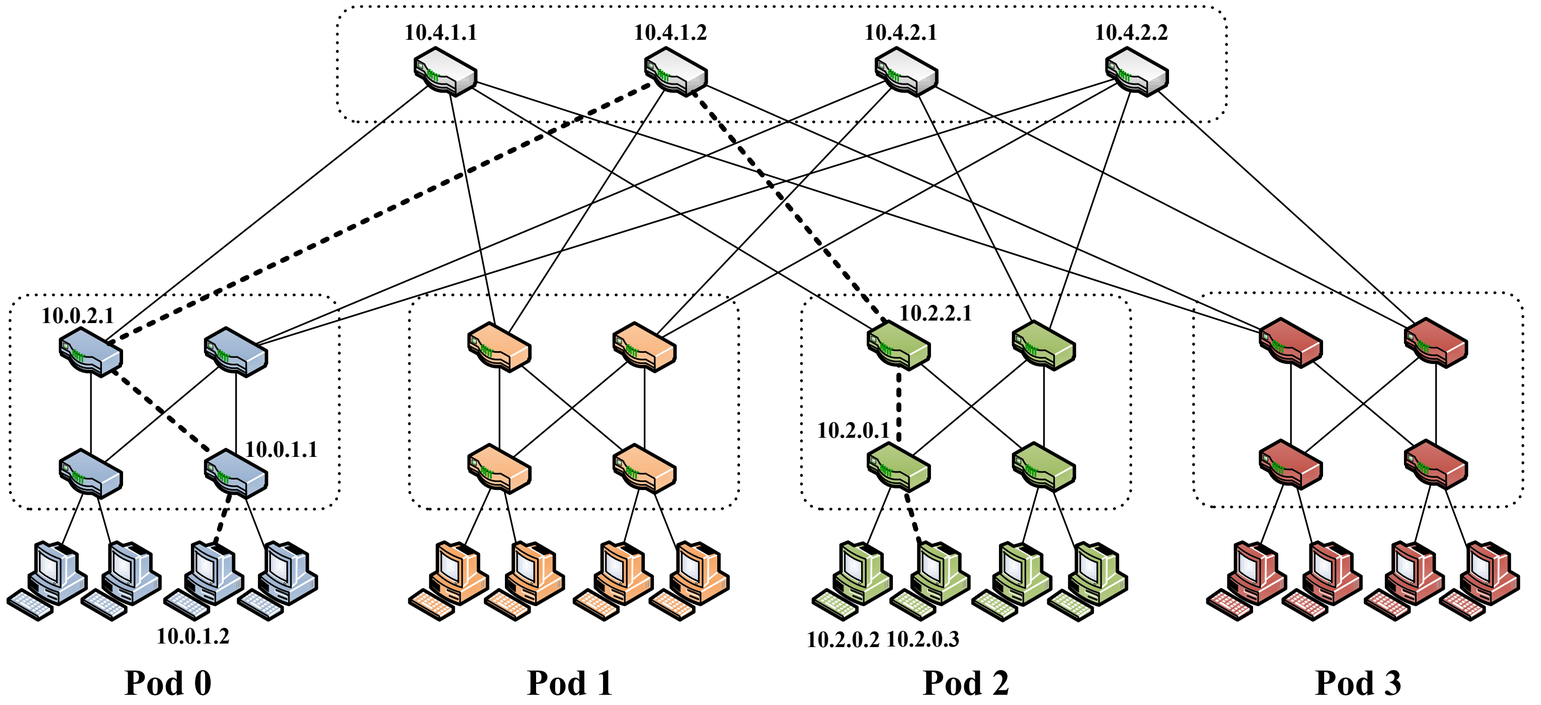
Brocken Inaglory [CC BY-SA 3.0] via Wikimedia

Mega DCs vs. many smaller sites

- Latency?
- Application complexity?
- Management complexity?
- Multiplexing?
- Need a “cache-hierarchy” structure?

Next lecture …

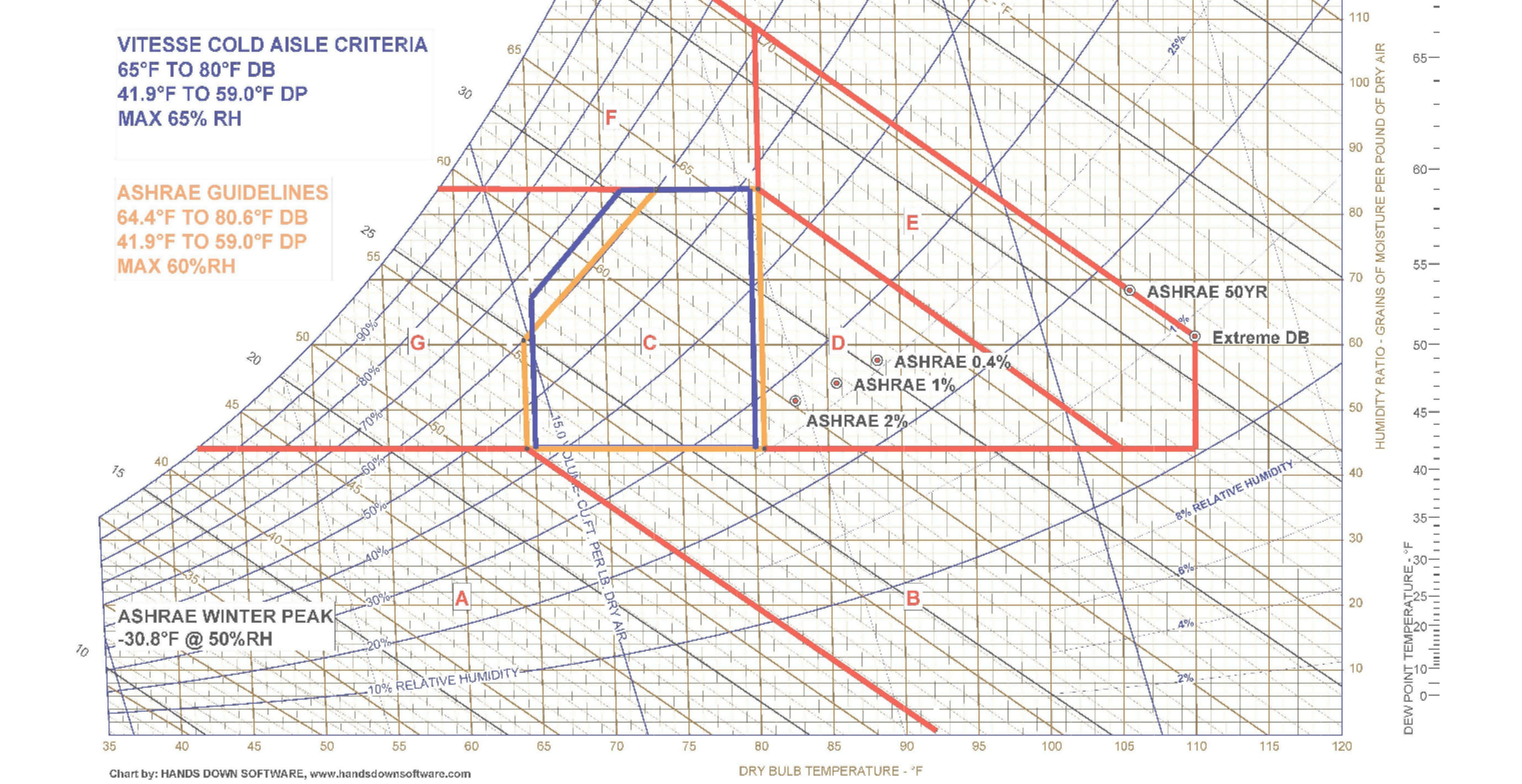
Weekly reading: fat-trees



ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

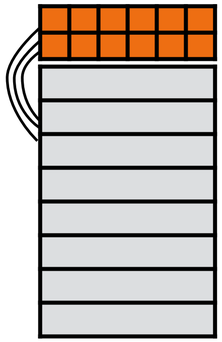
Mohammad Al-Fares Alexander Loukissas Amin Vahdat



[Data Center v1.0, Open Compute Project

[Trower, NASA]

A server rack

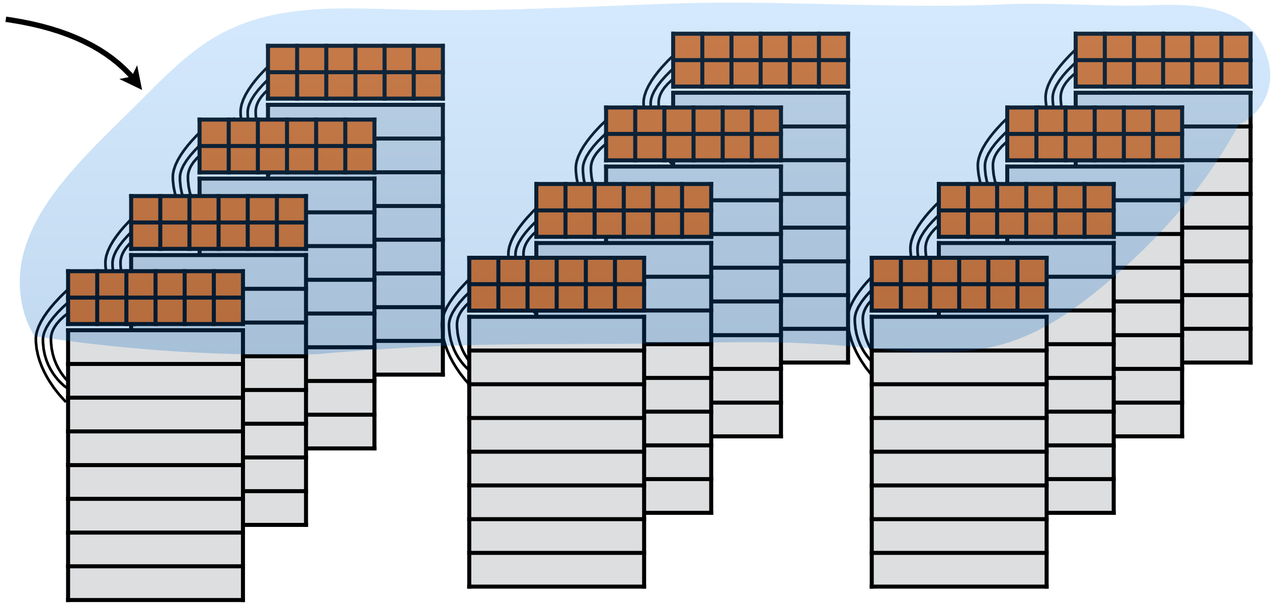


A top-of-rack switch

A rack of servers

Lots of racks

How to network the racks?

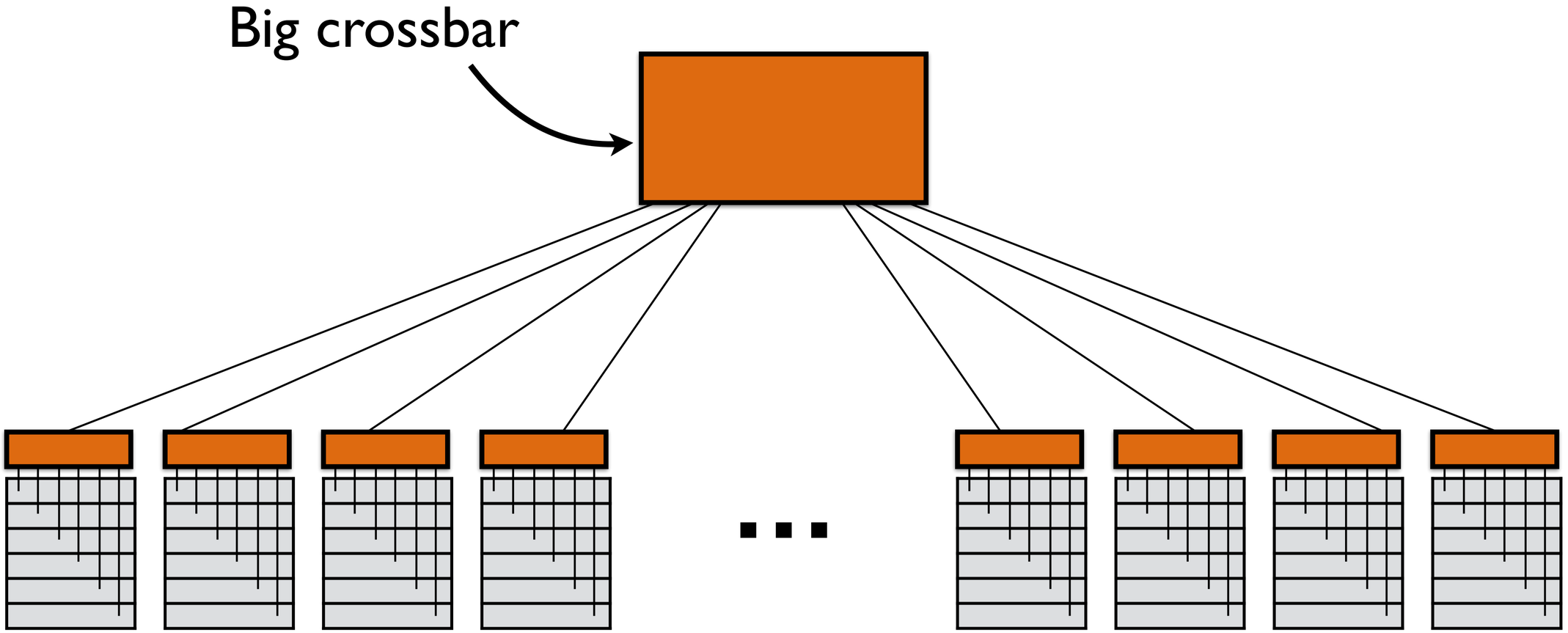


Facebook: machine-machine traffic "doubling at an interval of less than a year"

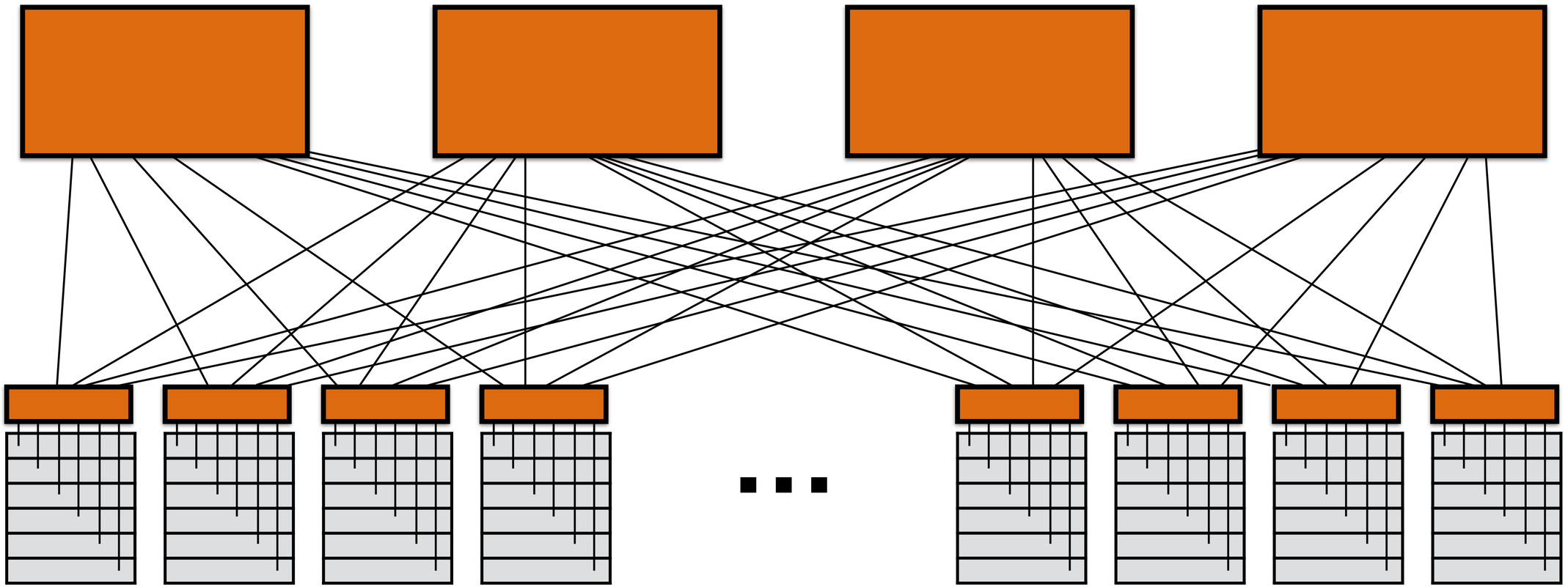
Need high throughput networks to ...

- Support big data analytics
- Ease virtual machine placement

“Big switch” approach

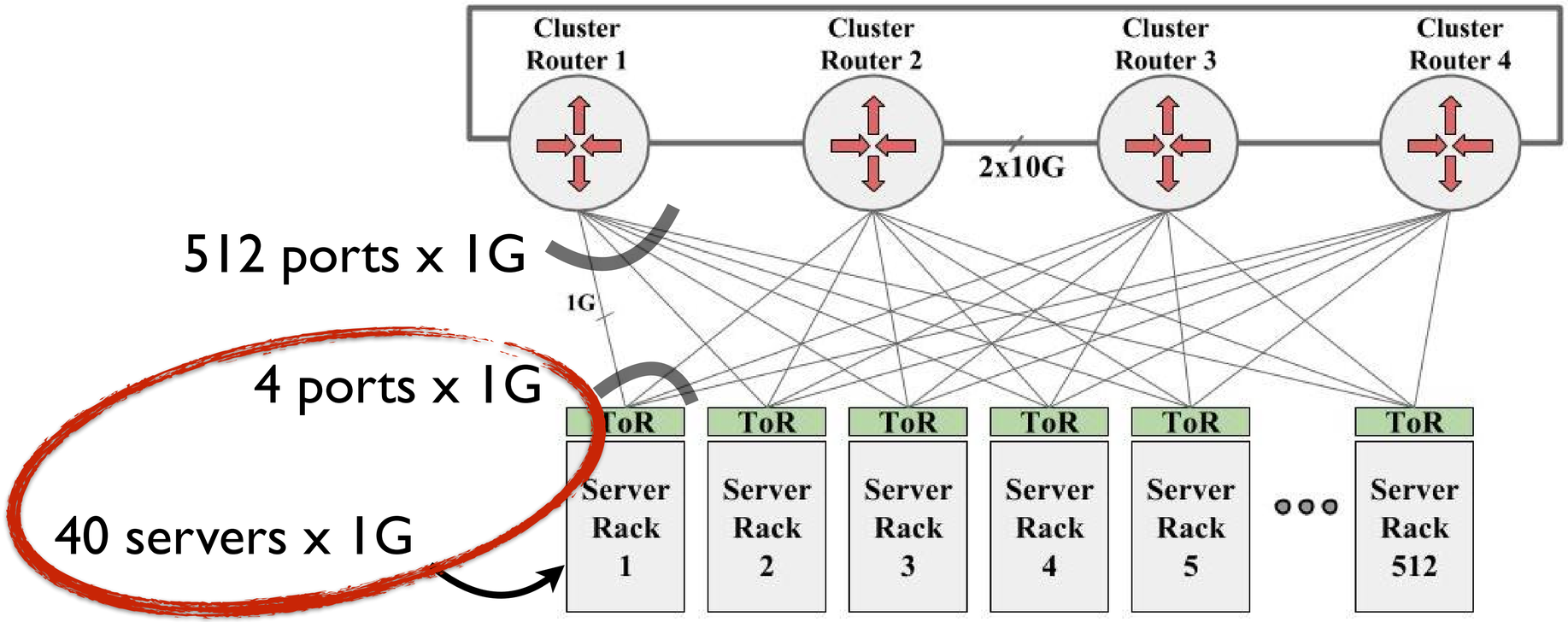


“Big switch” approach

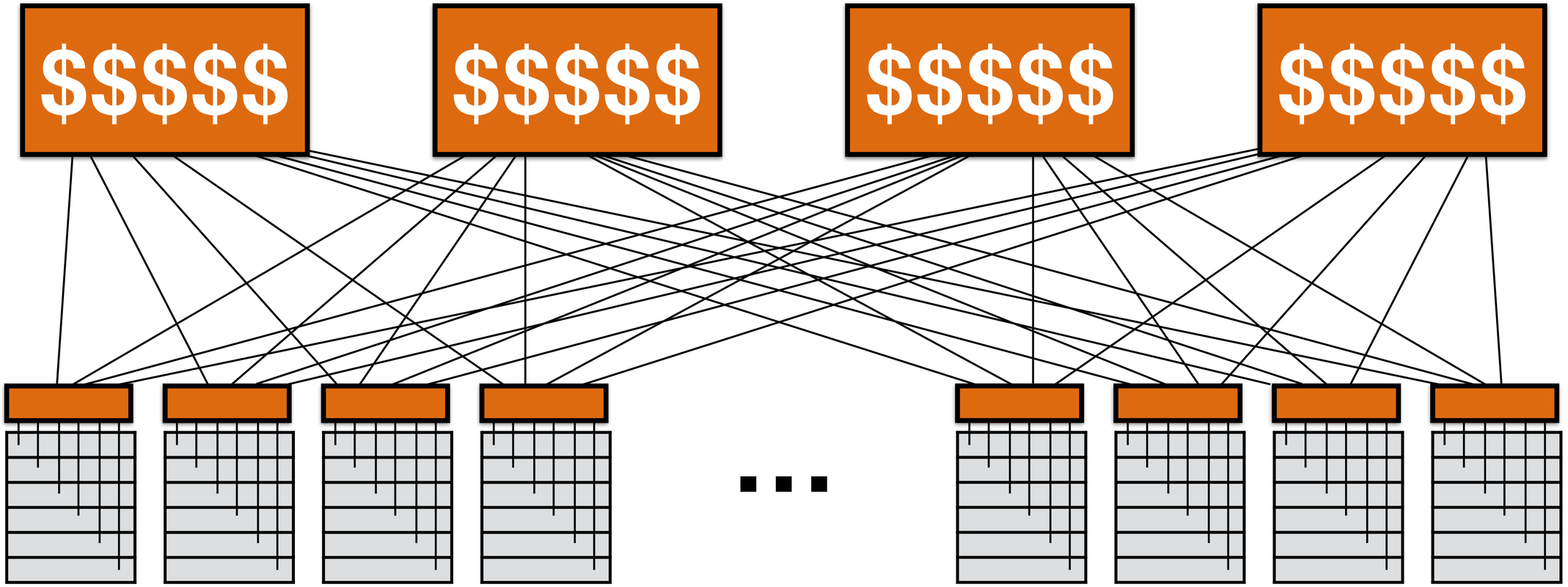


Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network

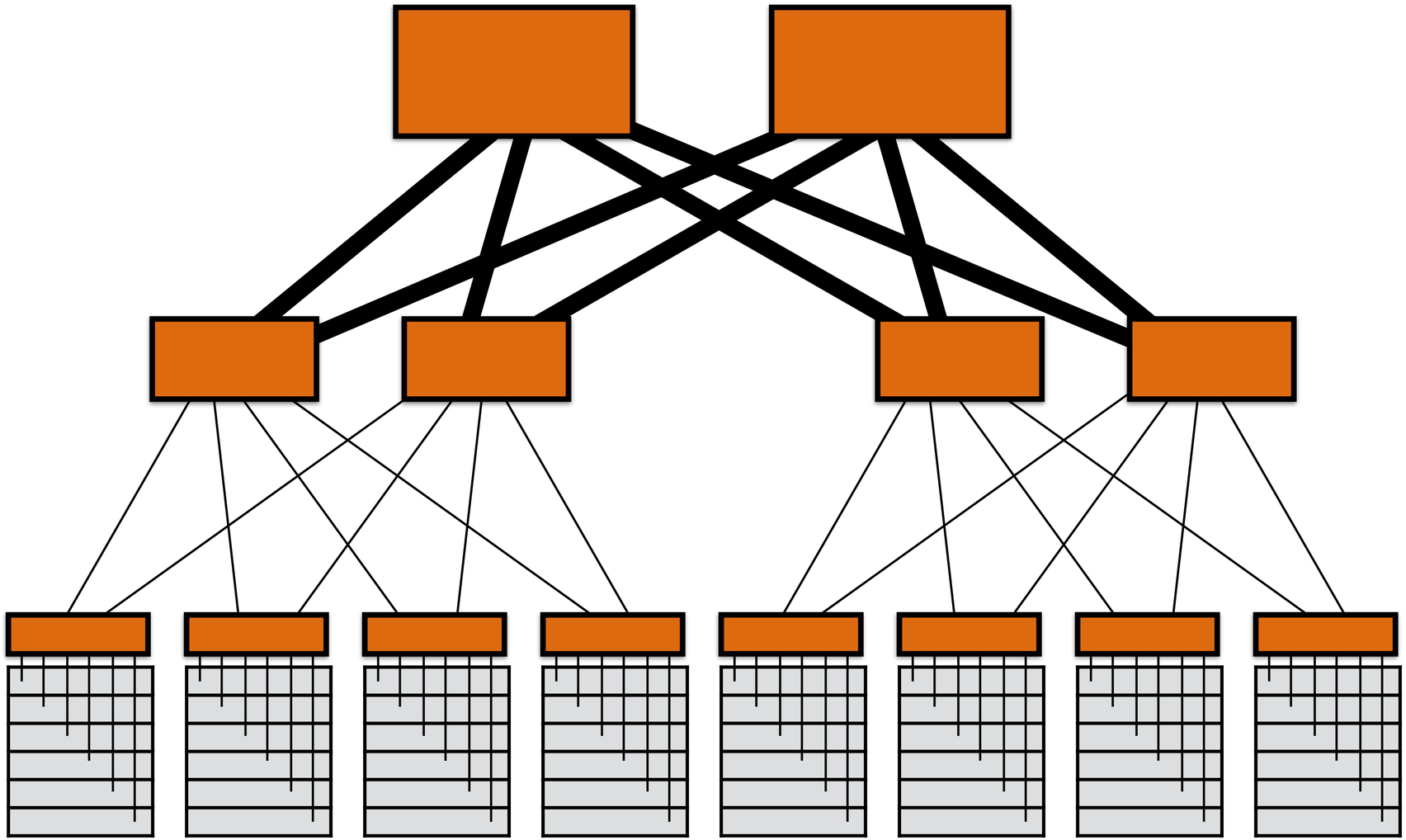
Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, and Amin Vahdat
Google, Inc.



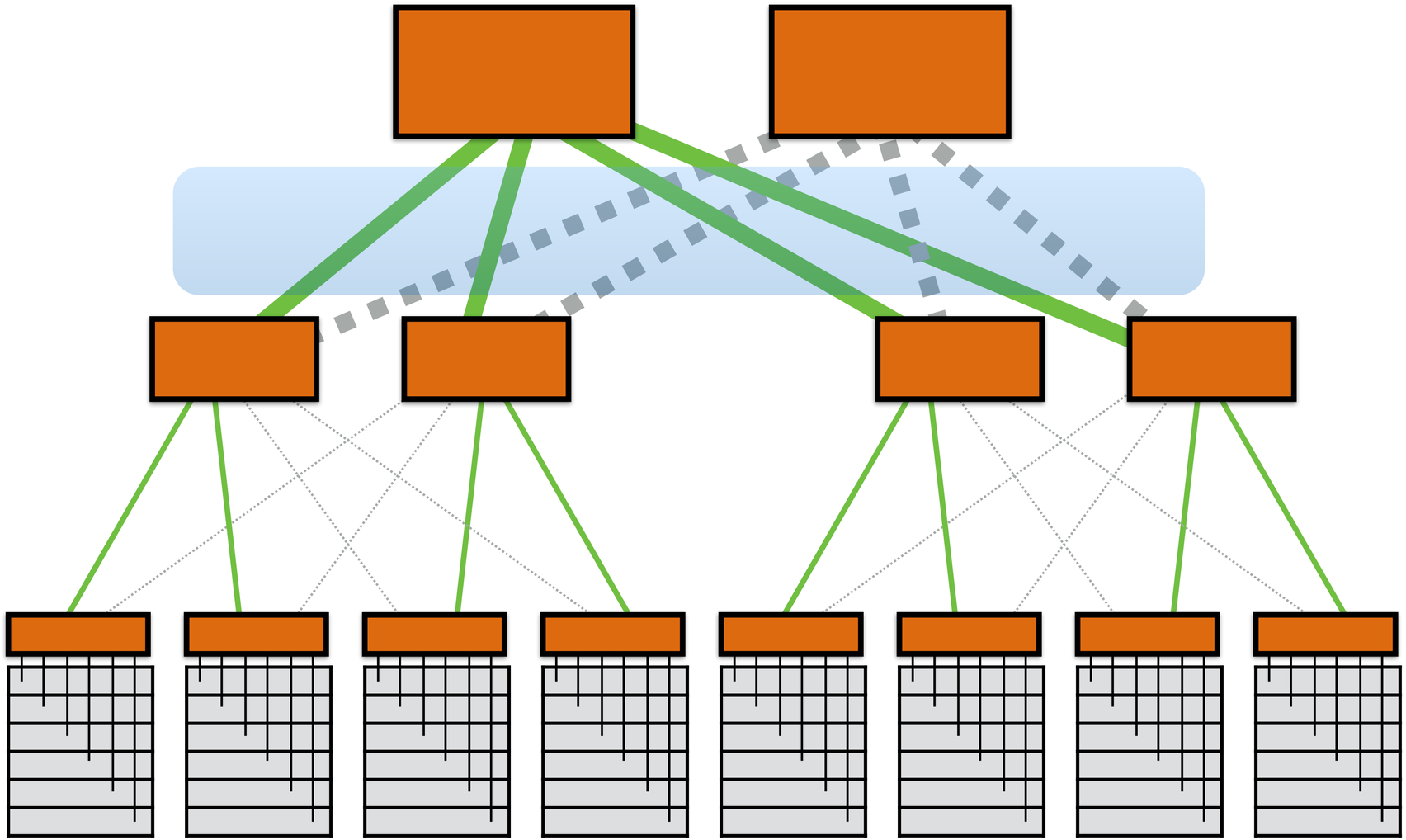
“Big switch” approach



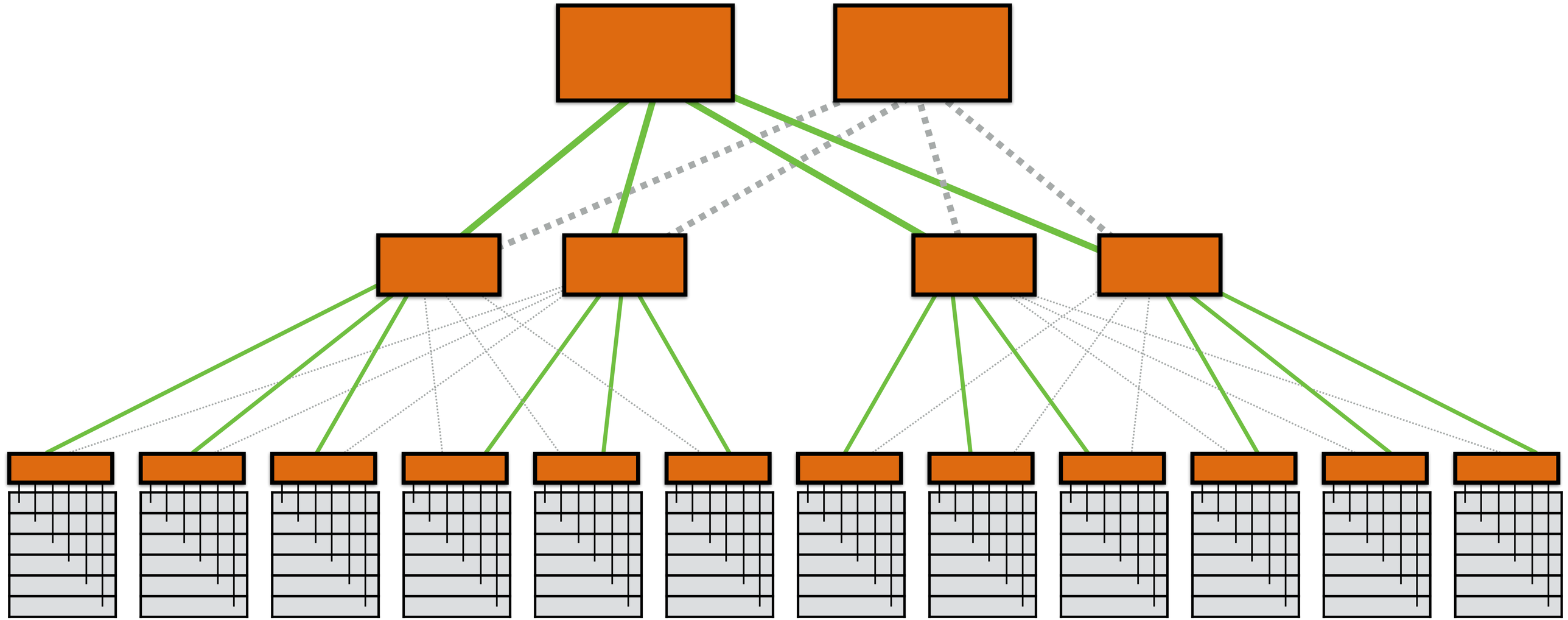
Alternative: tree network



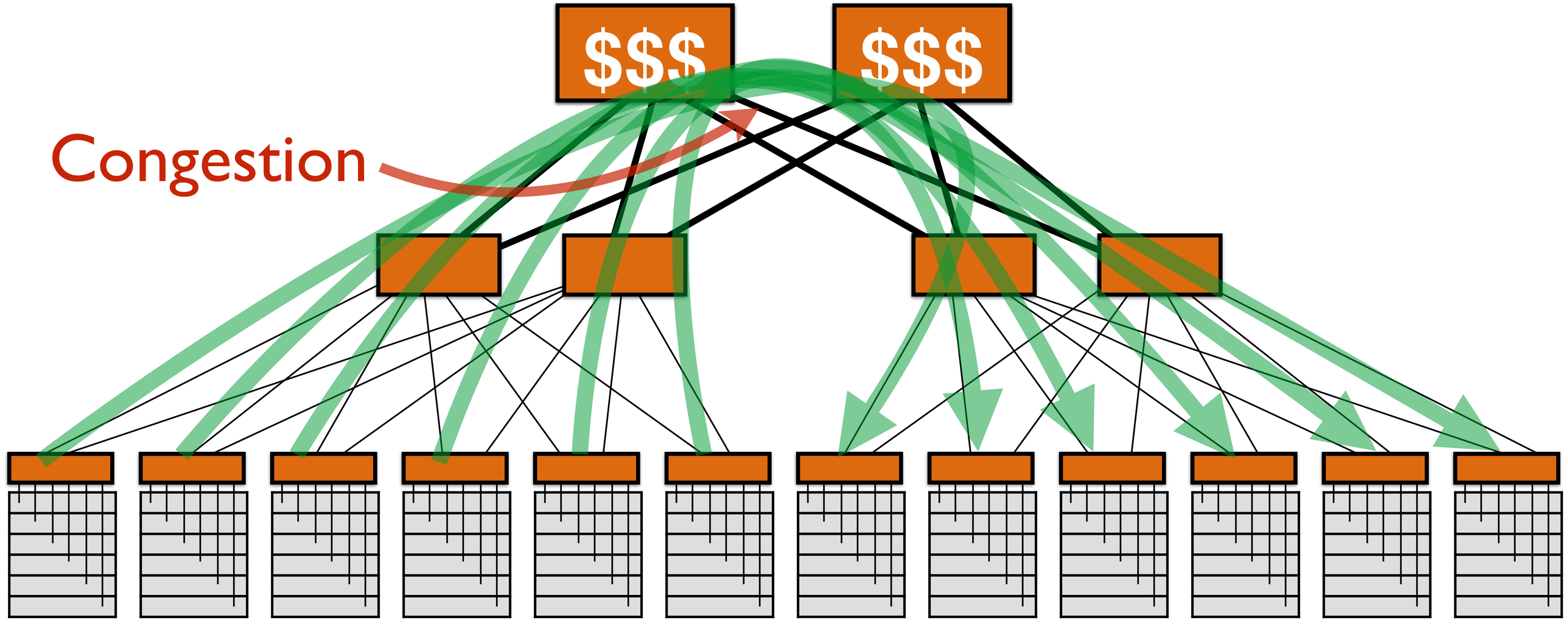
Alternative: tree network



Alternative: tree network



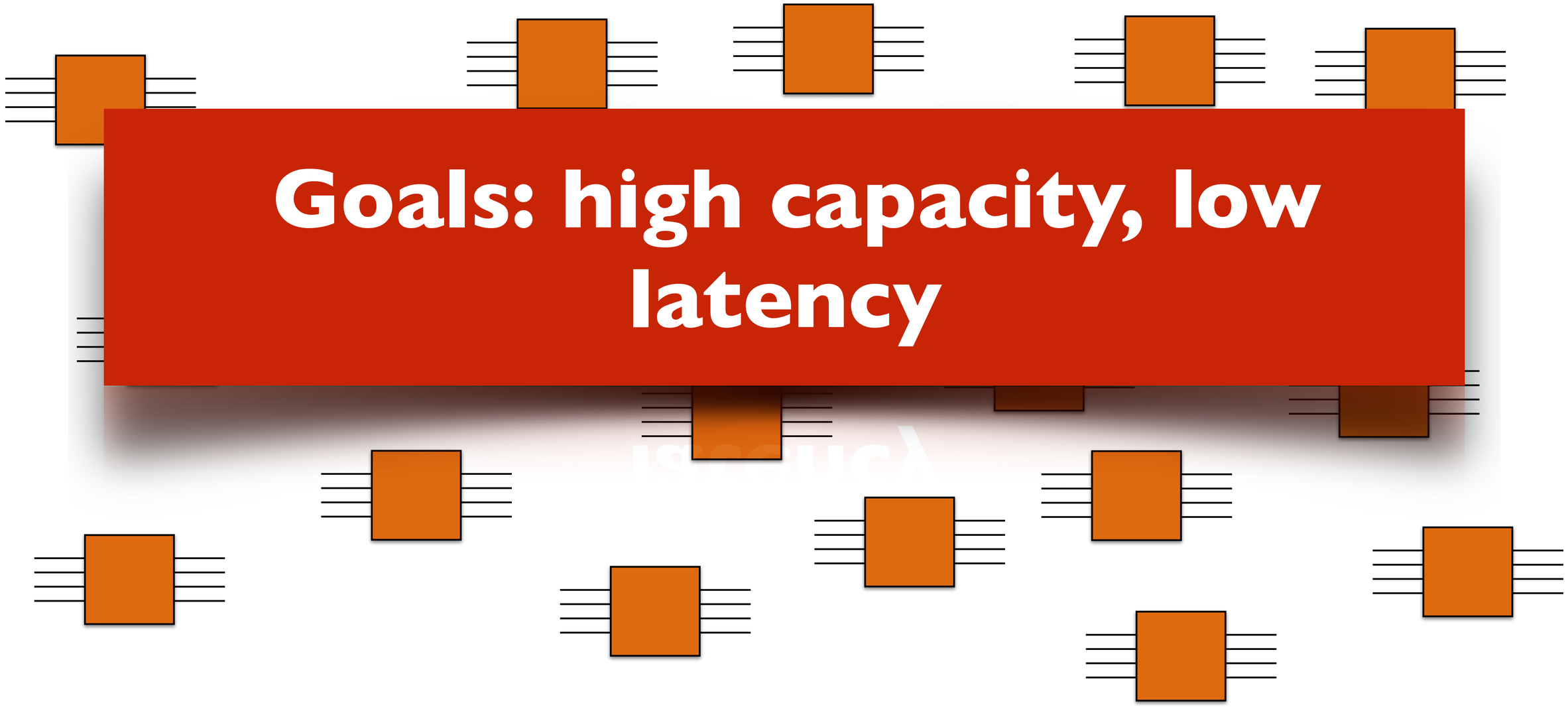
Alternative: tree network



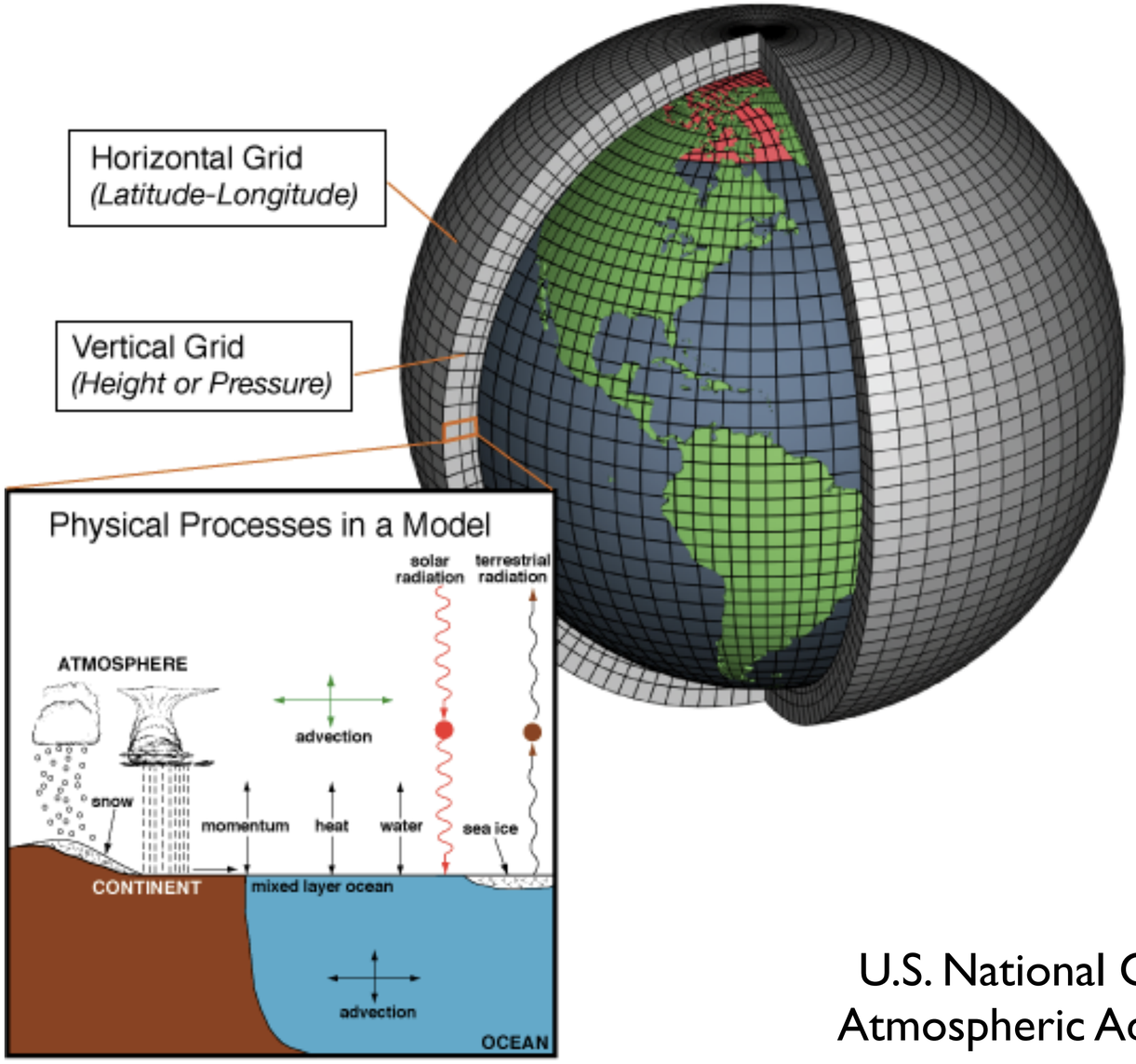
Build with identical switches throughout?

“Scaling out” vs. “scaling up”

How to connect large numbers of switches?

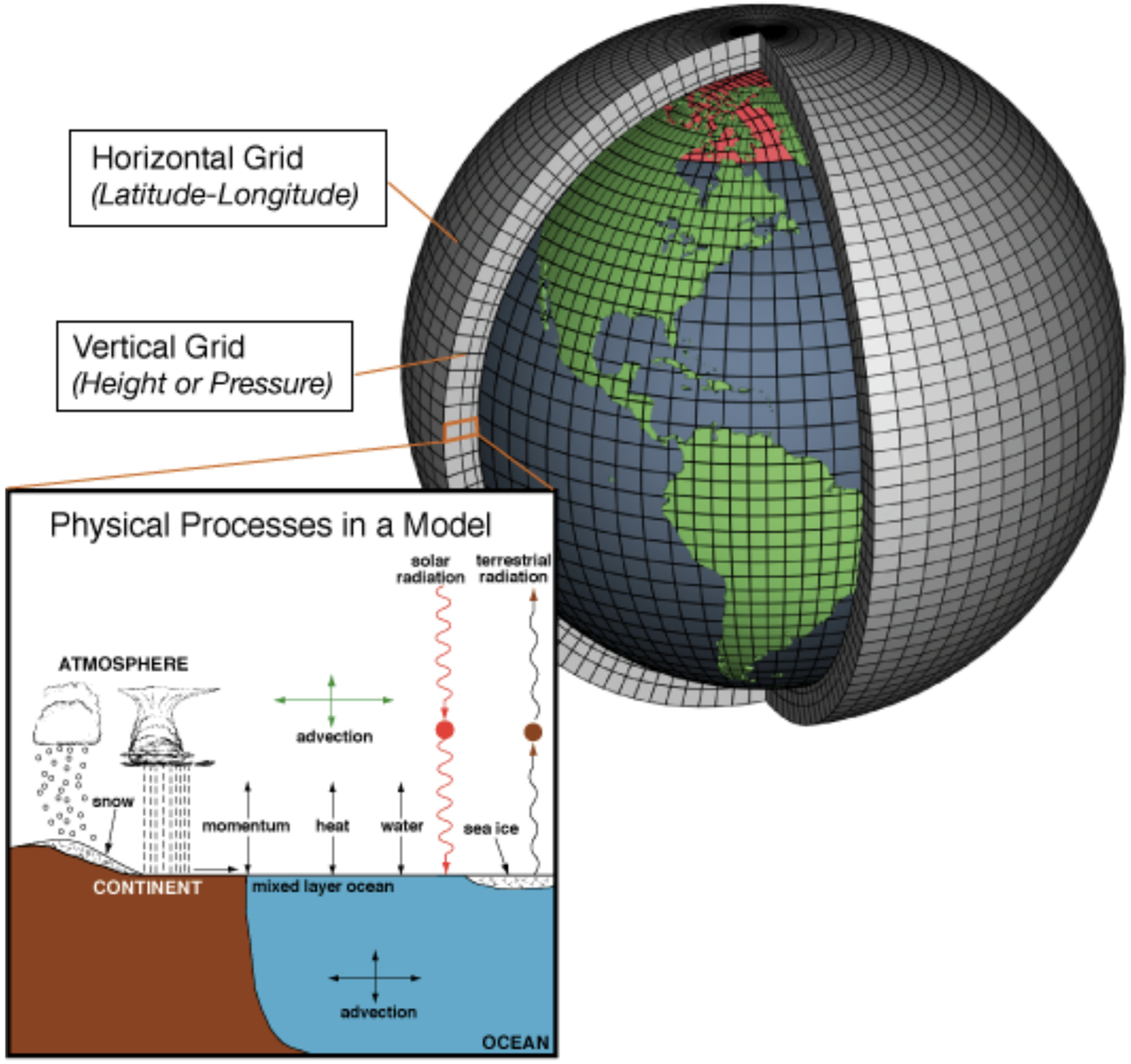


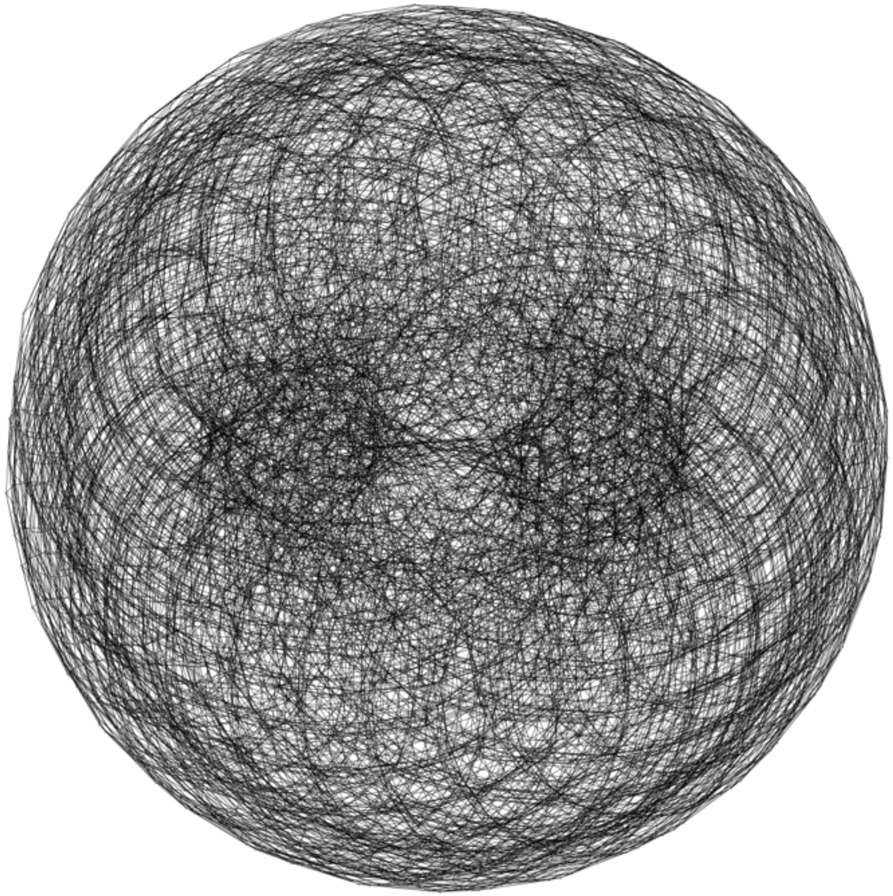
If you know your application …



U.S. National Oceanic and Atmospheric Administration

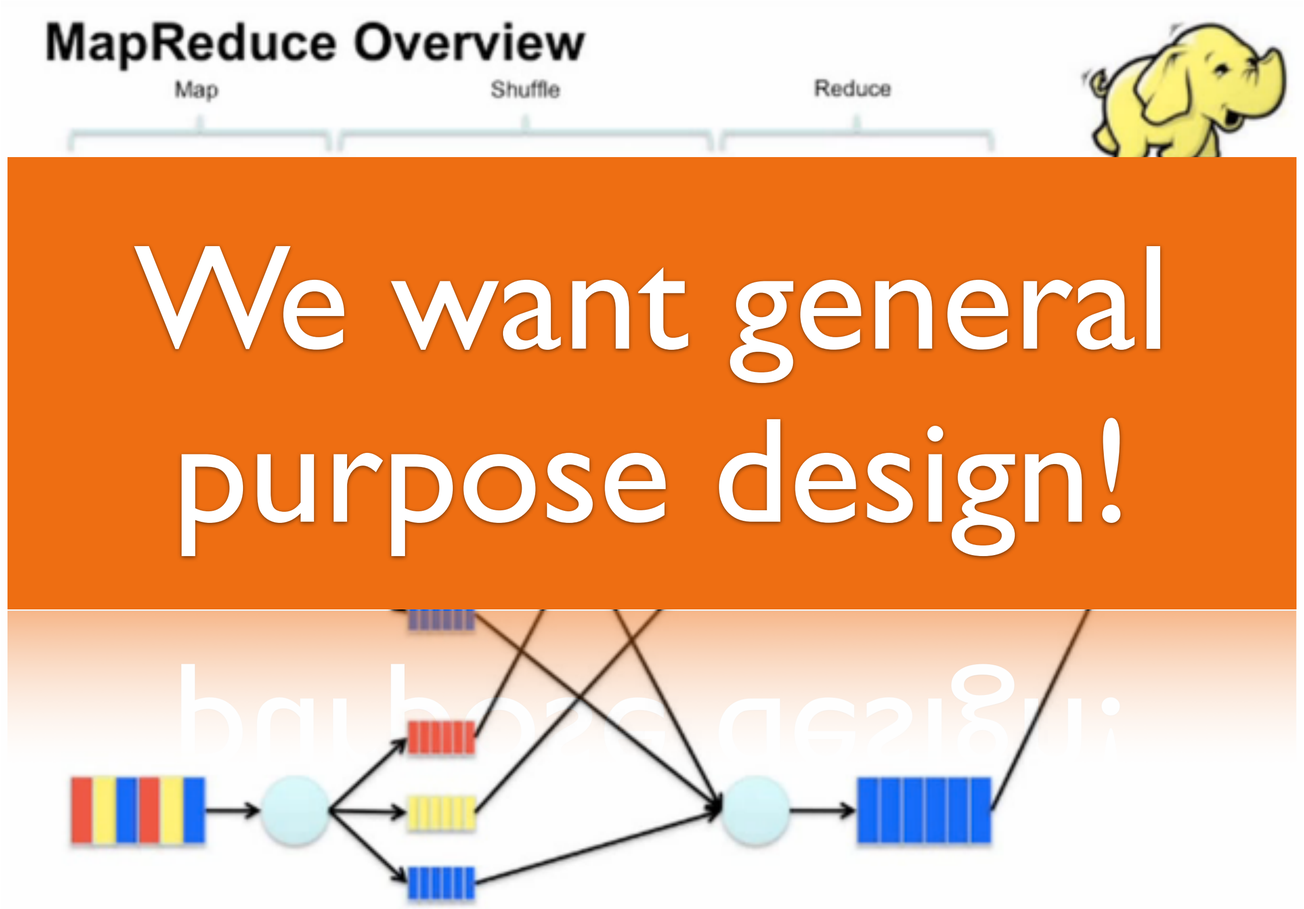
… design for it



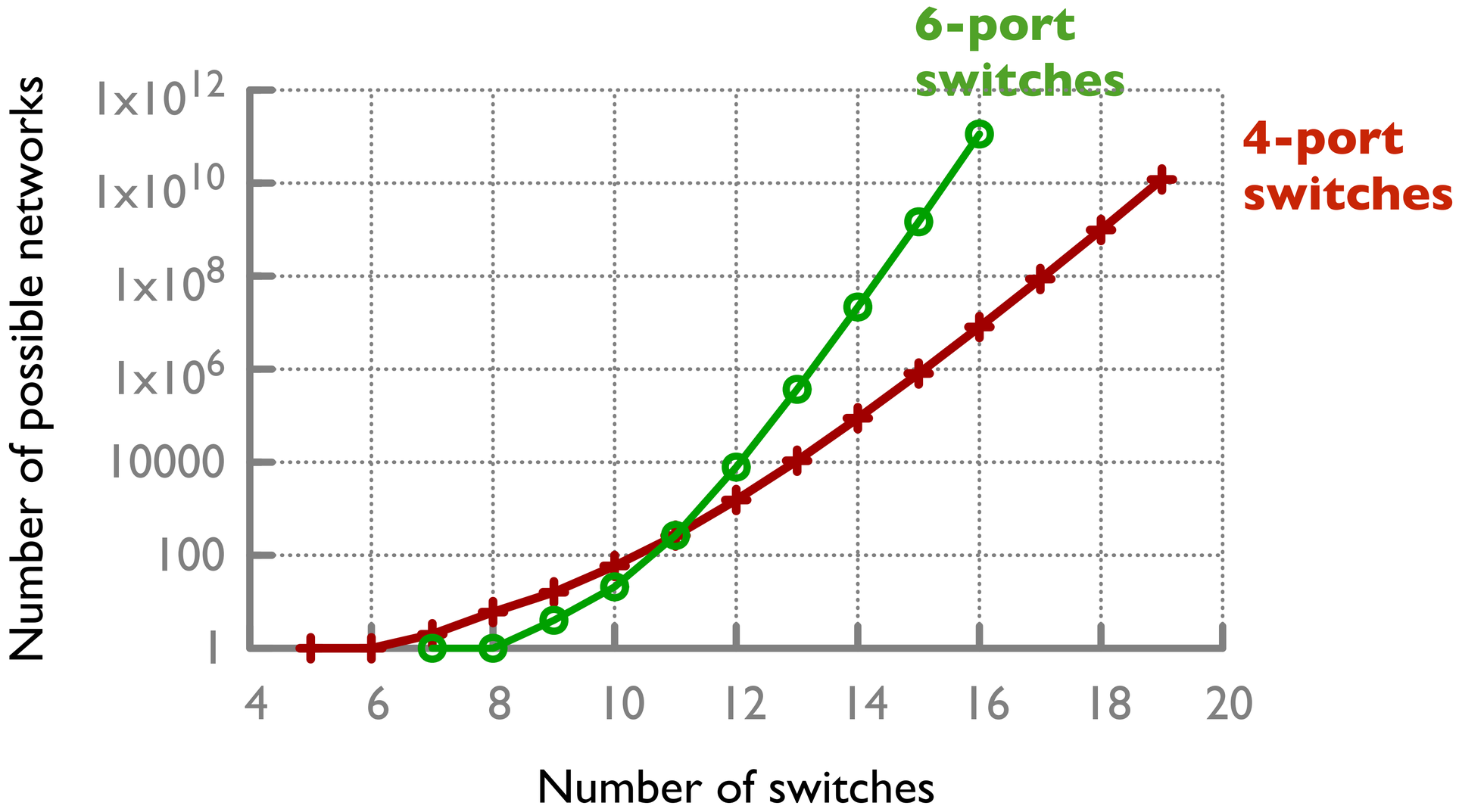


"Hopper", NERSC

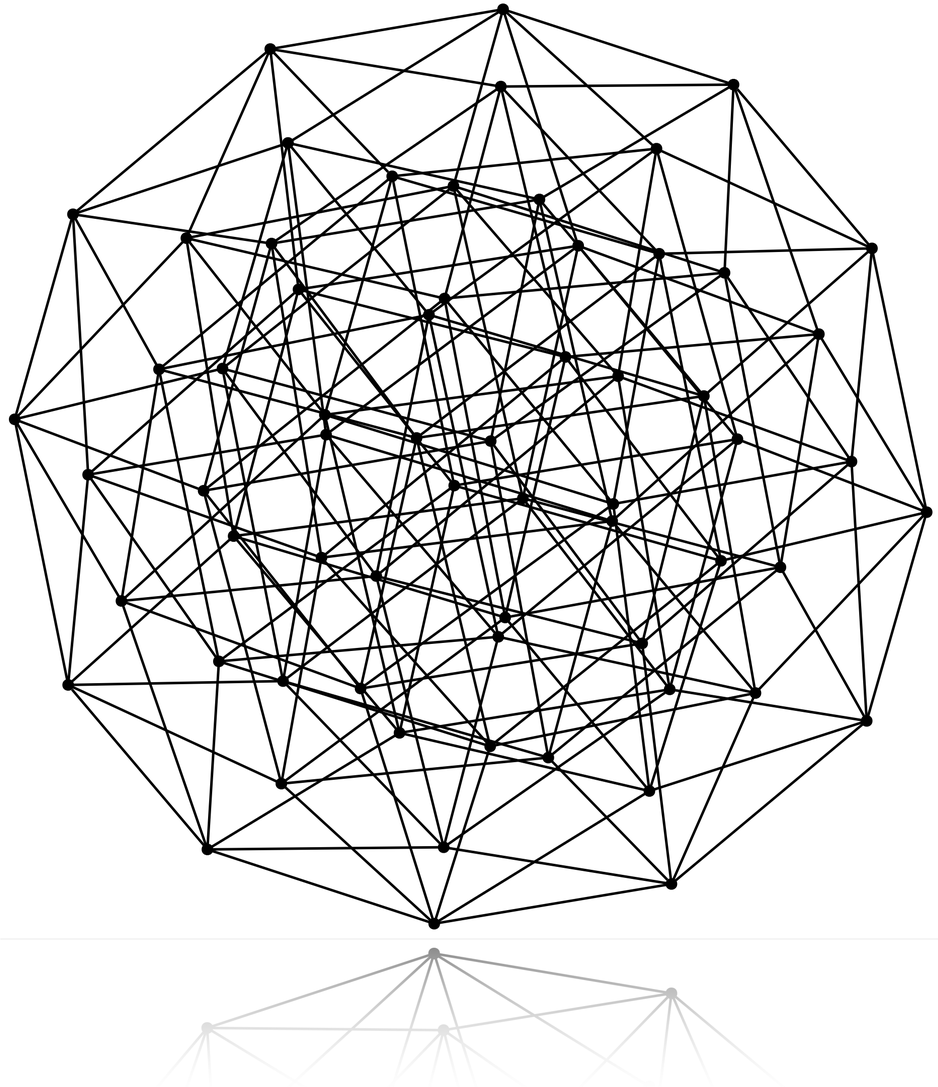
But, other apps may not work well …



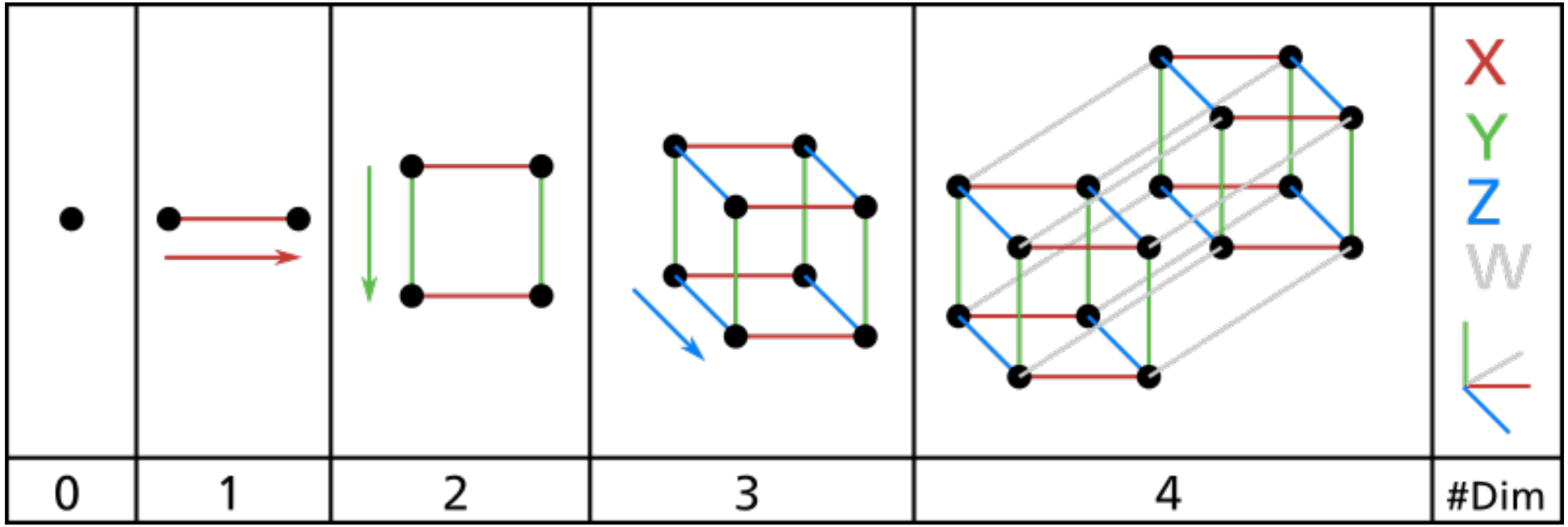
What's so hard about this?



So people pick “known good candidates”

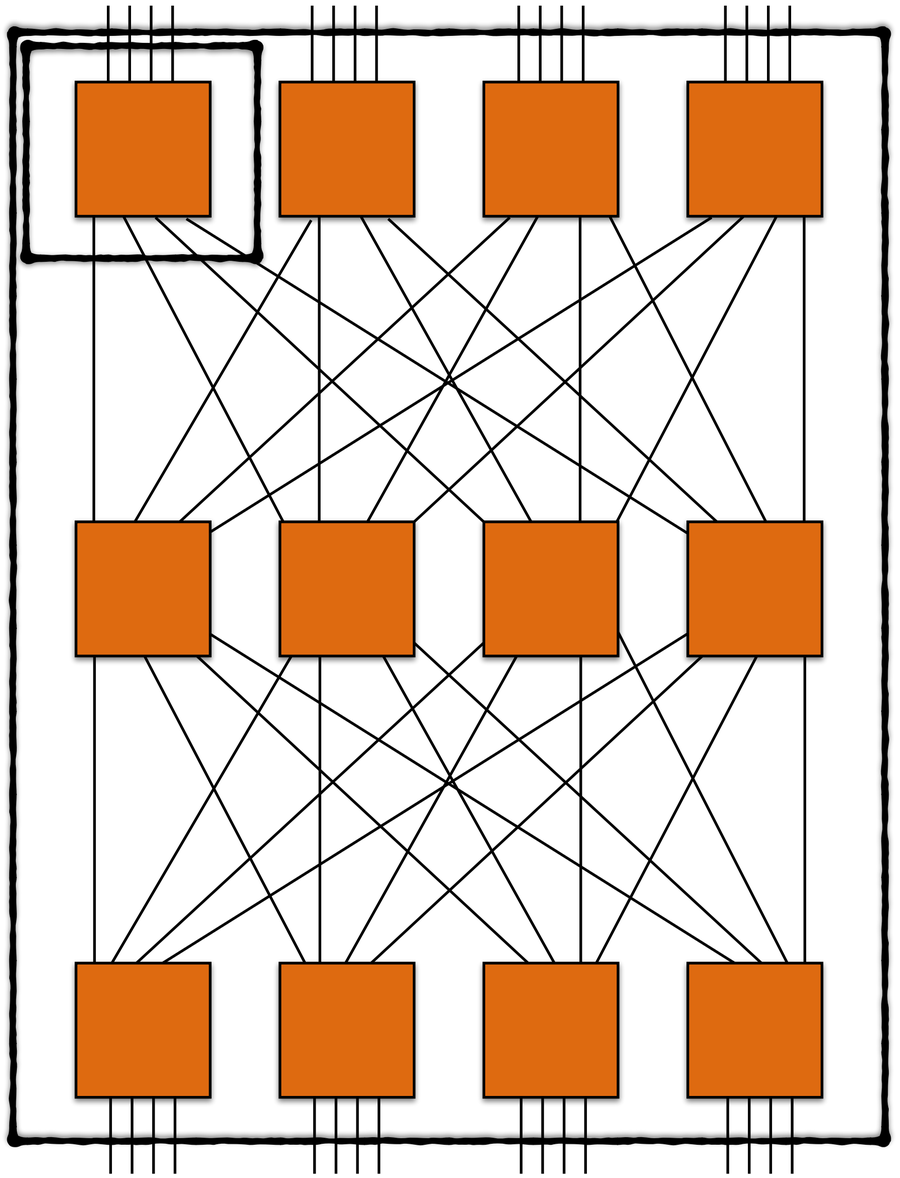


Hypercube



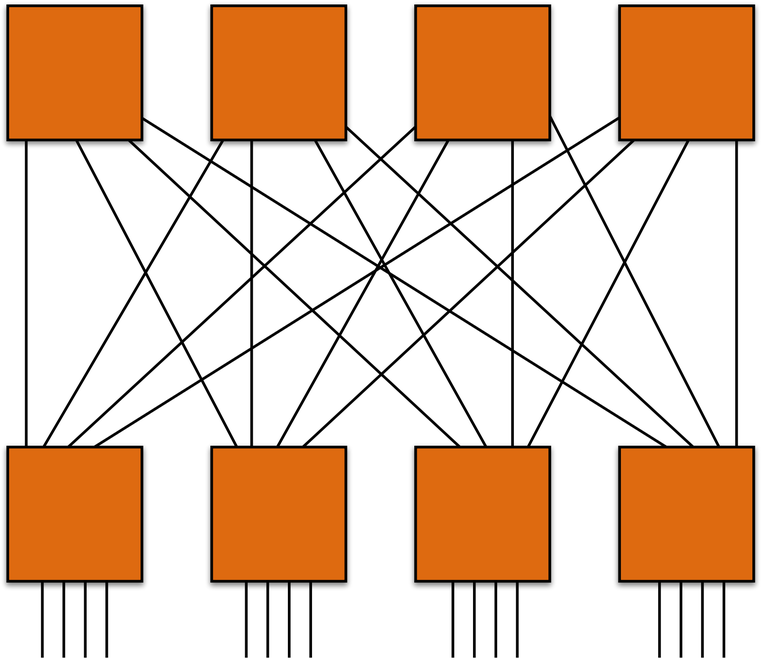
NerdBoy1392 [CC BY-SA 3.0], via Wikimedia

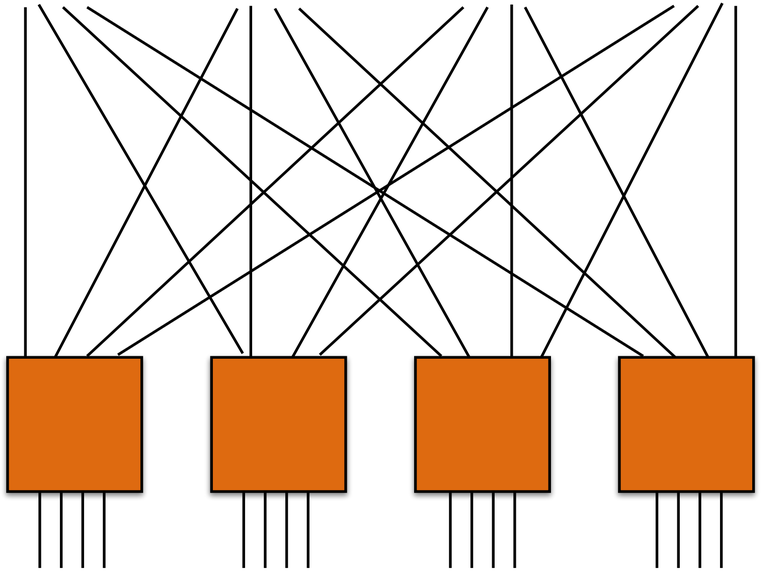
Clos networks



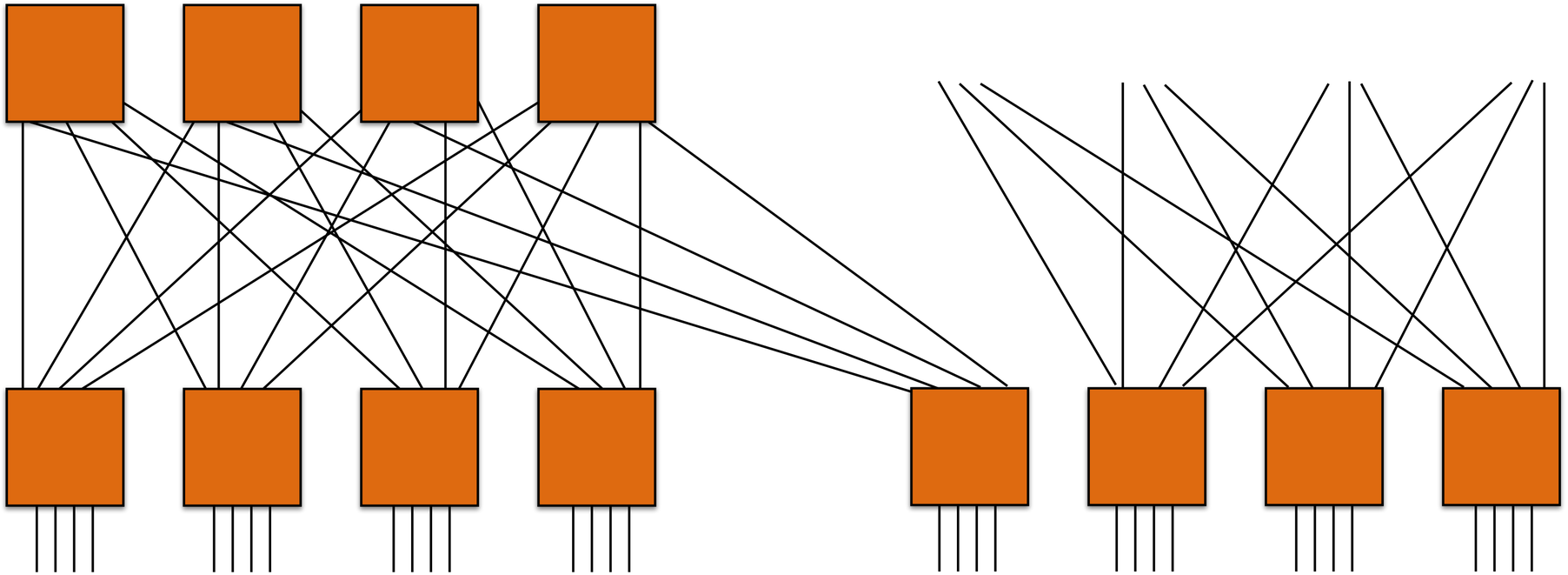
Use small, cheap elemen to build large networks!!

Folded-Clos?



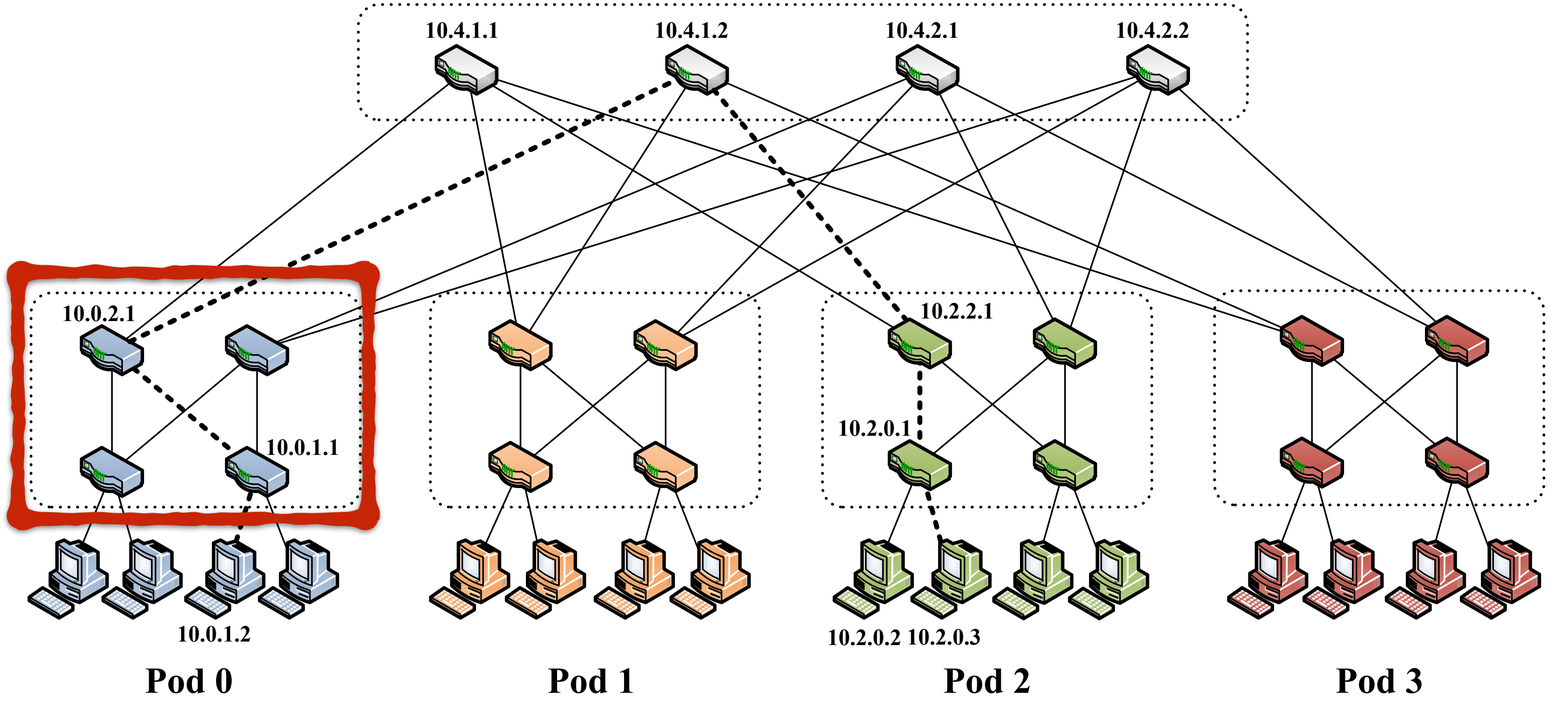


Folded-Clos?



Wikipedia user
Nachoman-au

Fat-tree network



ACM SIGCOMM, 2008

A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares Alexander Loukissas Amin Vahdat