

# Data centers: apps & traffic

Ankit Singla

ETH Zürich Spring 2018

# This lecture ...

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



**coursera**



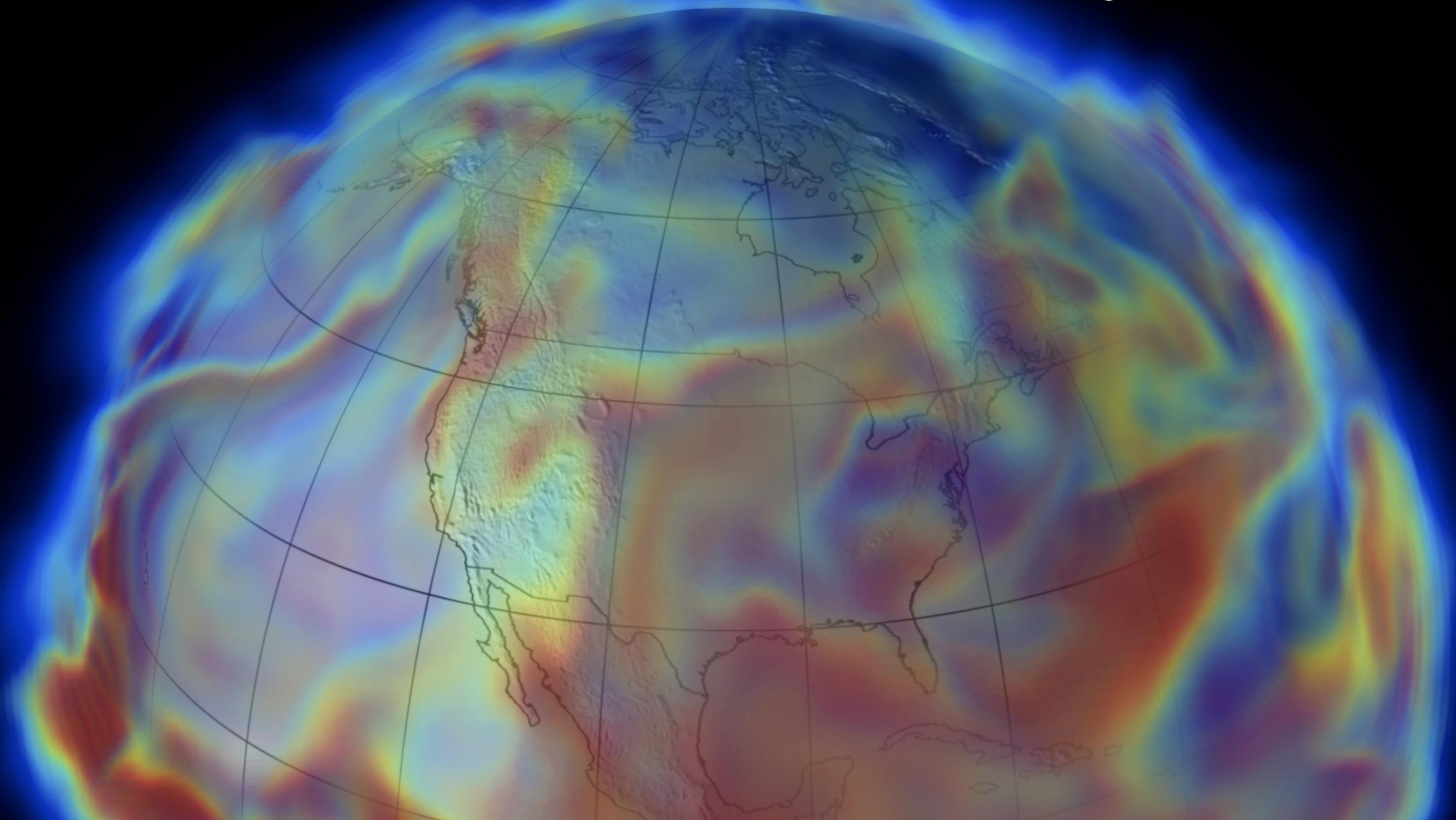
**amazon**

**Google**

**bing**

**NETFLIX**

[Image: NASA/Goddard/UMBC]



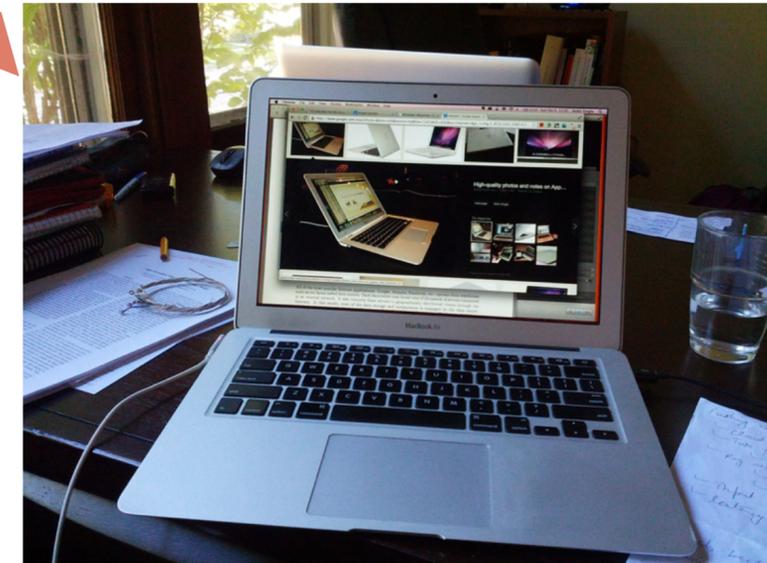
# How a Web search works



# 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”*

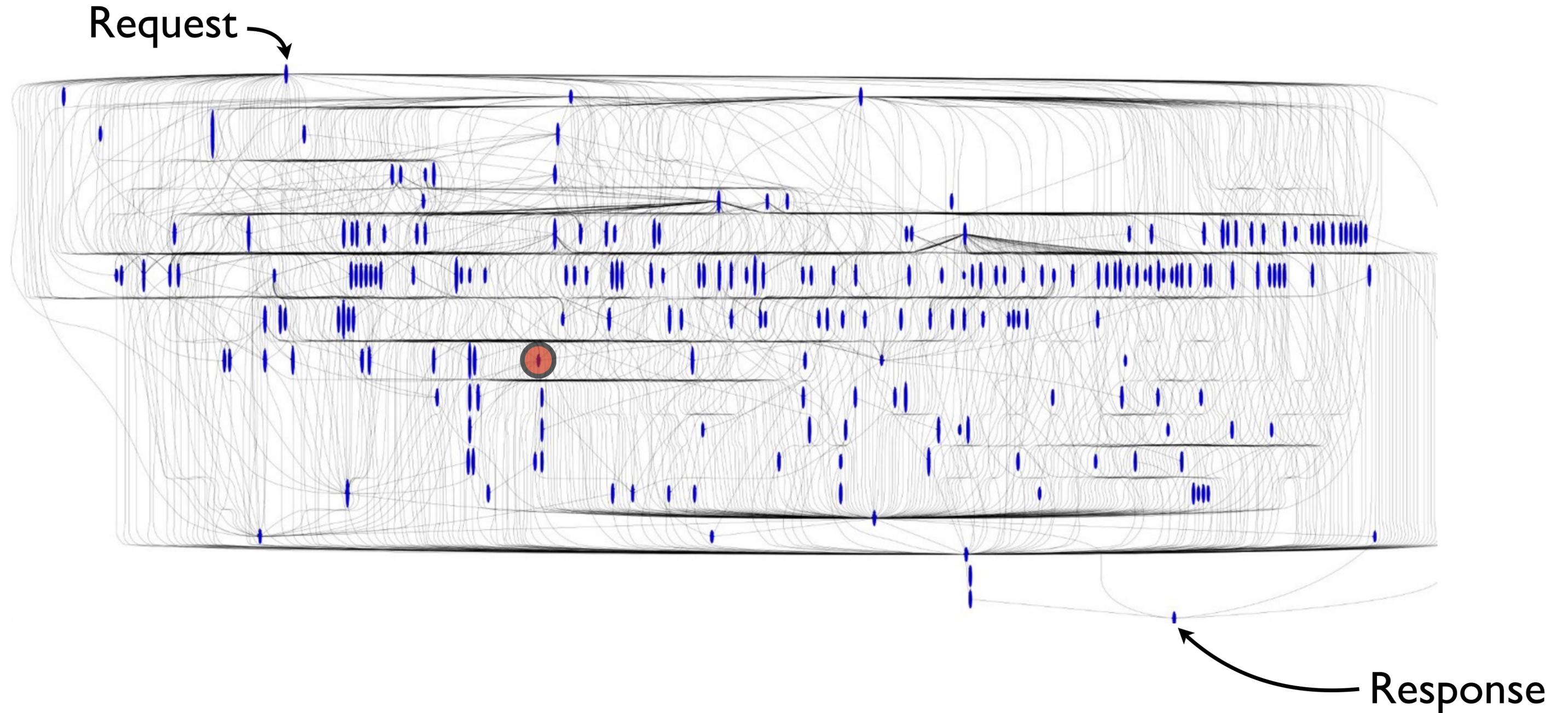


Image source: Talk on “Speeding up Distributed Request-Response Workflows”  
by Virajith Jalaparti at ACM SIGCOMM’13



# Other Web application traffic

**USENIX NSDI, 2013**

## **Scaling Memcache at Facebook**

Rajesh Nishtala, Hans Fugal, Steven Grimm, Marc Kwiatkowski, Herman Lee, Harry C. Li,  
Ryan McElroy, Mike Paleczny, Daniel Peek, Paul Saab, David Stafford, Tony Tung,  
Venkateshwaran Venkataramani  
{rajeshn,hans}@fb.com, {sgrimm, marc}@facebook.com, {herman, hcli, rm, mpal, dpeek, ps, dstaff, ttung, veeve}@fb.com  
*Facebook Inc.*

One popular page loaded  $\Rightarrow$  average of **521** distinct memcache fetches  
95th percentile: **1740** distinct memcache fetches

# Big data analytics

Hadoop

Spark

Database *joins*

■  
■  
■



What does data center  
traffic look like?

# Data center applications and traffic

ACM SIGCOMM,

## Inside the Social Network's (Datacenter) Network

Arjun Roy, Hongyi Zeng<sup>†</sup>, Jasmeet Bagga<sup>†</sup>, George Porter, and Alex C. Snoeren

- 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  $\Rightarrow$  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<sup>‡</sup>, Mihai Dobrescu<sup>†</sup>, Jianqing Du<sup>†</sup>, Katerina Argyraki<sup>†</sup>, Byung-Gon Chun<sup>§</sup>, Kevin Fall<sup>§</sup>,  
Gianluca Iannaccone<sup>§</sup>, Allan Knies<sup>§</sup>, Maziar Manesh<sup>§</sup>, Laurent Mathy<sup>‡</sup>, Sylvia Ratnasamy<sup>§</sup>  
<sup>§</sup> Intel Research, <sup>†</sup> EPFL, <sup>‡</sup> Lancaster University

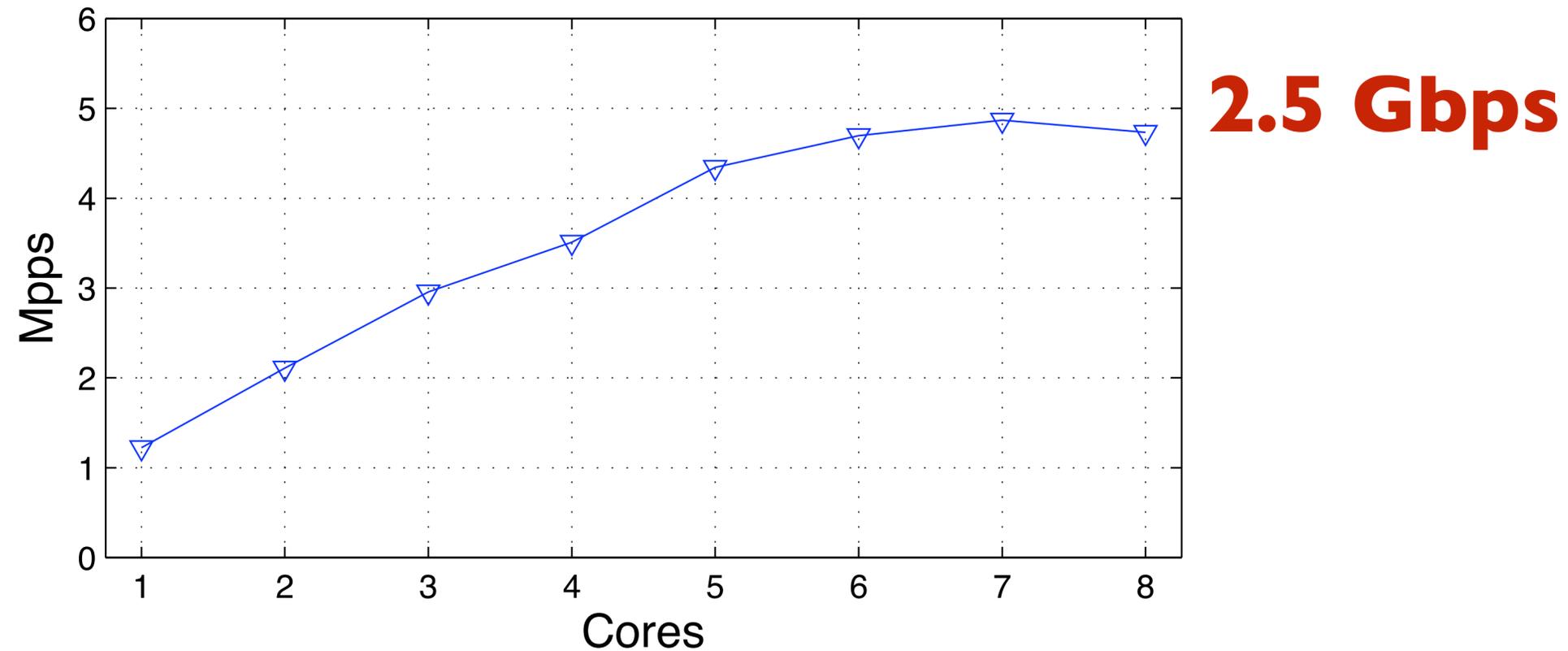


Figure 14: *Per-flow monitoring performance*

# 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.

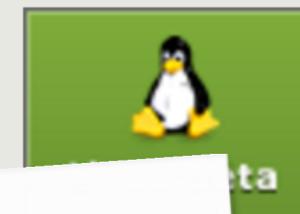
Page Contents:

- What is protein folding?
- Why is protein folding important?
- Foldit and protein folding
- News
- Rosemary
- Comments
- Let's play
- Instructions
- Terms of Service
- Credits

# Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks

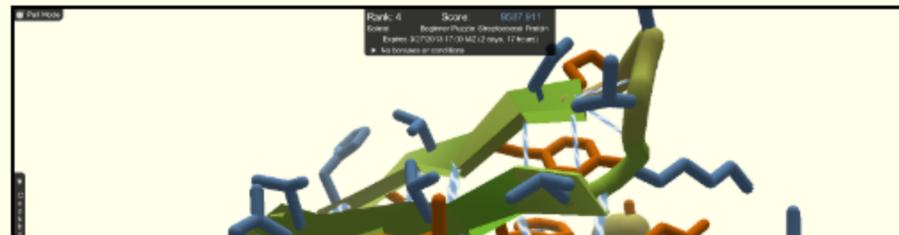
The online game poses protein-folding puzzles, and participants provided insights recently that solved the structure of an enzyme involved in reproduction of HIV

### GET STARTED: DOWNLOAD



## 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.



### RECOMMEND FOLDIT

### USER LOGIN

  Only search fold.it

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

[Home](#) | [News & Comment](#) | [Research](#) | [Careers & Jobs](#) | [Current Issue](#) | [Archive](#) | [Audio & Video](#) | [For](#)

[News & Comment](#)

[News](#)

[2017](#)

[February](#)

[Article](#)



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 unfunited 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

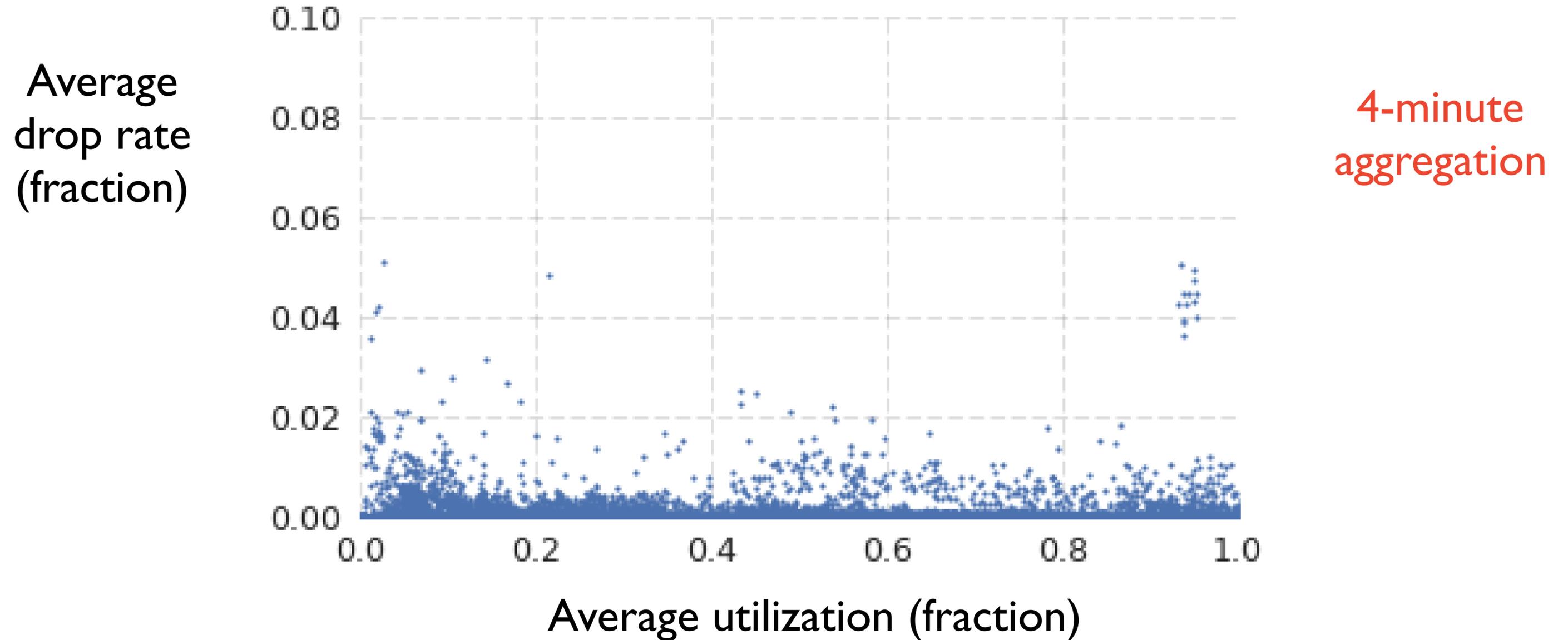


**Facebook: “machine to machine” traffic is several orders of magnitude larger than what goes out to the Internet**



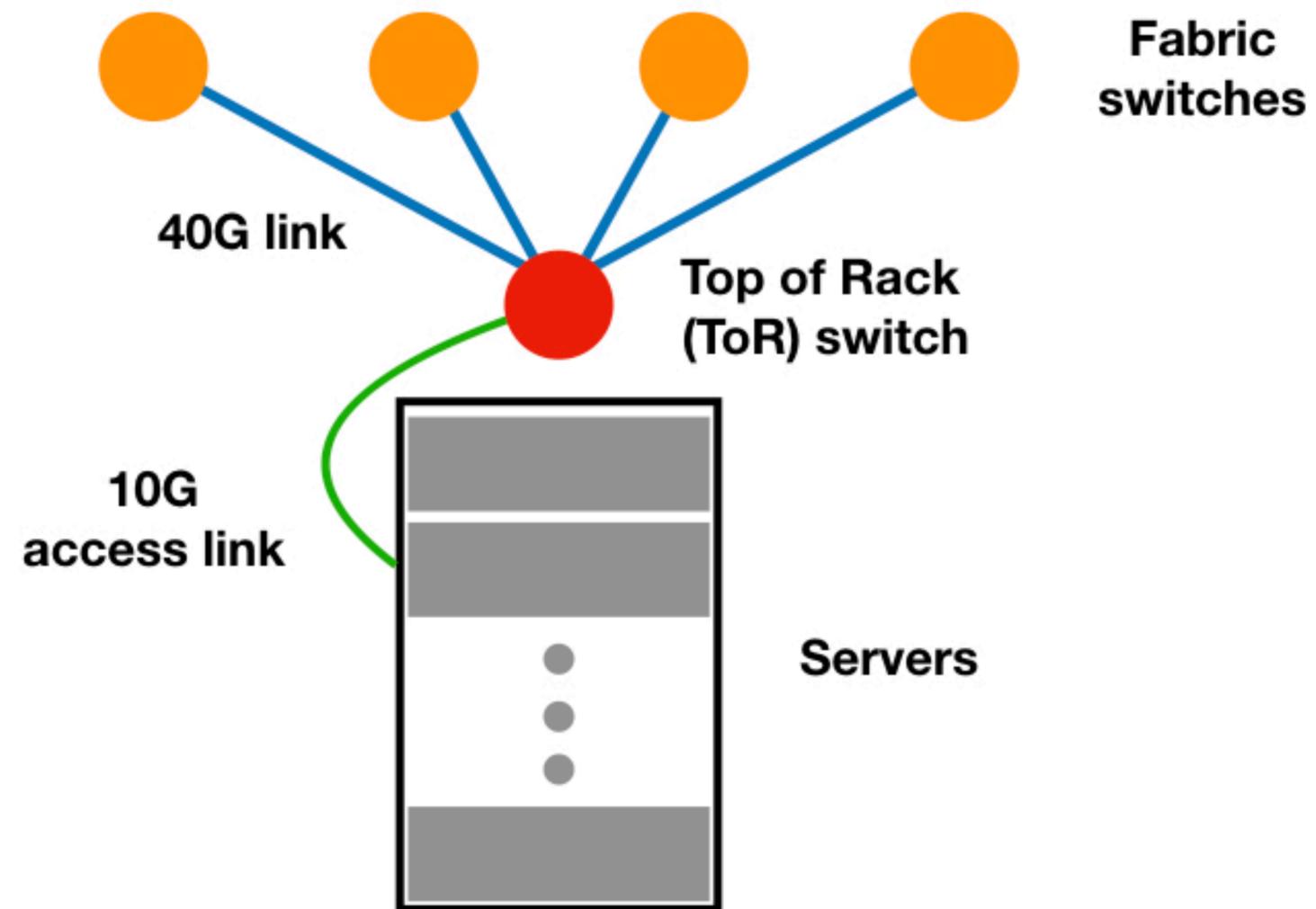
“Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google’s Datacenter Network”, Arjun Singh et al. @ **Google**, ACM SIGCOMM’15

# Where and how does congestion happen?



“High-Resolution Measurement of Data Center Microbursts”,  
Zhang et al., ACM IMC’17 — data from Facebook

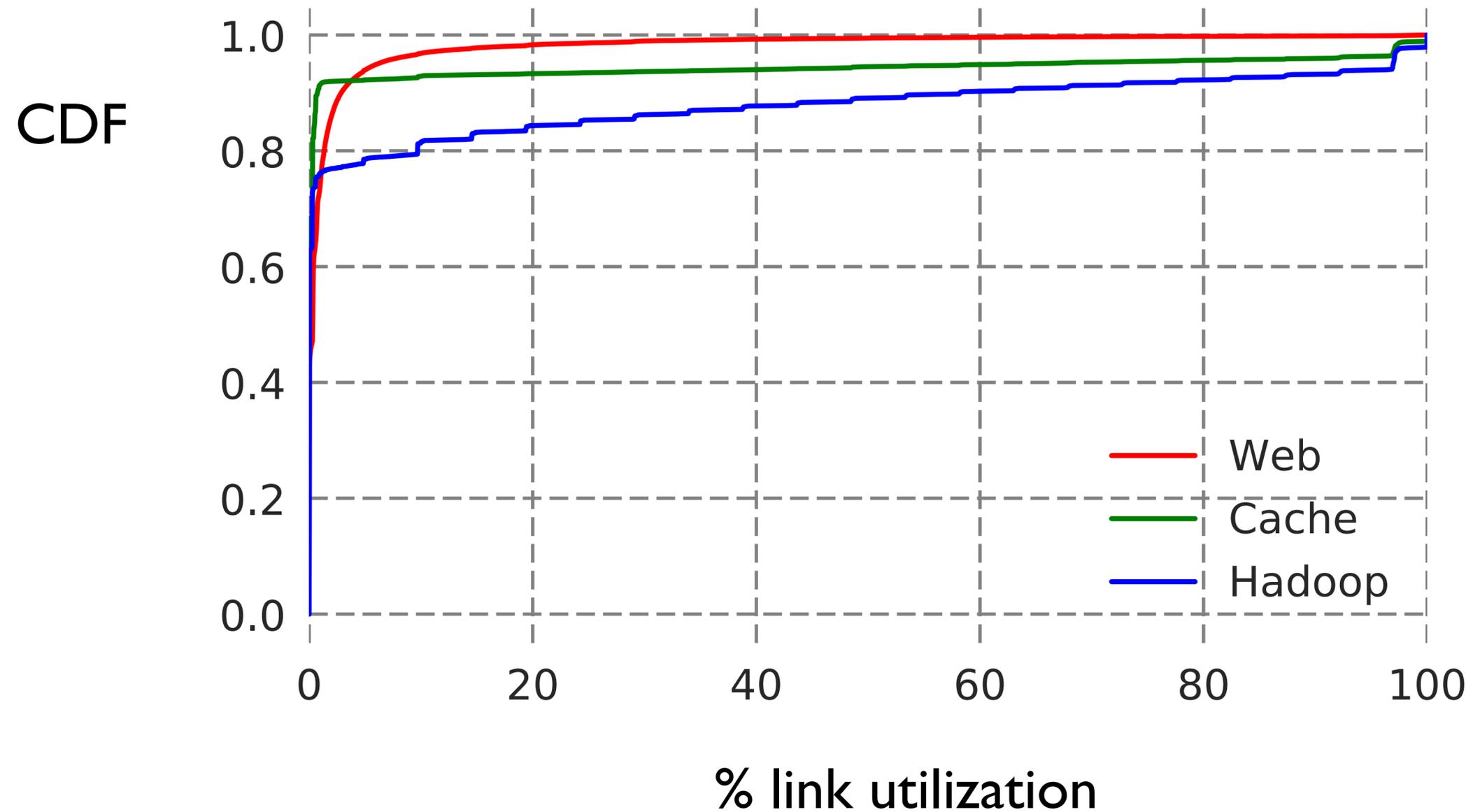
# Congestion: a closer look



Measure at ToR,  
**25  $\mu$ s granularity**

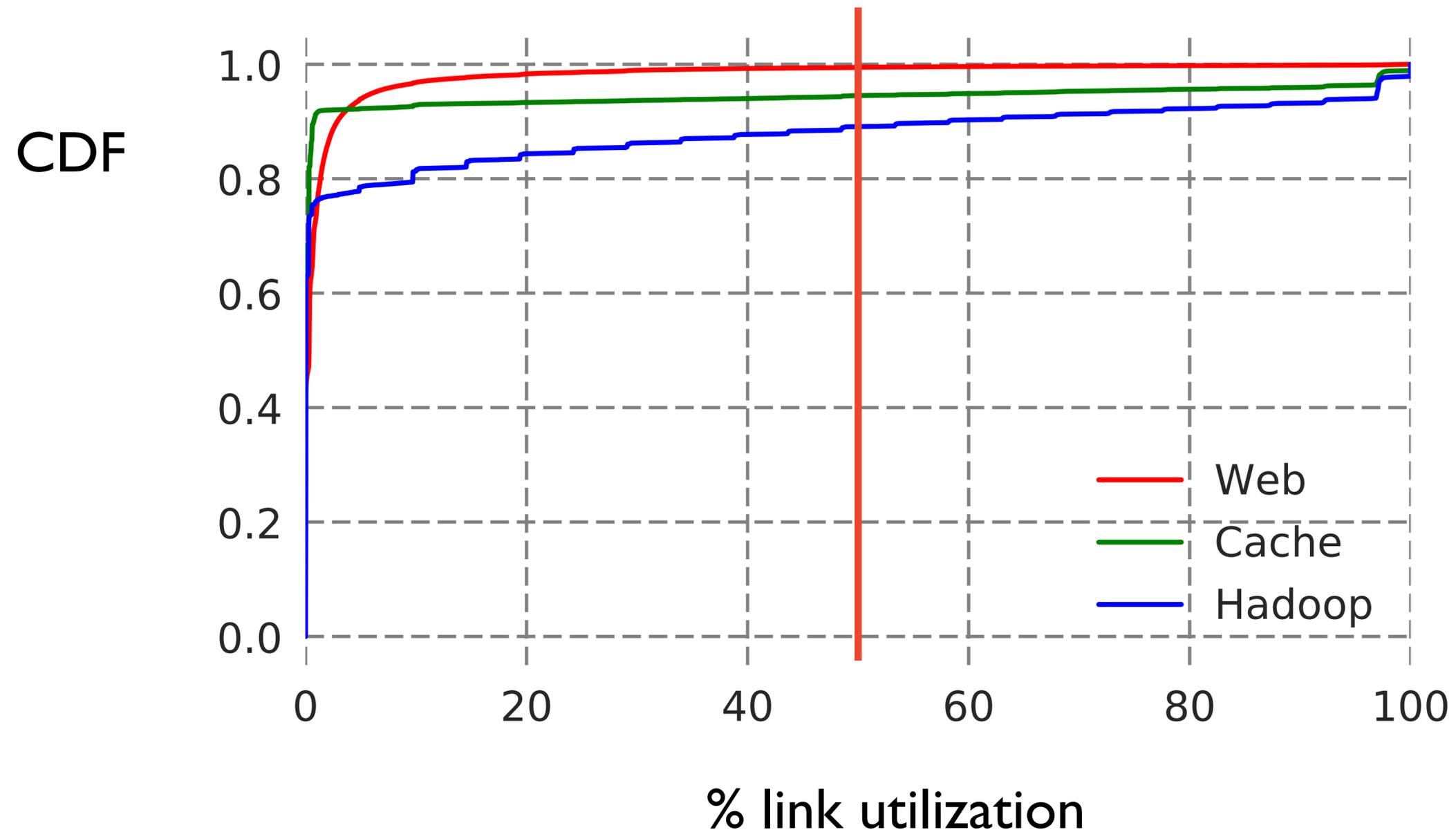
“High-Resolution Measurement of Data Center Microbursts”,  
Zhang et al., ACM IMC’17 — data from Facebook

# Low overall utilization, with some bursts



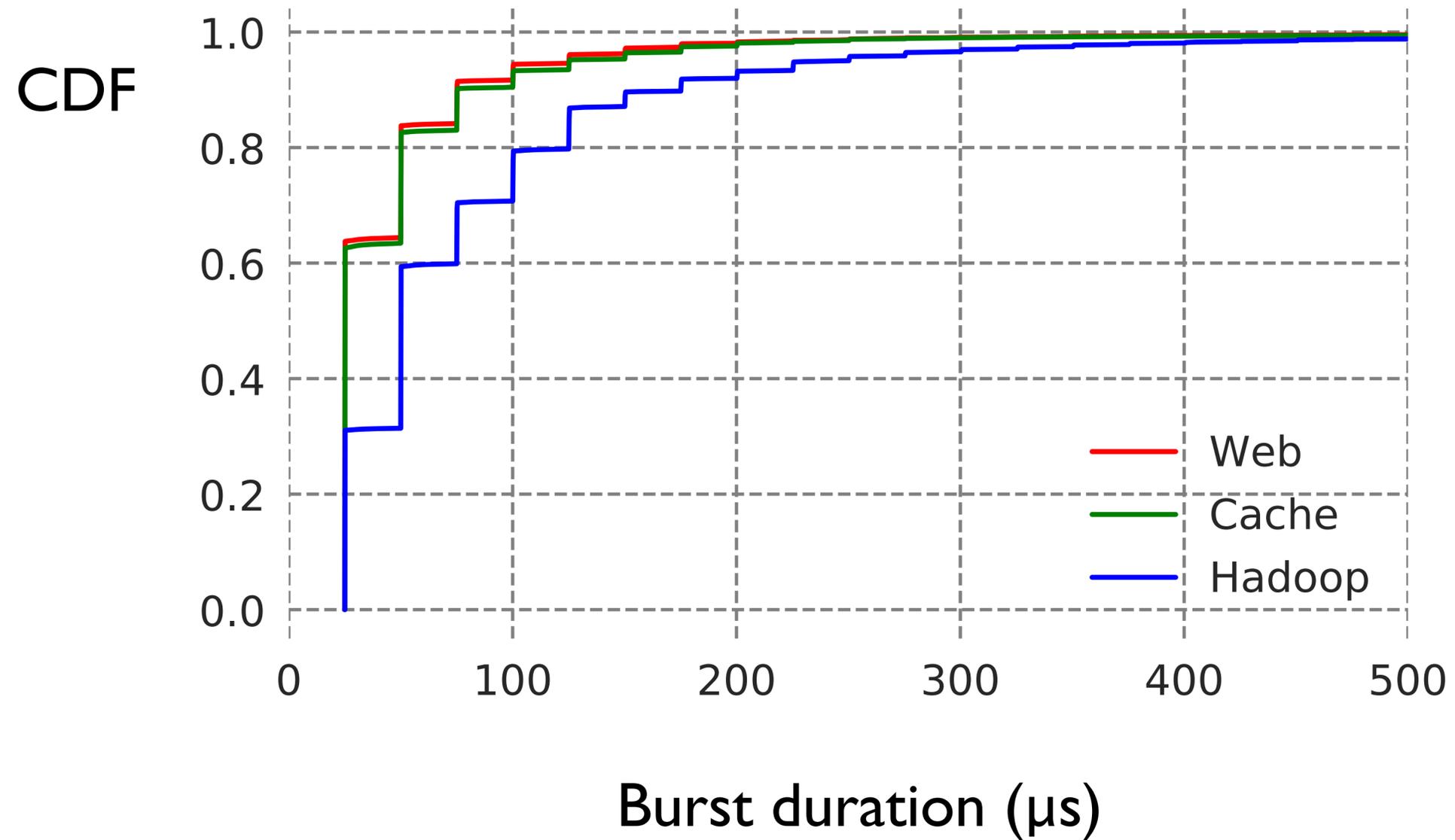
“High-Resolution Measurement of Data Center Microbursts”,  
Zhang et al., ACM IMC’17 — data from Facebook

# “Bursts”: short high-utilization periods



“High-Resolution Measurement of Data Center Microbursts”,  
Zhang et al., ACM IMC’17 — data from Facebook

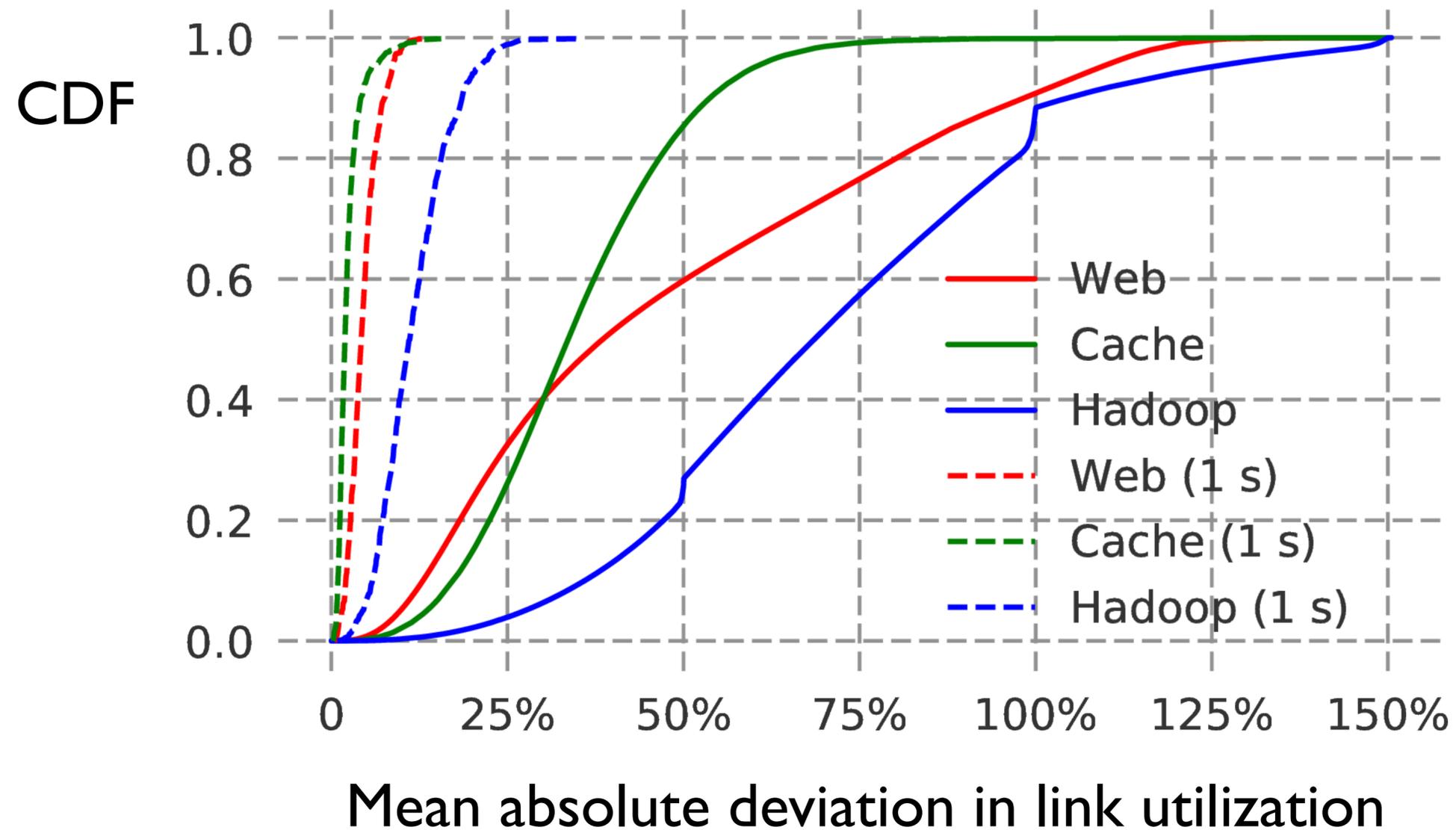
# “Bursts”: continuous 25 $\mu\text{s}$ intervals with $>50\%$



$\mu$ -bursts cause most drops!

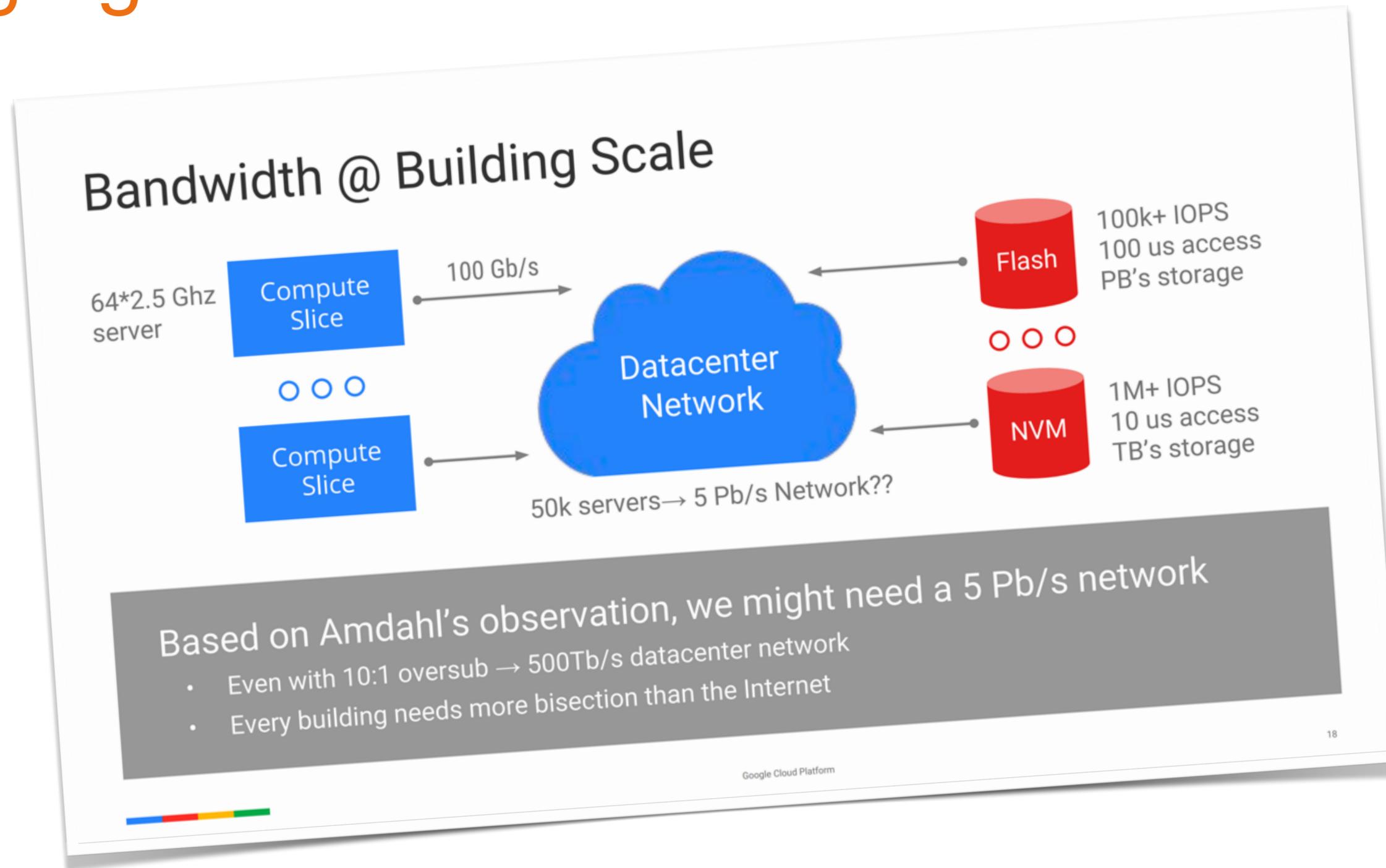
“High-Resolution Measurement of Data Center Microbursts”,  
Zhang et al., ACM IMC’17 — data from Facebook

# Link utilization skewed at small timescales



“High-Resolution Measurement of Data Center Microbursts”,  
Zhang et al., ACM IMC’17 — data from Facebook

# “Disaggregated data centers”



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

# Traffic characteristics: rack locality

## Facebook

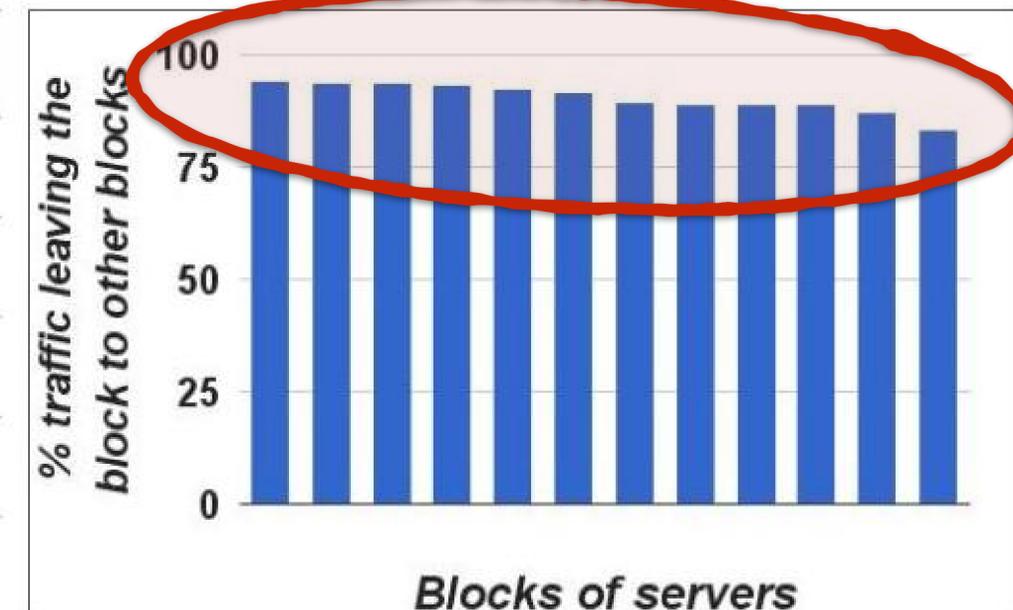
“Inside the Social Network’s (Datacenter) Network”  
Arjun Roy et al., ACM SIGCOMM’15

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

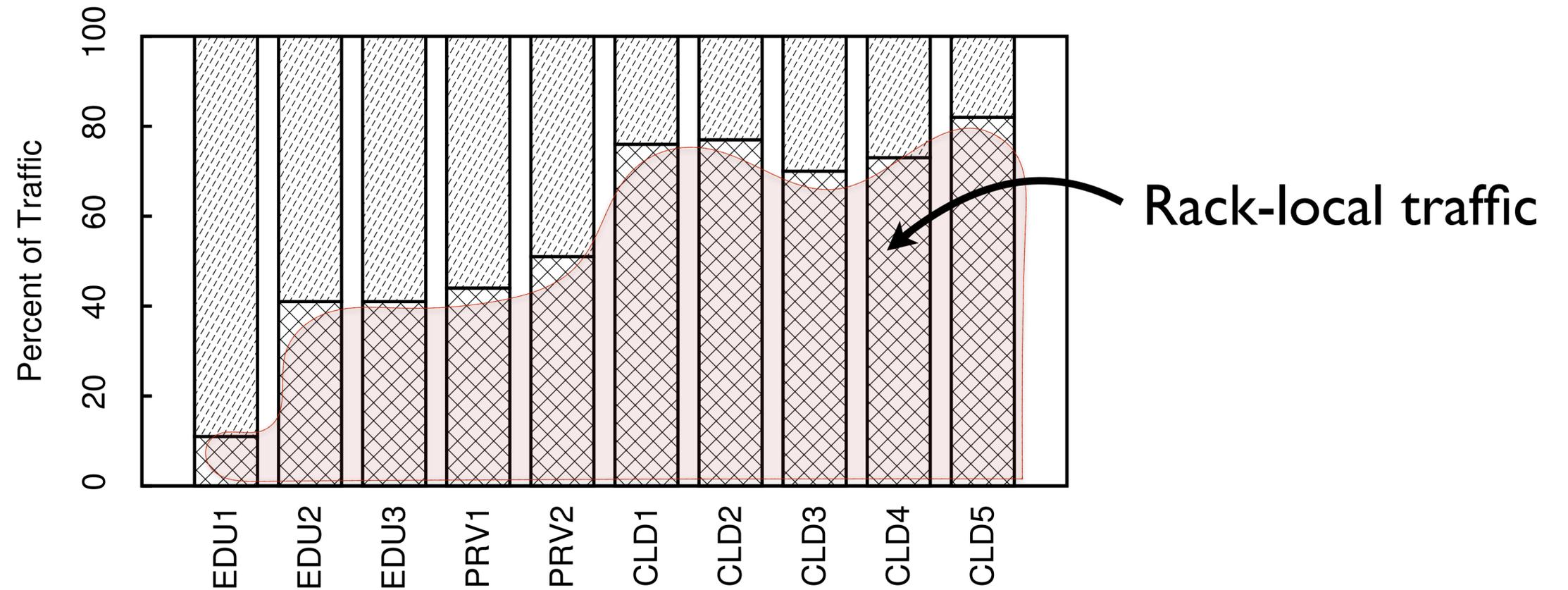
## Google

“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  $\approx$  median flow, not persistent

## 1500 server cluster @ ??

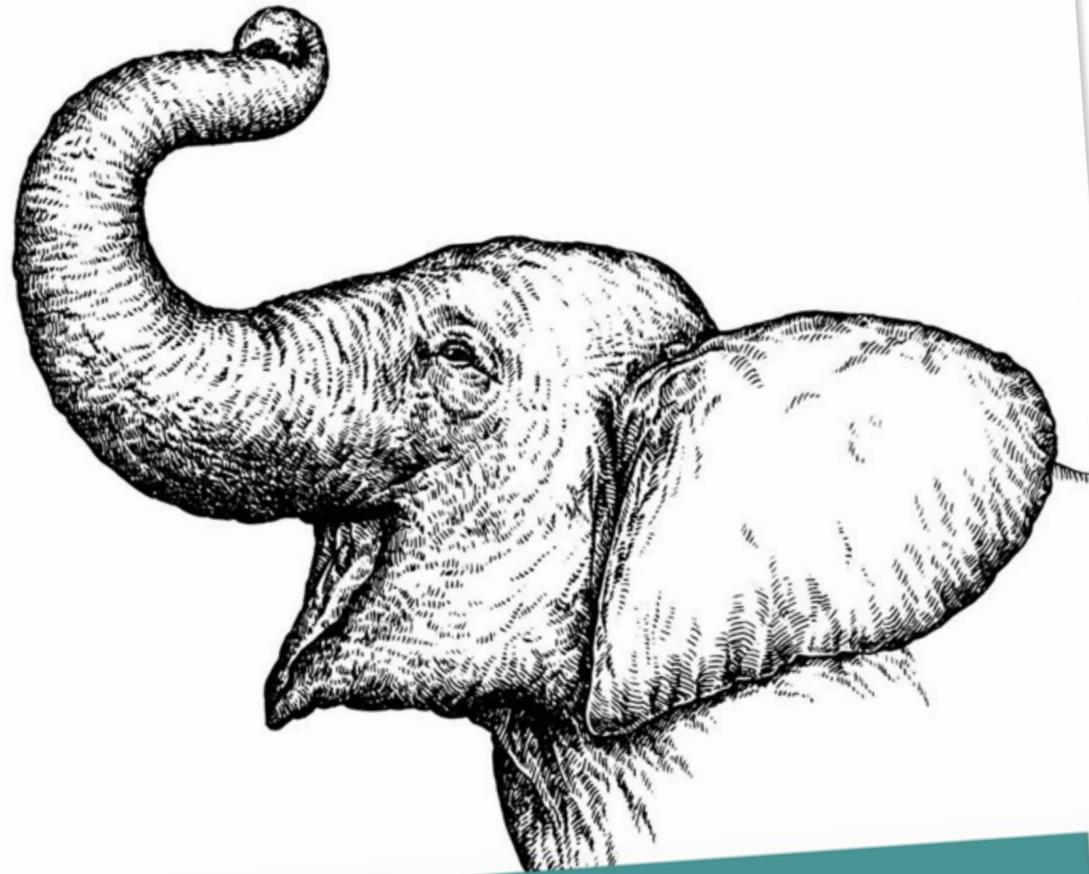
“The Nature of Datacenter Traffic: Measurements & Analysis”  
Srikanth Kandula et al. (Microsoft Research), ACM IMC’09

> 80% of the flows last < 10sec

> 50% bytes are in flows lasting less < 25sec

So ... what does data  
center traffic look like?

*The answer to every programming question ever conceived*



# It Depends

*The Definitive Guide*

O RLY?

@ThePracticalDev

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

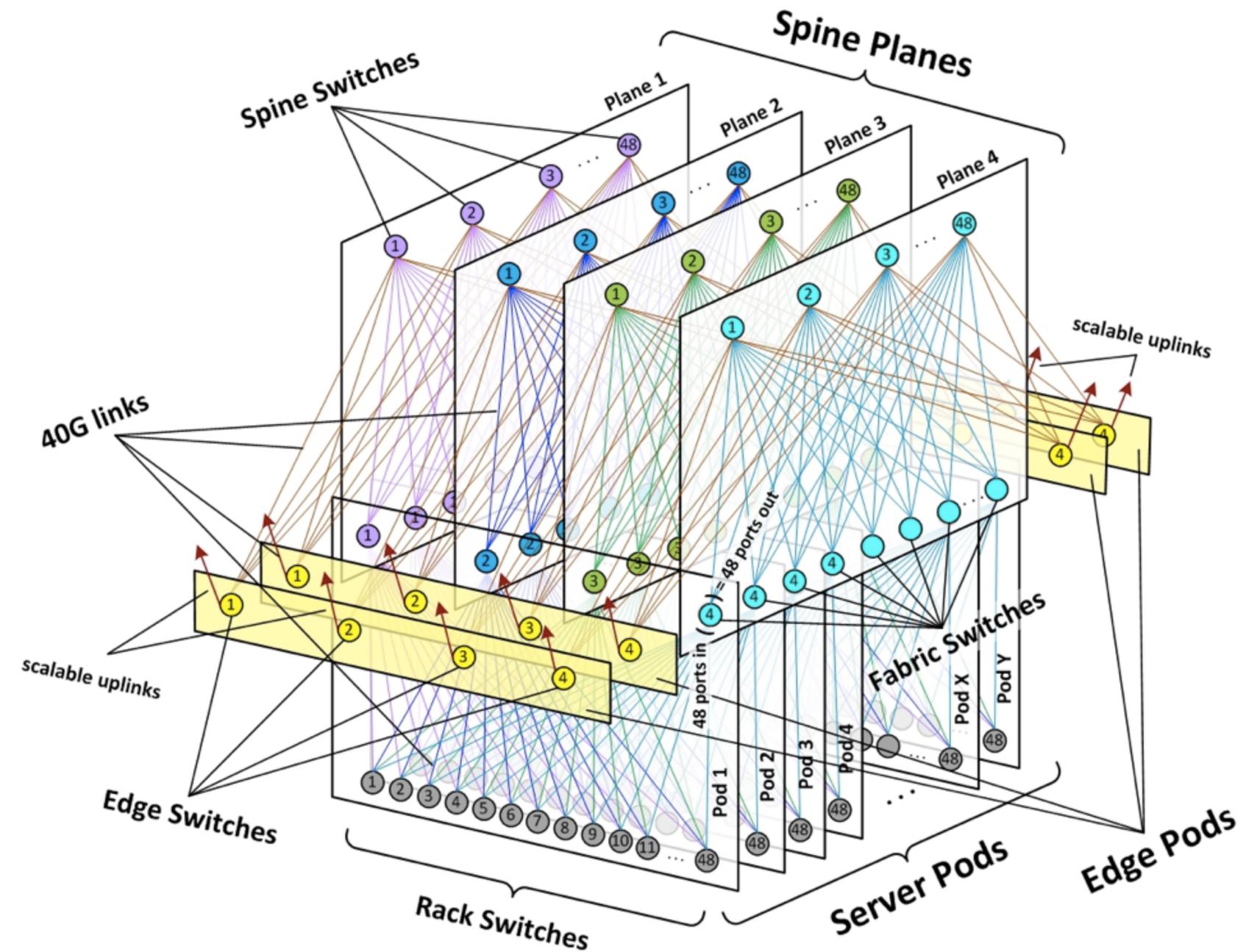
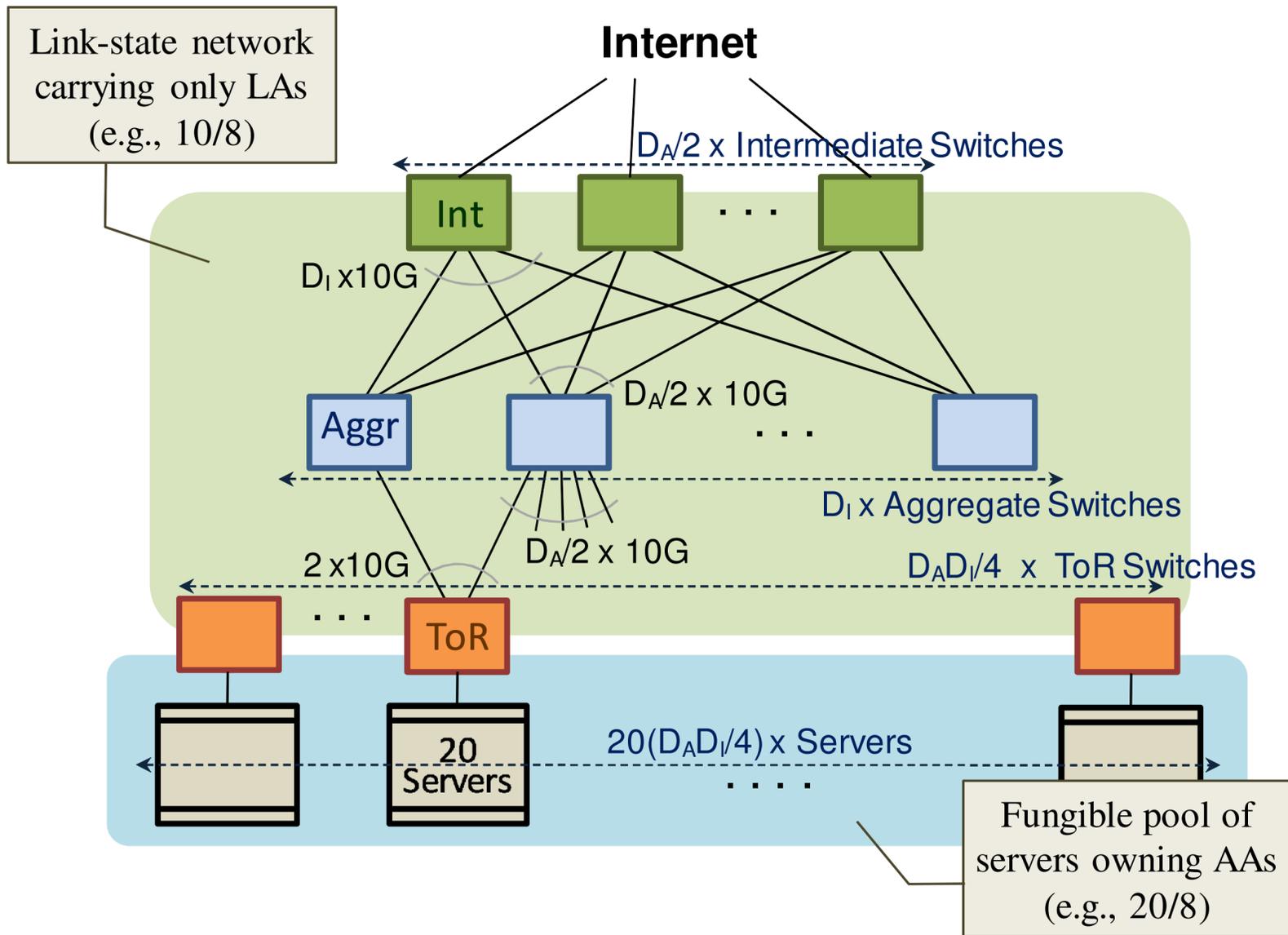
50x Traffic generated by servers in our datacenters

**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?



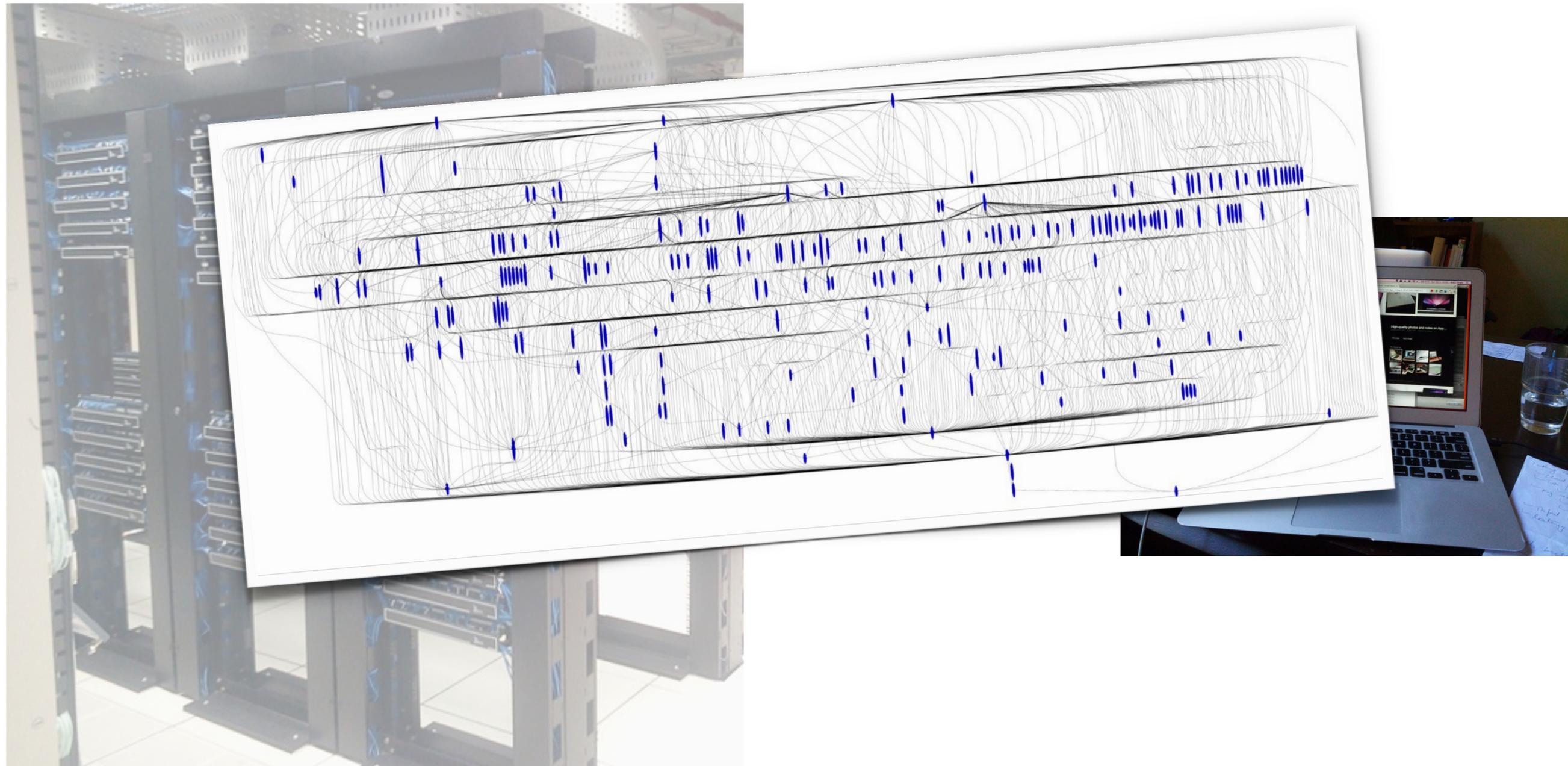
VL2 @ **Microsoft**, ACM SIGCOMM'09

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

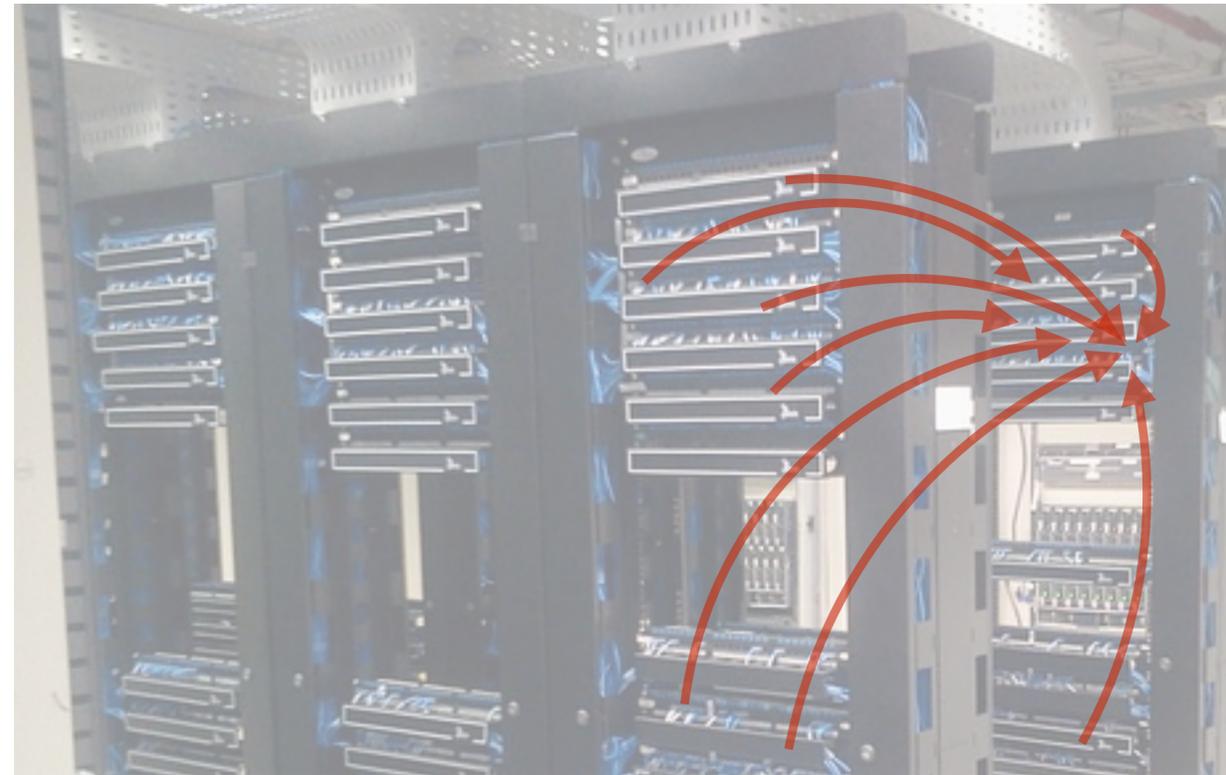
Suppose: server response-time is 10ms for 99% of requests; 1s for 1%

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

**Need to reduce variability and tolerate some variation**

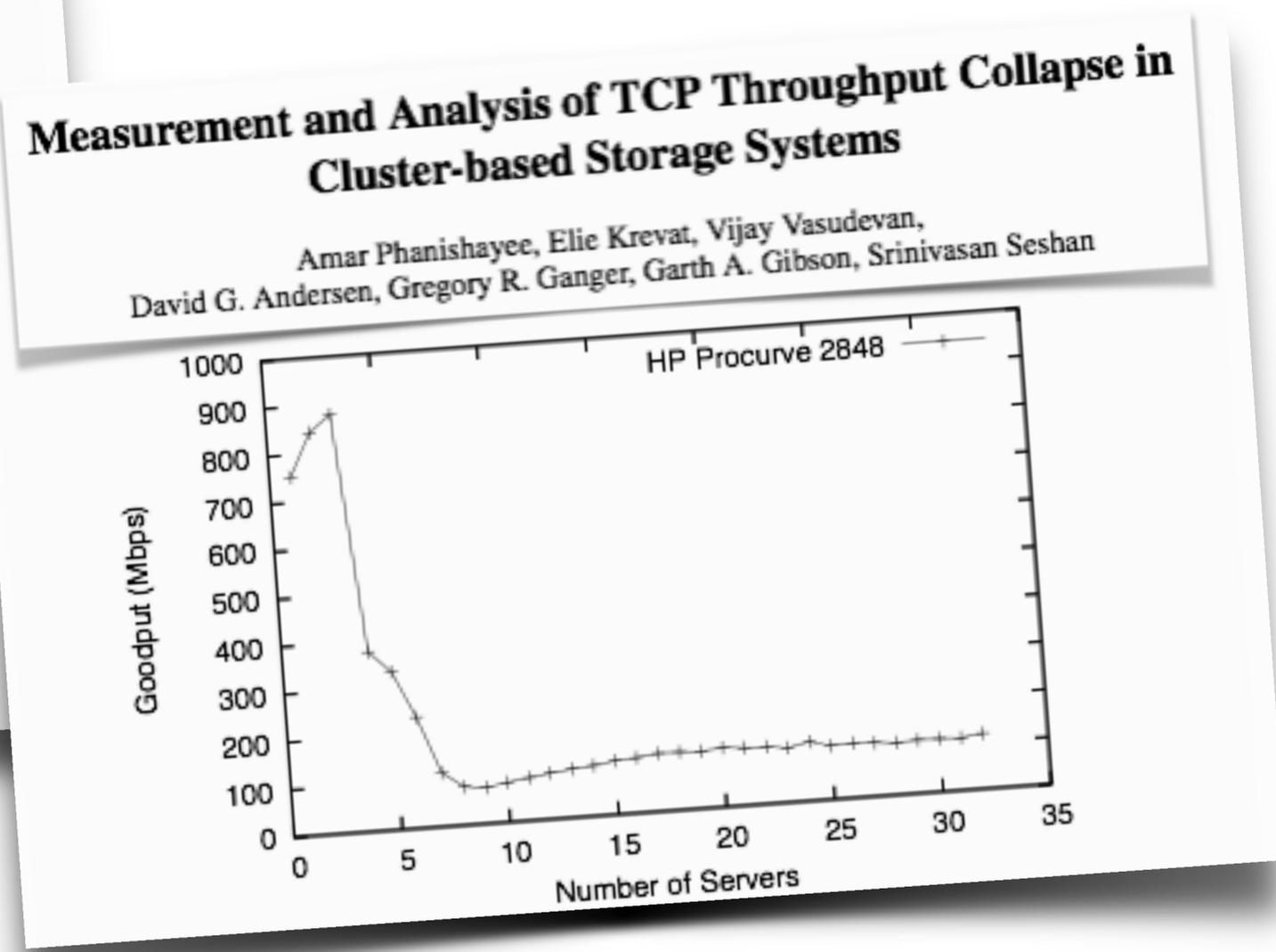
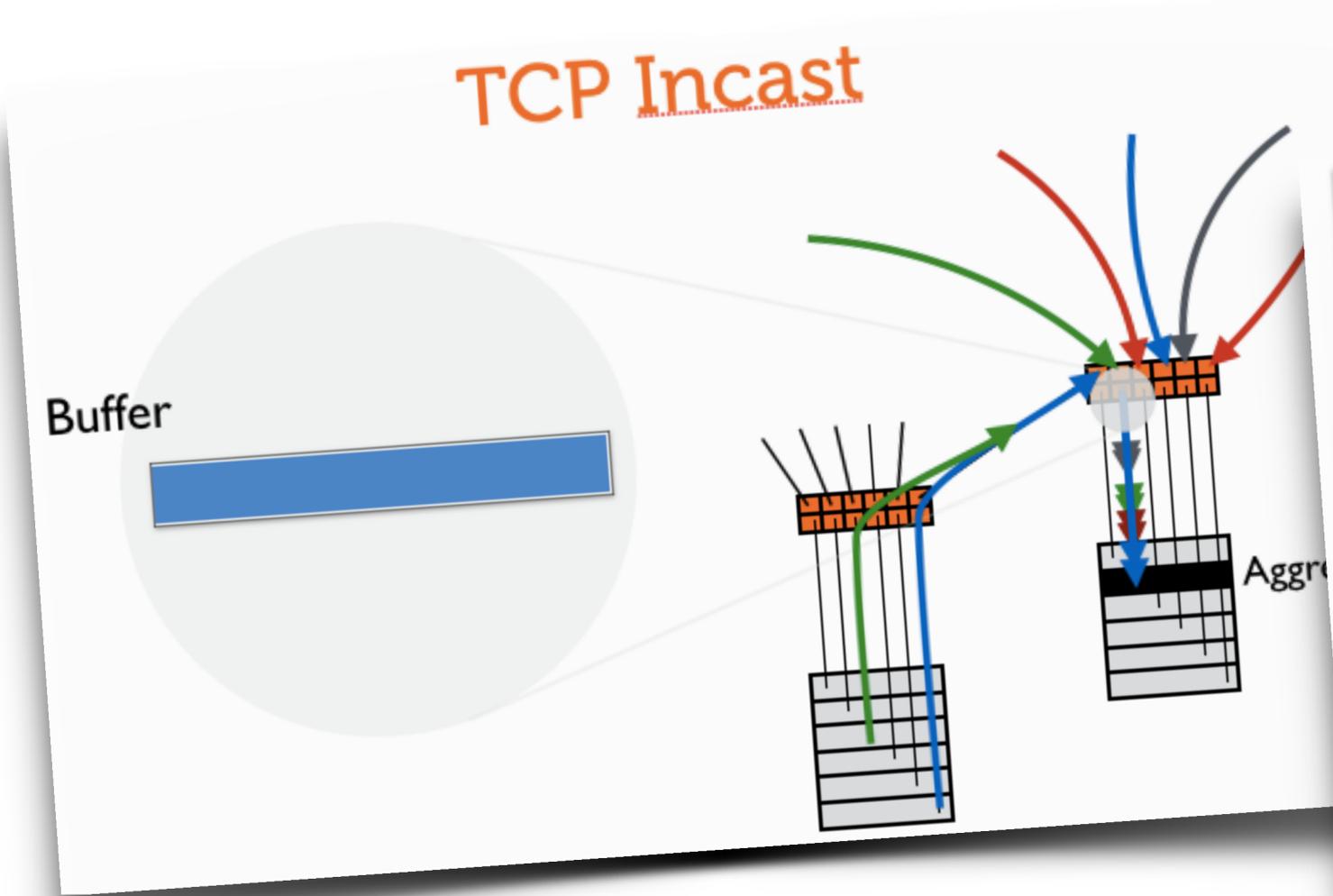
# Implications for networking

## 3 Congestion and TCP incast



**TCP does not work very well**

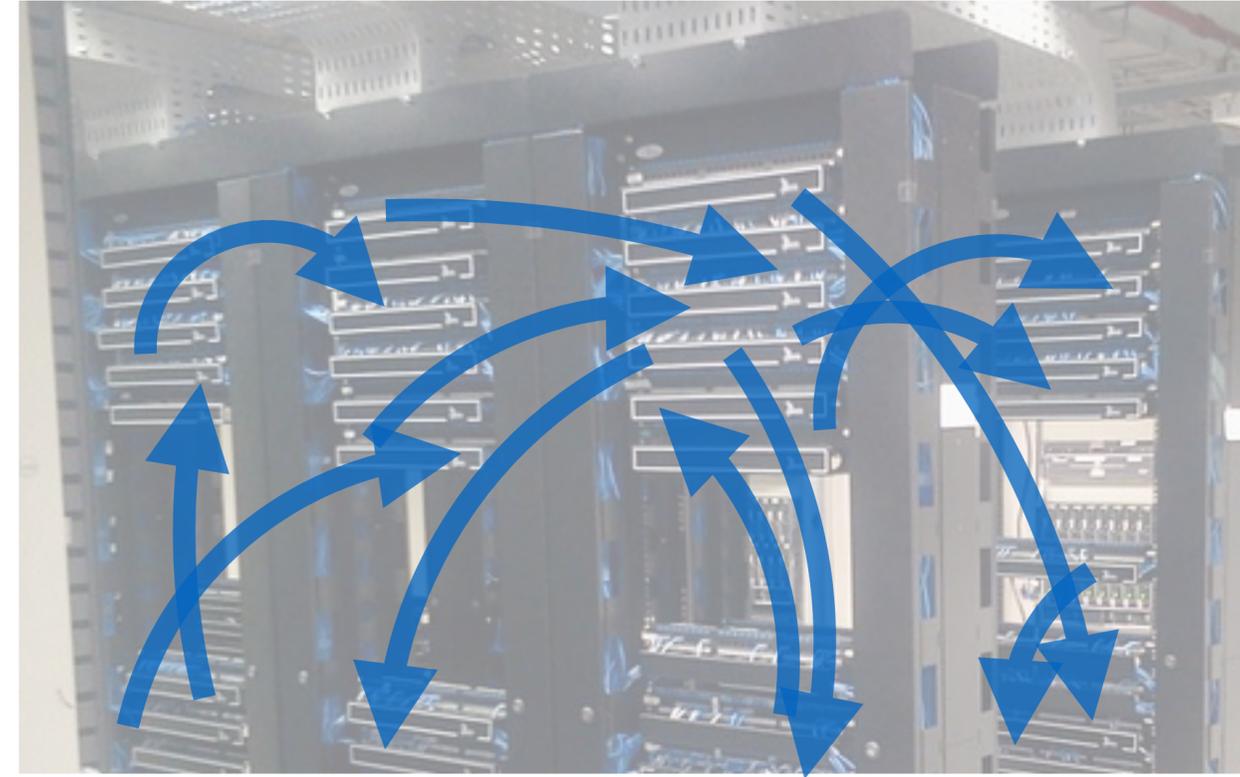
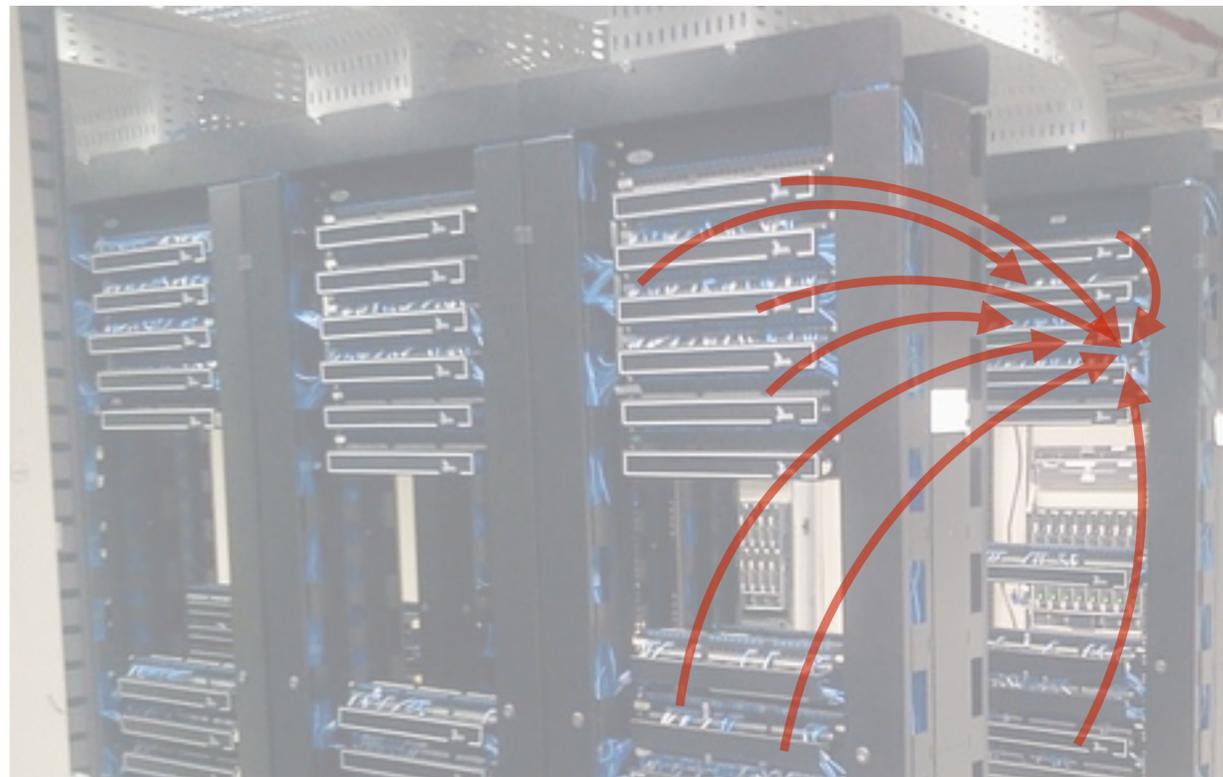
# How do we use networks best?



Congestion control with  $cwnd \ll 1$ ?

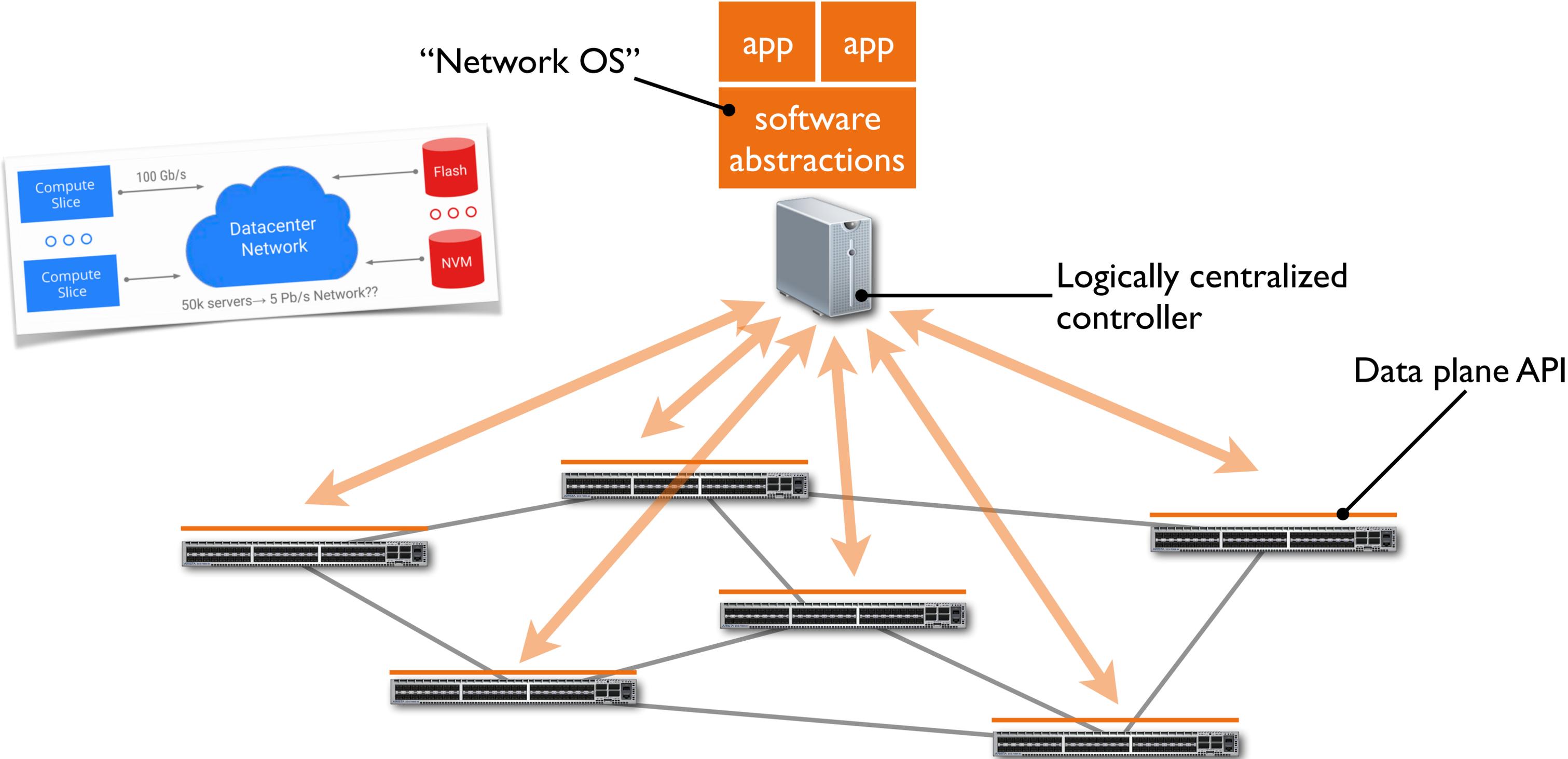
# Implications for networking

## 4 Complex network shared by applications

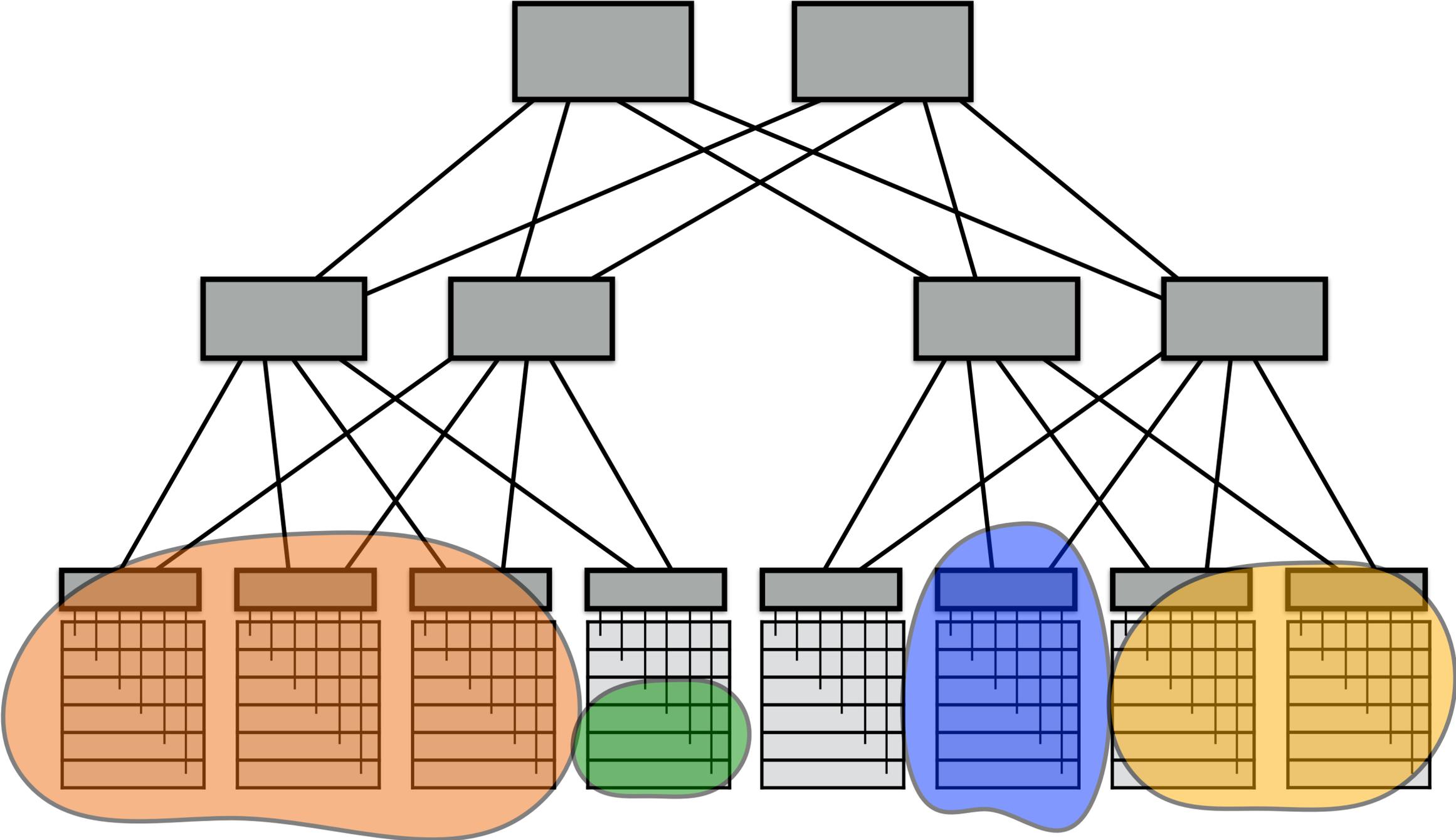


**Applications with different objectives sharing the network**

# How do we manage sharing?



# How do we manage sharing?



# Implications for networking

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

**Traffic characteristics: flow sizes**

**Facebook**  
“Inside the Social Network’s (Datacenter) Network”  
Arjun Roy et al., ACM SIGCOMM’15

**1500 server cluster @ ??**  
“The Nature of Datacenter Traffic: Measurements & Analysis”  
Srikanth Kandula et al. (Microsoft Research), ACM IMC’09

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

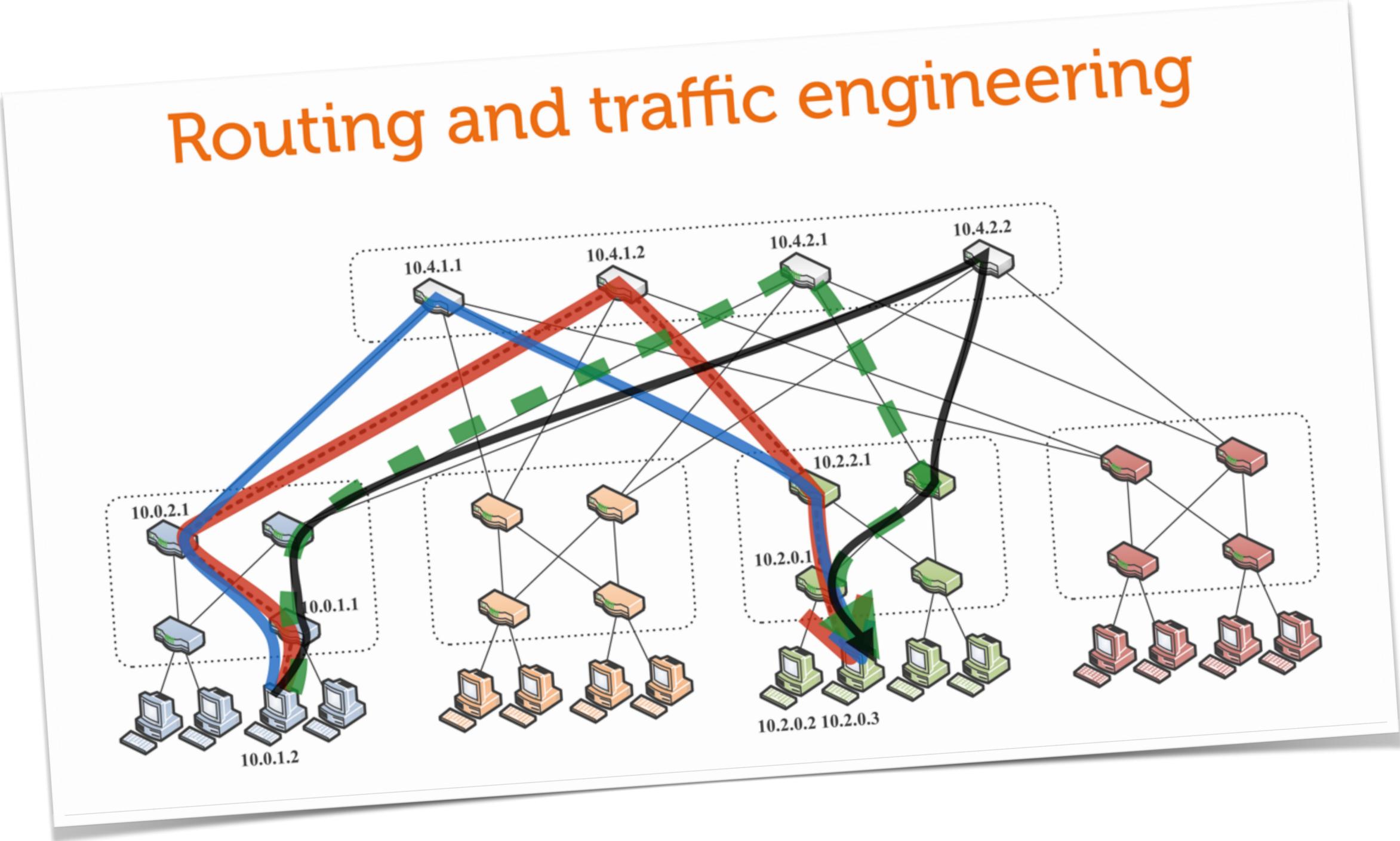
Caching: most flows are long-lived  
... but bursty internally

Heavy-hitters  $\approx$  median flow, not persistent

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

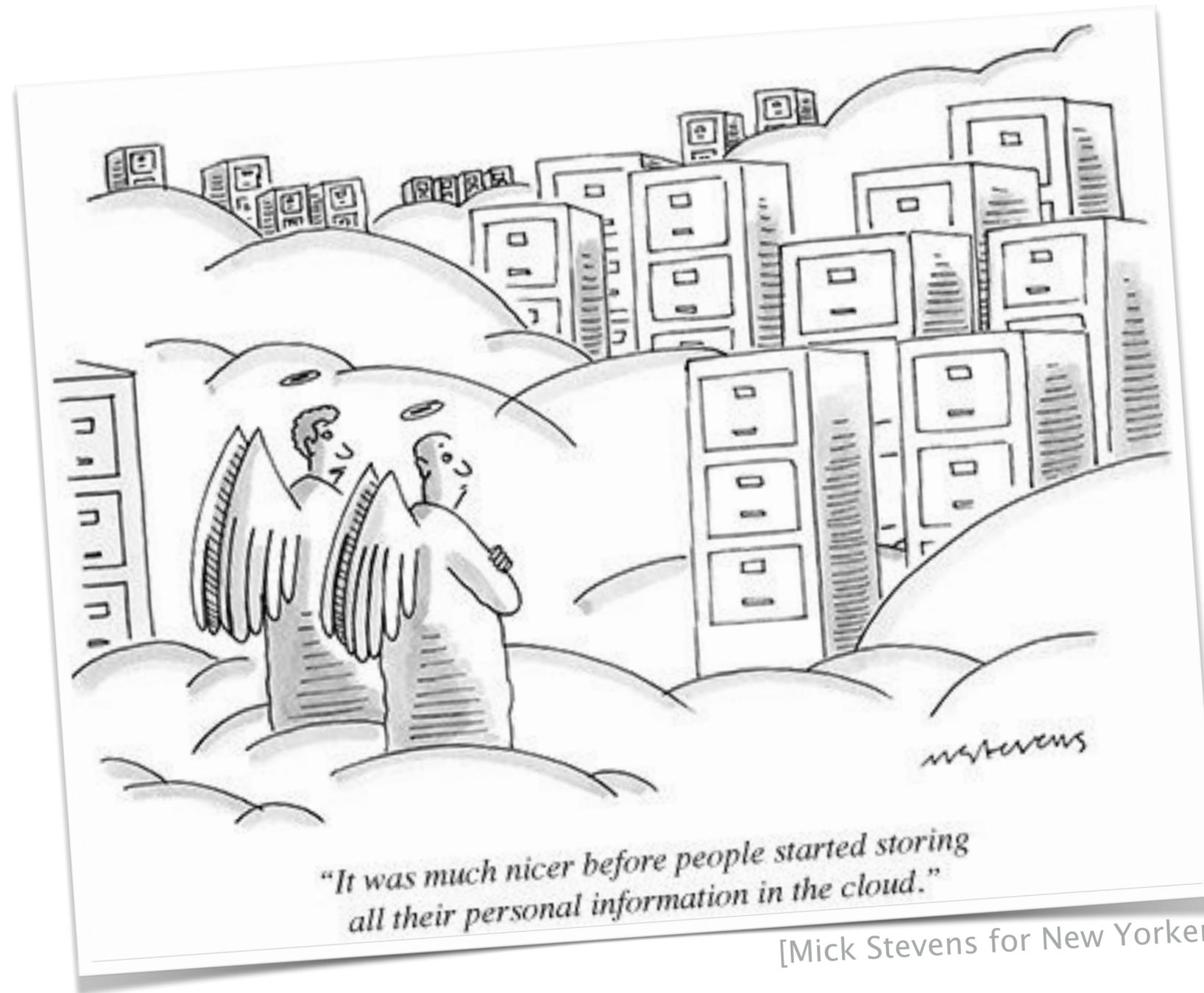
**Distributed control, perhaps with some centralized tinkering?**

# How do we use networks best?



Food for thought ...

# Are DCs even the right way?





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 ...

