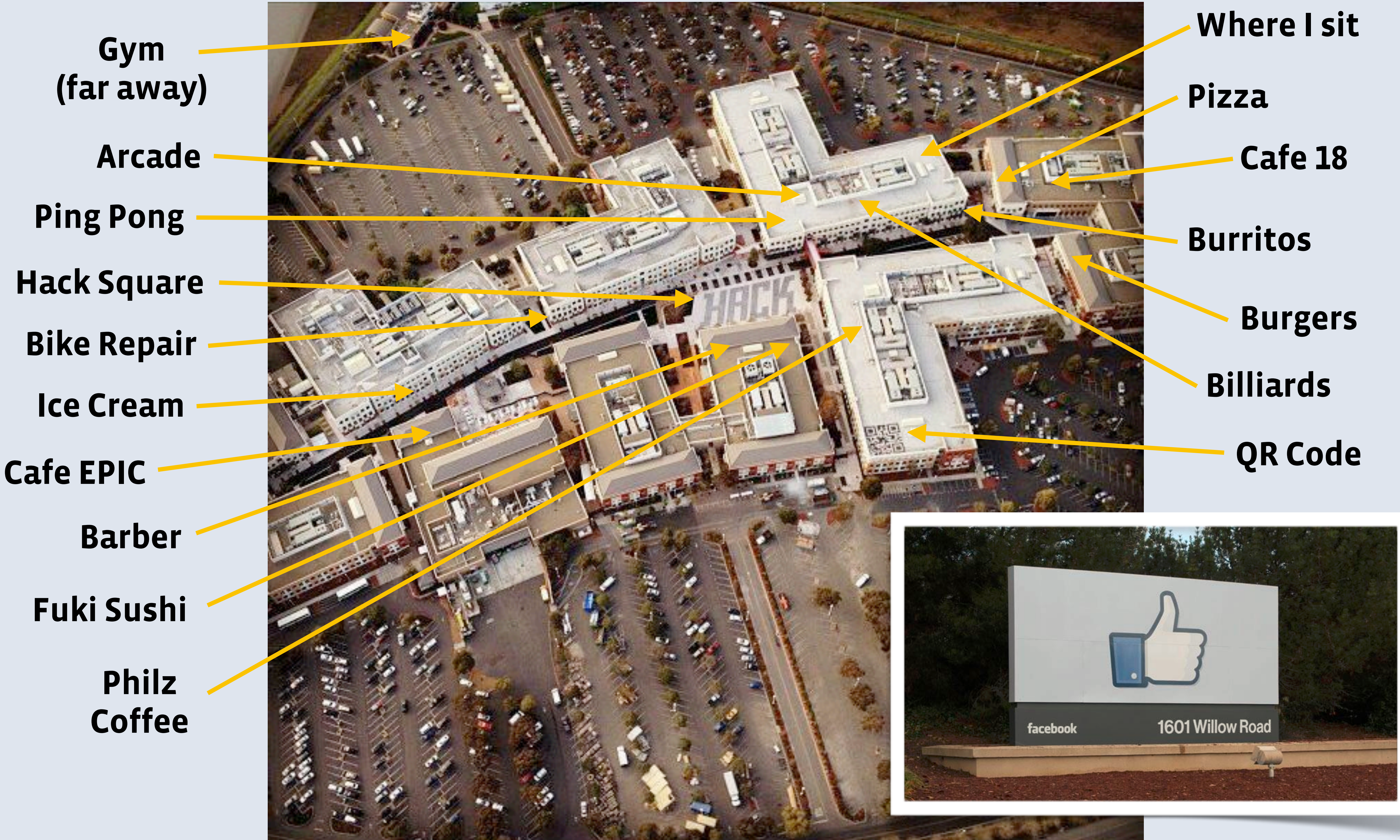
facebook

facebook

Facebook's Data Center Network Architecture

Nathan Farrington
Data Center Network Engineer

2013-05-07 (IEEE Optical Interconnects Conference, Santa Fe, New Mexico)


Gym
(far away)
Arcade
Ping Pong
Hack Square
Bike Repair
Ice Cream
Cafe EPIC
Barber
Fuki Sushi
Philz
Coffee
Where I sit
Pizza
Cafe 18
Burritos
Burgers
Billiards
QR Code
HACK
facebook
1601 Willow Road

Mission

Make the world more open and connected

Values

- Focus on Impact
- Move Fast
- Be Bold
- Be Open
- Build Social Value

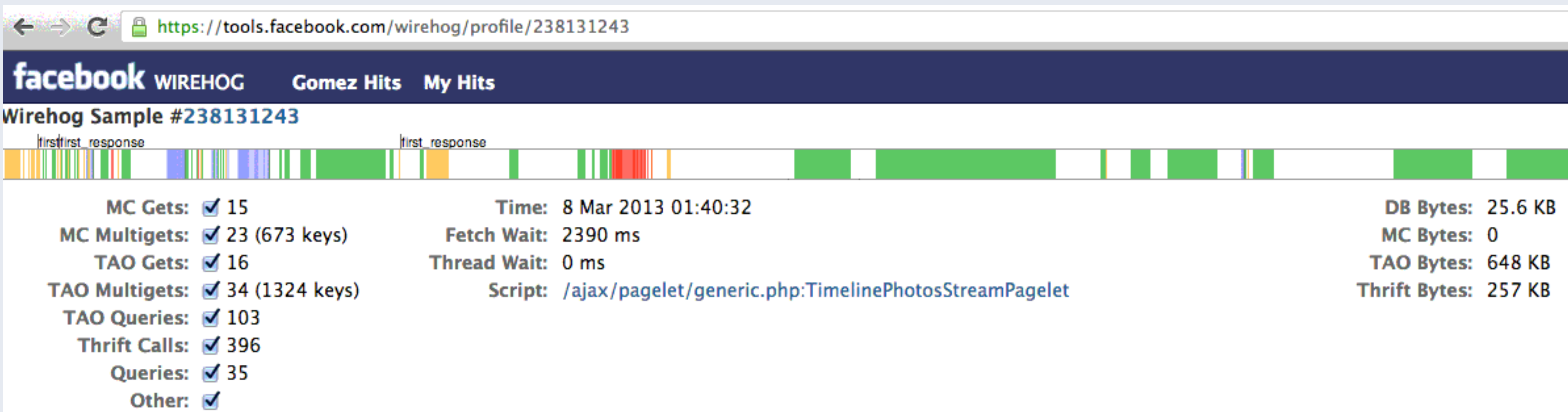
OPEN

Compute Project

Network measurements

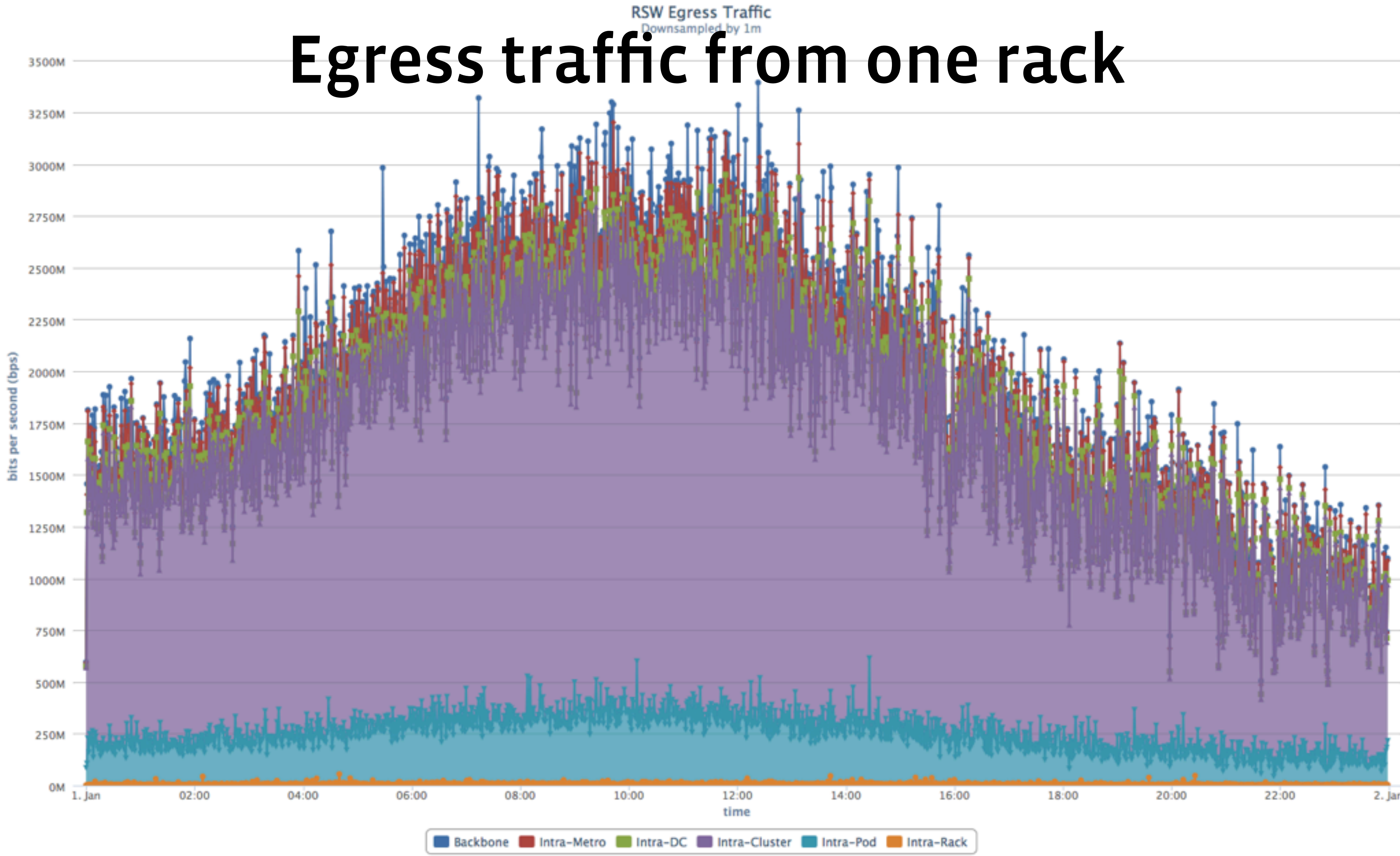
HTTP request amplification

This 1 KB HTTP request generated 930 KB of internal network traffic

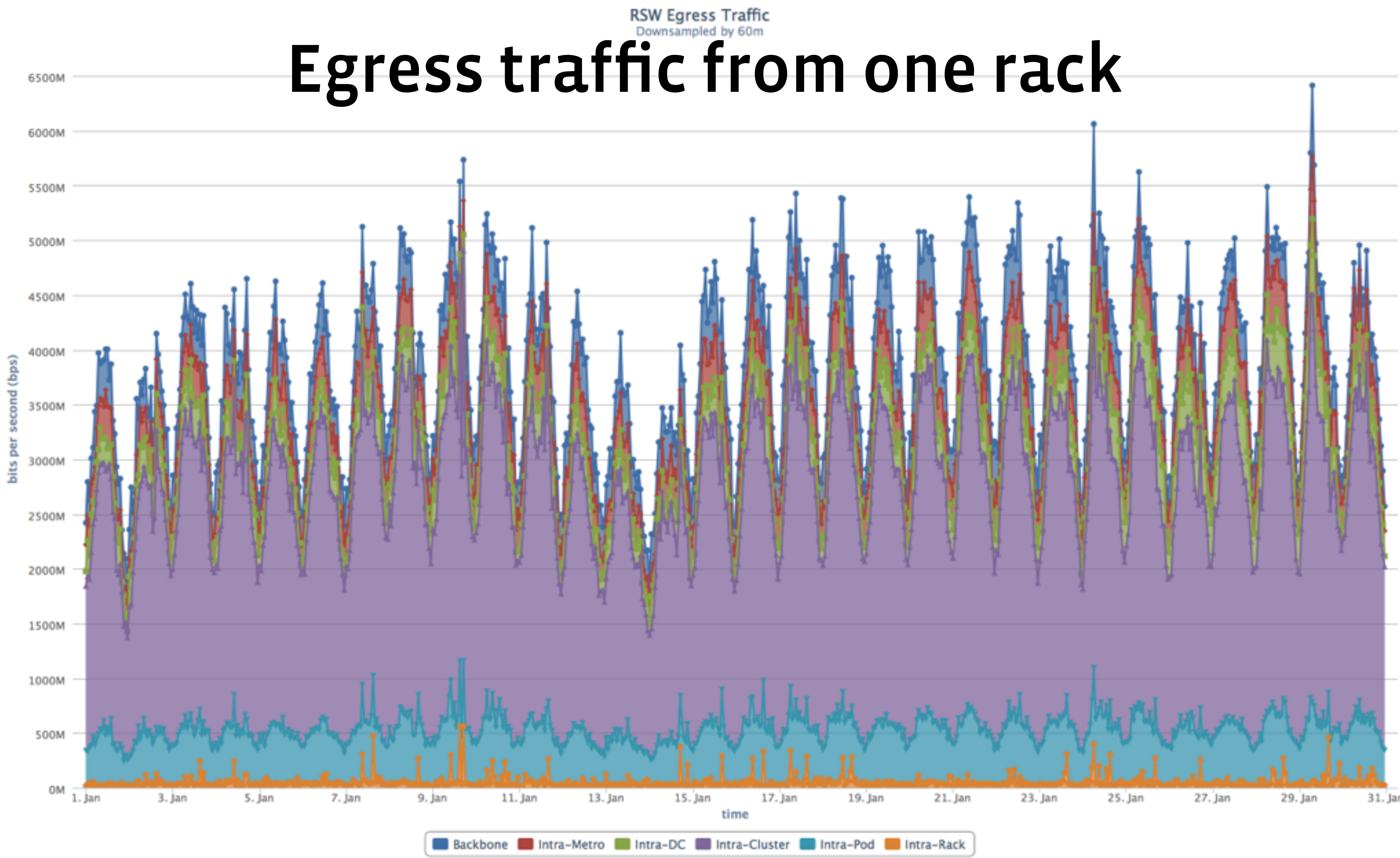


Not necessarily representative of all traffic

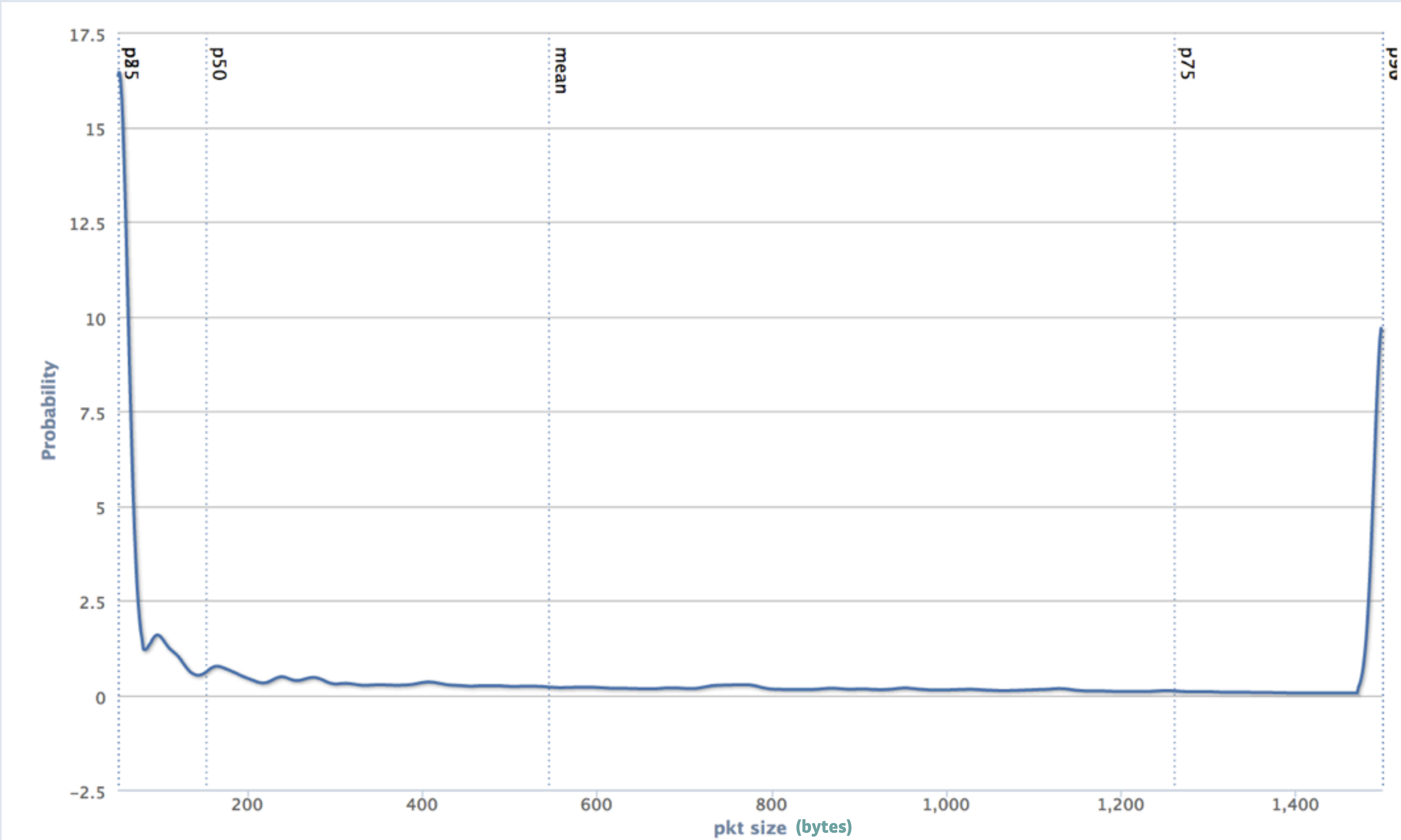
Egress traffic from one rack



Egress traffic from one rack



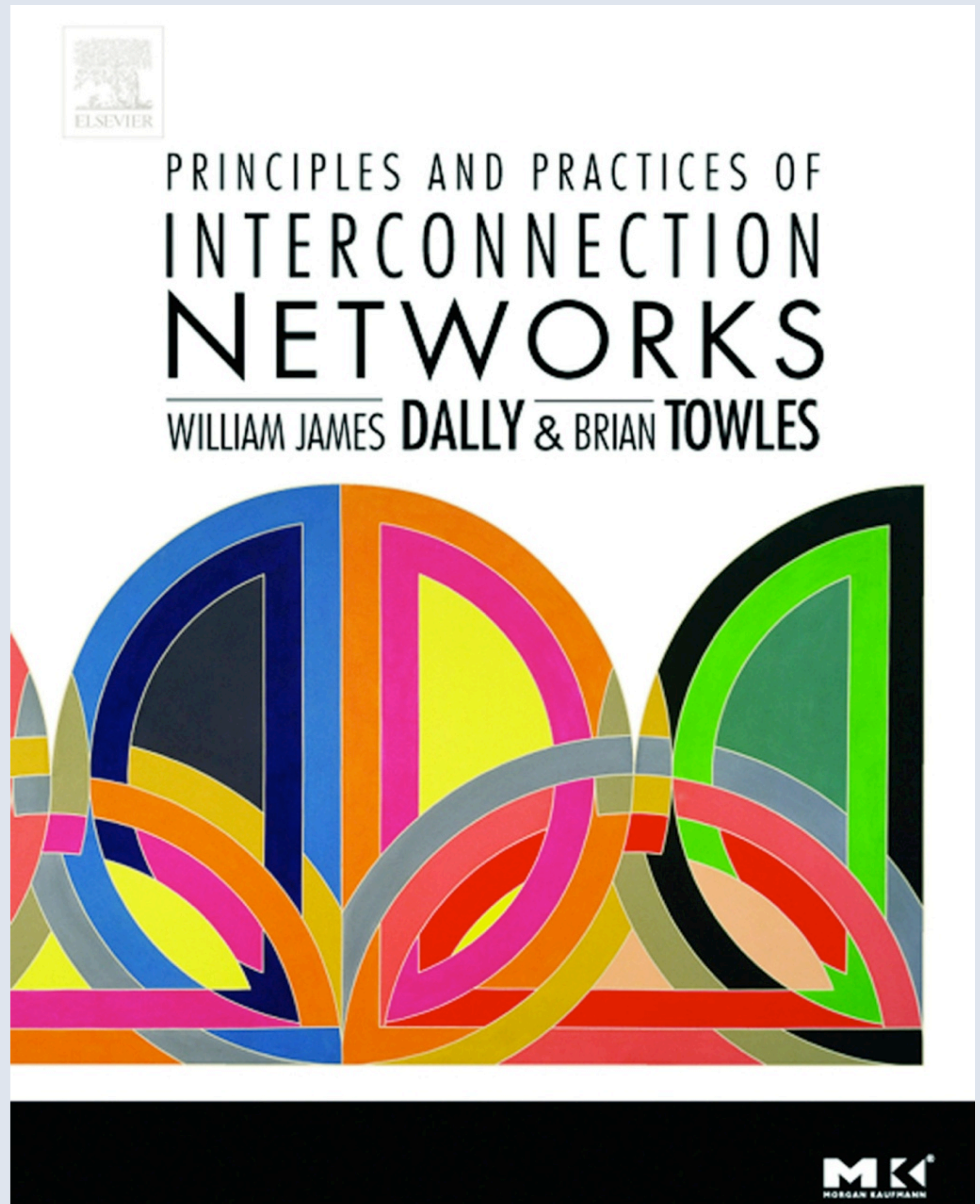
Internet-facing packet size distribution



Network topologies

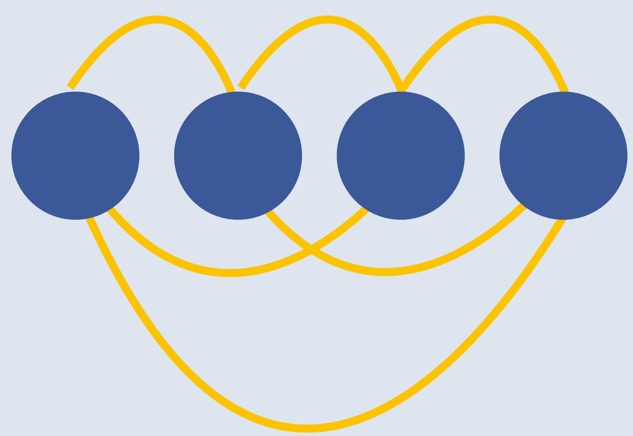
Topologies are chosen for religious reasons.

The Bible


ELSEVIER
PRINCIPLES AND PRACTICES OF
INTERCONNECTION
NETWORKS
WILLIAM JAMES DALLY & BRIAN TOWLES
MK
MORGAN KAUFMANN

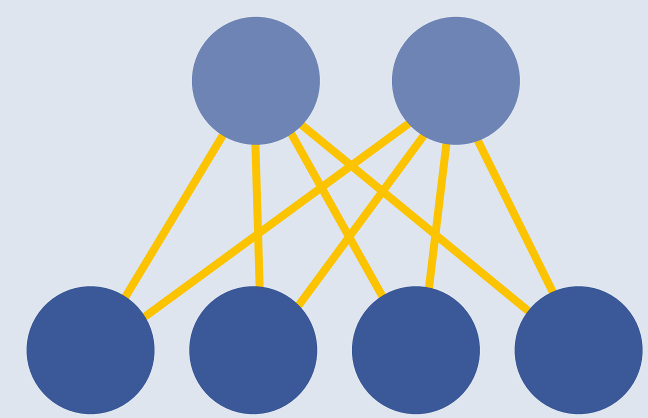
There are only 2 kinds of Topologies

East-West and North-South



Torus/Mesh/Hypercube

Direct



Tree/Clos

Indirect

There are only 2 kinds of Topologies

East-West and North-South

All other topologies are recursively composed of these two.

Topologies

When to use Direct Networks: Torus/Mesh/Hypercube

- Known, unchanging communication pattern that maps very well to physical topology
- Need low latency (nanoseconds)
- Need application-level control of packet routing

Typical Application: HPC Interconnects

ORNL Titan, #1 Supercomputer (Nov 2012)

Cray Gemini 3D Torus: 11.96 Pb/s; 9,344 switches; 56,064 links

https://www.olcf.ornl.gov/wp-content/uploads/2012/12/titan_lores.png

Topologies

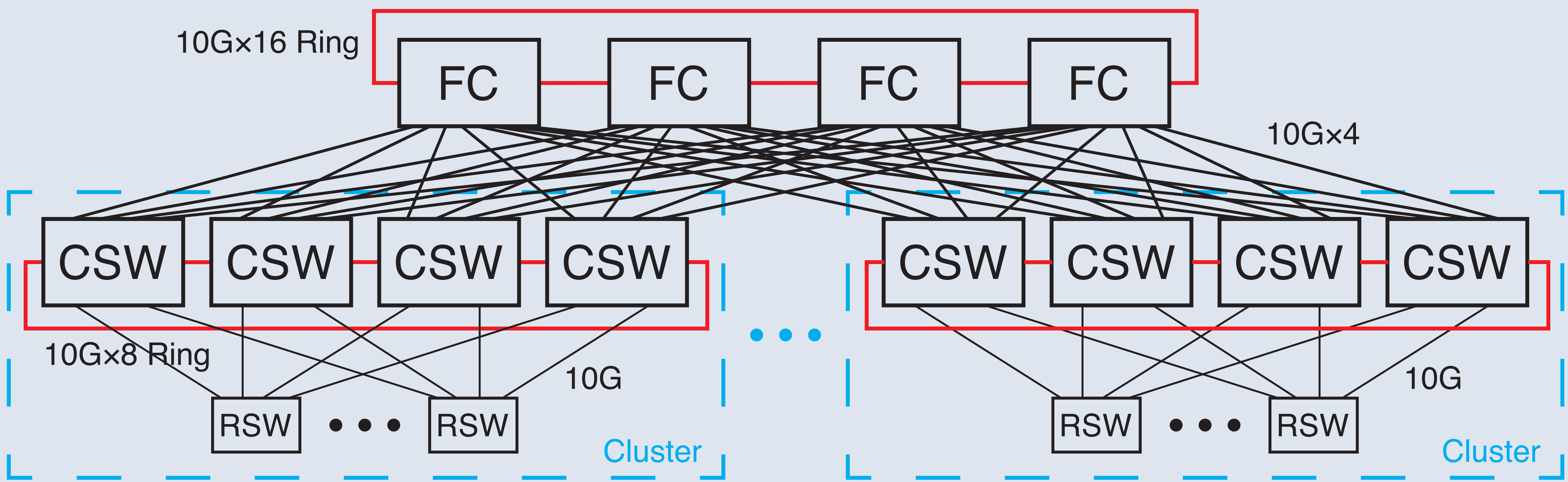
When to use Indirect Networks: Tree/Clos

- Unknown or changing communication patterns
- Latency not as important (microseconds)
- Multiple uncoordinated applications sharing same network
- Need high throughput

Typical Application: Datacenter Networks

Facebook “4-post” Architecture

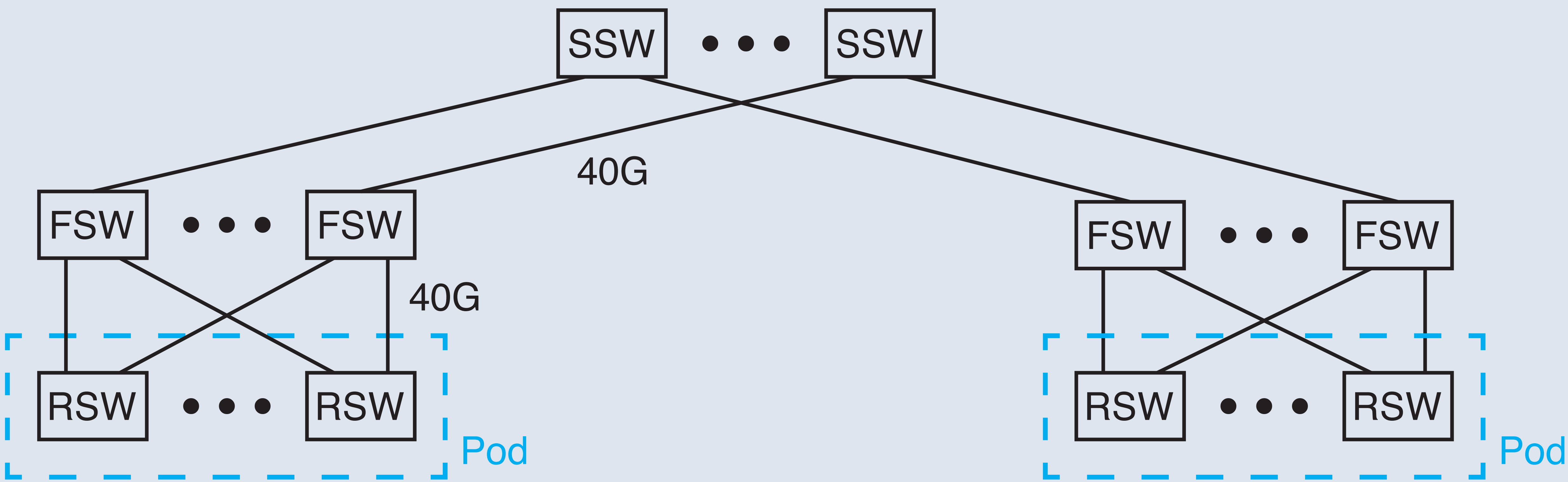
25% blast radius, a few large clusters



Hypothetical 5-stage Folded-Clos

small blast radius, lots of small clusters (pods), commodity

Challenge: cables and optics



Ethernet link rates

Less is More: 25G vs 40G Ethernet

Ethernet Link Rate	# of 10G SERDES Lanes	# of 25G SERDES Lanes	# of 50G SERDES Lanes
1G	1	1	1
2.5G	1	1	1
10G	1	1	1
25G	3	1	1
40G	4	2	1
50G	5	2	1
100G	10	4	2

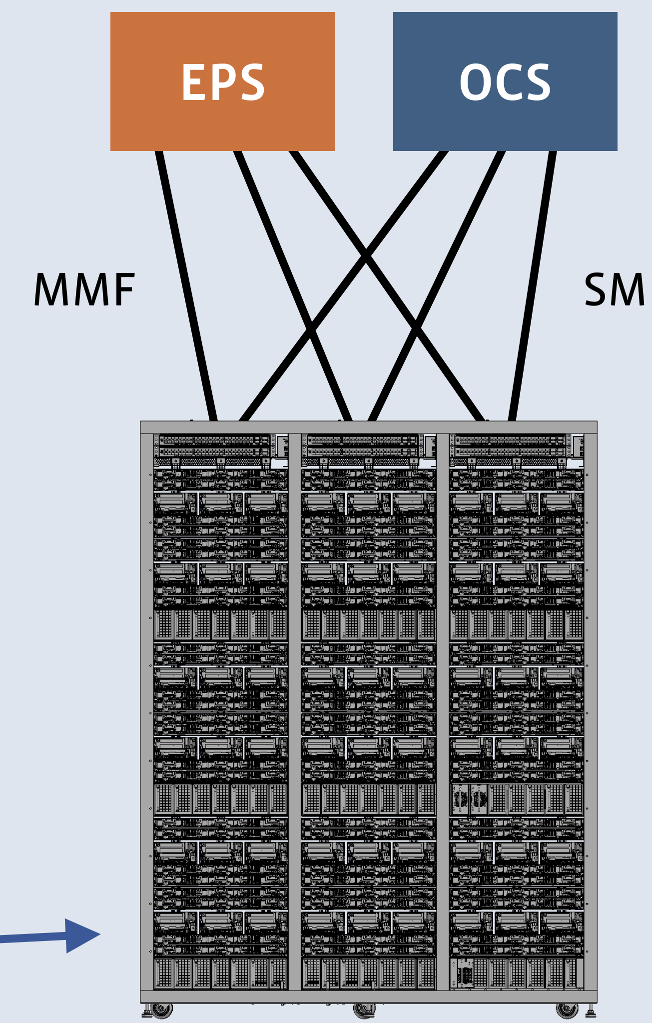
Optical circuit switching

Barriers to deploying OCS in the datacenter

- MMF vs SMF
  - 10GBASE-SR: $5/Gb/s, 100mW/Gb/s, MMF
  - 10GBASE-LR: $25/Gb/s, 100mW/Gb/s, SMF
  - 40GBASE-LR4: $37/Gb/s, 87.5mW/Gb/s, SMF
- New “Cisco” protocol vs SDN
- Where does the OCS go?
  - Between regions? (longhaul)
  - Between buildings? (metro)
  - Between clusters (intra-datacenter)
  - Between racks (intra-cluster)
  - Between servers (intra-rack)

Note: prices shown are industry estimates

“Helios/c-Through” model



This example

How to remove those barriers

- Make a cost competitive transceiver for SMF
  - Then MMF will disappear
  - Silicon photonics promises reduced CAPEX and smaller packaging
- Develop mature SDN technologies
  - In the switch
  - In the operating system
  - In the hypervisor
  - In the traffic controller
- Develop mature workload placement technologies
- Develop mature bulk traffic scheduling technologies

From OEM to ODM: a story of SDN

Facebook currently deploys OEM gear

- Past OEM suppliers: Cisco, Arista, Juniper
- Buy gear, recruit operators trained to use that gear . . . win!
- OEMs have a one-size fits all business model
  - Lots of features (we only use a few, e.g. BGP, ECMP, MPLS-TE, ...)
  - Millions of lines of code
  - Modular architecture (because some people really want FibreChannel)
- Optics/cables typically bundled as part of switch/router purchase
  - PRO: guaranteed transceiver compatibility & supply chain
  - CON: higher CAPEX

Possible Future #1

Stay with OEM, use more "SDN" features

- Cisco onePK [1], Arista EOS [2]
- Allows easier monitoring and measurement collection
- PRO: Use existing infrastructure, no need to qualify new hardware
- CON: Closed source
  - [1] Cisco BRKCDN-1969 (2012)
  - [2] http://www.aristanetworks.com/media/system/pdf/EOSWhitepaper.pdf

Possible Future #2

Move to ODM, use 3rd-party software stack

- Merchant silicon: Broadcom, Intel, Marvell, Mellanox, Gnodal, …
- Lots of contract manufacturers: Quanta, Foxconn, Celestica, …
- Closed-source software stacks: Broadcom FastPath, WindRiver ONS, …
  - Open source: Quagga, ExaBGP, …
- Who do you go to when something breaks?
  - Similar argument made against Linux >10 years ago

Possible Future #3

Move to ODM, write our own software stack

- Same hardware choices as Possible Future #2
- We own the software
  - BGP Route Disaggregation/Reaggregation
  - Weighted Multipath Routing
- PRO: Flexibility and reliability
- CON: I come in to work at 3:00 AM when something breaks

Beyond SDN

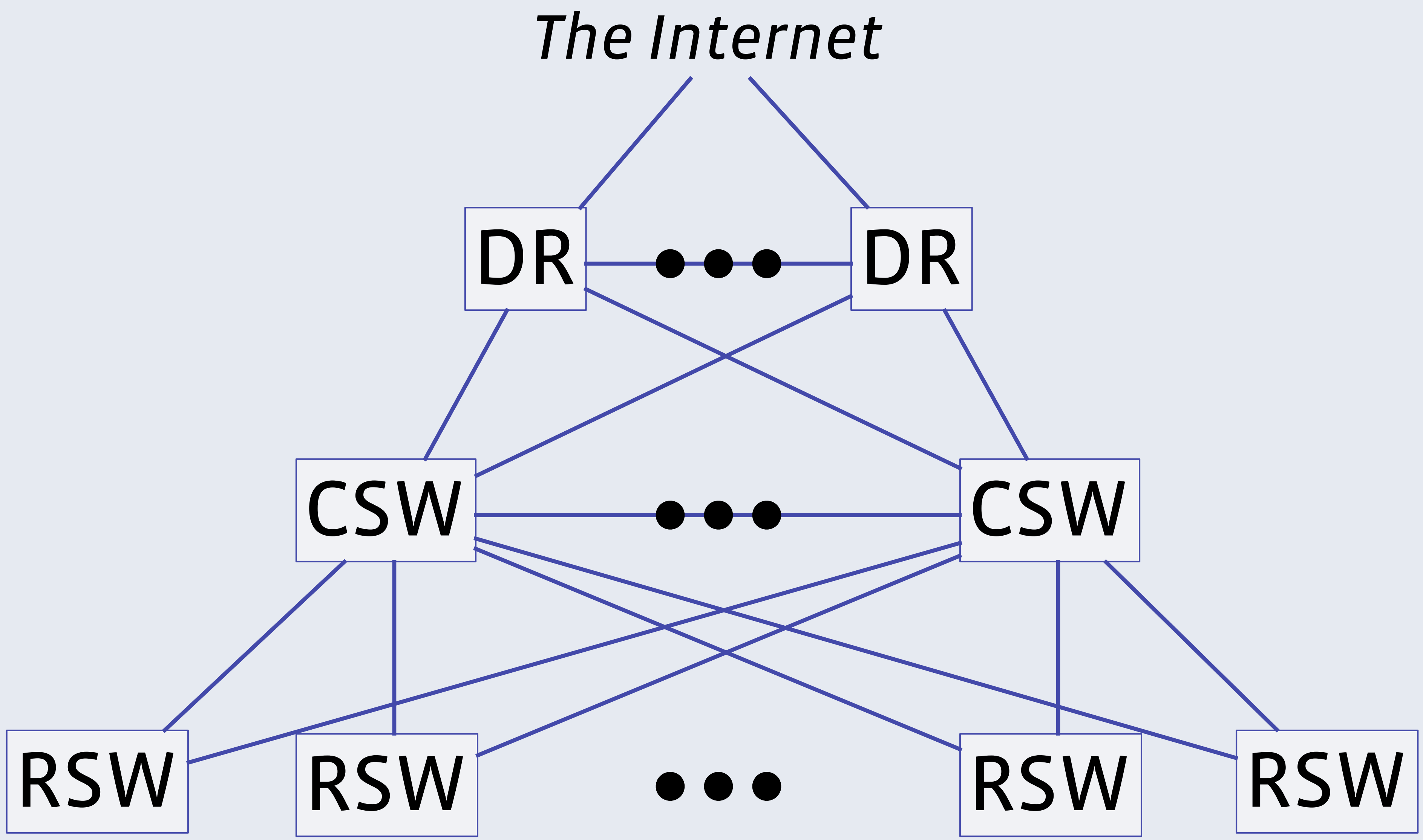
HDN: Hardware Defined Networking

- Perspective: currently we all have CDN: Cisco Defined Networking
- SDN is too slow for some important things, like
  - Detecting link failures and rerouting
  - Load balancing, load balancing in the presence of failures
  - Congestion control, traffic engineering
- Examples of HDN from HPC:
  - Adaptive load balancing
  - Credit-based flow control
- David Zats, Tathagata Das, Prashanth Mohan, Dhruba Borthakur, Randy Katz, "**DeTail: Reducing the Flow Completion Time Tail in Datacenter Networks**," in SIGCOMM 2012.

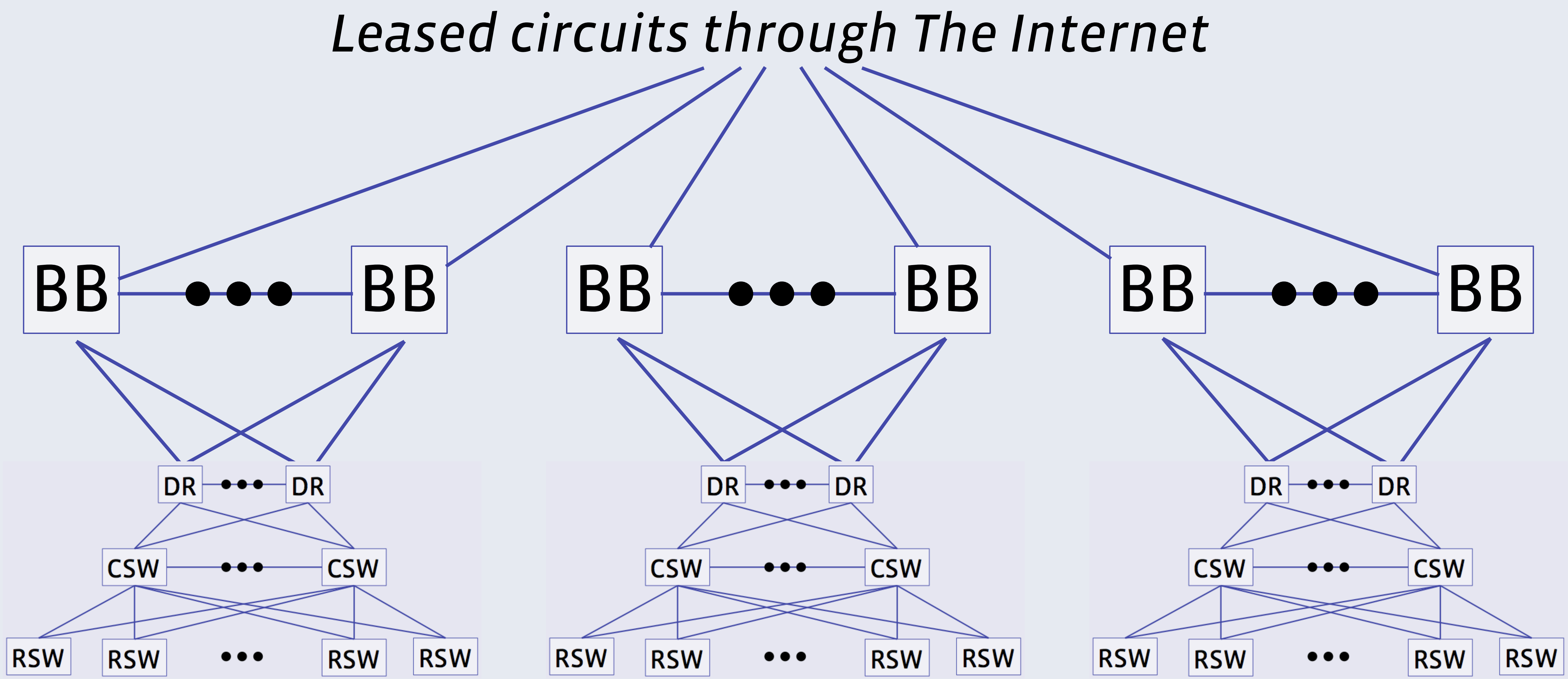
command
option
alt
THIS JOURNEY
1% FINISHED

Facebook’s datacenter network architecture

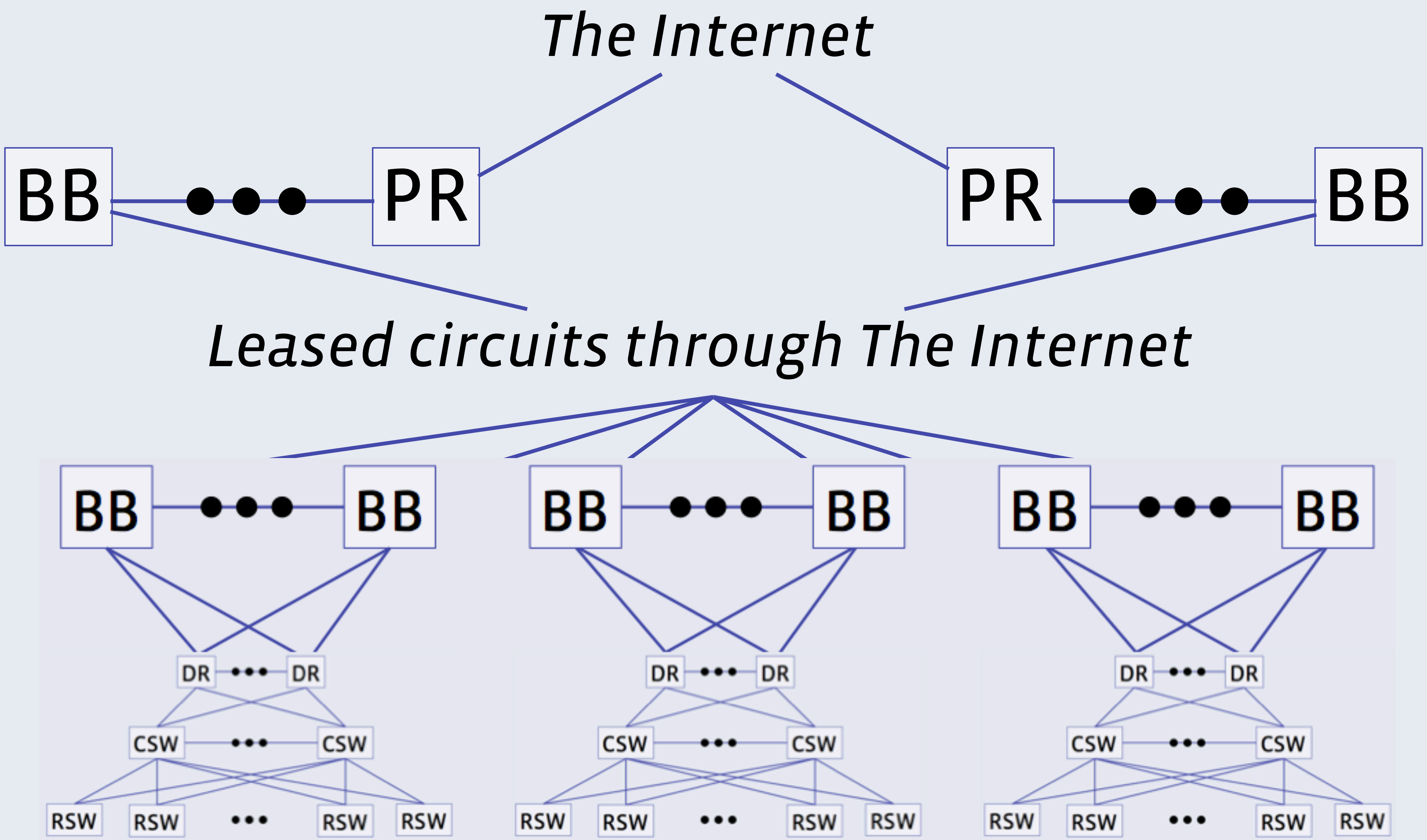
1. Capacity & redundancy



2. Backbone for predictable performance



3. POPs to reduce latency

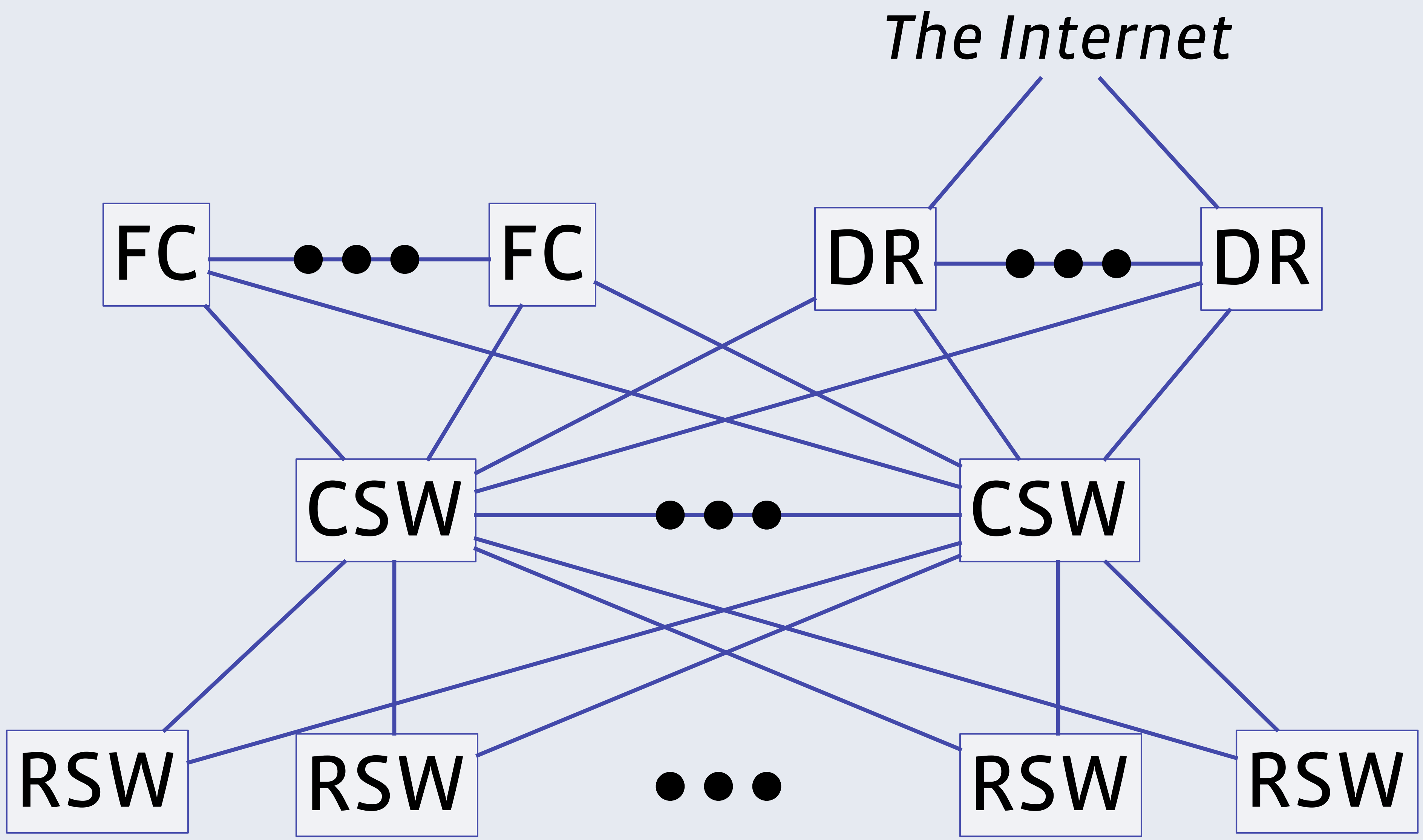


Main point of entry

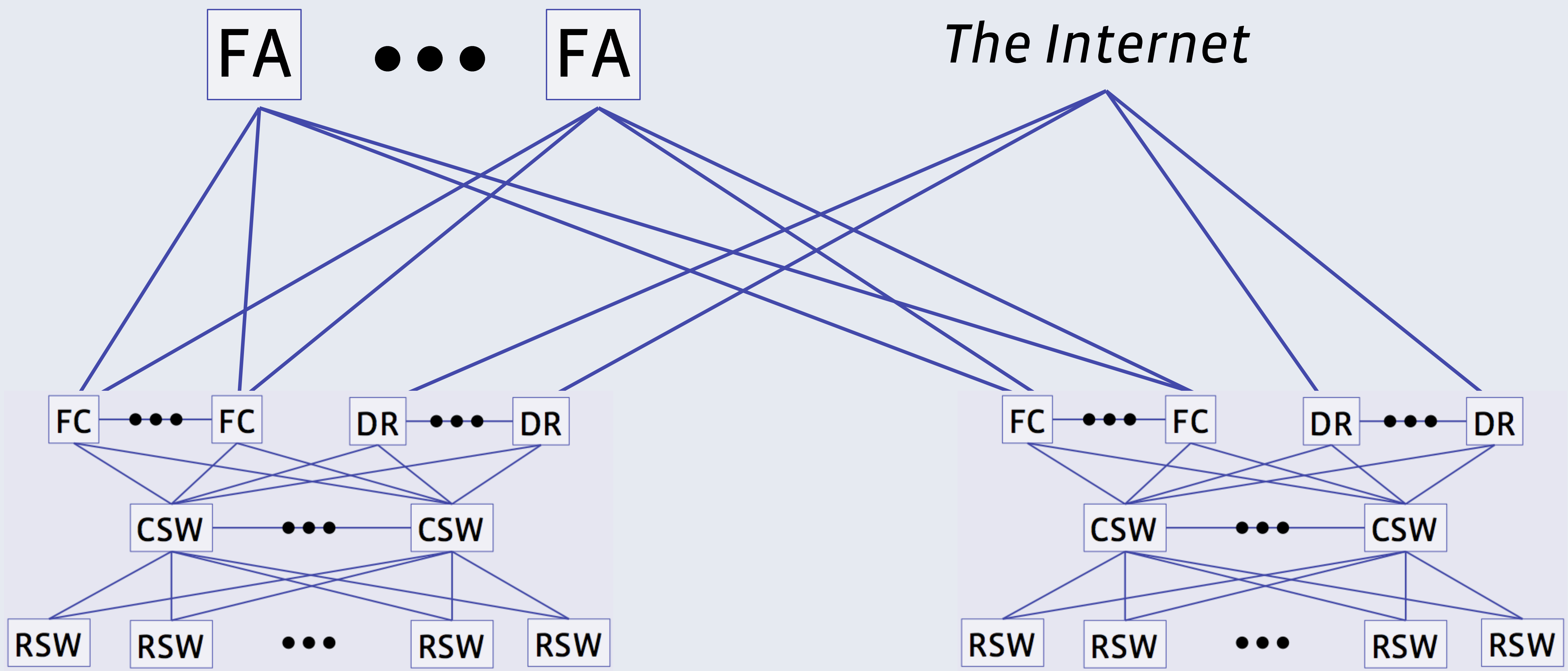
A few overhead cable trays

Challenge: big data

4. Datacenter as one computer



5. Multiple datacenters as one computer



facebook