Hybrid Datacenter Networks

George Papen
Department of Electrical and Computer Engineering
University of California at San Diego

Hot Interconnects 2013

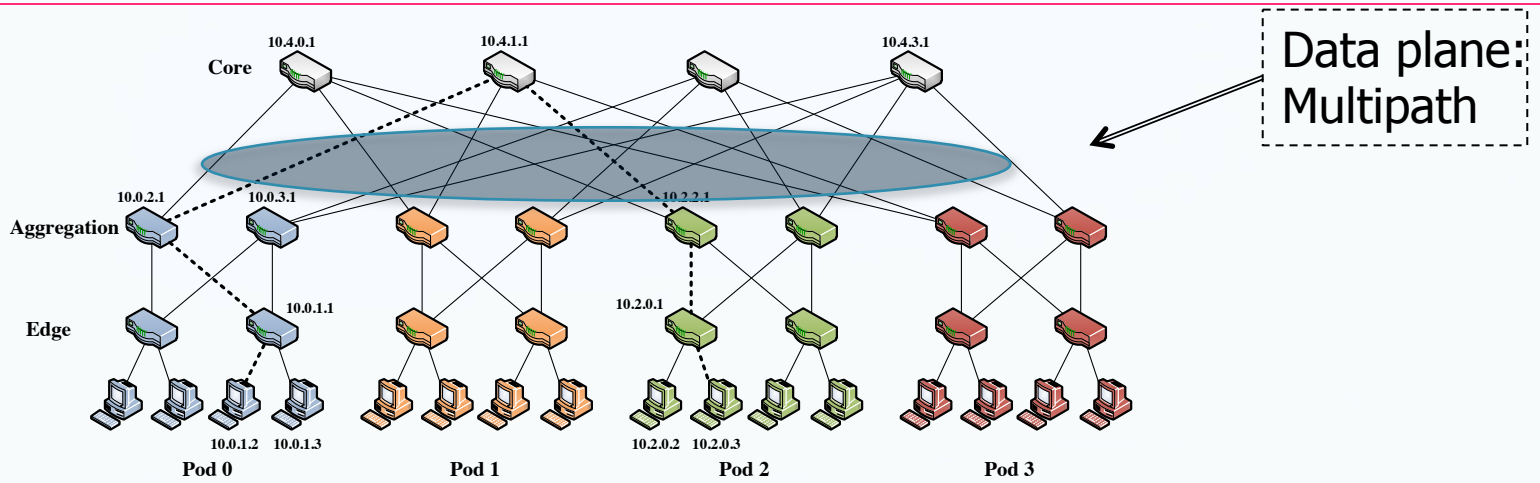
Research teams

- REACToR
  - *Students*: He Liu, Feng Lu, Rishi Kapoor, Malveeka Tewari, Alex Forencich
  - *Senior Researchers*: Stefan Savage, Geoff Voelker, George Papen, Alex C. Snoeren, George Porter
- Mordia
  - *Students*: Nathan Farrington, Alex Forencich, Richard Strong
  - *Senior Researchers*: Joe Ford, Pang Chen-Sun,Tajana Rosing, Yeshaiahu Fainman, George Papen, George Porter, Amin Vahdat

Outline

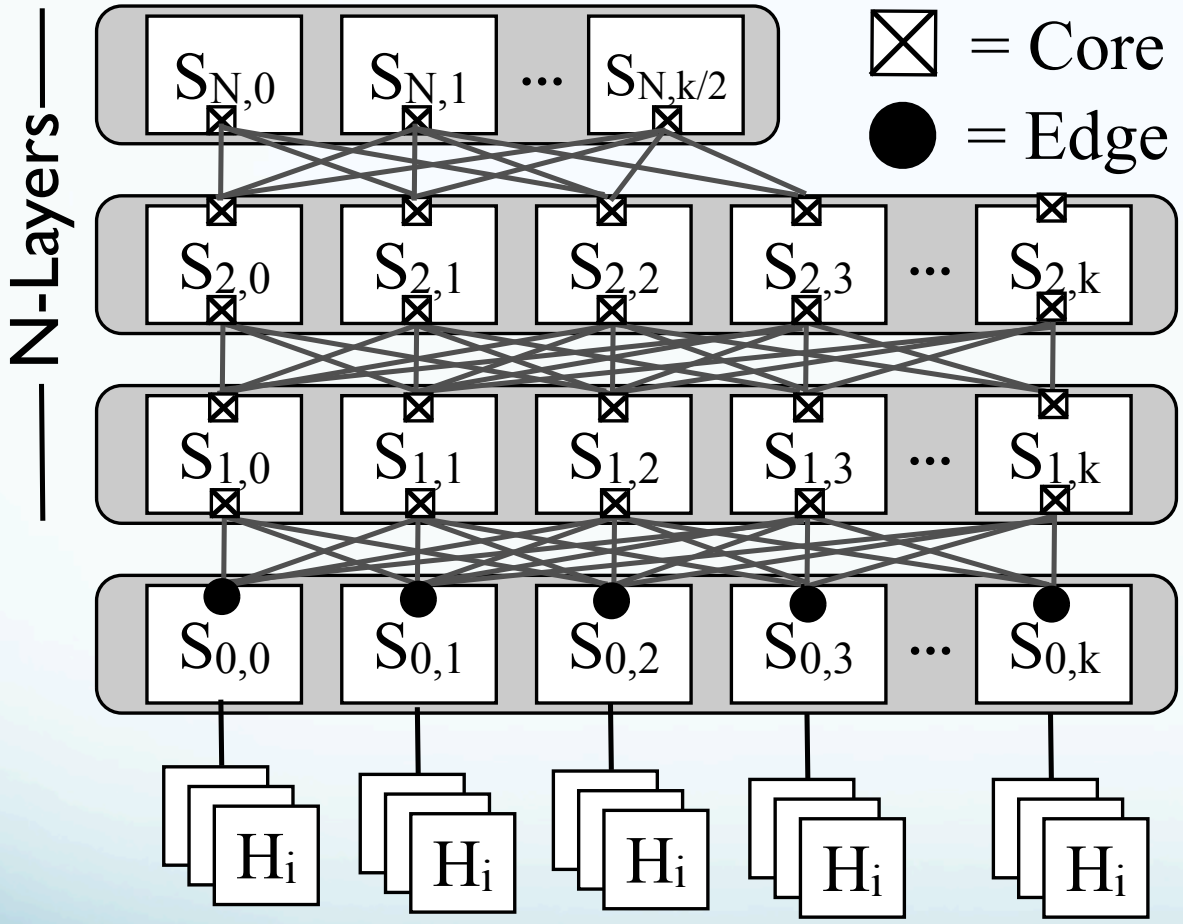
- Motivation and Background
  - Scale-out Datacenters
  - Hybrid Networks
- Research Issues
  - Circuits in a Packet-based World
  - Burstiness of Traffic
  - Scheduling
  - Optical Circuit Switching
- Conclusions and Acknowledgements

Scale-out data centers



- Massive scale; large number of nodes (100k+)
- Applications are different, unpredictable, uncoordinated
- *Bi-section bandwidth*, *low latency, and reliability* critical
- Scale-out designs [VL2,FatTree,...]:
  - ✓ No oversubscription
  - Cost, Power, Complexity

Scalability → High costs



⊠ = Core transceiver

● = Edge transceiver

Number of components Is strong function of # layers

Sources of cost, power, and complexity

Network	# nodes	# levels	Switch radix	Core Transceivers	Core Transceivers per host
10G	27,648	3	48	138,240	**5**
10G	65,536	3	64	393,216	**5**
40G, redundancy	15,552	5	12 (effective)	139,968	**9**

Faster links
Fault tolerance

} Smaller switch radix }

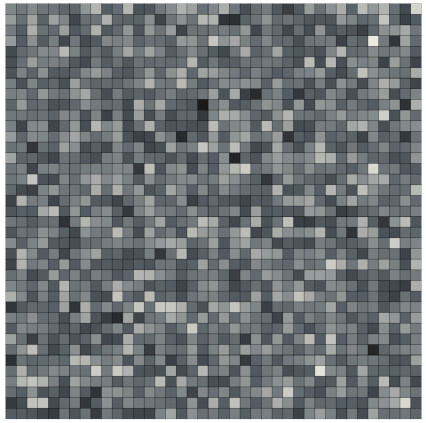
More layers
More cost

Network Demand

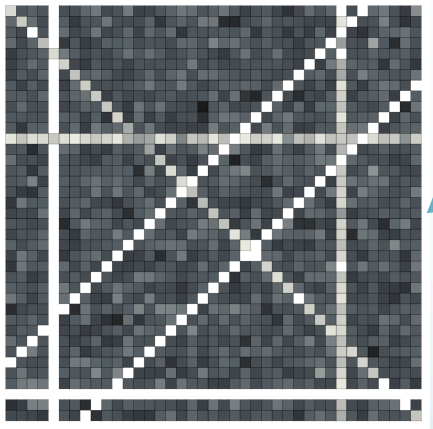
More "network-centric"

Datacenter workloads can vary across this space

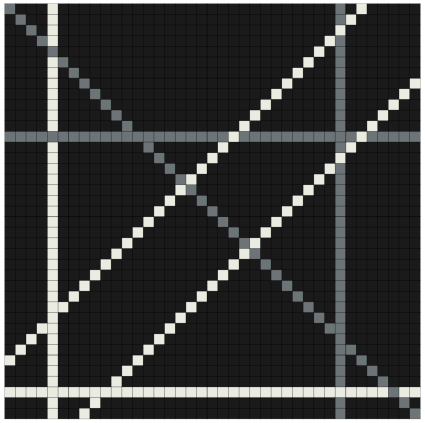
More "processor-centric"



Random traffic



Many datacenter traffic patterns are a mixture



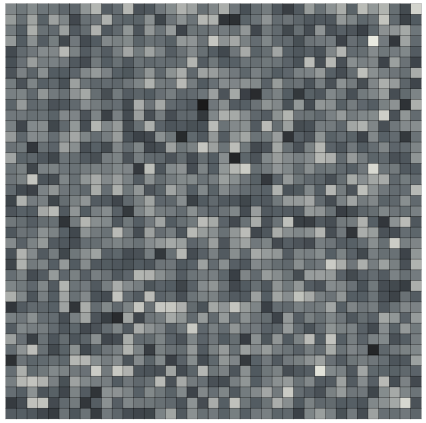
Correlated traffic

Network Demand

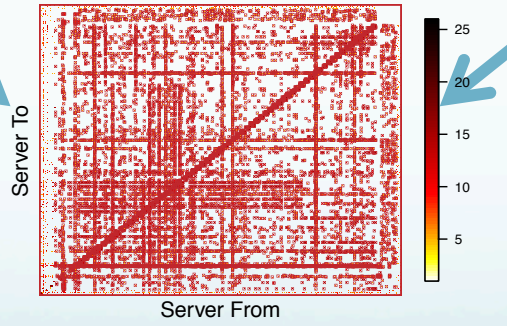
More "network-centric"

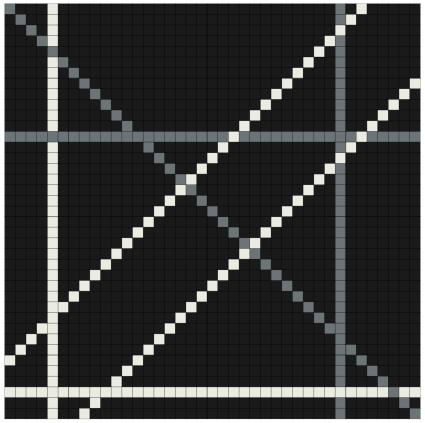
Datacenter workloads can vary across this space

More "processor-centric"



Random traffic

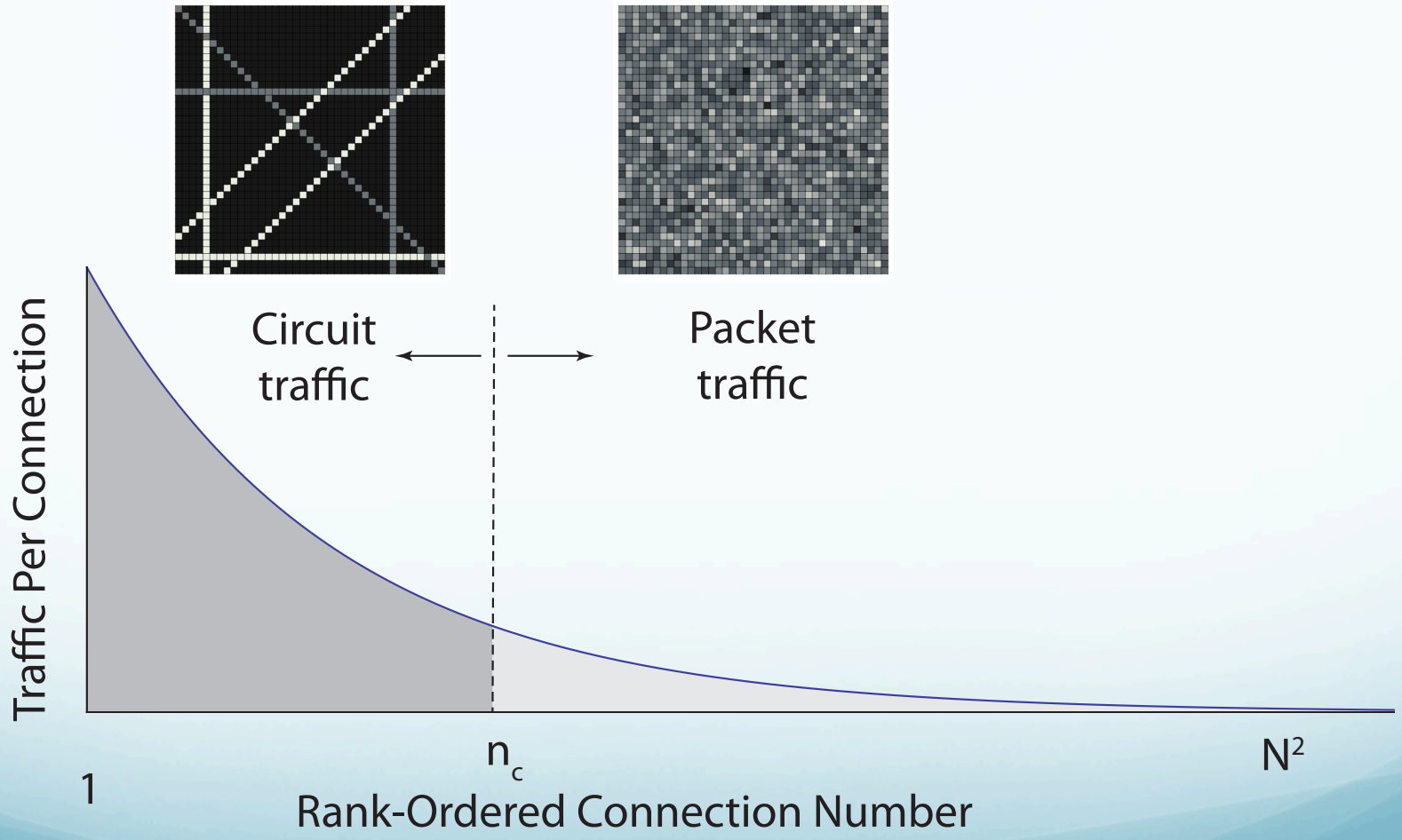




Correlated traffic

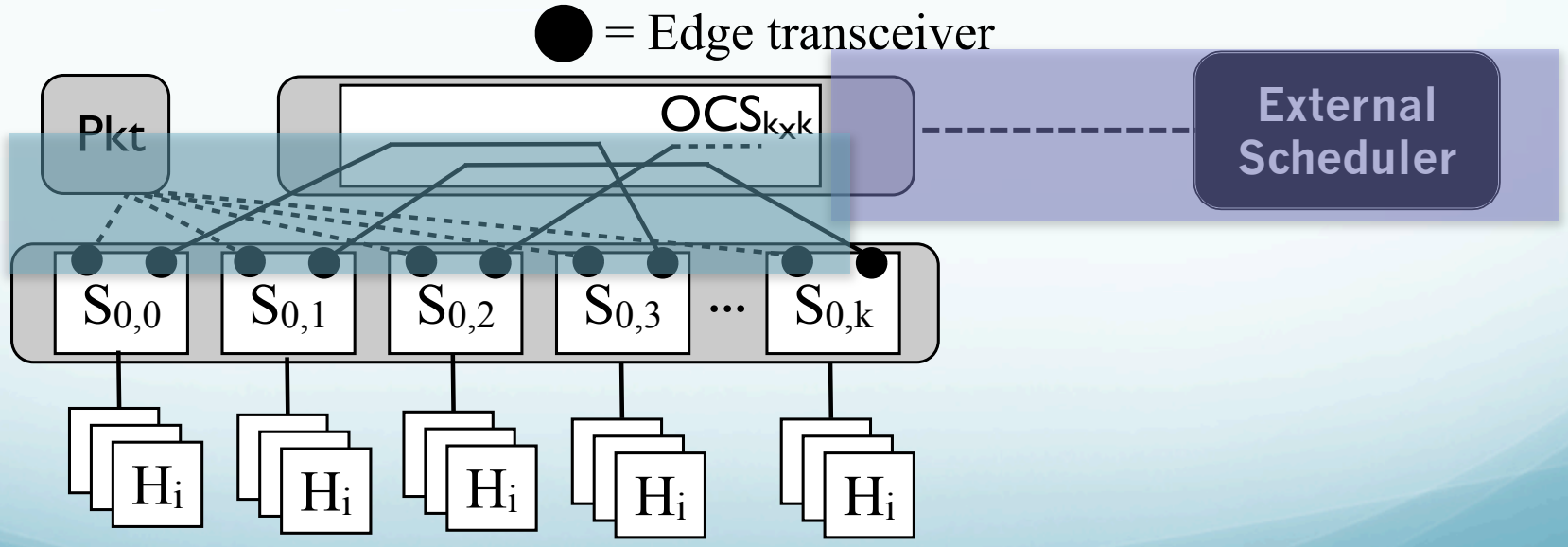
Measured datacenter traffic
(data from Microsoft - Kandula 2009)

Rank-Ordered Demand

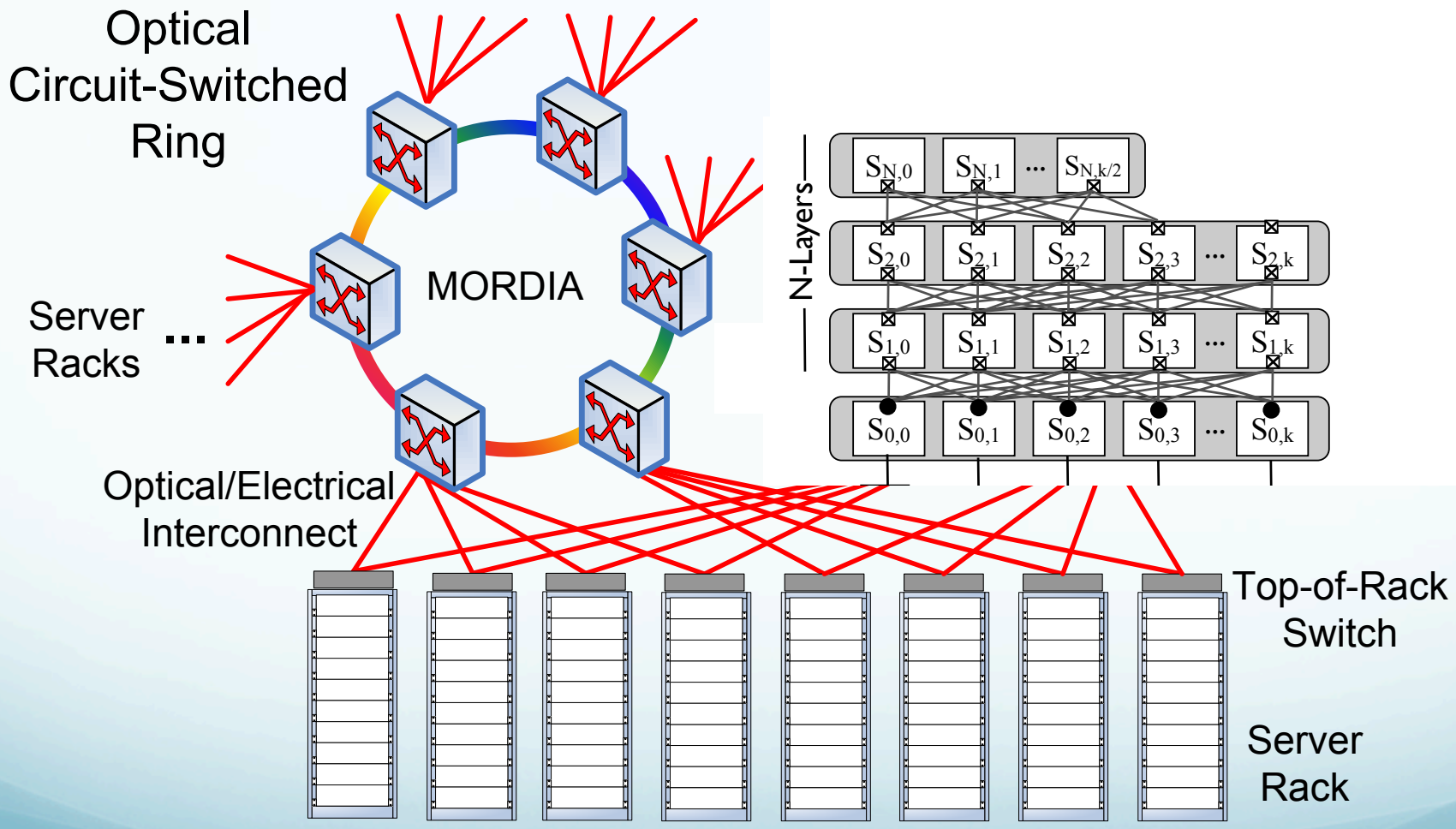


Hybrid Electrical/Optical Networks

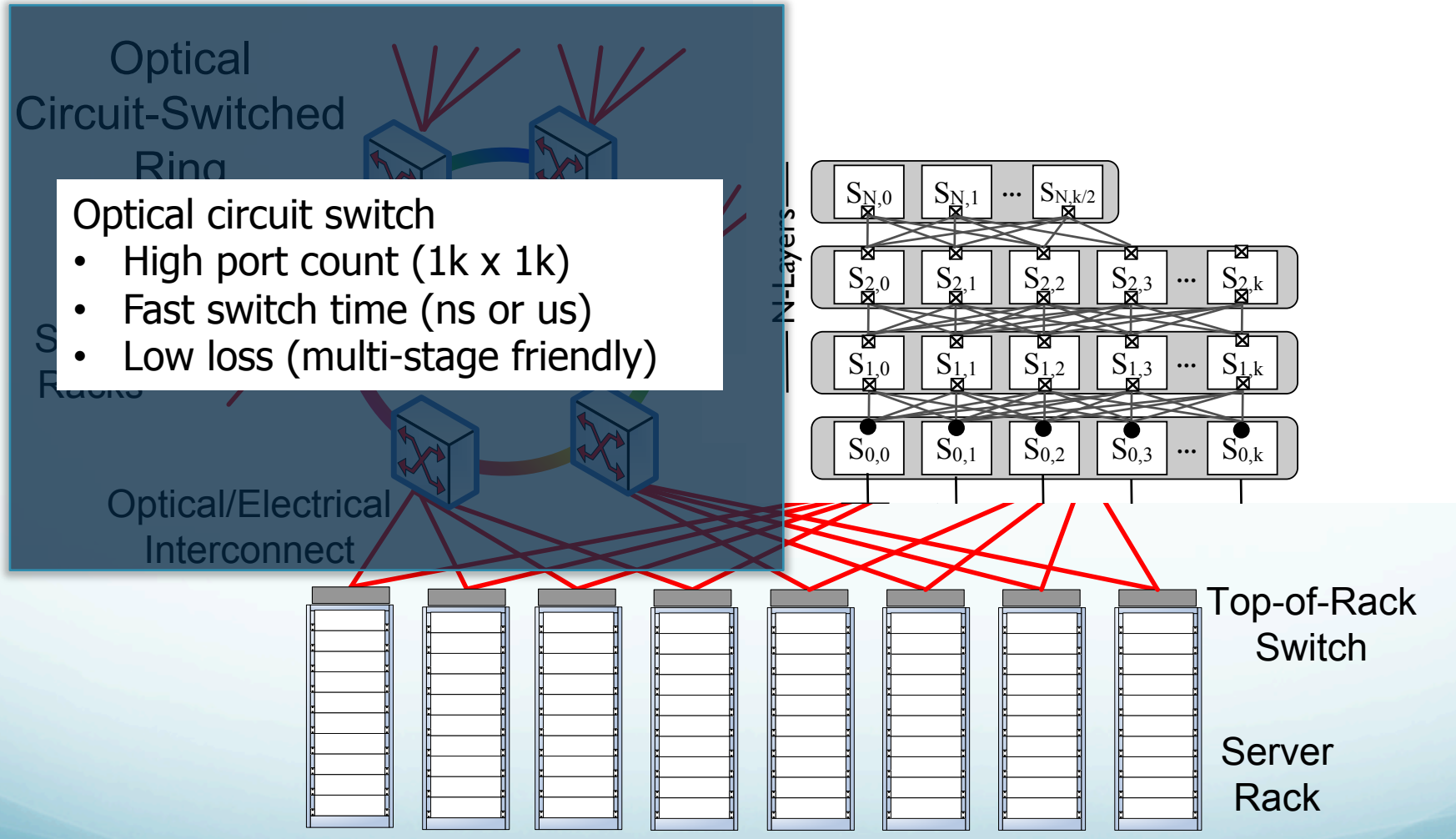
- Circuit switching
  - Decouple line rate from speed of control plane
  - Used for persistent high-data rate traffic – must be scheduled
- Packet switching
  - Handle 'tail' of traffic demand
  - Can correct for errors in circuit schedule



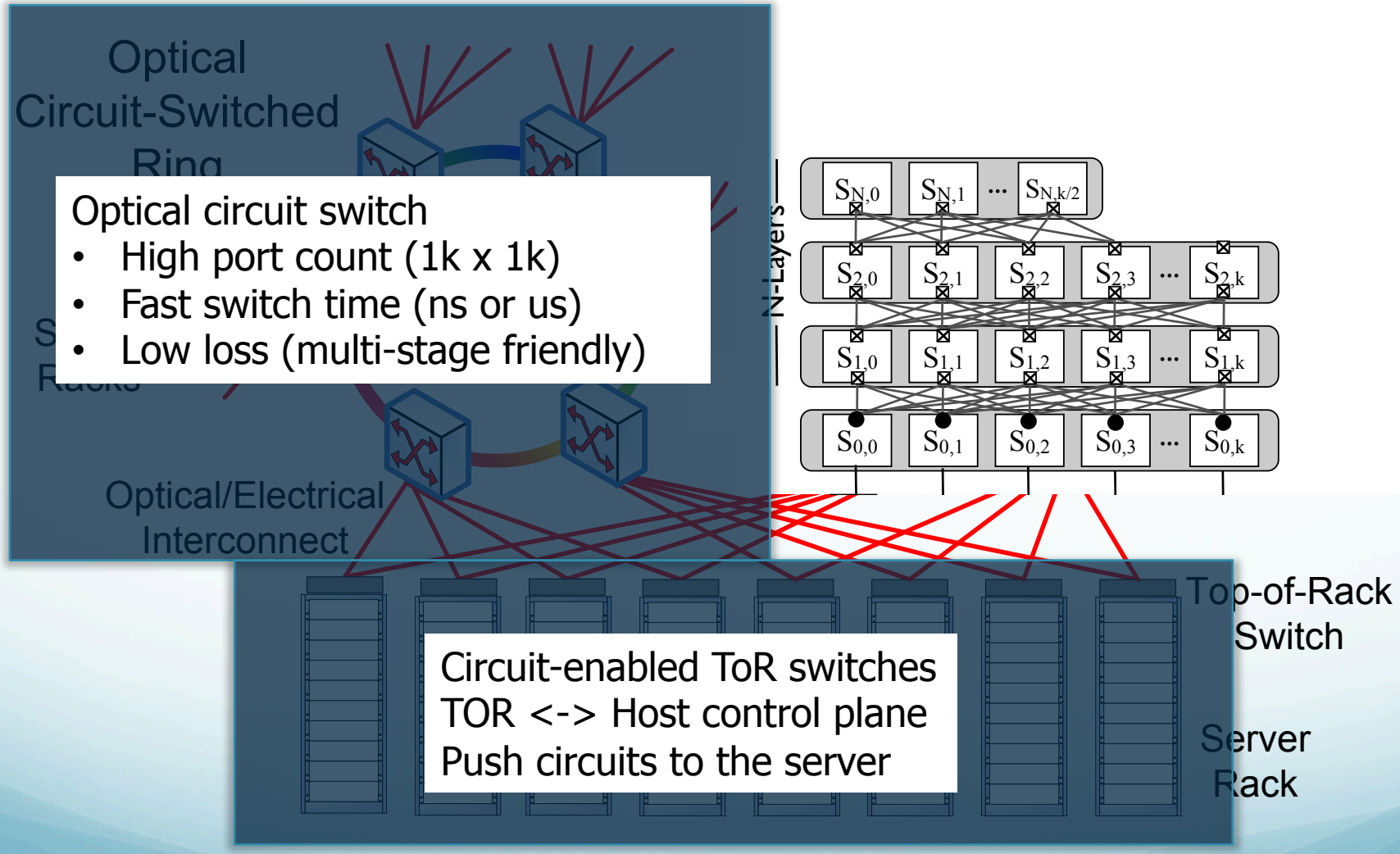
Hybrid Switching Architecture for Data Centers



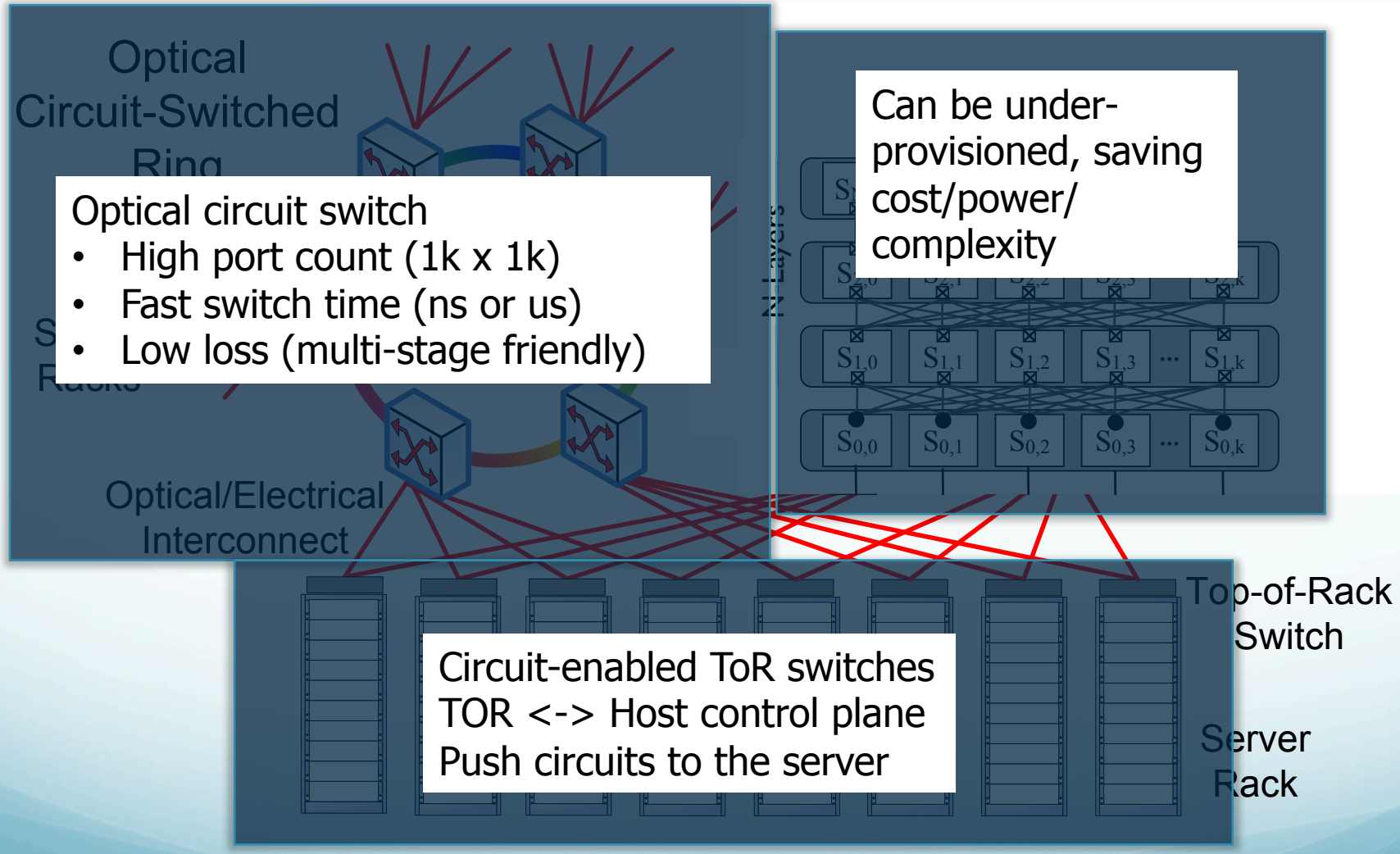
Hybrid Switching Architecture for Data Centers



Future Hybrid Needs



Future Hybrid Needs

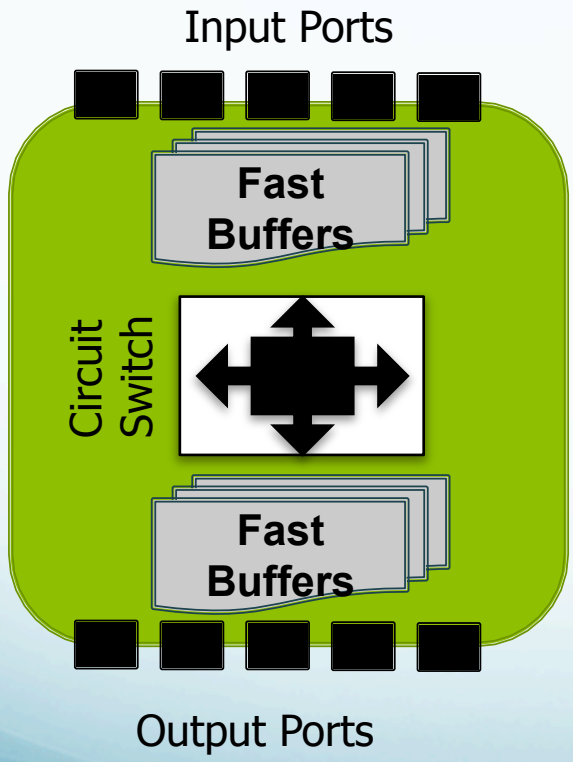


Outline

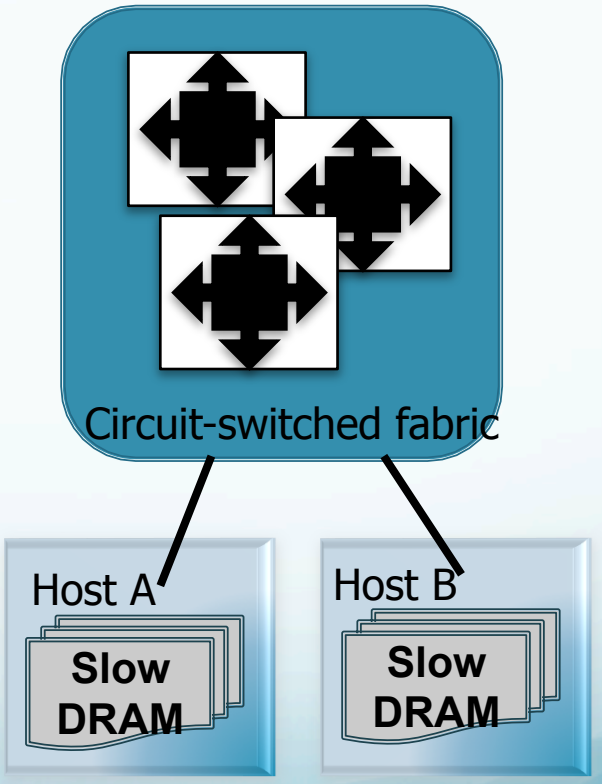
- Motivation and Background
  - Scale-out datacenters
  - Distributed vs. Centralized Network Control
- Research Issues
  - Circuits in a Packet-based World
  - Burstiness of Traffic
  - Scheduling
  - Optical Circuit Switching
- Conclusions and Acknowledgements

High-Level Diagram of REACToR

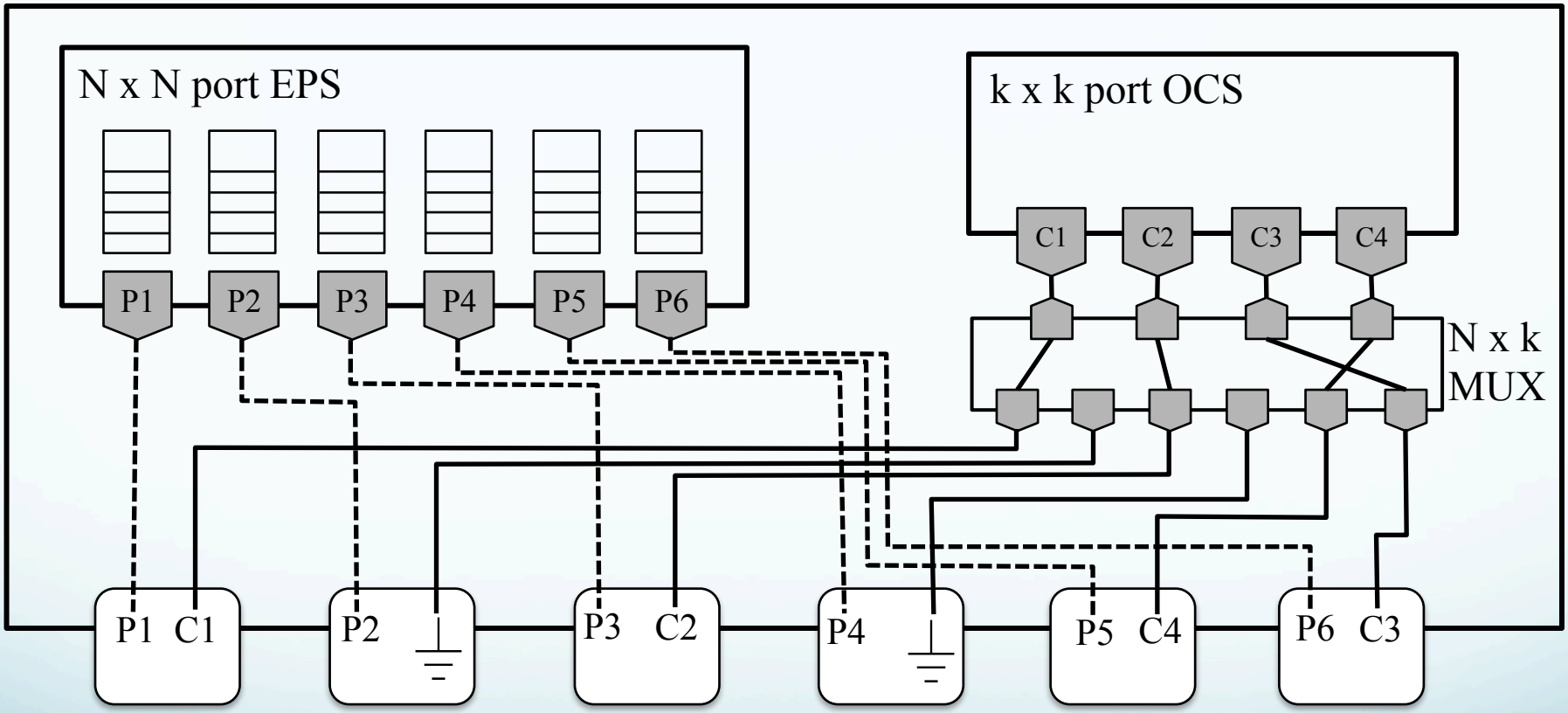
Standard Switch



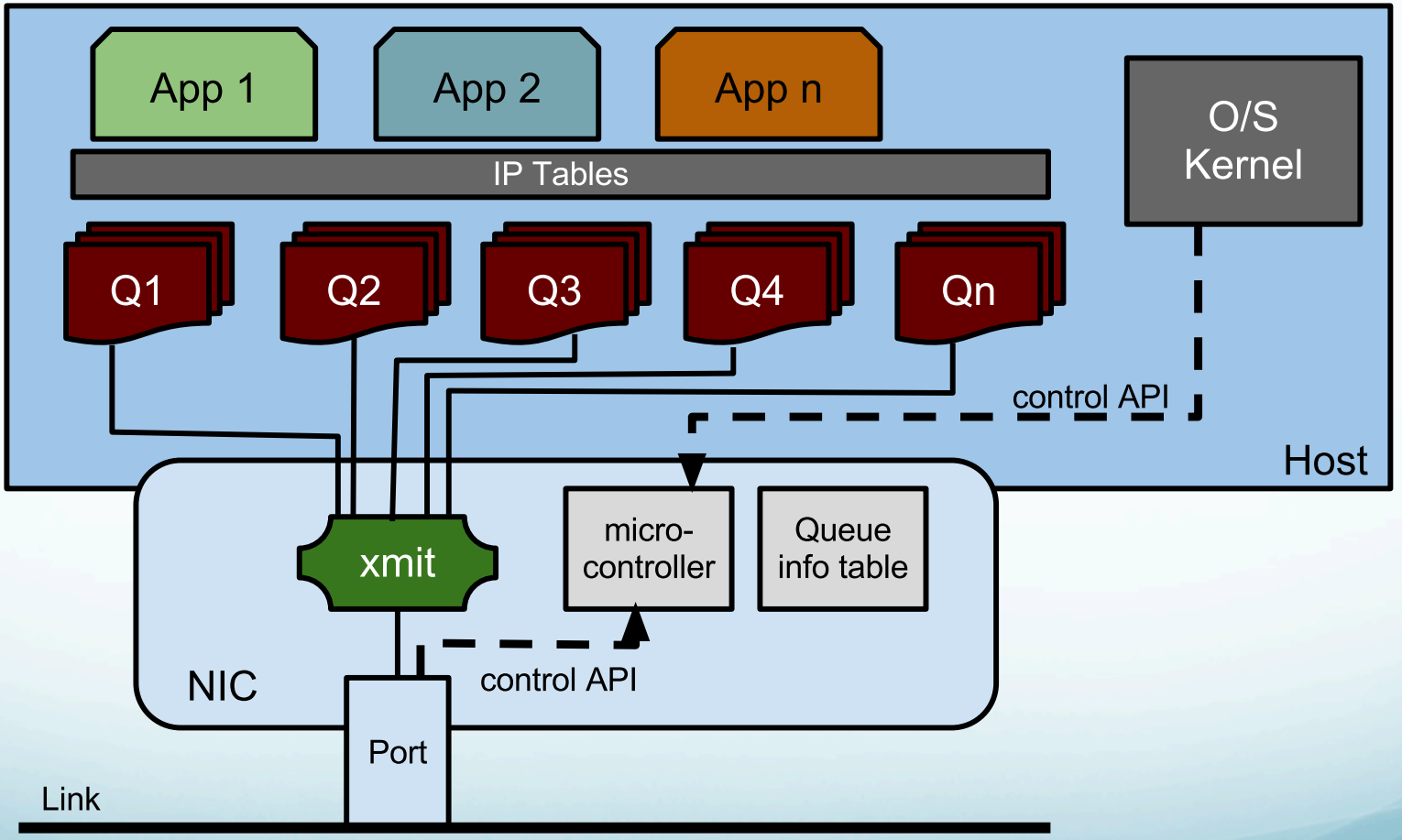
REACToR



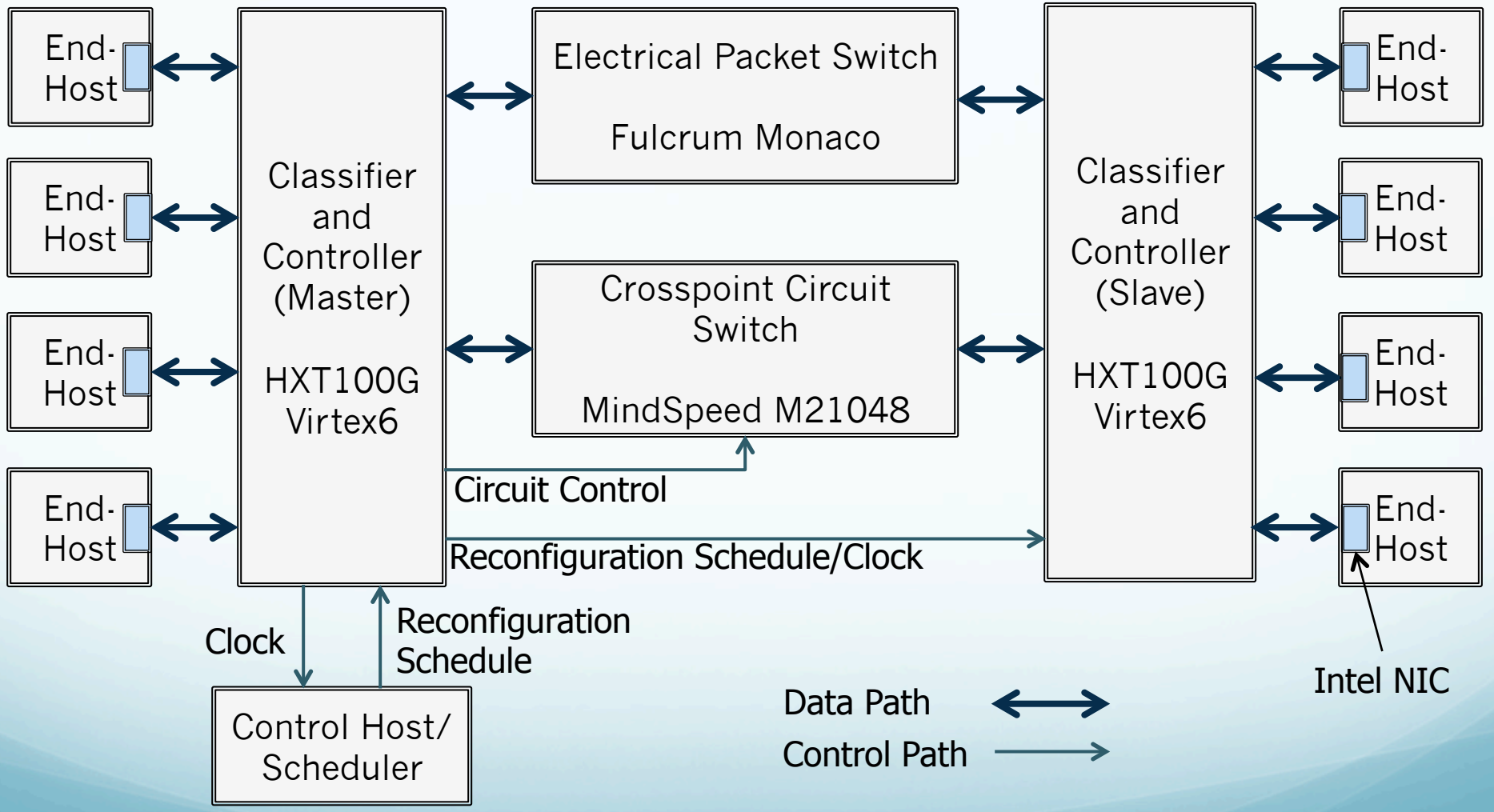
REACToR Architecture

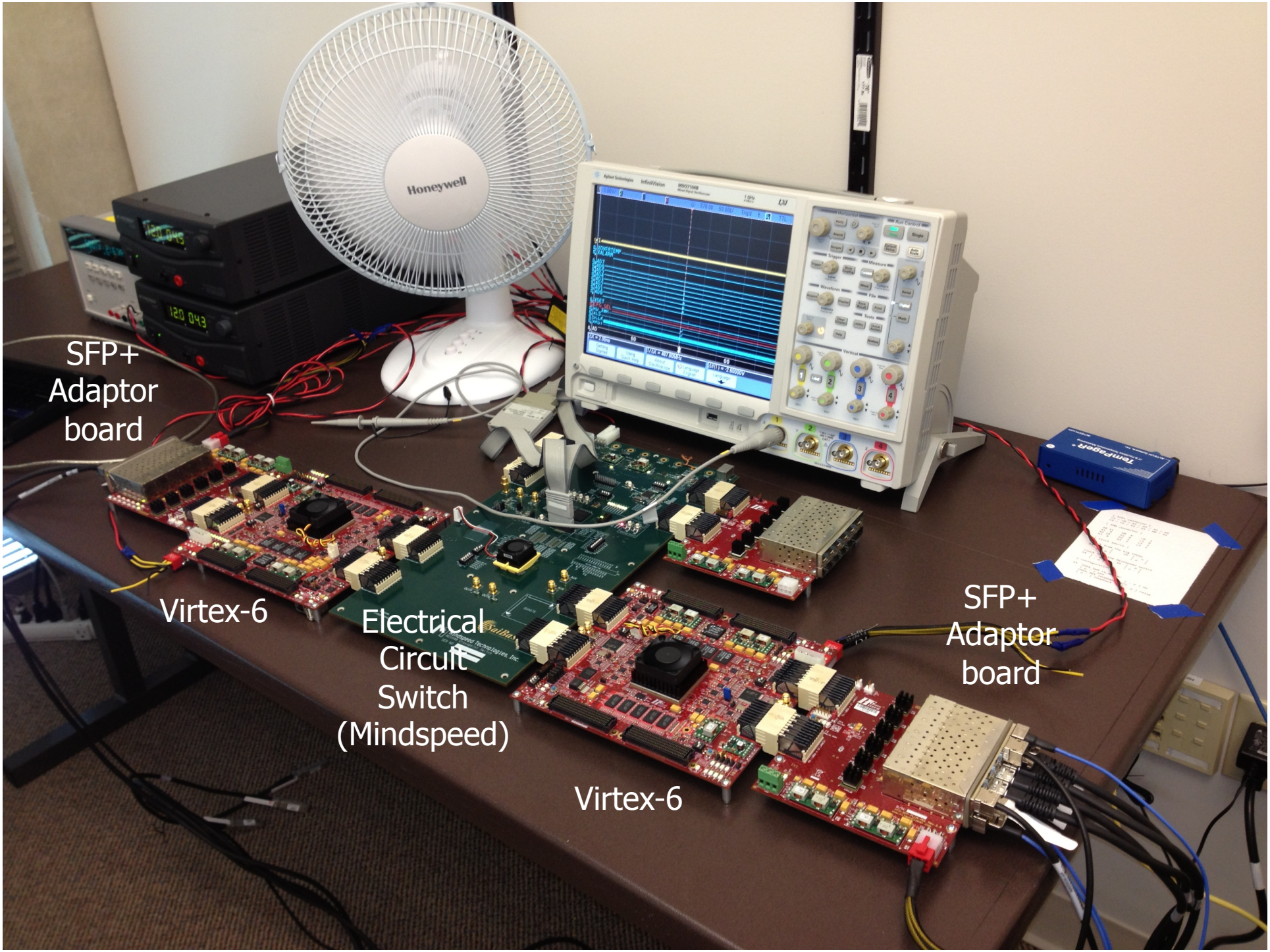


Circuit-friendly NIC – More Queues



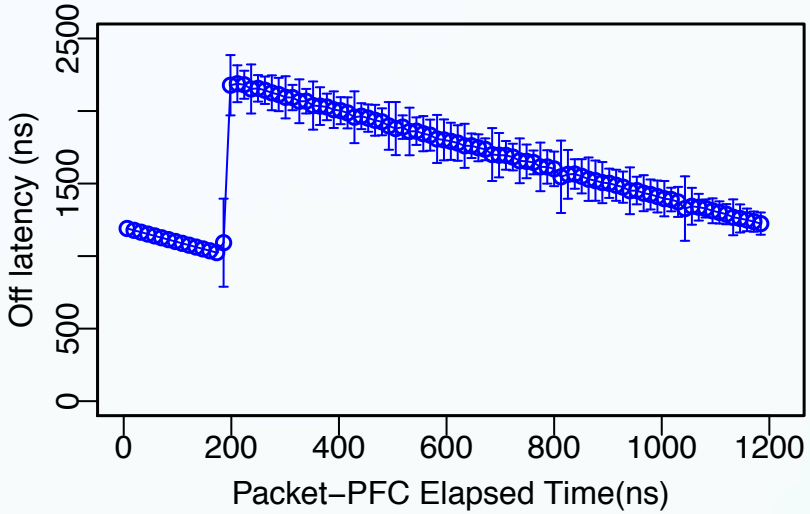
Testbed Components

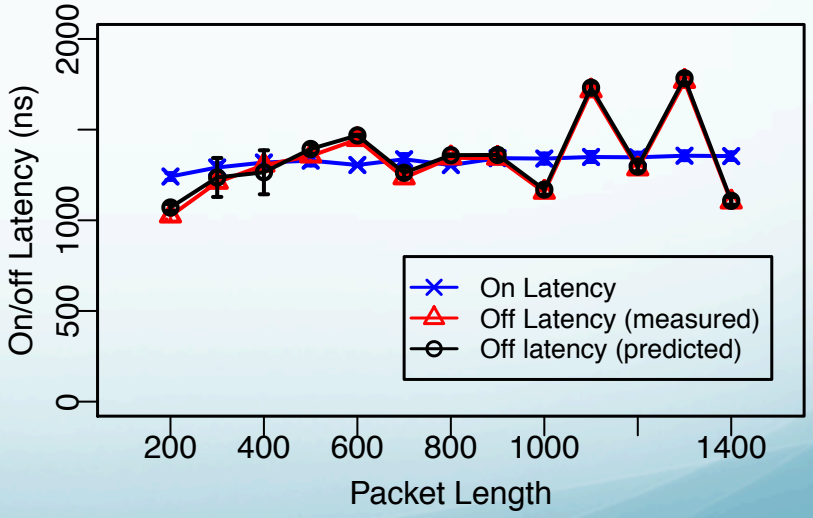



SFP+
Adaptor
board
Virtex-6
Electrical
Circuit
Switch
(Mindspeed)
Virtex-6
SFP+
Adaptor
board

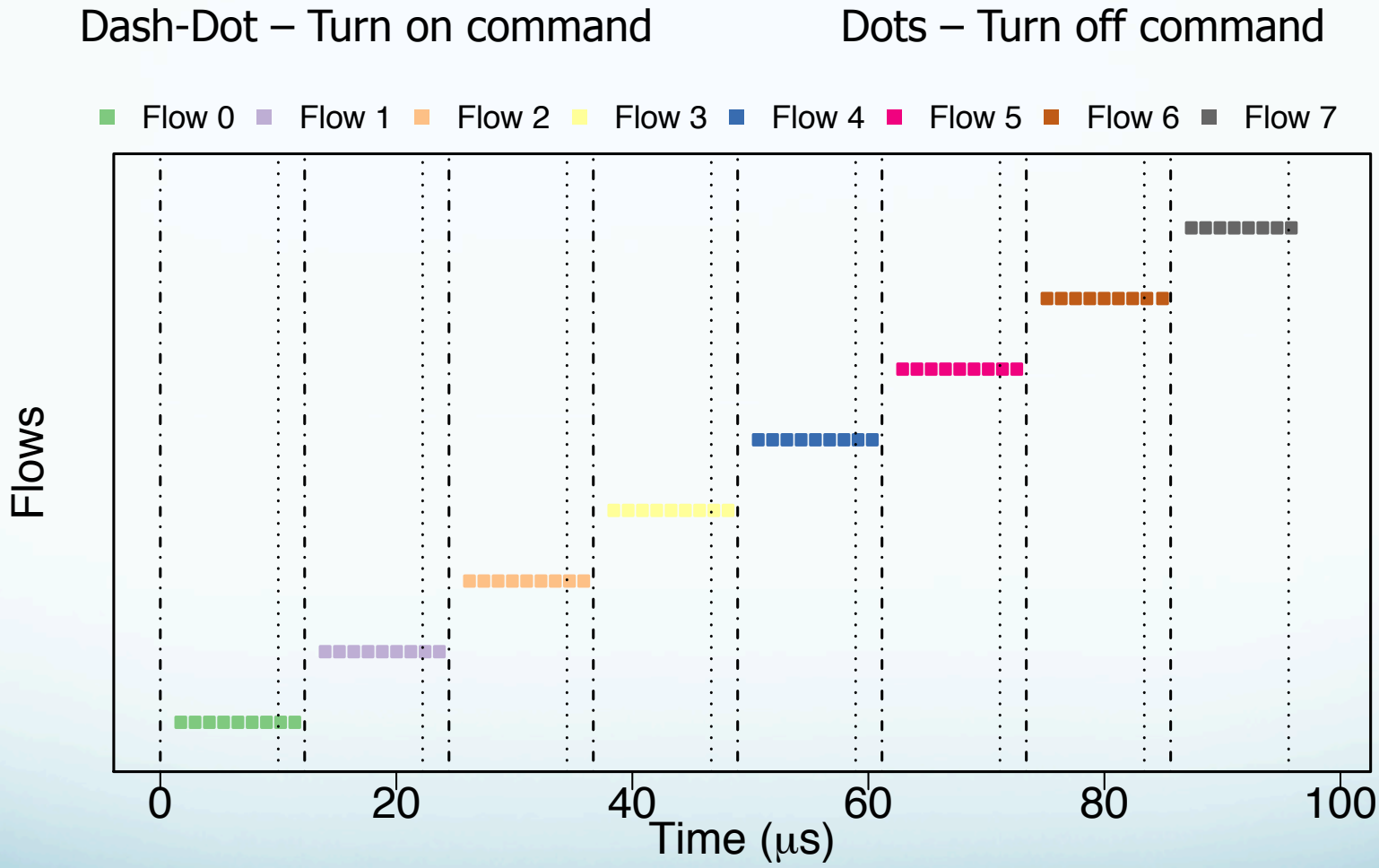
Pausing and synchronization

- Relying on 802.1Qbb
  - aka Priority Flow Control
- Eight endhost queues
  - Maintained by O/S
  - Non-realtime
- Queues "paused" by control packets from Hybrid TOR
  - Maintained by NIC
  - Real-time





Pausing and synchronization



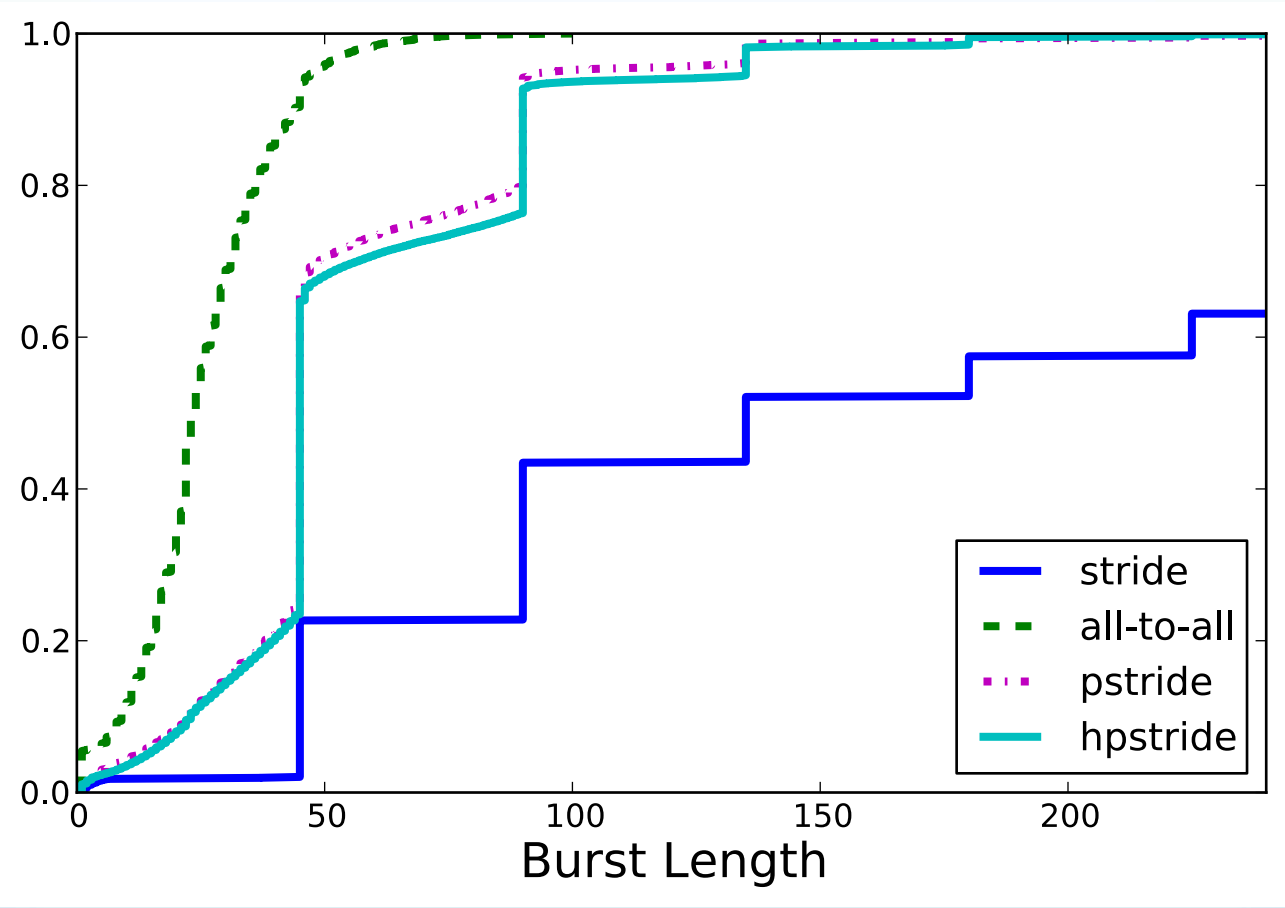
Outline

- Motivation and Background
  - Scale-out datacenters
  - Distributed vs. Centralized Network Control
- Research Issues
  - Using Circuits in a Packet-based Environment
  - Burstiness of Traffic
  - Scheduling
  - Optical Circuits
- Discussion

Burstiness of Traffic

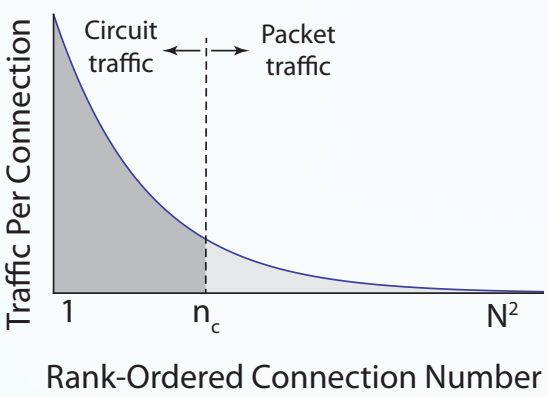
- How fast does circuit switch need to be?
  - Try to match to the burstiness of traffic
- What determines burstiness?
  - Where the circuit switch is deployed (ToR vs. core)
  - Application dependent
    - All-to-all vs. highly coherent (traffic matrix coherence)
  - Also the O/S, TCP, the NIC (e.g., TCP offloading)

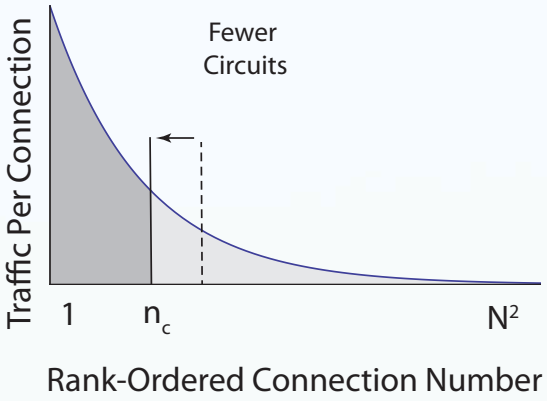
How fast is fast enough?

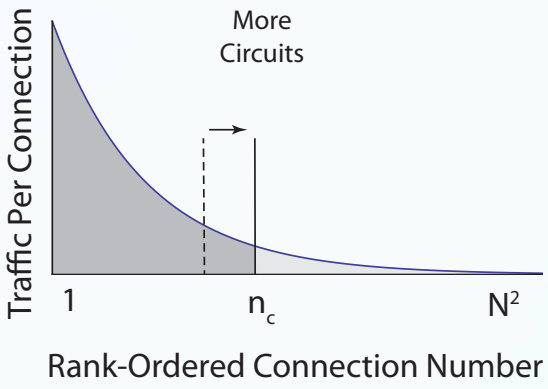


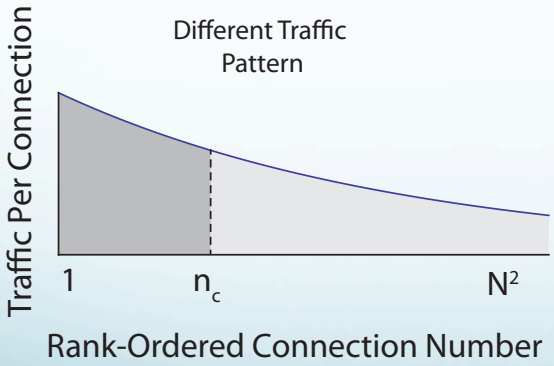
Burst behavior of Intel 82599 NIC with workloads taken from [Helios,Sigcomm10]

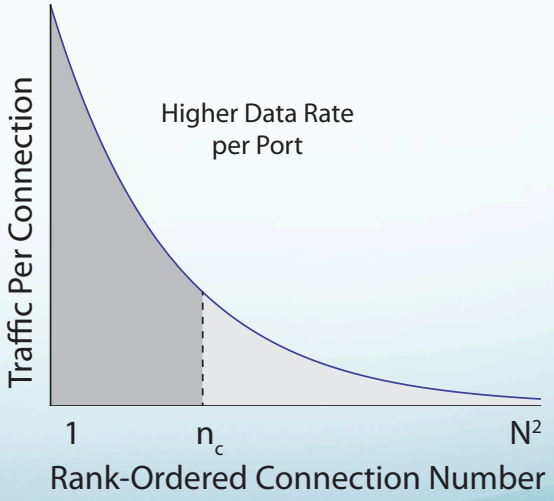
Traffic matrix coherence







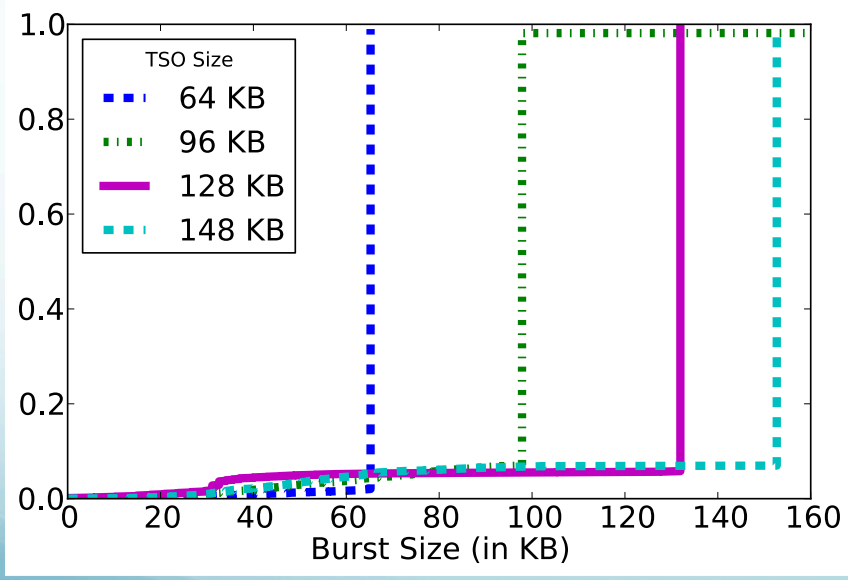




Traffic conditioning

100us circuit = 75 MTU-sized frames per circuit

- Natural correlation
  - Macro: Multi-packet data objects
  - Micro: TCP segmentation offloading (TSO)
- Induced correlation
  - Example: sort
  - Coordinate shuffle phase to create skew at small timescales



Bullet Trains: A Study of NIC Burst Behavior at Microsecond Timescales

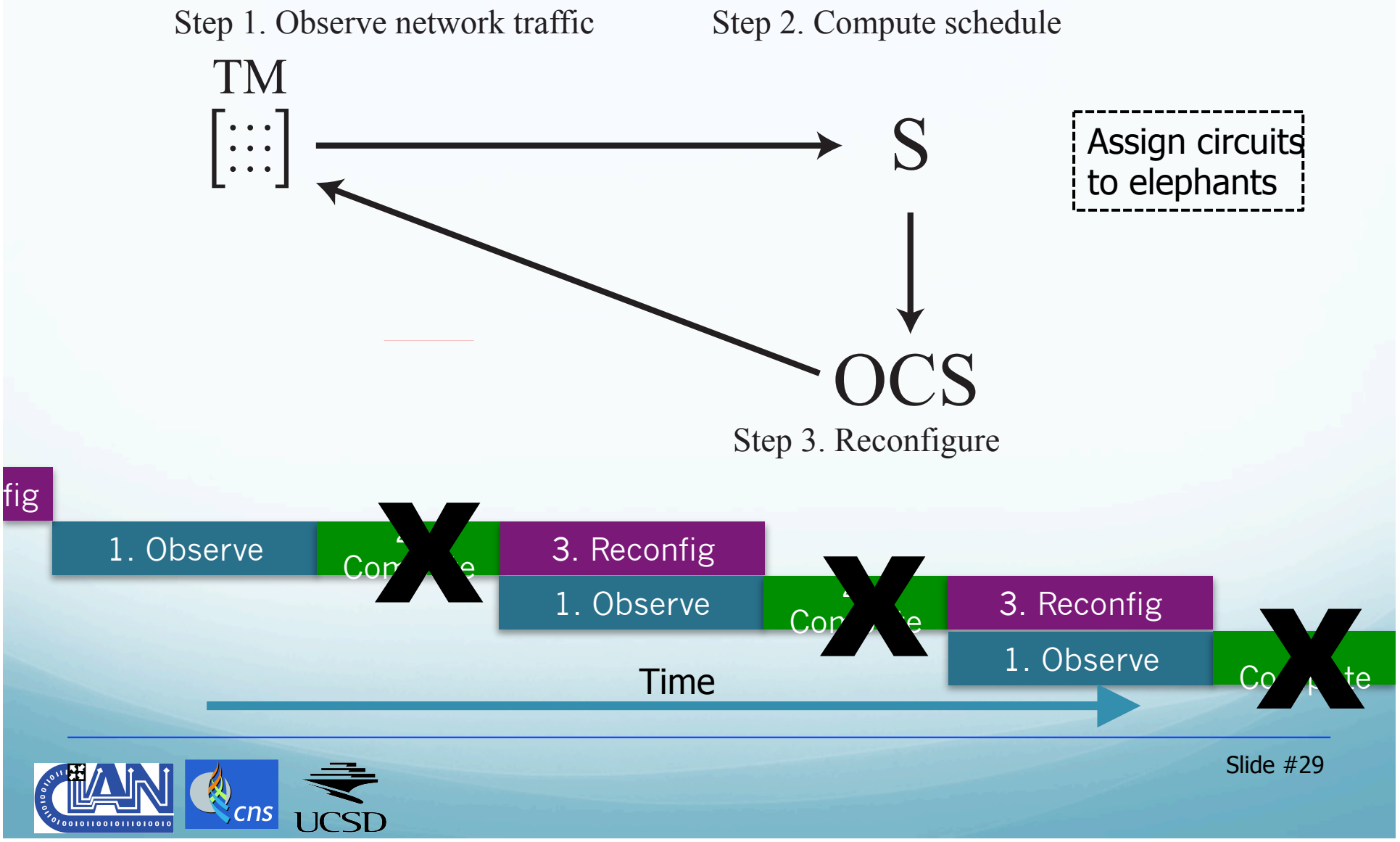
Rishi Kapoor, Alex C. Snoeren, Geoffrey M. Voelker, George Porter

(In submission to ACM CoNEXT'13)

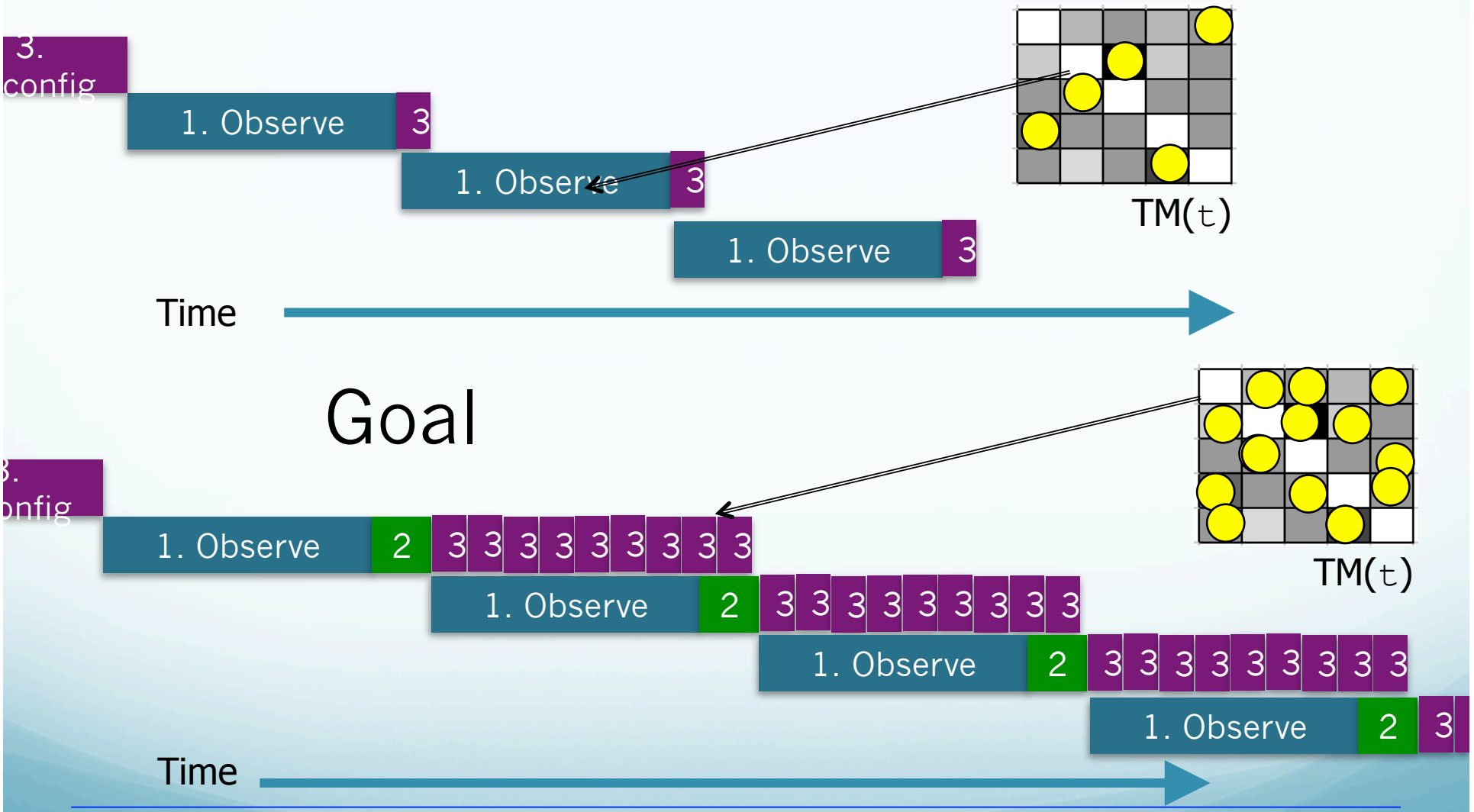
Outline

- Motivation and Background
  - Scale-out datacenters
  - Distributed vs. Centralized Network Control
- Research Issues
  - Using Circuits in a Packet-based Environment
  - Persistence of Traffic
  - Scheduling
  - Hardware for Optical Circuits
- Discussion

Previous approaches: Hotspot Scheduling



Limitations of Hotspot Scheduling



Choosing a schedule

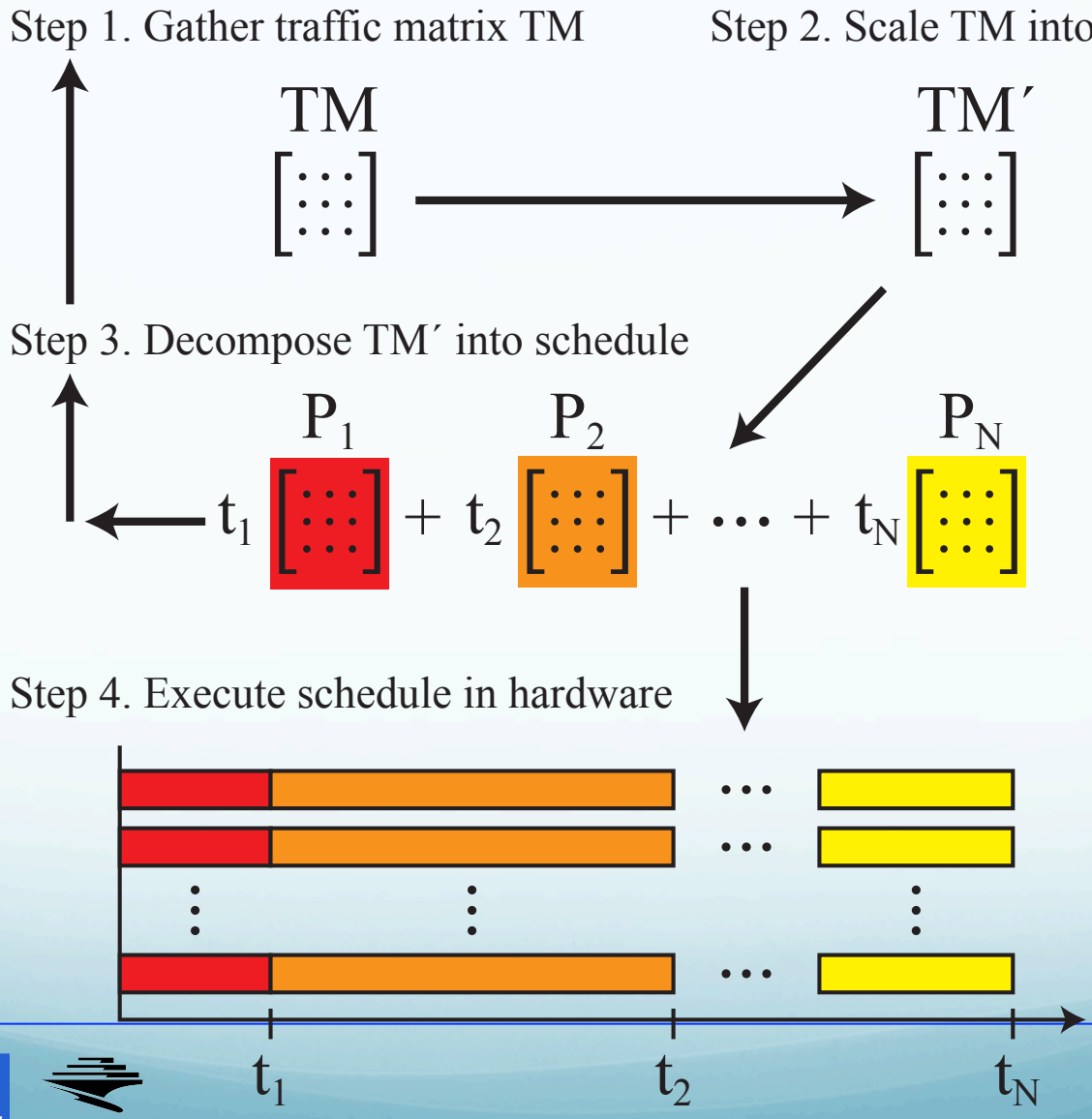
1) TM $\left[\begin{smallmatrix}\cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\end{smallmatrix}\right]$ → Maximal matching ( ) → P_i

P_i is a permutation matrix:

	1	2	3	4
1	0	1	0	0
2	0	0	0	1
3	1	0	0	0
4	0	0	1	0

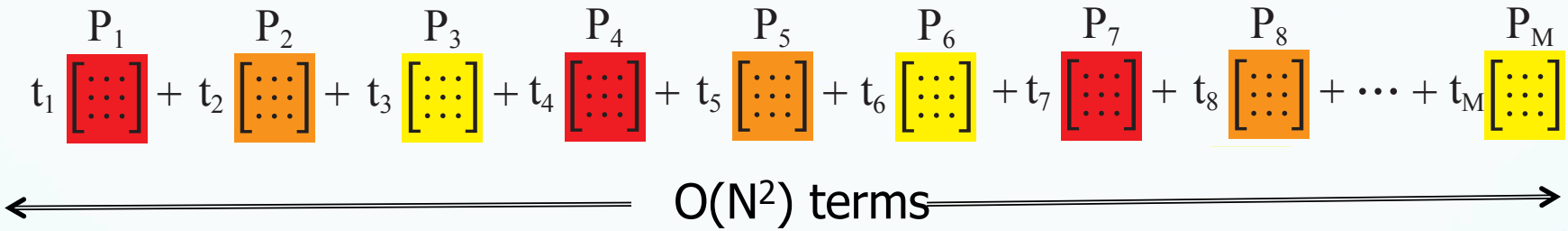
2) TM $\left[\begin{smallmatrix}\cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\end{smallmatrix}\right]$ → Matrix decomposition ( ) → $TM = \sum_{i}^{N} t_i P_i$

Traffic Matrix Scheduling

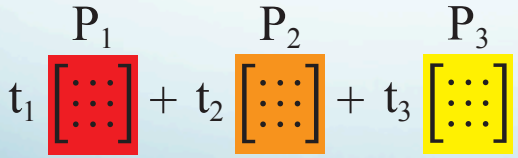


Evaluating Schedules

- Perfect decomposition
  - E.g., Birkhoff-von Neumann: “equal”



- Longest-time slot first: “approximate”



Remainder routed using packet switch

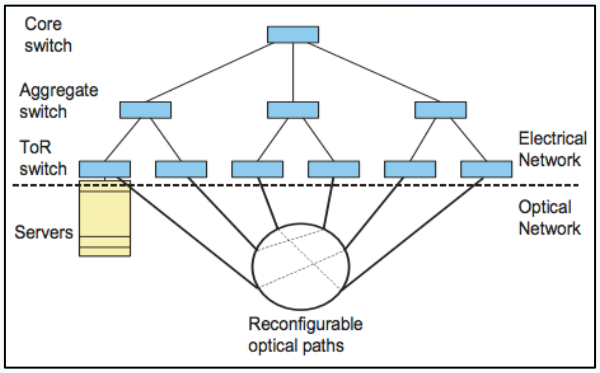
Scheduling Summary

- The faster the circuit switch, the more of the overall matrix can be scheduled using circuits.
- Schedule for circuit switch does not need to be perfect – rest is routed over packet switch.

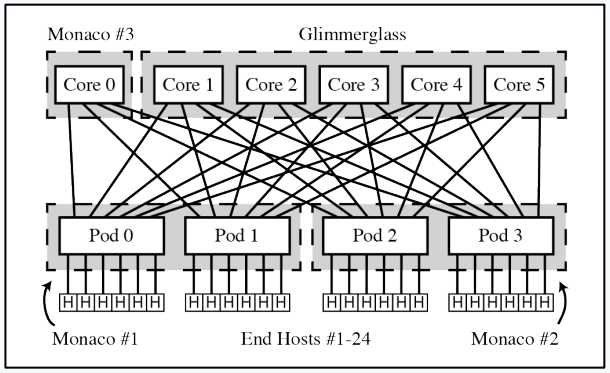
Outline

- Motivation and Background
  - Scale-out datacenters
  - Hybrid Networks
- Research Issues
  - Circuits in a Packet-based World
  - Burtiness of Traffic
  - Scheduling
  - Optical Circuit Switches
- Conclusions and Acknowledgements

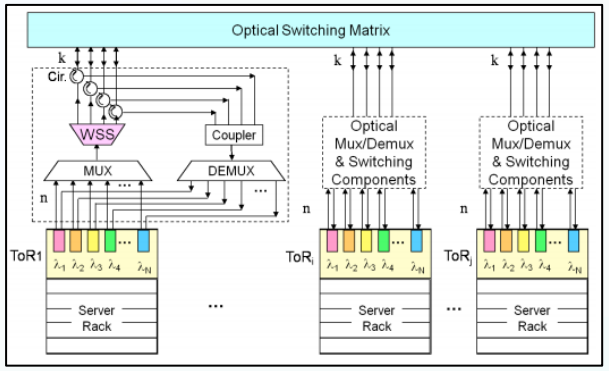
Circuit switching in data centers



c-Through
[SIGCOMM'10]



Helios
[SIGCOMM'10]

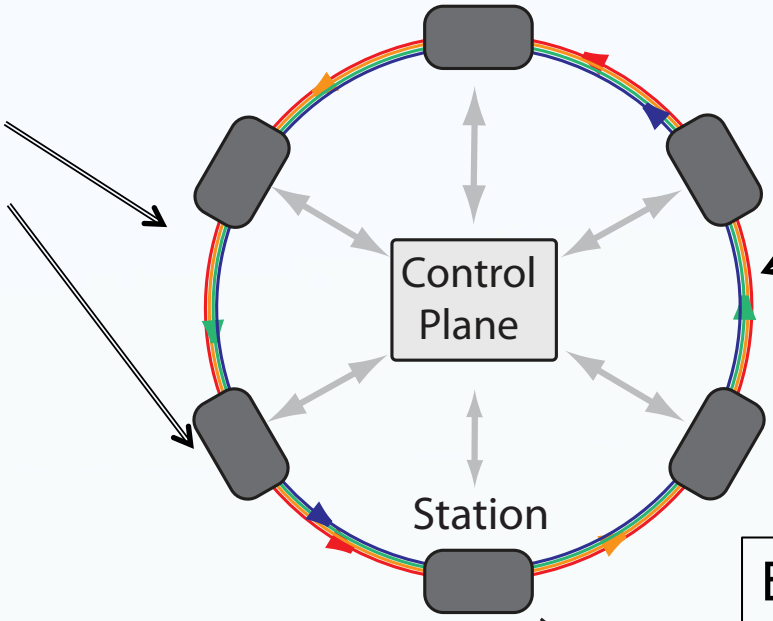


OSA
[NSDI'12]

- Speed of optical circuits (~ 10 of ms) means that can be used at aggregation level or higher
- May be sufficient for some kinds of networks such as Google's B4 network

Fast Optical Circuit Switch- Mordia

Each station:
1. Inserts signals into ring
2. "Tunes into" subset of signals



Data plane: Physical ring w/ 24 DWDM signals

Each device transmits on its own frequency

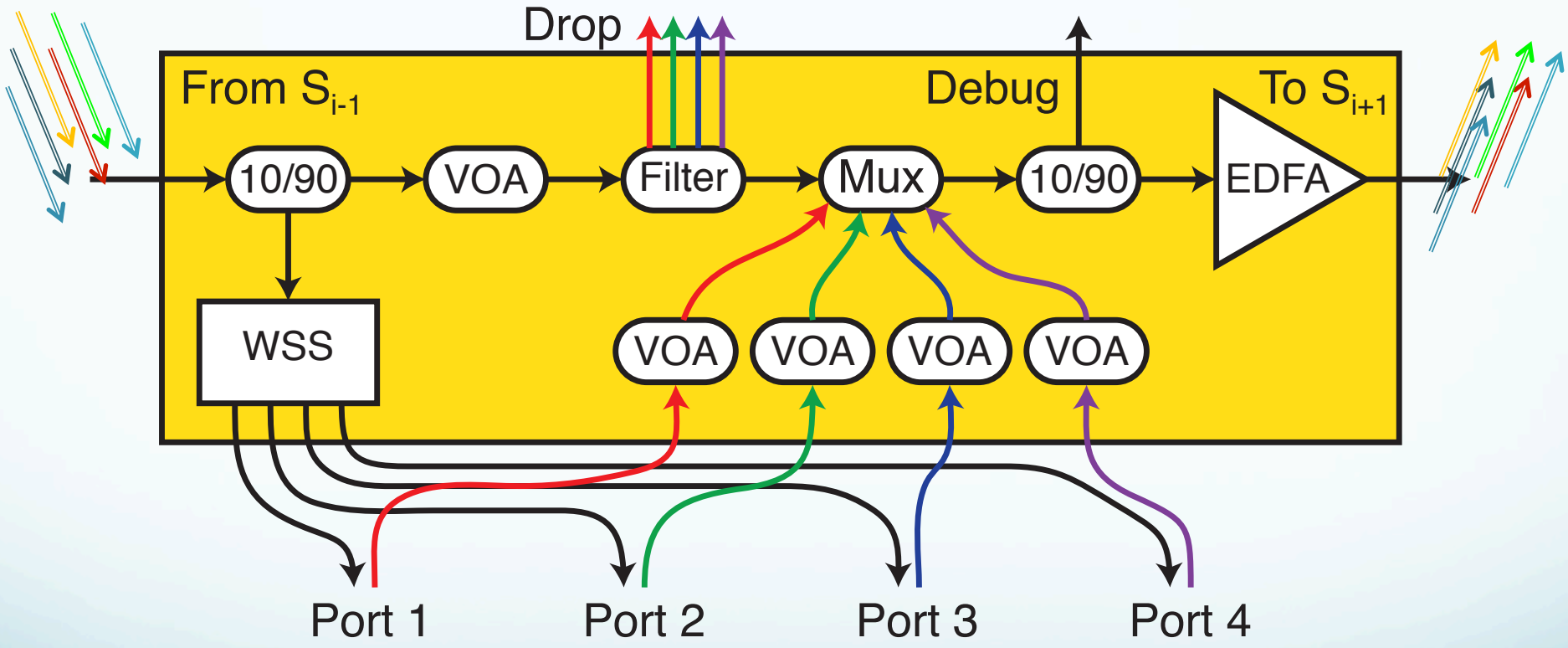
Top of Rack switch

OR

Hosts

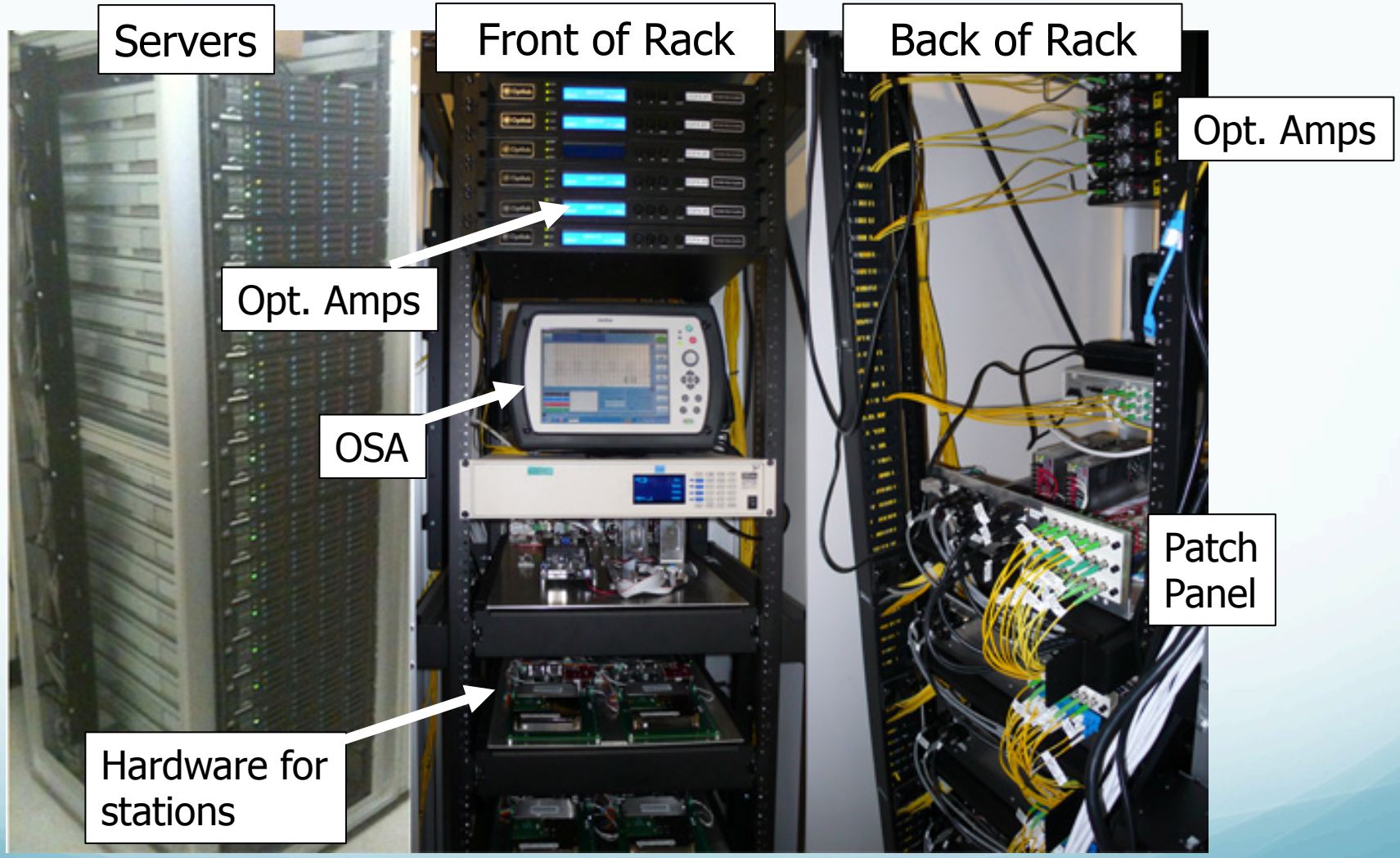
Photonics Technology Letters *Vol.* 25(16) 1589-1592, 2013

Mordia Station Architecture



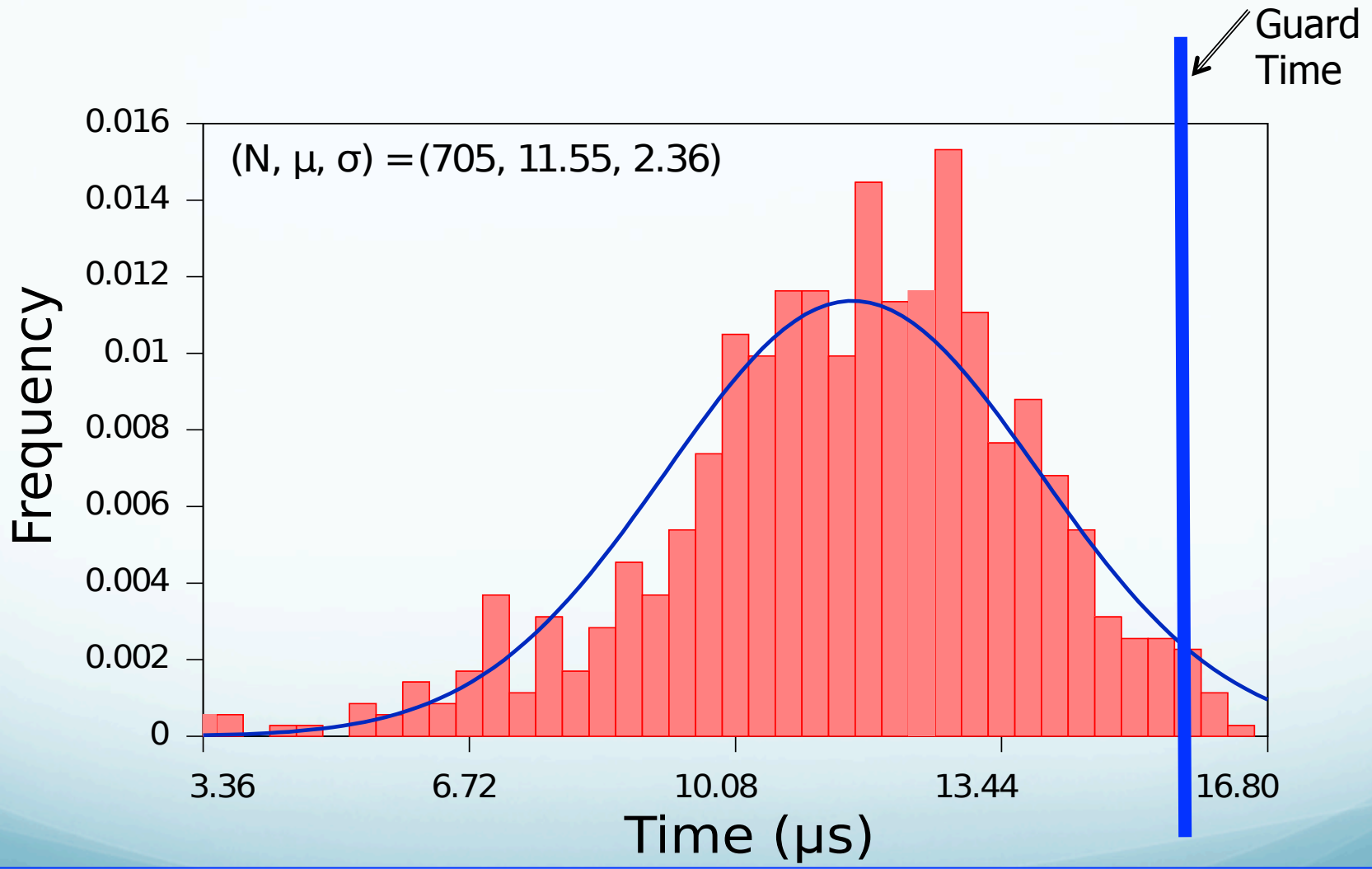
WSS – Wavelength selective switch VOA – variable optical attenuator
EDFA – Optical amplifier

Mordia Hardware



Measured 10G Data through WSS

End-to-end reconfiguration time



Conclusions

- Hybrid networking has potential to reduce cost, power, and complexity
- Complements current trends towards SDNs
- Where in the network it is deployed depends on the persistence of the traffic matrix and the applications
- Deployment at the ToR level is an open research topic.

Thank you!

Acknowledgements

- This work has been supported by the following sources:
  - NSF Engineering Research Center for Integrated Access Networks (CIAN)
  - Center for Networked Systems (CNS)
  - Google
  - Cisco
  - Corning
  - Mindspeed