SIGCOMM Preview Session: Data Center Networking (DCN)

George Porter, UC San Diego

2015

“The cloud”

“The cloud”

Google

amazon.com

NETFLIX

Microsoft Office 365

Spotify

Google docs

bing

salesforce

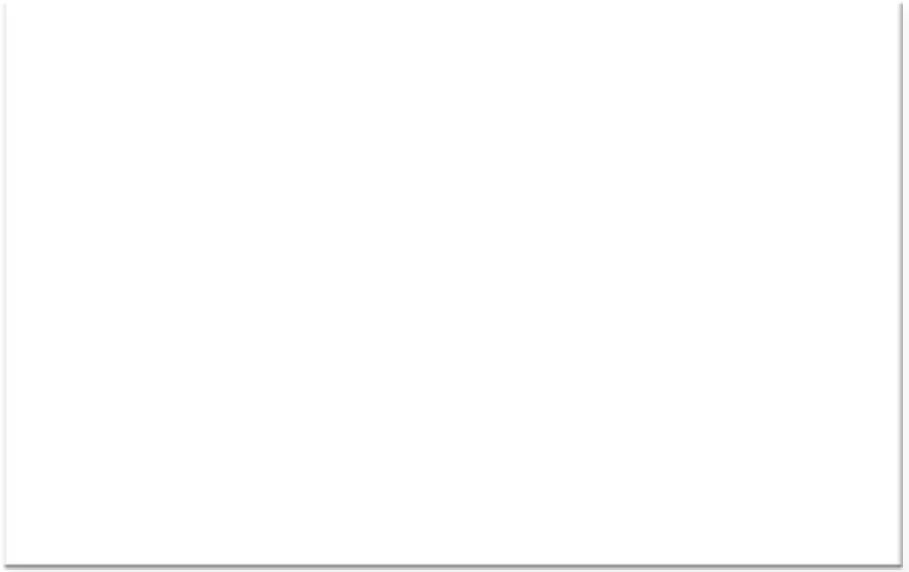
facebook

“The Cloud” = Lots of computing and data

+ amazon.com =

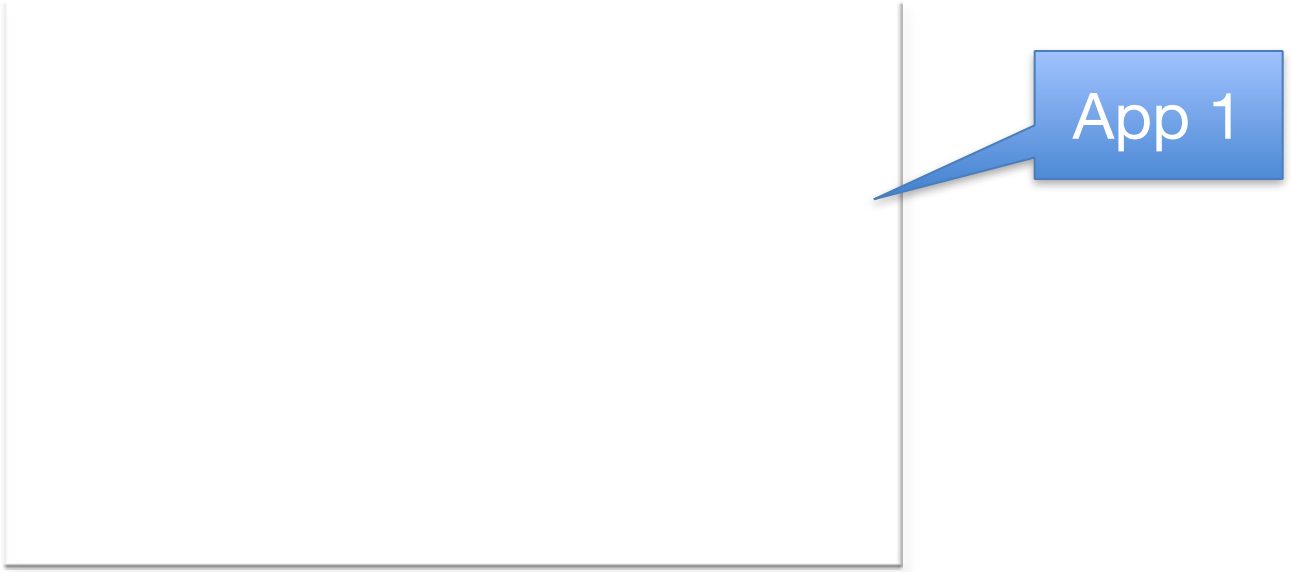
“The Cloud” = Lots of computing and data

+ amazon.com =



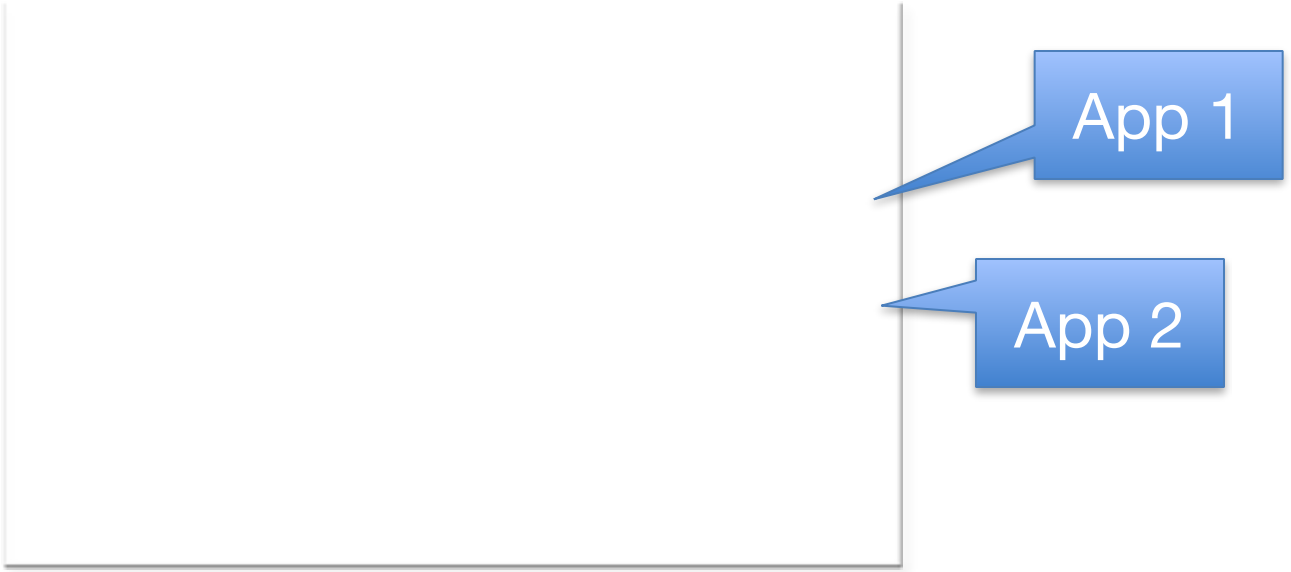
“The Cloud” = Lots of computing and data

+ amazon.com =

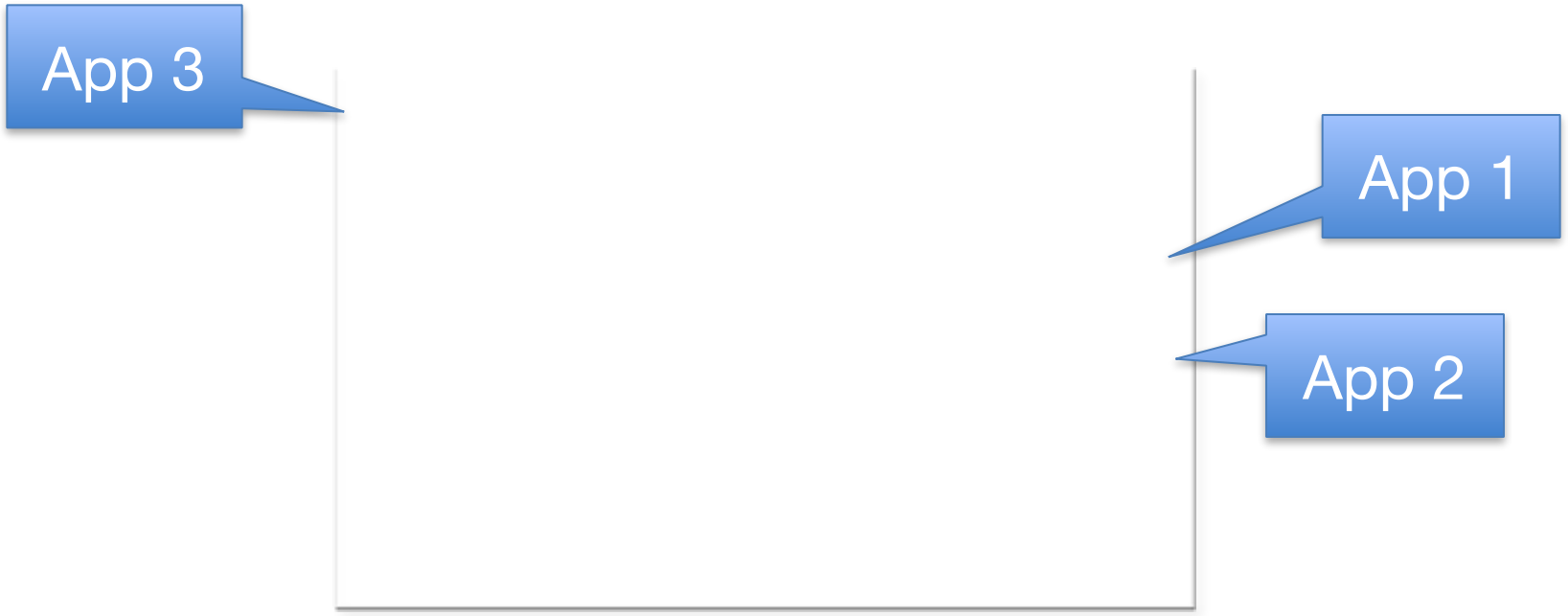


“The Cloud” = Lots of computing and data

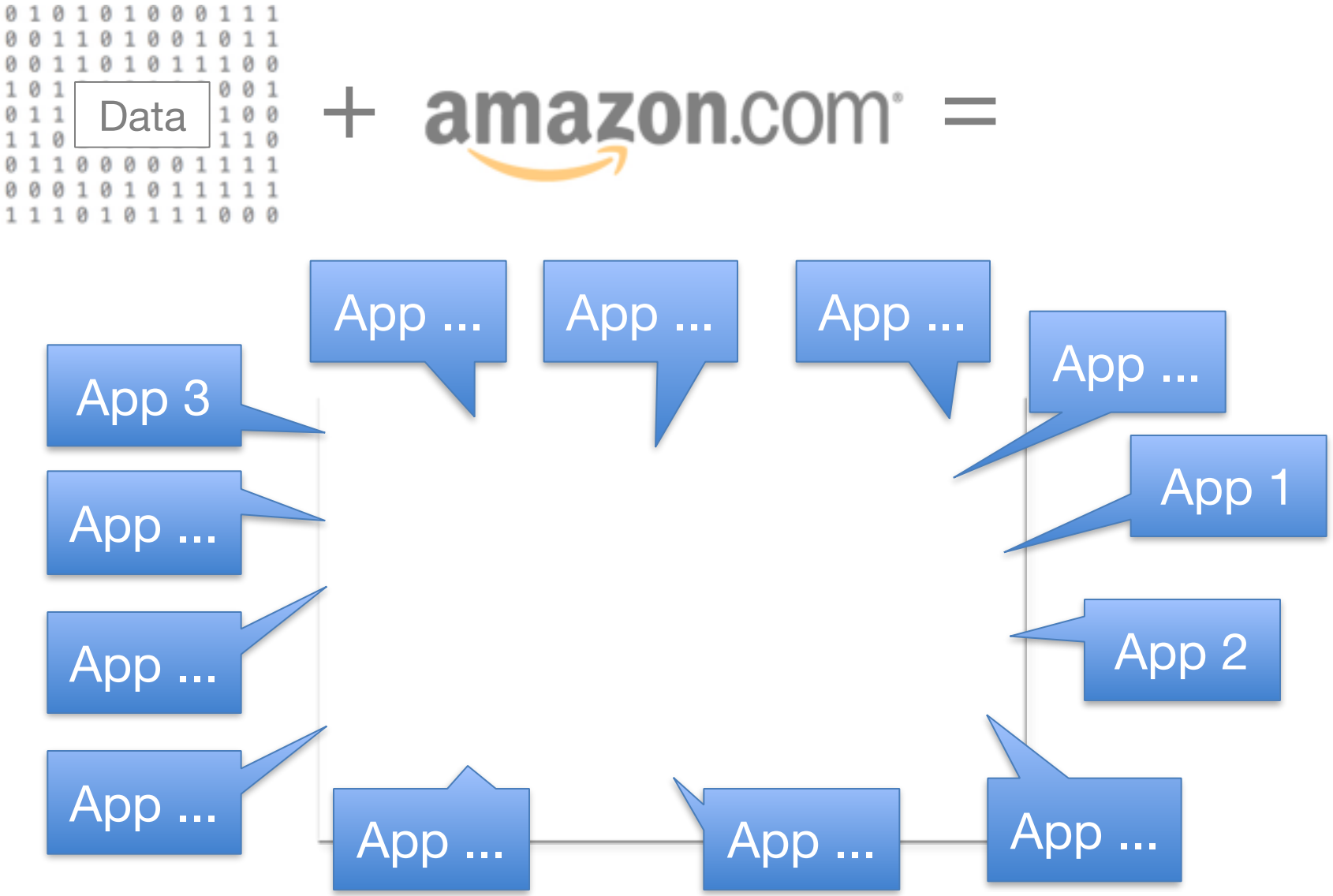
+ amazon.com =



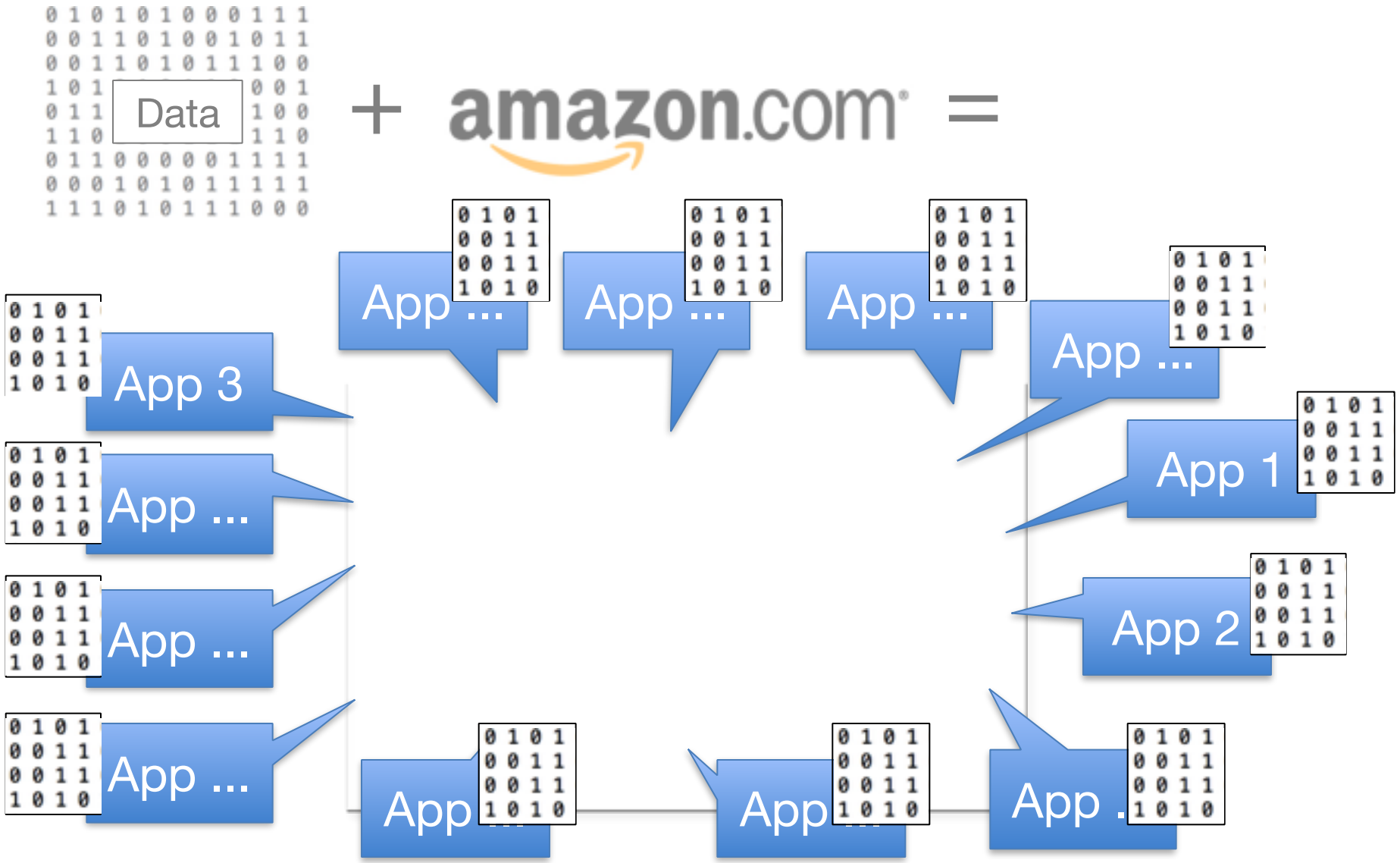
“The Cloud” = Lots of computing and data



“The Cloud” = Lots of computing and data



“The Cloud” = Lots of computing and data



Computing and data has to live somewhere…

Inside a data center

Inside a data center

- 10s or 100s of thousands of servers

Inside a data center

- 10s or 100s of thousands of servers
- Petabytes of data storage

Inside a data center

- 10s or 100s of thousands of servers
- Petabytes of data storage
- Single “applications” spread across many thousands of servers (e.g., Amazon.com)
  - Application components such as caches, web servers, data bases, distributed file servers, …
  - Each component is “scaled” to meet needs of millions of users

Why study DCNs?

Why study DCNs?

- Scale
  - Google: 0 to 1B users in ~15 years
  - Facebook: 0 to 1B users in ~10 years
  - *Must operate at the scale of O(1M+) users*

Why study DCNs?

- Scale
  - Google: 0 to 1B users in ~15 years
  - Facebook: 0 to 1B users in ~10 years
  - *Must operate at the scale of O(1M+) users*
- Cost:
  - To build: Google ($3B/year), MSFT ($15B/total)
  - To operate: 1-2% of global energy consumption*
  - *Must deliver apps using efficient HW/SW footprint*

* LBNL, 2013.

What defines a data center network?

The Internet	Data Center Network (DCN)

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	***One administrative domain***

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	***One administrative domain***
Distributed control/routing	***Centralized control and route selection***

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	*One administrative domain*
Distributed control/routing	*Centralized control and route selection*
Single shortest-path routing	*Many paths from source to destination*

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	*One administrative domain*
Distributed control/routing	*Centralized control and route selection*
Single shortest-path routing	*Many paths from source to destination*
Hard to measure	*Easy to measure, but lots of data…*

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	*One administrative domain*
Distributed control/routing	*Centralized control and route selection*
Single shortest-path routing	*Many paths from source to destination*
Hard to measure	*Easy to measure, but lots of data…*
Standardized transport (TCP and UDP)	*Many transports (DCTCP, pFabric, …)*

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	*One administrative domain*
Distributed control/routing	*Centralized control and route selection*
Single shortest-path routing	*Many paths from source to destination*
Hard to measure	*Easy to measure, but lots of data…*
Standardized transport (TCP and UDP)	*Many transports (DCTCP, pFabric, …)*
Innovation requires consensus (IETF)	*Single company can innovate*

What defines a data center network?

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	*One administrative domain*
Distributed control/routing	*Centralized control and route selection*
Single shortest-path routing	*Many paths from source to destination*
Hard to measure	*Easy to measure, but lots of data…*
Standardized transport (TCP and UDP)	*Many transports (DCTCP, pFabric, …)*
Innovation requires consensus (IETF)	*Single company can innovate*
"Network of networks"	*"Backplane of giant supercomputer"*

DCN research “cheat sheet”

DCN research “cheat sheet”

- How would you design a network to support 1M endpoints?

DCN research “cheat sheet”

- How would you design a network to support 1M endpoints?
- If you could…
  - Control all the endpoints and the network?
  - Violate layering, end-to-end principle?
  - Build custom hardware?
  - Assume common OS, dataplane functions?

DCN research “cheat sheet”

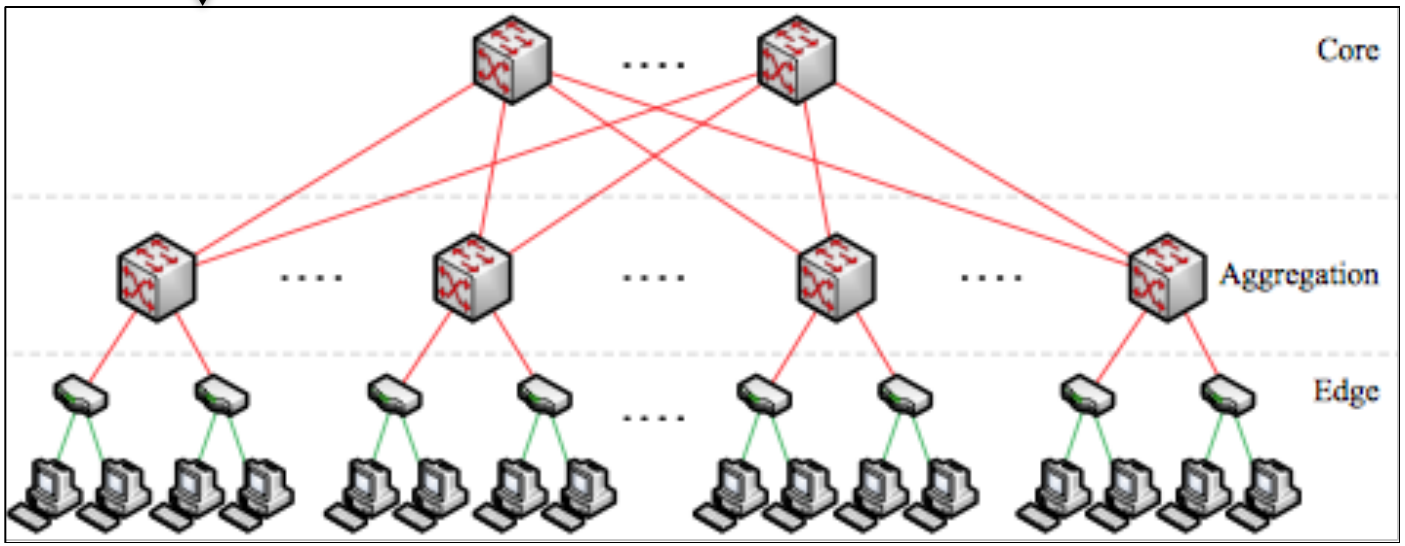
- How would you design a network to support 1M endpoints?
- If you could…
  - Control all the endpoints and the network?
  - Violate layering, end-to-end principle?
  - Build custom hardware?
  - Assume common OS, dataplane functions?

Top-to-bottom rethinking of the network

Paper previews: Topologies

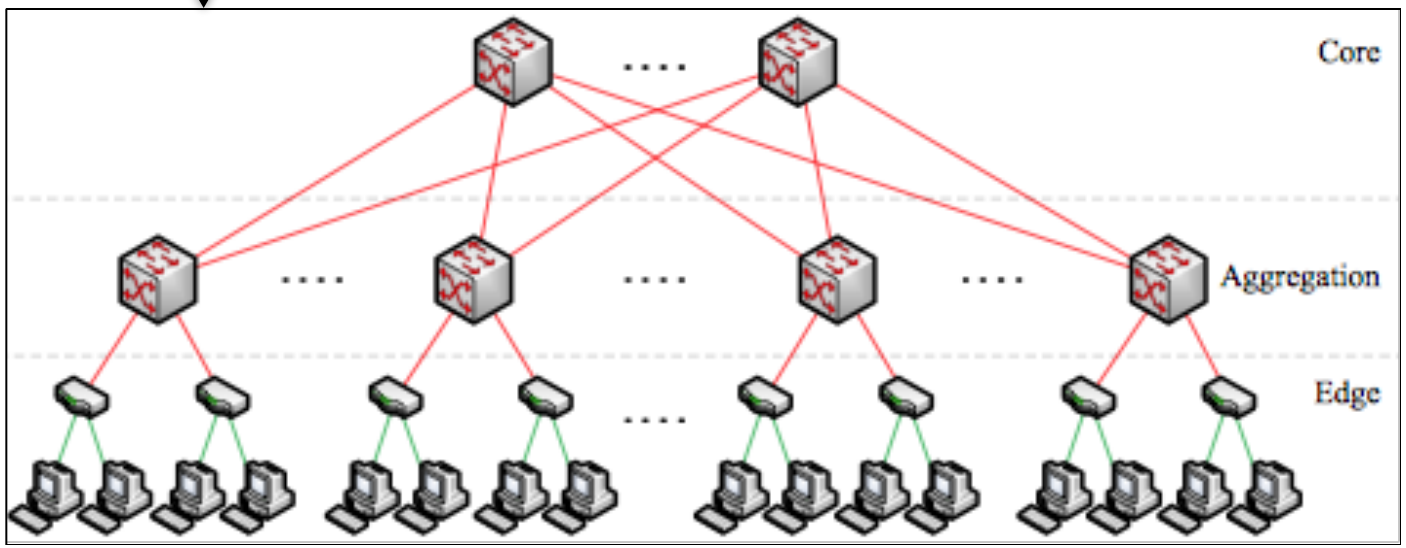
Tree-based network topologies

1993 1997 2001 2005 2009 2013



Tree-based network topologies

1993 1997 2001 2005 2009 2013

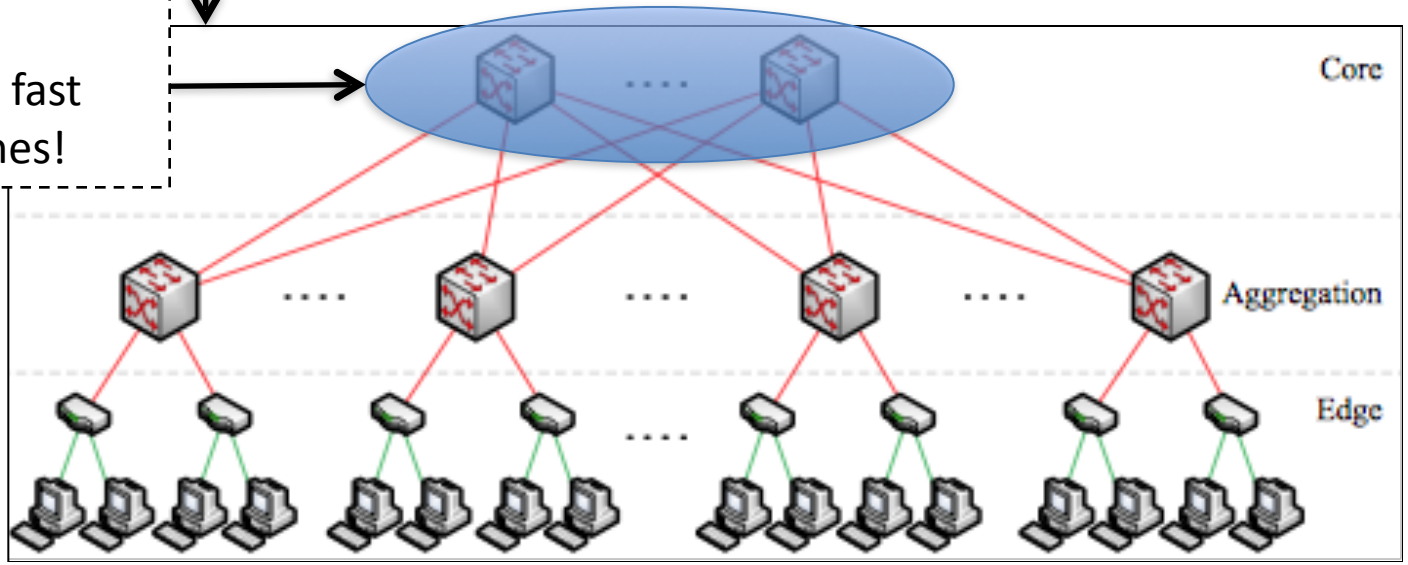


100,000 x 10 Gb/s = 1 Pb/s

Tree-based network topologies

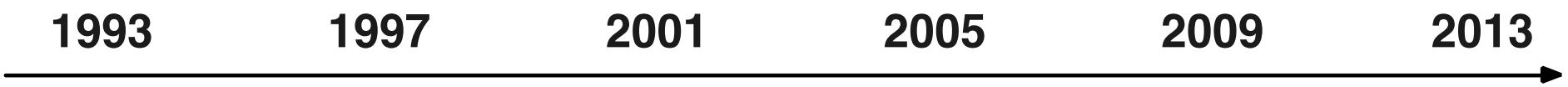
1993 1997 2001 2005 2009 2013

Can't buy sufficiently fast core switches!

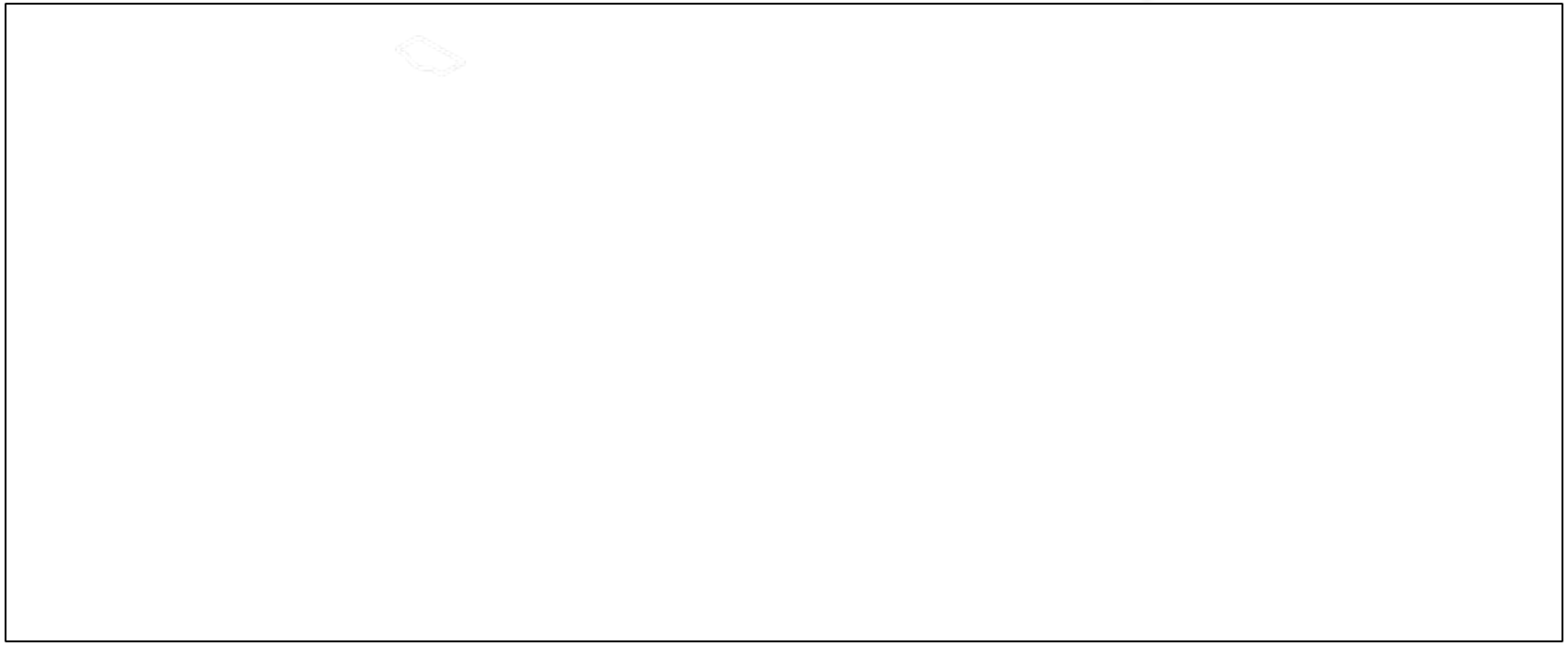


100,000 x 10 Gb/s = 1 Pb/s

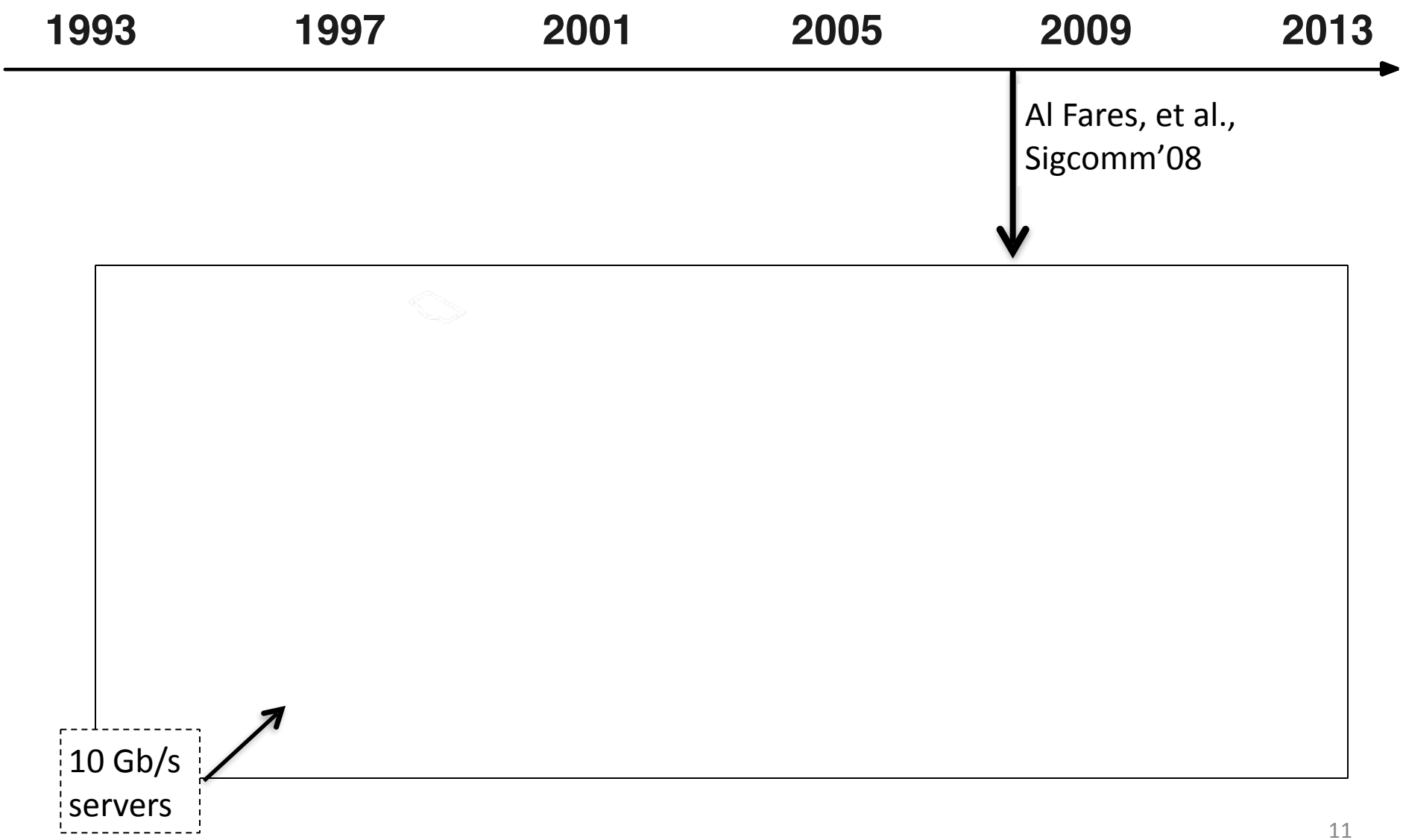
Folded-Clos multi-rooted trees



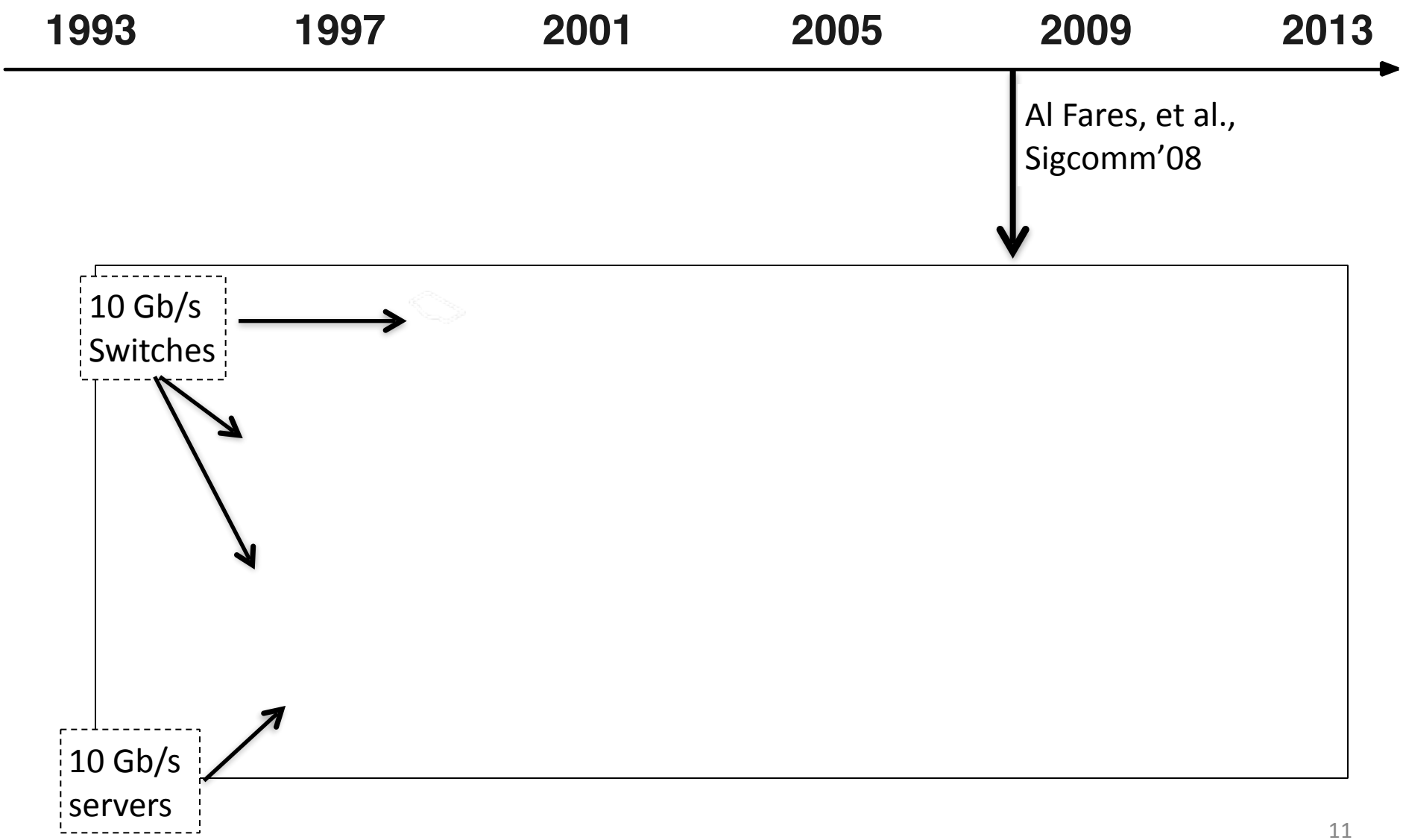
Folded-Clos multi-rooted trees



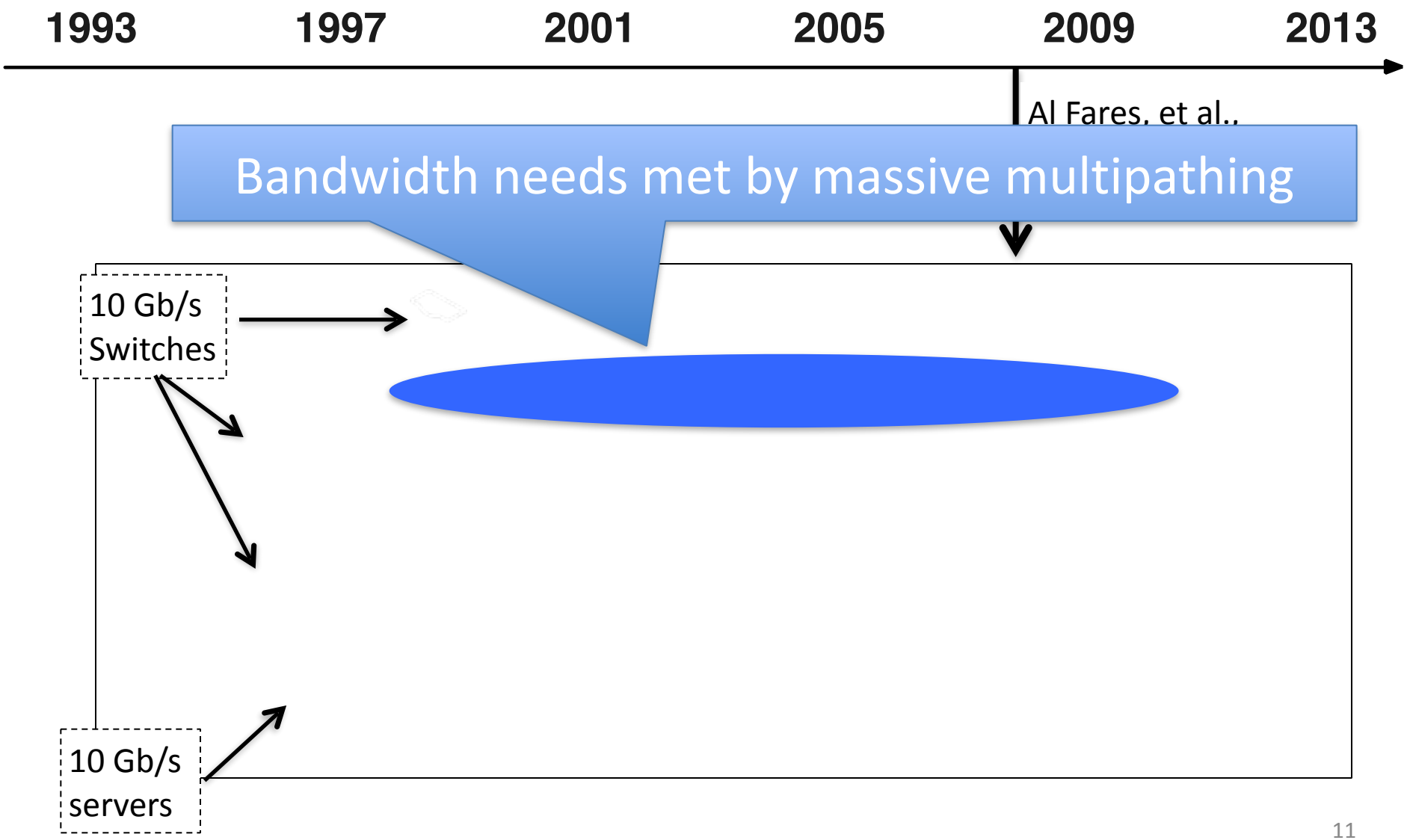
Folded-Clos multi-rooted trees



Folded-Clos multi-rooted trees



Folded-Clos multi-rooted trees



Paper previews: Topologies

- *Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network (Singh et al.)*
  - Tu 5pm-6:15pm Session 3.2: Experience Track: 2
  - 10 year retrospective on Google's experiences building large-scale networks

- *Condor: Better Topologies through Declarative Design (Schlinker et al.)*
  - Th 8:50am - 10:30am Session 8: Data center networking
  - Describing and reasoning about the network structure

Paper previews: Measurement

Network measurement

Network measurement

- Measuring the Internet:
  - No central vantage point, only indirect access to certain portions, multiple ASes hiding information…

Network measurement

- Measuring the Internet:
  - No central vantage point, only indirect access to certain portions, multiple ASes hiding information…
- Measuring data centers:
  - Need low latency
  - Need fine-grained precision (milli- or microsecond)
  - An enormous amount of data to collect
  - Hard to publish findings (proprietary data sets)

Paper previews: Measurement (1/2)

- *Inside the Social Network's (Datacenter) Network (Roy et al.)*
  - Tu 4pm-4:50pm Session 3.1: Experience Track 1
  - Measurement study of Facebook's data center
- *Pingmesh: A Large-Scale System for Data Center Network Latency Measurement and Analysis (Guo et al.)*
  - Tu 4pm-4:50pm Session 3.1: Experience Track 1
  - Experience paper on Microsoft's system for collecting inter-server ping times at scale

Paper previews: Measurement (2/2)

- *Packet-Level Telemetry in Large Datacenter Networks (Zhu et al.)*
  - Th 8:50am - 10:30am Session 8: Data center networking
  - Packet tracing system deployed at Microsoft designed for finding network faults

Paper previews: Packet/flow handling

Packet and flow handling

Packet and flow handling

- Internet service model:
  - Best-effort, “end-to-end principle”, generally just one path to a destination

Packet and flow handling

- Internet service model:
  - Best-effort, “end-to-end principle”, generally just one path to a destination
- Data center networks:
  - *Load balancing*: how to effectively use all the many paths to a given destination?
  - *Better than best-effort*: how to prioritize, rate-limit, adjust relative sending rates…

Paper previews: Packet/flow handling

- *Presto: Edge-based Load Balancing for Fast Datacenter Networks (He et al.)*
  - Th 8:50am - 10:30am Session 8: Data center networking
  - Choosing paths for packets with help from endhosts

- *Enabling End-Host Network Functions (Ballani et al.)*
  - Th 8:50am - 10:30am Session 8: Data center networking
  - Providing better than best-effort handling of packets with help from endhosts

In closing

- DCN is an exciting, fun research area
- While many papers are from Microsoft, Google, Facebook, …
  - YOU have the ability to have enormous impact
  - Many projects are open-source
    - E.g., http://opencompute.org
- Rethink the entire network stack!
  - Hardware, software, protocols, OS, NIC, …