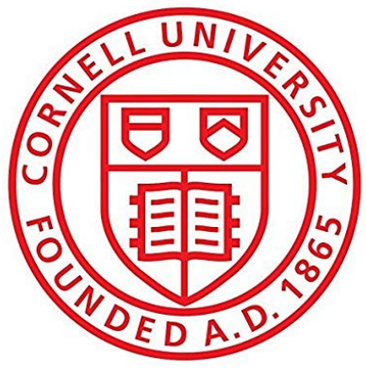
Building High-speed Datacenter Networks in the Post-Moore's Law Era

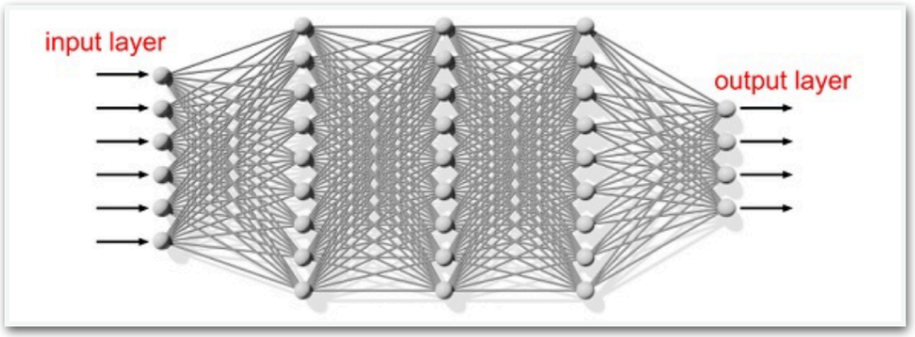
Vishal Shrivastav
Cornell University



Datacenters : Workhorses of Modern Internet

Network Intensive Applications

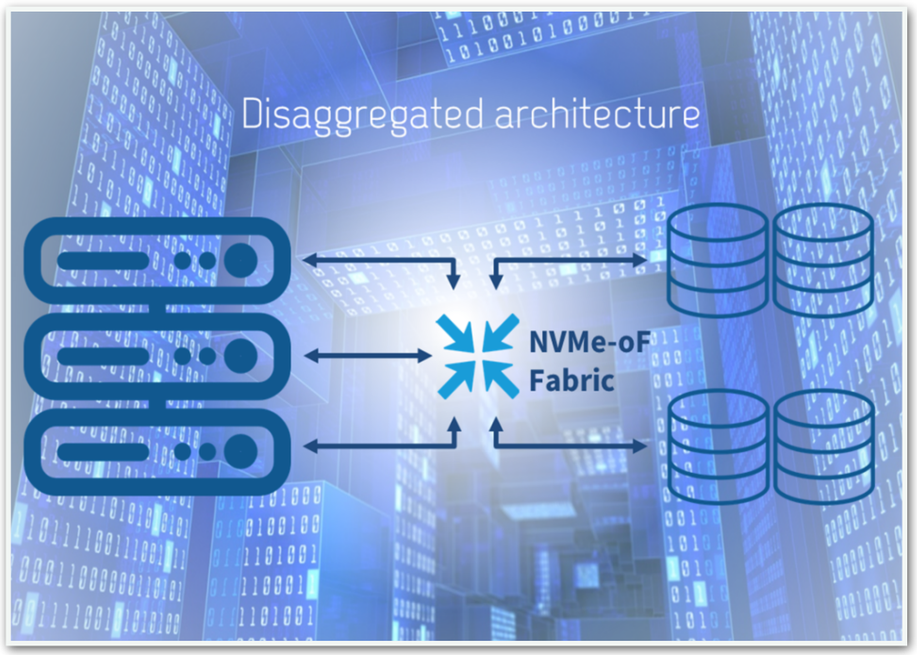
Applications getting more distributed



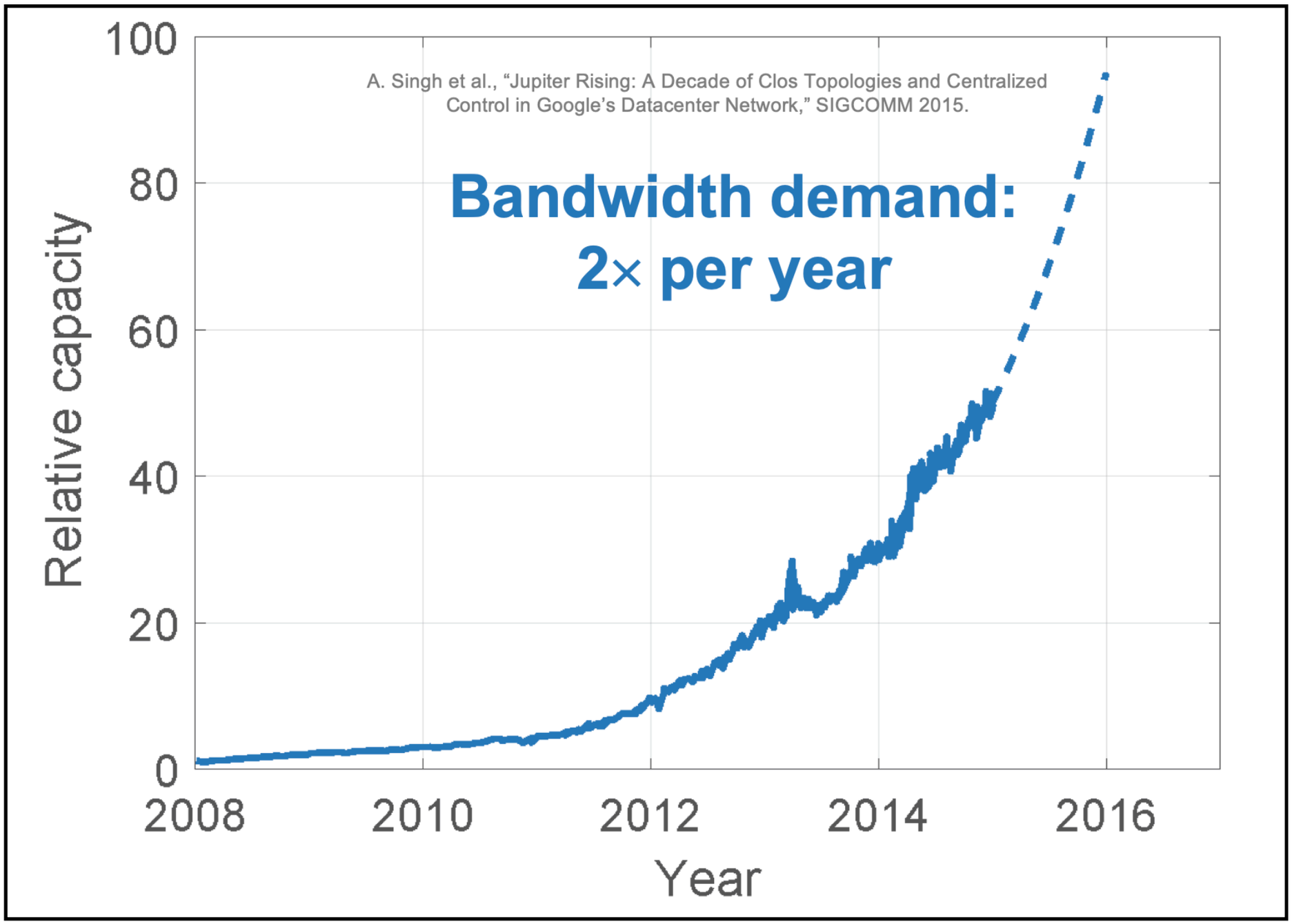
GraphLab

APACHE Spark™

Resources getting disaggregated

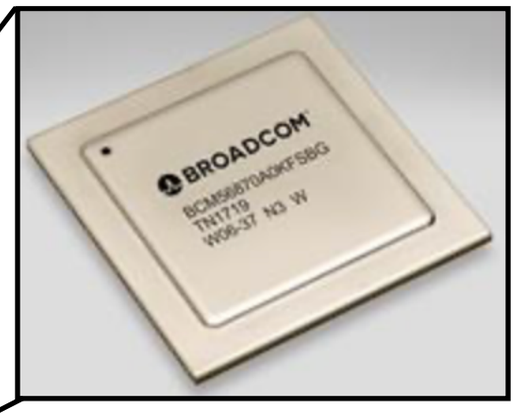


Bandwidth Demand

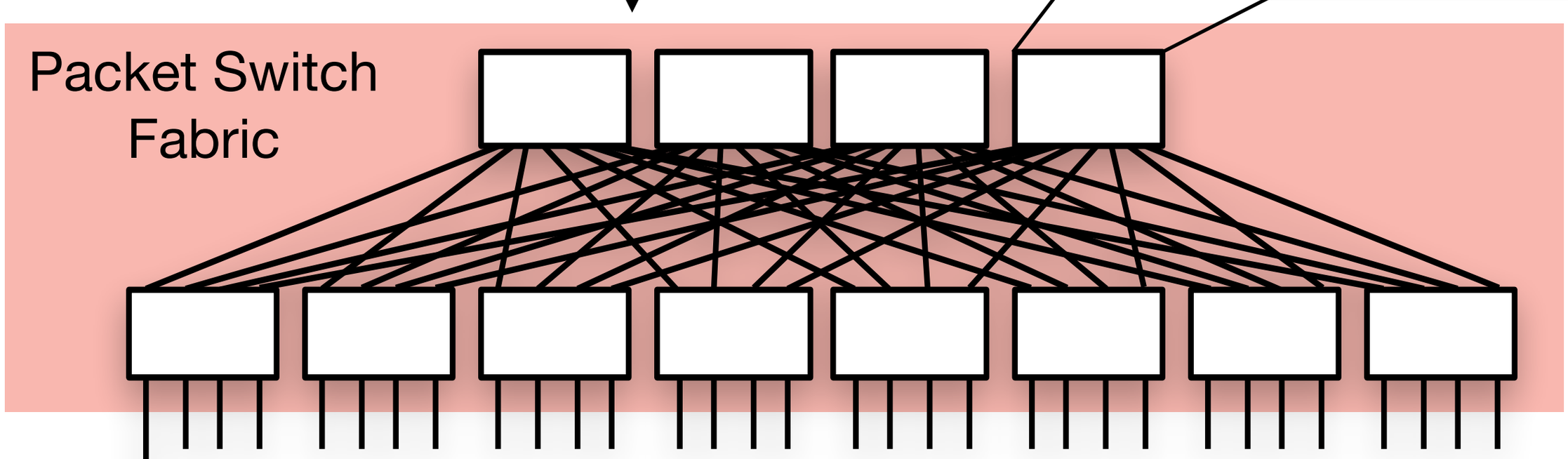


Today's Datacenter Networks

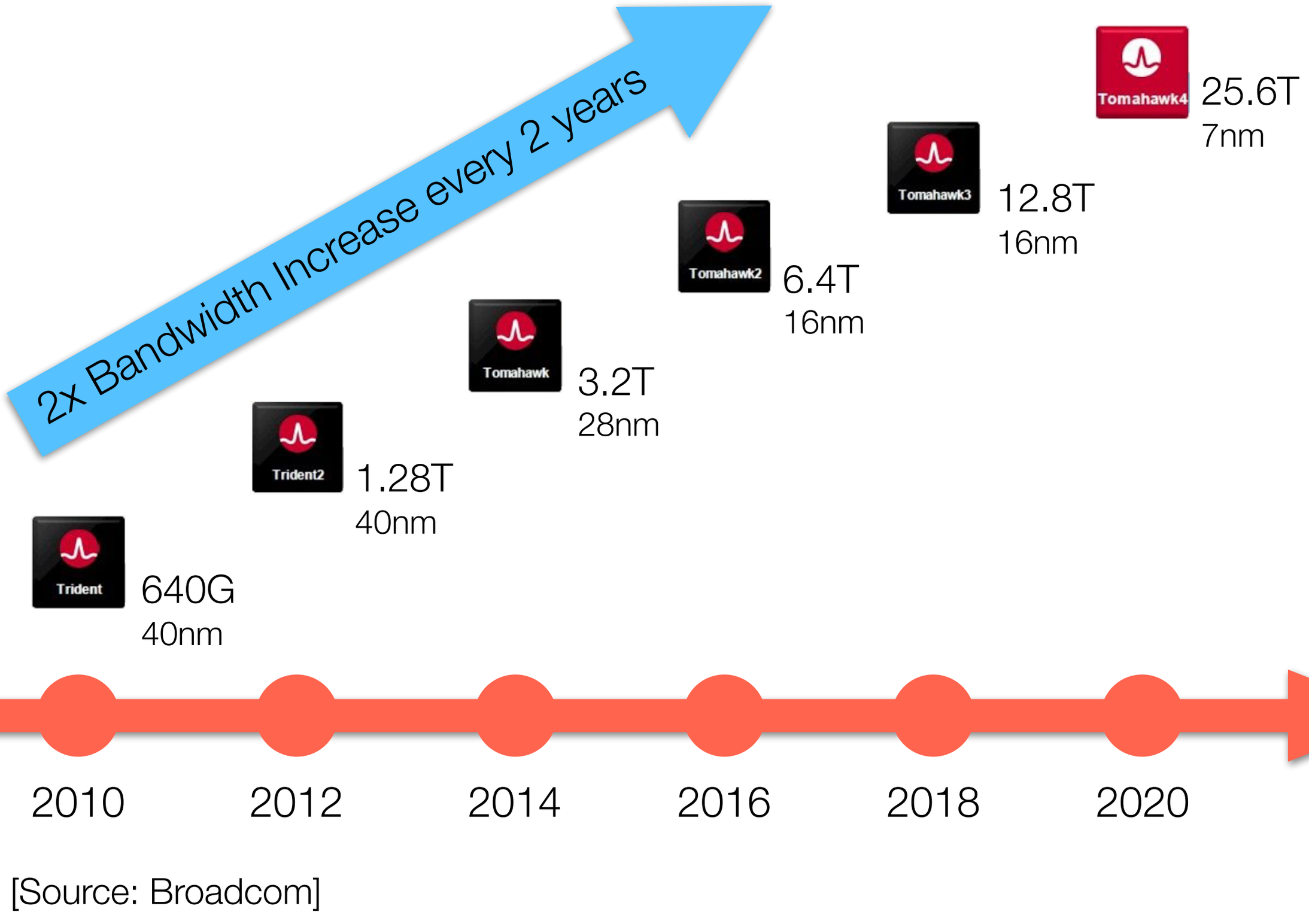
Switching Chip



Folded Clos Topology

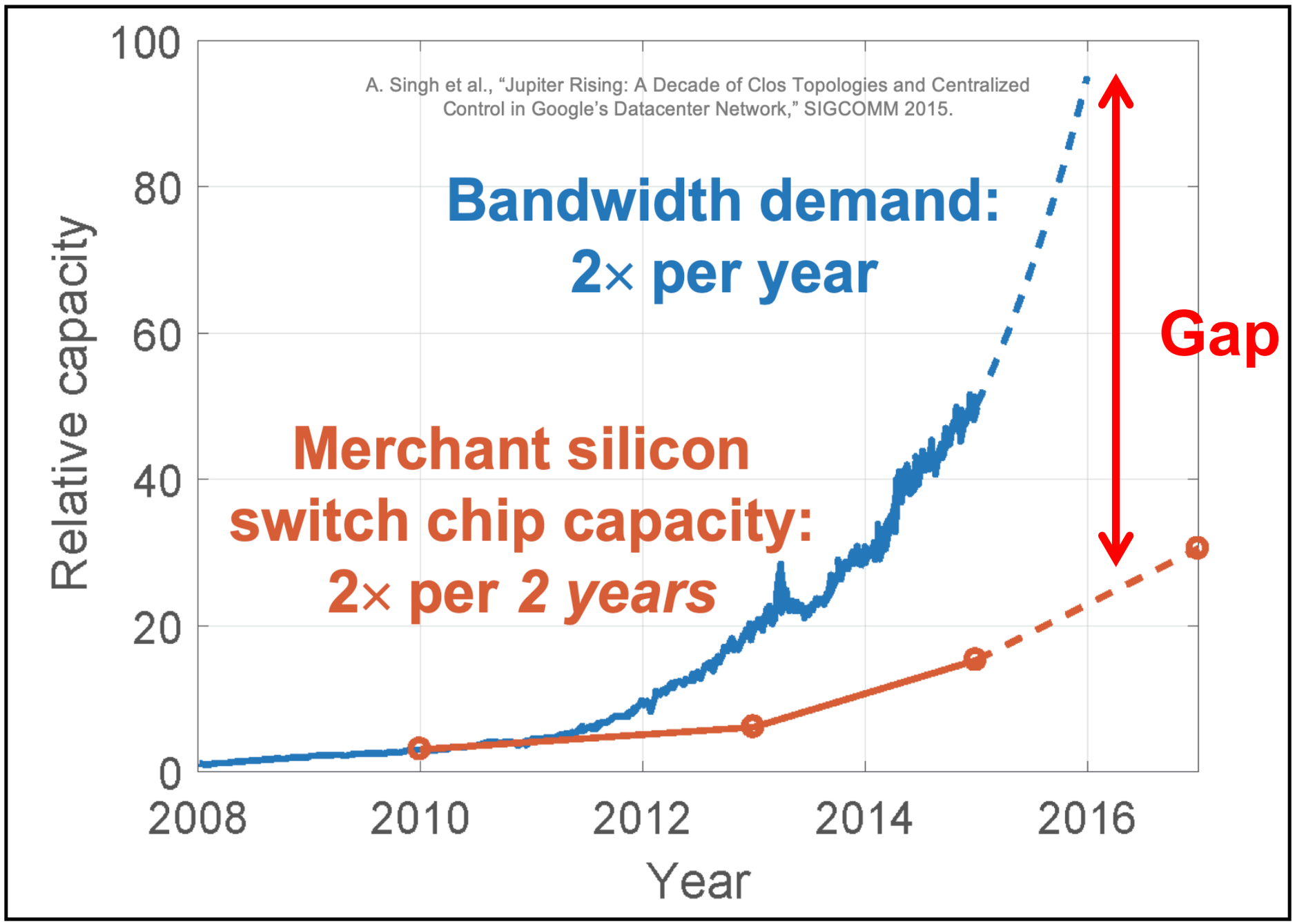


Switching Capacity



[Source: Broadcom]

Growing Gap


A. Singh et al., "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," SIGCOMM 2015.
Bandwidth demand:
2× per year
Merchant silicon
switch chip capacity:
2× per 2 years
Gap
Relative capacity
100
80
60
40
20
0
2008
2010
2012
2014
2016
Year

Scale-Out to the Rescue?

Fully Provisioned Clos Network with 8,192 hosts

Topology	# Hops	# Switch Chips	# Transceivers	# Wires
2 tiers	4	192	24K	12K
3 tiers	6	1,280	40K	20K

Goal

A **datacenter switching fabric**
that provides **unlimited bandwidth scaling**
at **low power**, **low cost**, **high performance**

Circuit Switching?

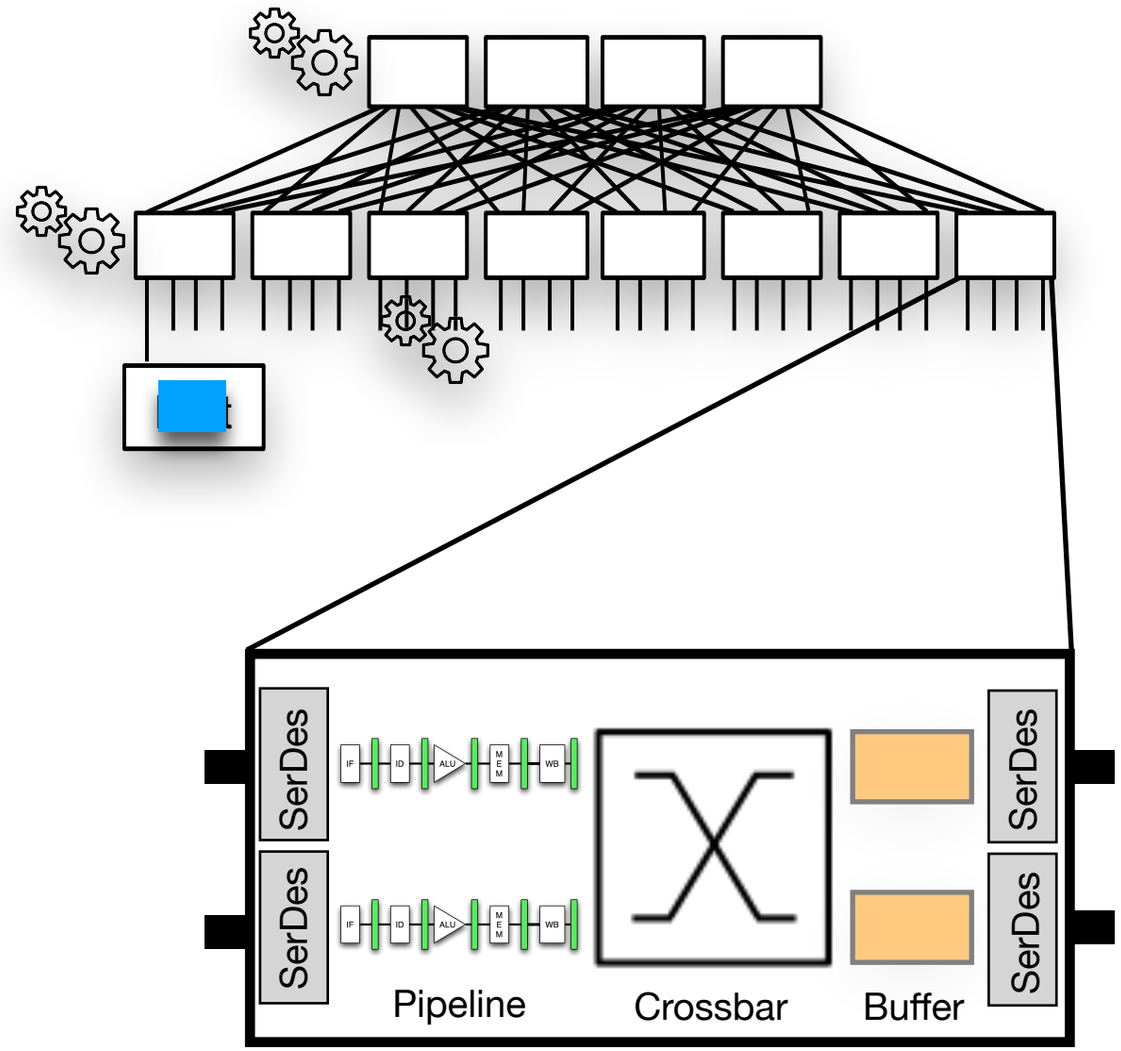
Circuit Switching: Opportunities & Challenges

Packet Switching

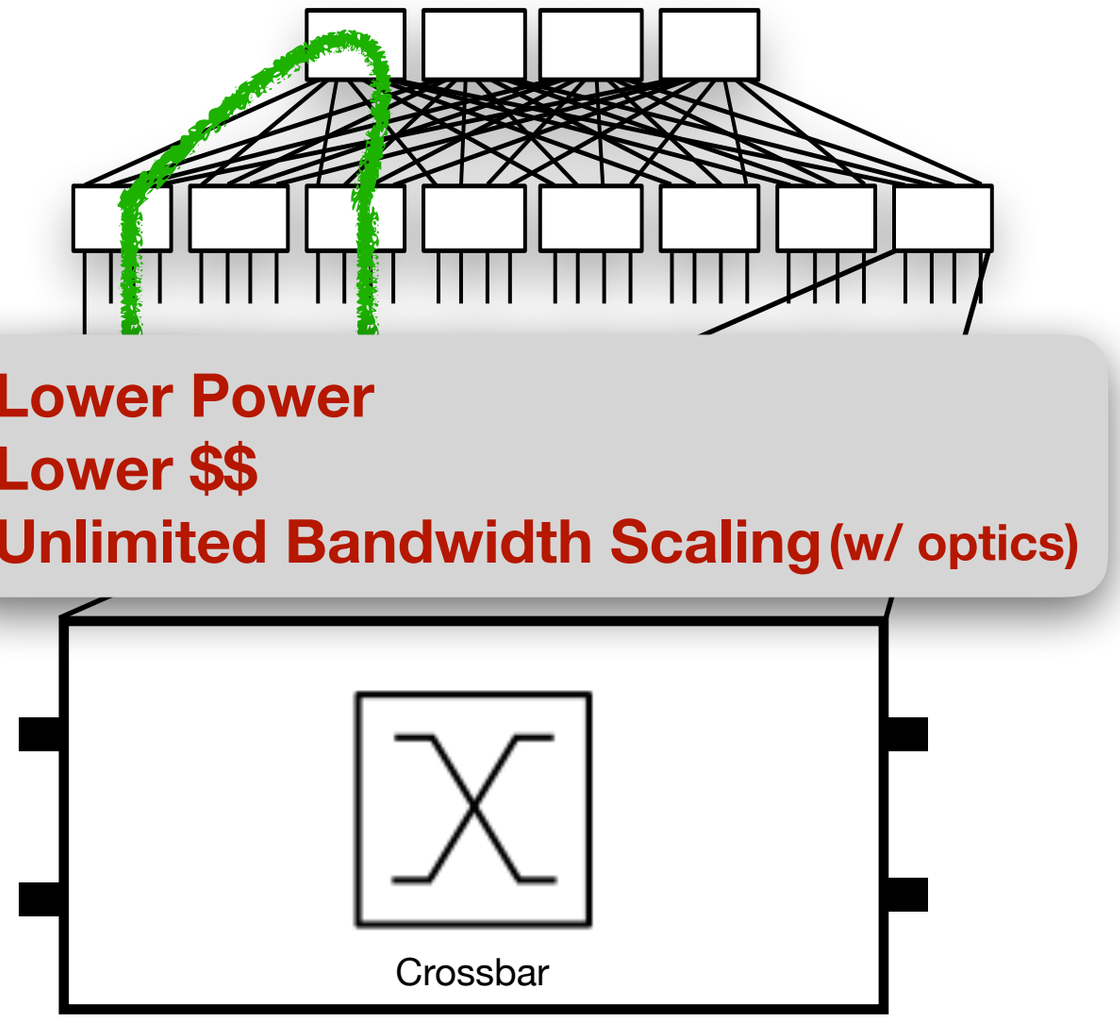
(Per packet, Per hop Decisions)

VS.

Circuit Switching



Packet Switch



Circuit Switch

Circuit Switching: Opportunities & Challenges

Circuit Configuration

Data Plane

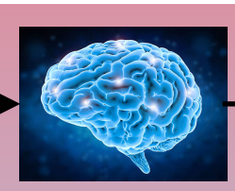
setting up physical circuits

milliseconds

nanoseconds

Control Plane

circuit scheduling

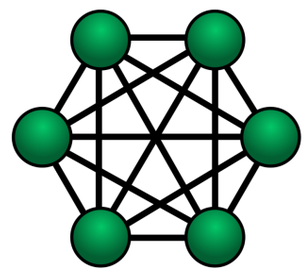
Traffic demand →  → Circuits

Centralized

???

ns-scale circuit scheduling with high performance?

Shoal



“Shoal: A Network Architecture for Disaggregated Racks”

Vishal Shrivastav, Asaf Valadarsky, Hitesh Ballani, Paolo Costa, Ki Suh Lee, Han Wang, Rachit Agarwal, Hakim Weatherspoon

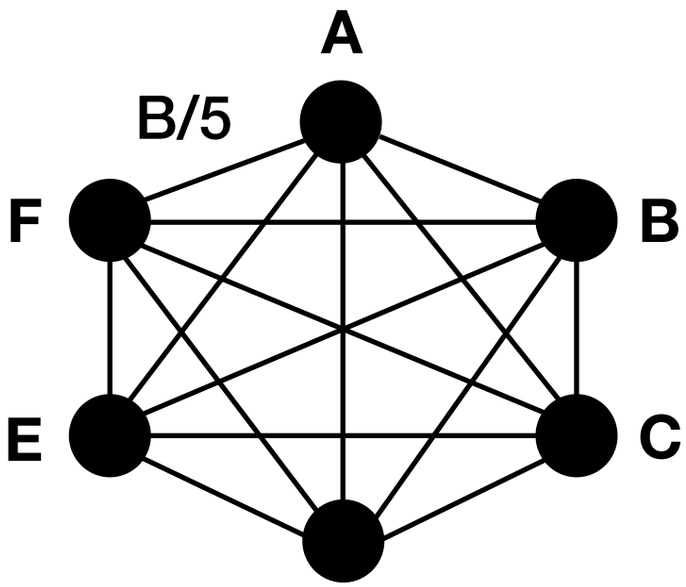
In USENIX NSDI 2019

Shoal is the first **network stack** and **switching fabric** design for a **fast circuit-switched** network

1. **Physical Layer :** Fast circuit scheduling mechanism
2. **Routing :** Bounded worst-case throughput
3. **Congestion Control :** Bounded worst-case queuing

Achieves comparable or better performance than several recent packet-switched network designs

Physical Layer



Full Mesh
Virtual Topology

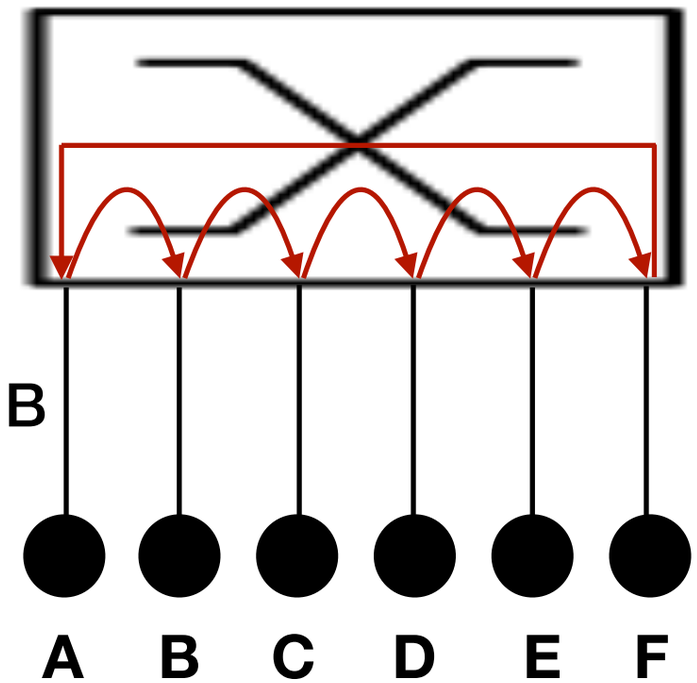
N-1 timeslots (epoch)

End-hosts	1	2	3	4	5
A	B	C	D	E	F
B	C	D	E	F	A
C	D	E	F	A	B
D	E	F	A	B	C
E	F	A	B	C	D
F	A	B	C	D	E

static, pre-defined schedule

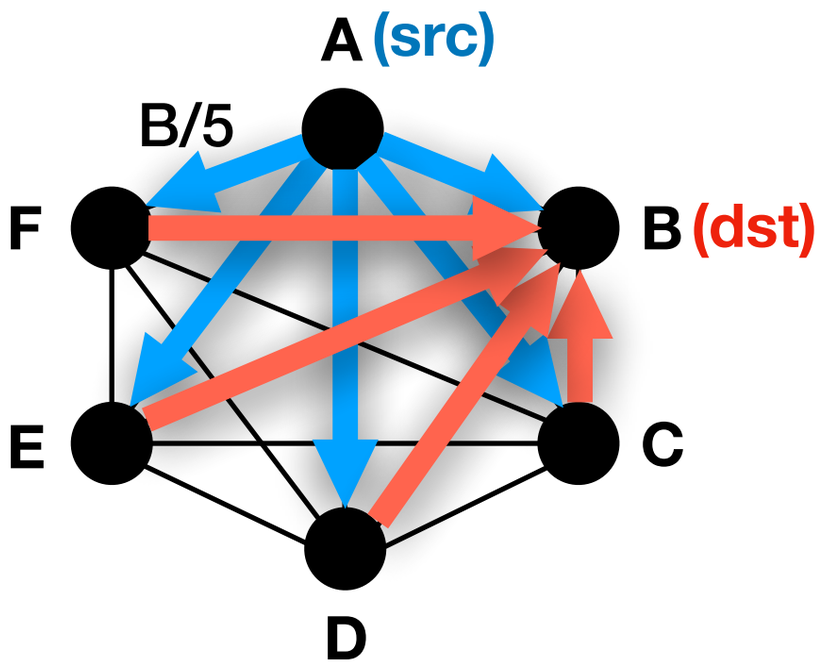
+

Synchronous System



Physical Topology

Routing



	Best-case throughput	Worst-case throughput	
Direct Routing	1	**1/(N-1)**	✗
Valiant LB	1	**0.5**	✓

Uniform load balancing → Forward to destination

Routing in Shoal

Practical Challenges

Shoal is a synchronous system

Requires ns-precision network-wide synchronization

“Globally Synchronized Time via Datacenter Networks”

Vishal Shrivastav, Ki Suh Lee, Han Wang, Hakim Weatherspoon

In Journal of Transactions on Networking (ToN) 2019

*** conference version in ACM SIGCOMM 2016**

Circuit set-up changes at ns-scale

Requires fast, ns-precision scheduling at end-hosts

“Fast, Scalable, and Programmable Packet Scheduler in Hardware”

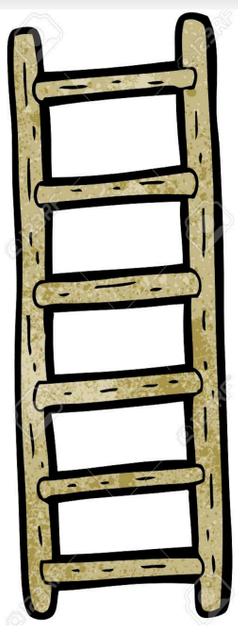
Vishal Shrivastav

In ACM SIGCOMM 2019

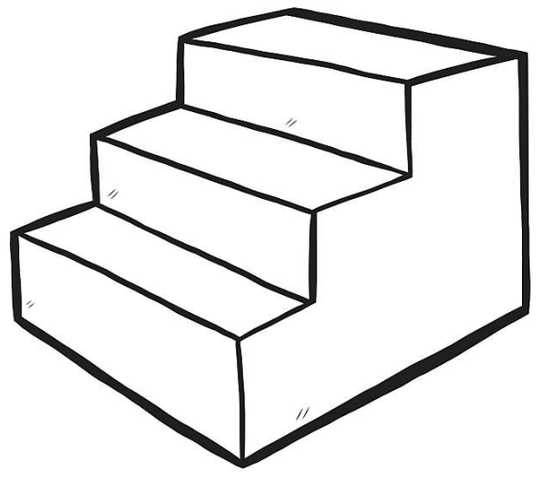
Conclusion

Unlimited
Bandwidth Scaling
at
Low Power, Low Cost, High Performance

Fast, High Performance
Circuit Scheduling



Shoal



Circuit Switching

Thank you!

Email: vishal@cs.cornell.edu

Webpage: http://www.cs.cornell.edu/~vishal/