Alternative Switching Technologies: Optical Circuit Switches

Hakim Weatherspoon

Assistant Professor, Dept of Computer Science

CS 5413: High Performance Systems and Networking

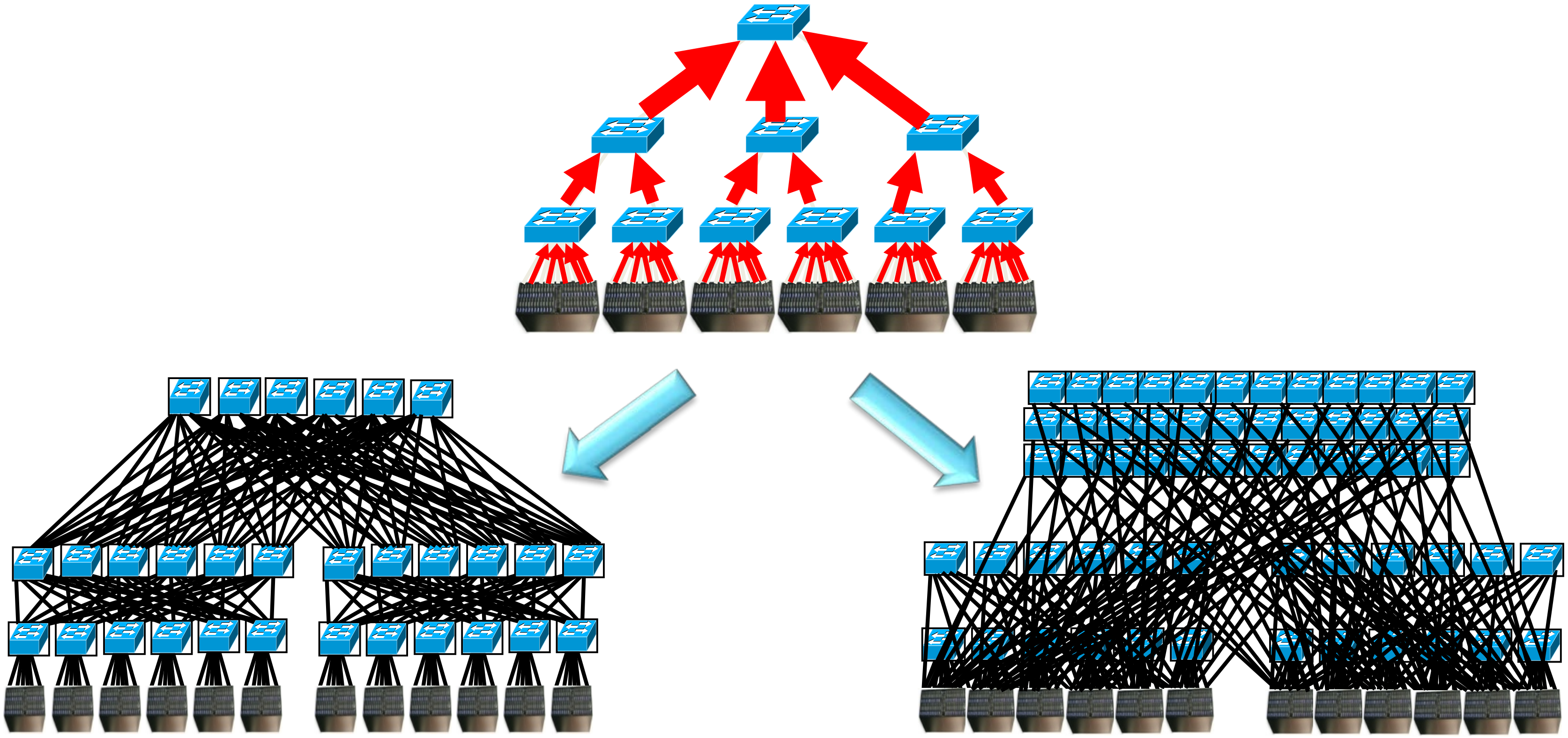
October 22, 2014

Slides from the “On the Feasibility of Completely Wireless Datacenters” at the ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), October 2012.

Goals for Today

- On the Feasibility of Completely Wireless Datacenters
  - J. Y. Shin, E. G. Sirer, H. Weatherspoon, and D. Kirovski, *IEEE/ACM Transactions on Networking (ToN)*, Volume 21, Issue 5 (October 2013), pages 1666-1680.

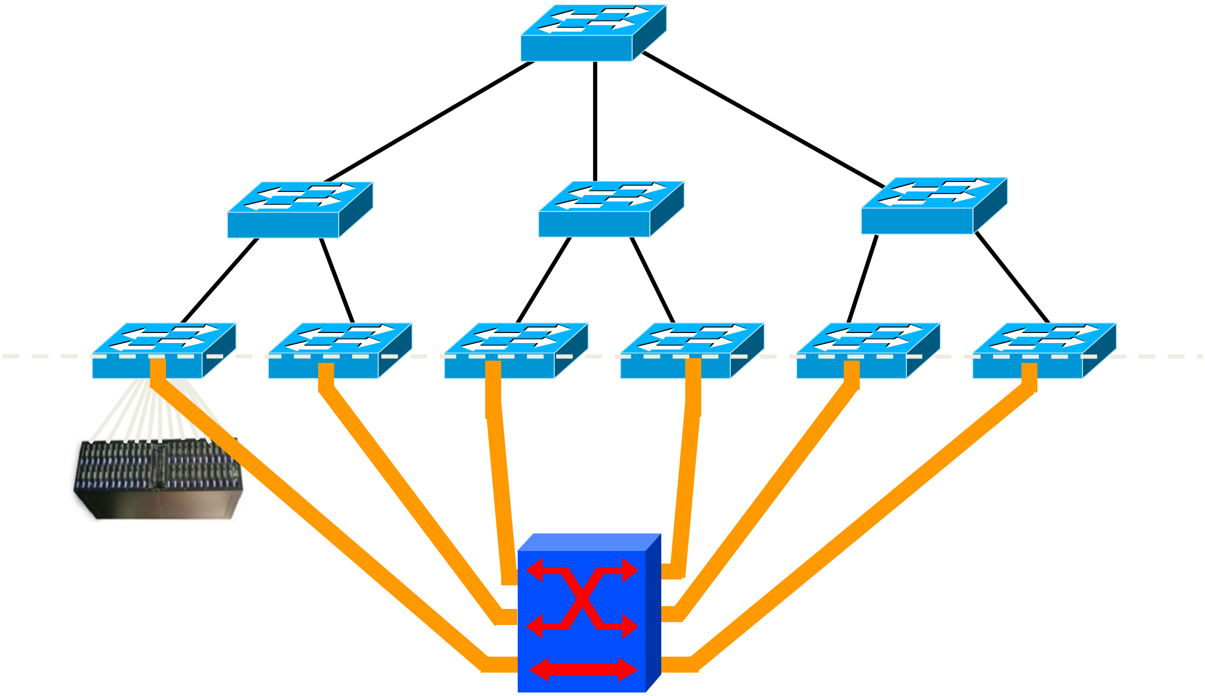
Current solutions for increasing data center network bandwidth



1. Hard to construct

2. Hard to expand

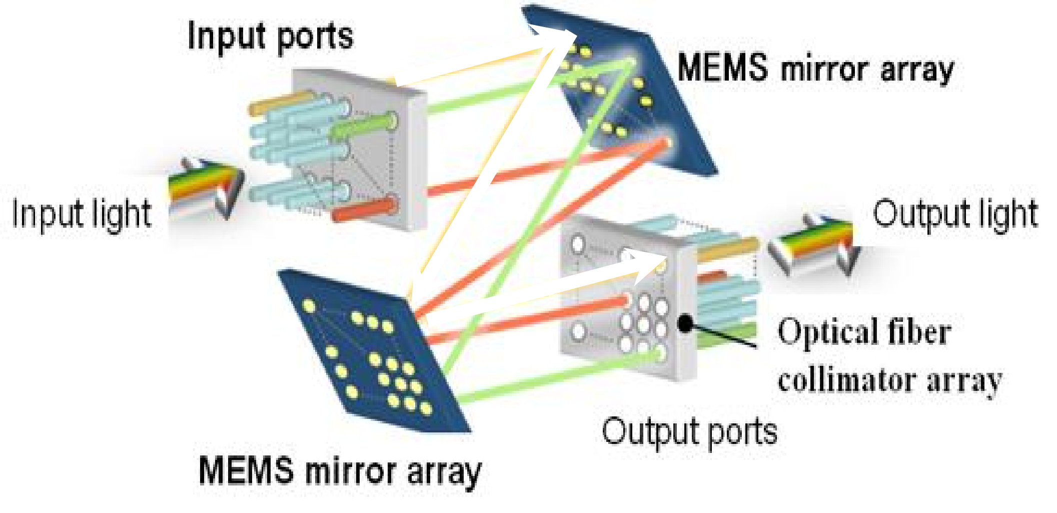
An alternative: hybrid packet/circuit switched data center network



- Goal of this work:
  - Feasibility: software design that enables efficient use of optical circuits
  - Applicability: application performance over a hybrid network

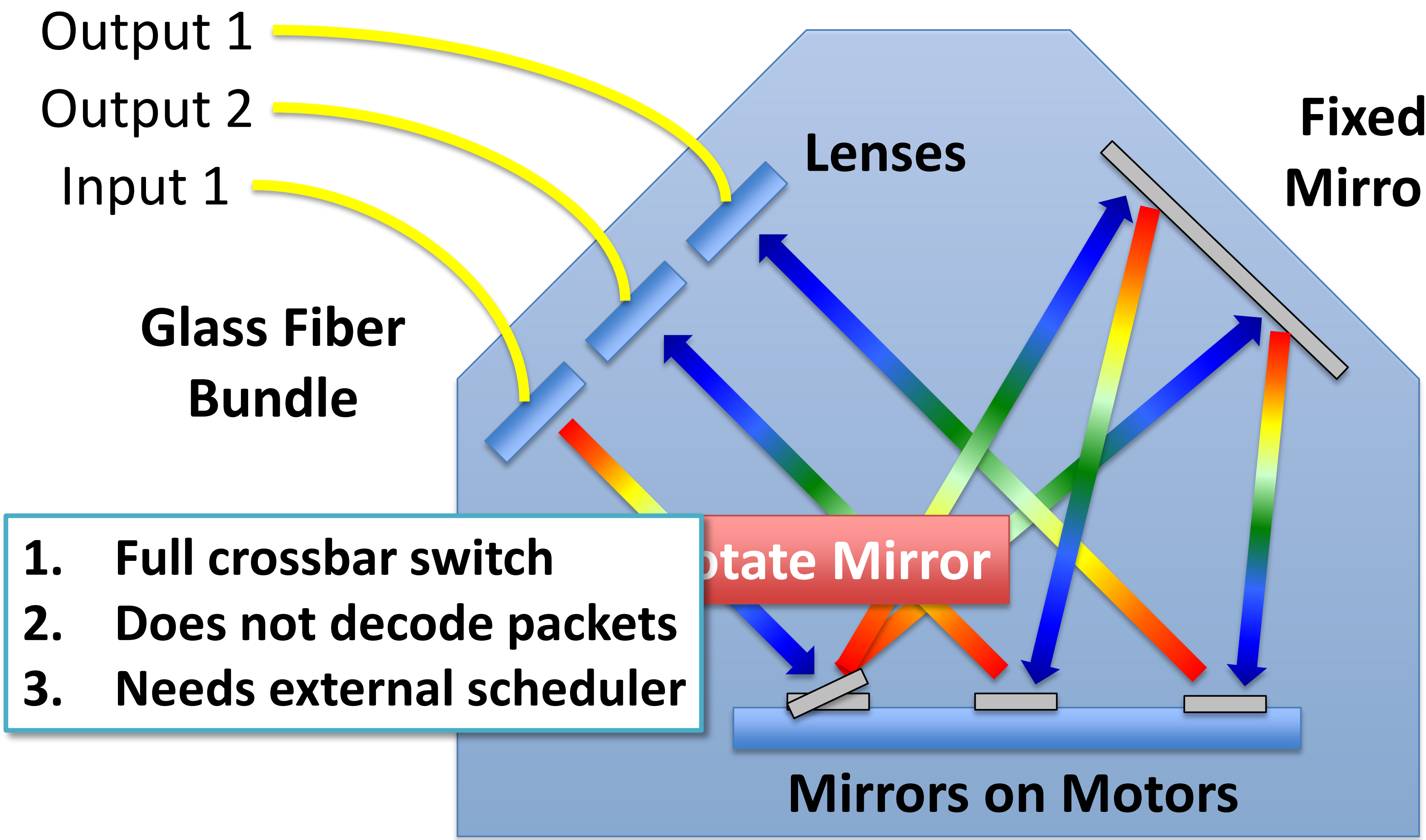
Optical circuit switching v.s. Electrical packet switching

	Electrical packet switching	Optical circuit switching
Switching technology	Store and forward	Circuit switching
Switching capacity	1 e	et, ect
Switching time	F	



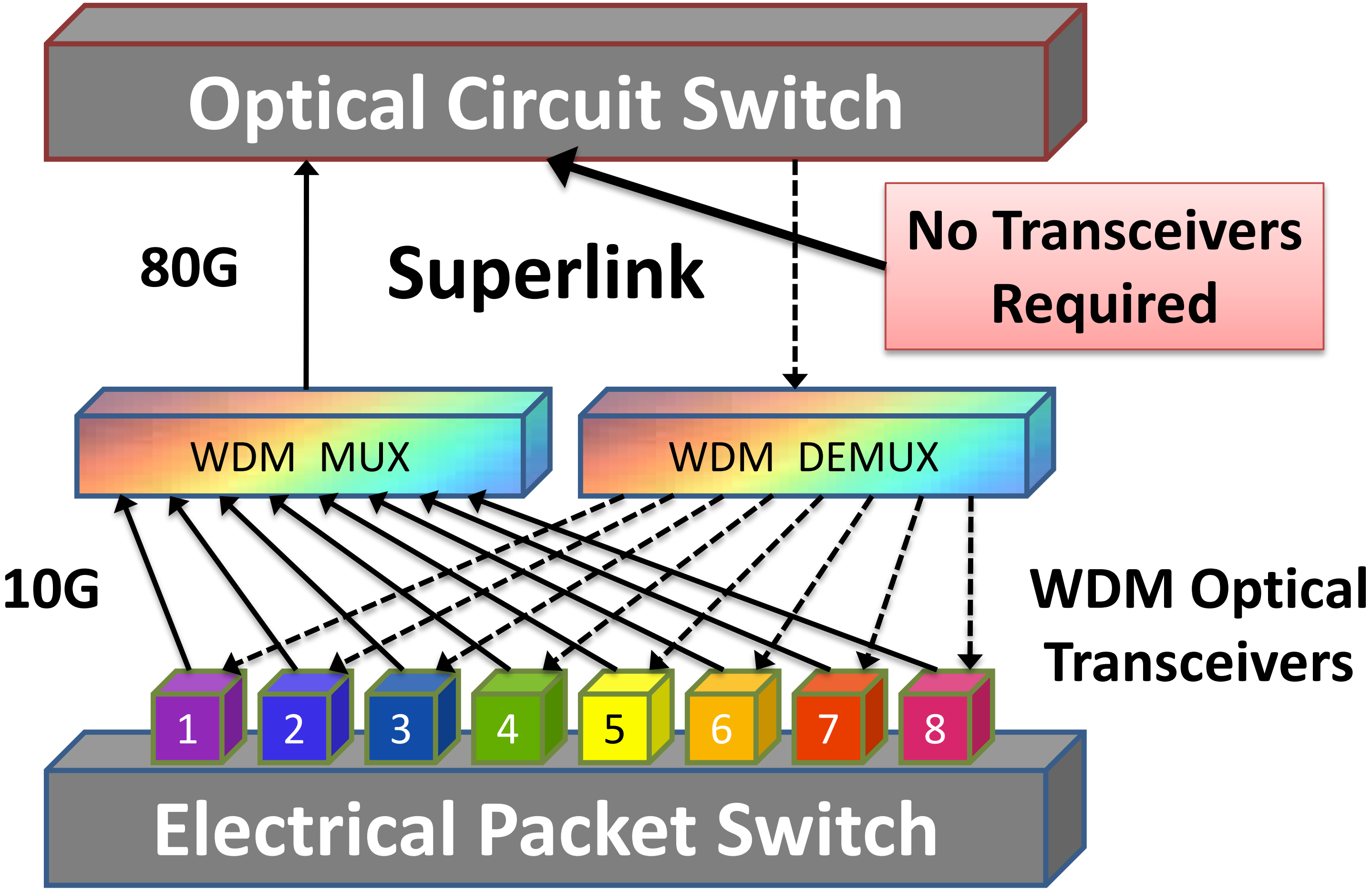
e.g. MEMS optical switch

Technology: Optical Circuit Switch



1. **Full crossbar switch**
2. **Does not decode packets**
3. **Needs external scheduler**

Wavelength Division Multiplexing



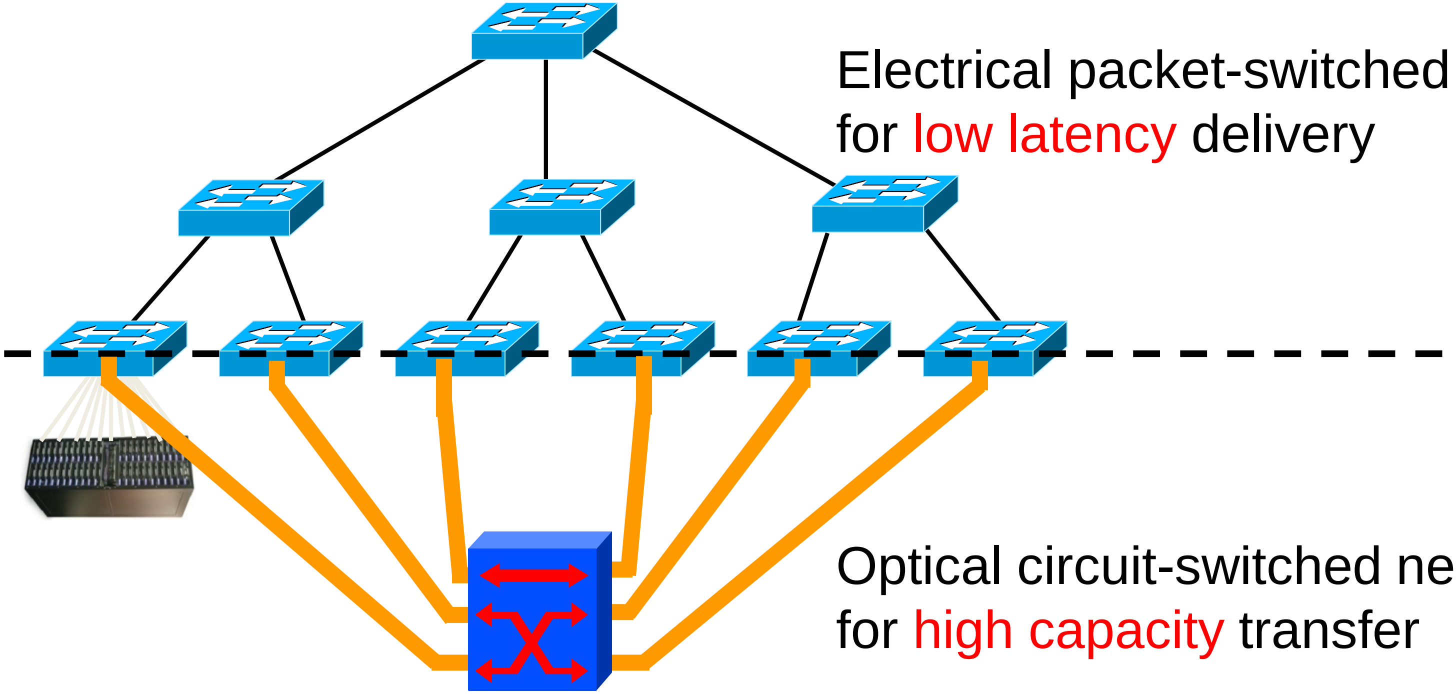
Optical circuit switching is promising despite slow switching time

- [IMC09][HotNets09]: “*Only a few ToRs are hot and most their traffic goes to a few other ToRs.* …”

- [WREN09]: “*…we find that traffic at the five edge switches exhibit an ON/OFF pattern…* ”

Full bisection bandwidth at packet granularity
may not be necessary

Hybrid packet/circuit switched network architecture

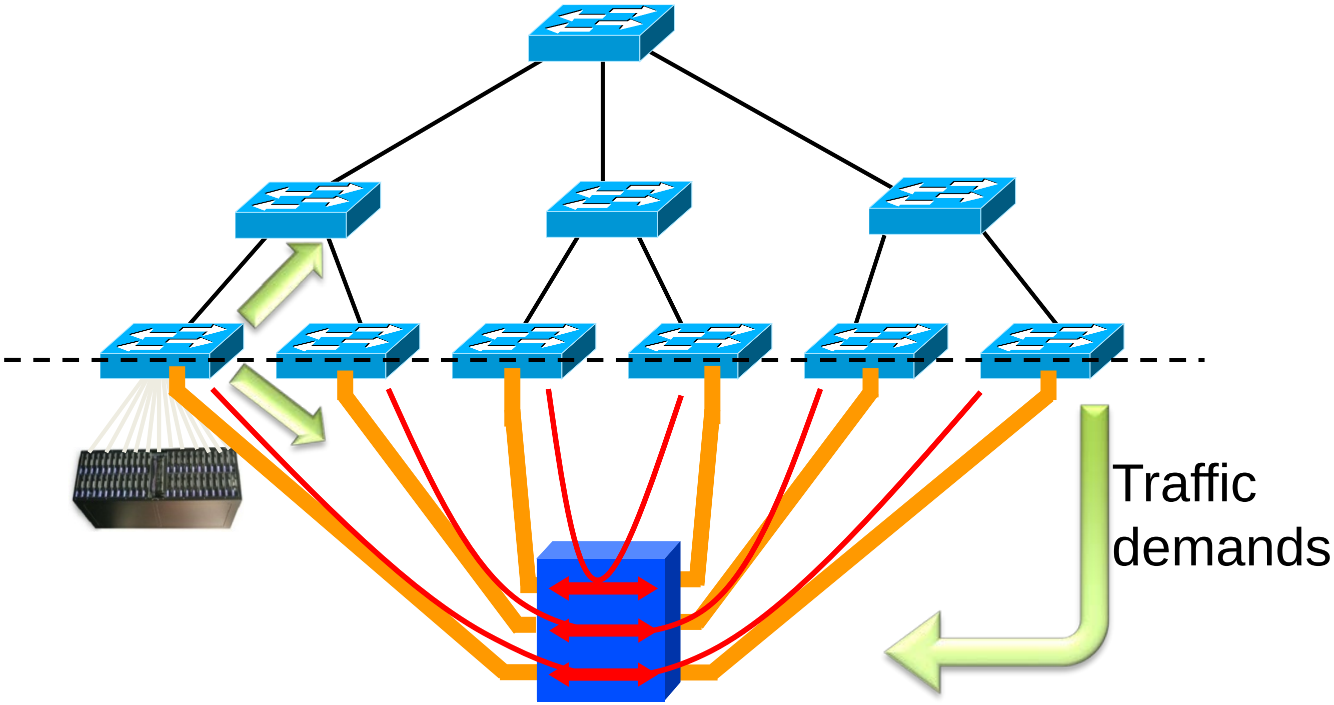


Electrical packet-switched network for low latency delivery

Optical circuit-switched network for high capacity transfer

- Optical paths are provisioned rack-to-rack
  - A simple and cost-effective choice
  - Aggregate traffic on per-rack basis to better utilize optical circuits

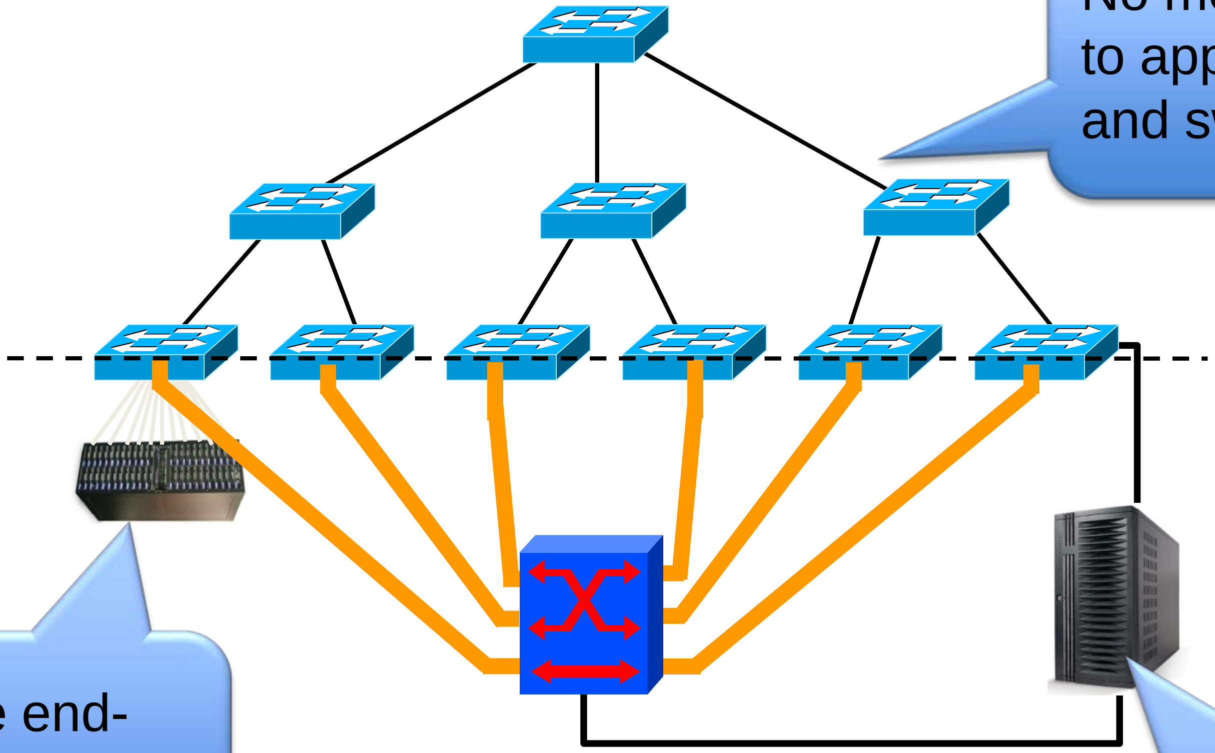
Design requirements



- Control plane:
  - Traffic demand estimation
  - Optical circuit configuration

- Data plane:
  - Dynamic traffic de-multiplexing
  - Optimizing circuit utilization (optional)

c-Through (a specific design)

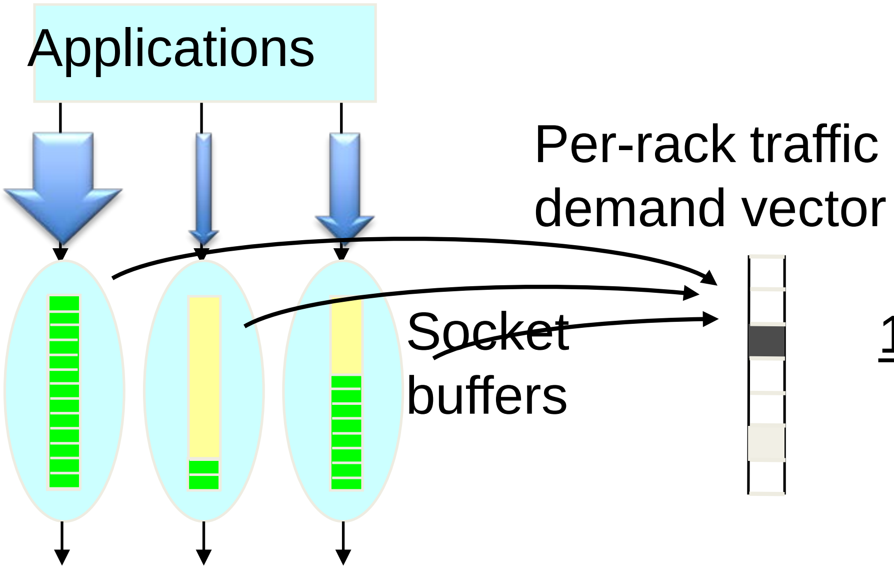


No modification to applications and switches

Leverage end-hosts for traffic management

Centralized control for circuit configuration

c-Through - traffic demand estimation and traffic batching



Applications

Per-rack traffic demand vector

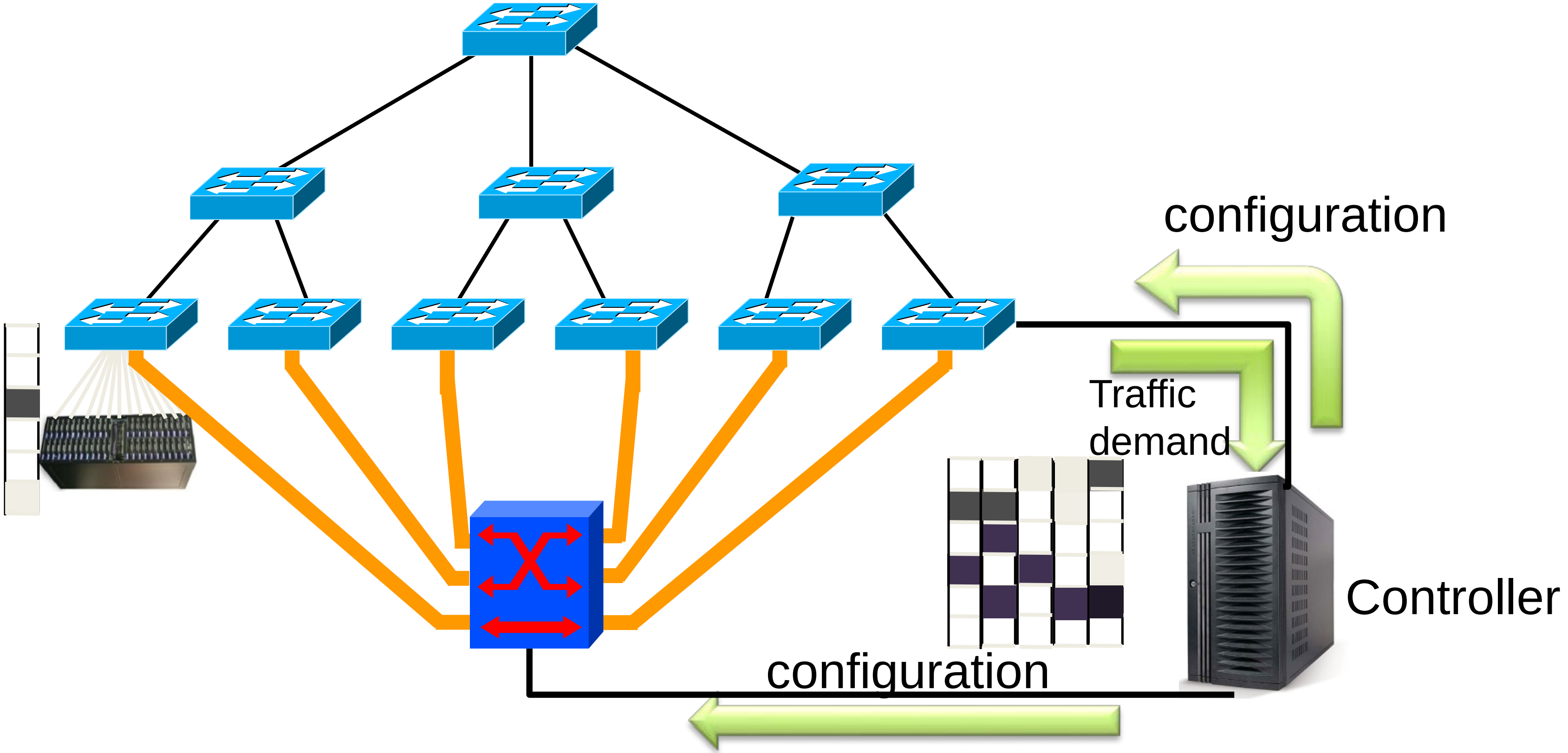
Socket buffers

1. Transparent to applications.

2. Packets are buffered per-flow to avoid HOL blocking.

- Accomplish two requirements:
  - Traffic demand estimation
  - Pre-batch data to improve optical circuit utilization

c-Through - optical circuit configuration

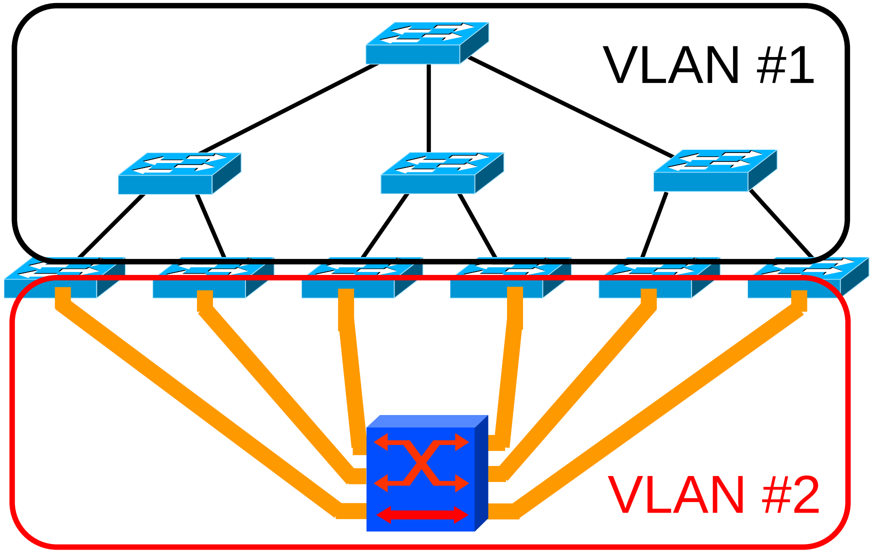


Use Edmonds' algorithm to compute optimal configuration

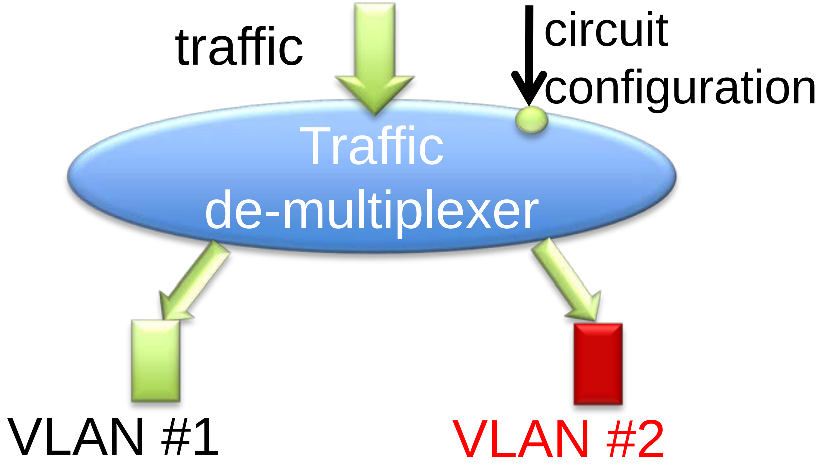
Many ways to reduce the control traffic overhead

c-Through - traffic de-multiplexing

- VLAN-based network isolation:
  - No need to modify switches
  - Avoid the instability caused by circuit reconfiguration

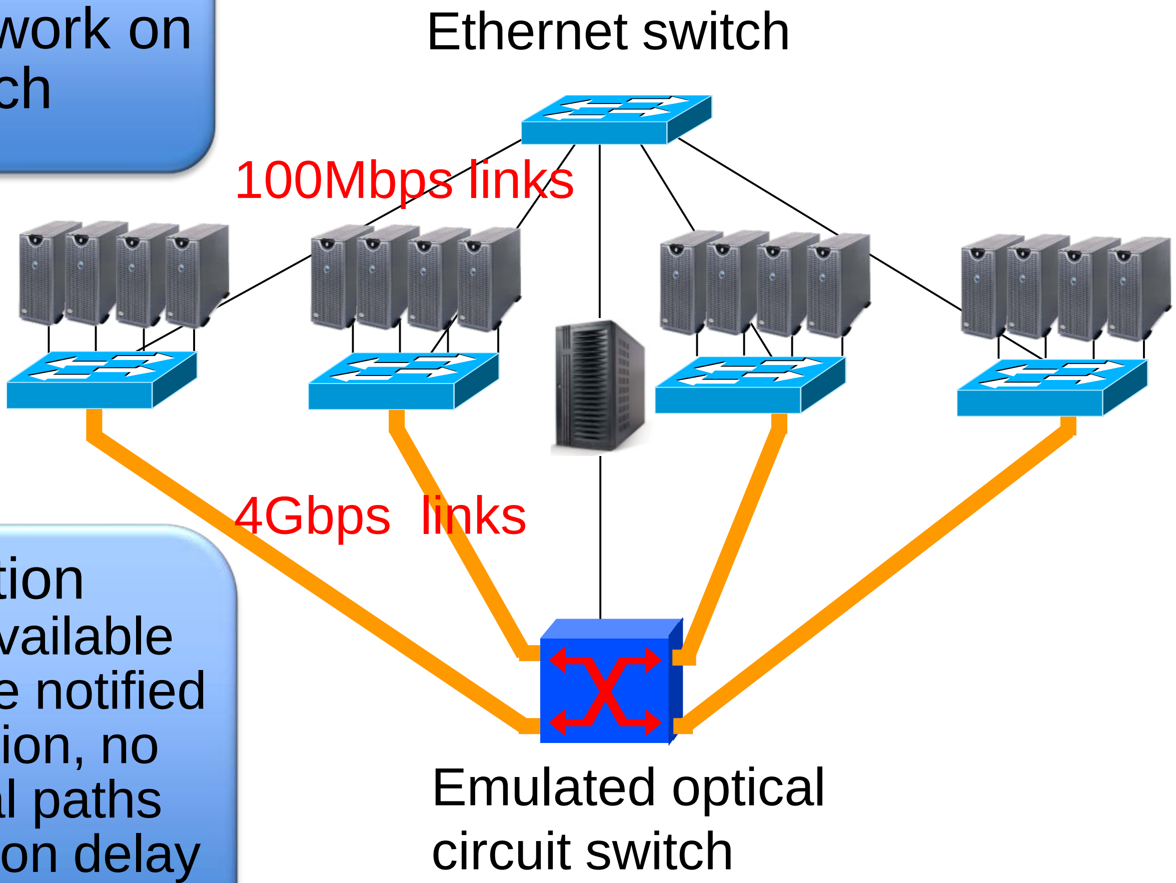


- Traffic control on hosts:
  - Controller informs hosts about the circuit configuration
  - End-hosts tag packets accordingly



Testbed setup

- 16 servers with 1Gbps NICs
- Emulate a hybrid network on 48-port Ethernet switch

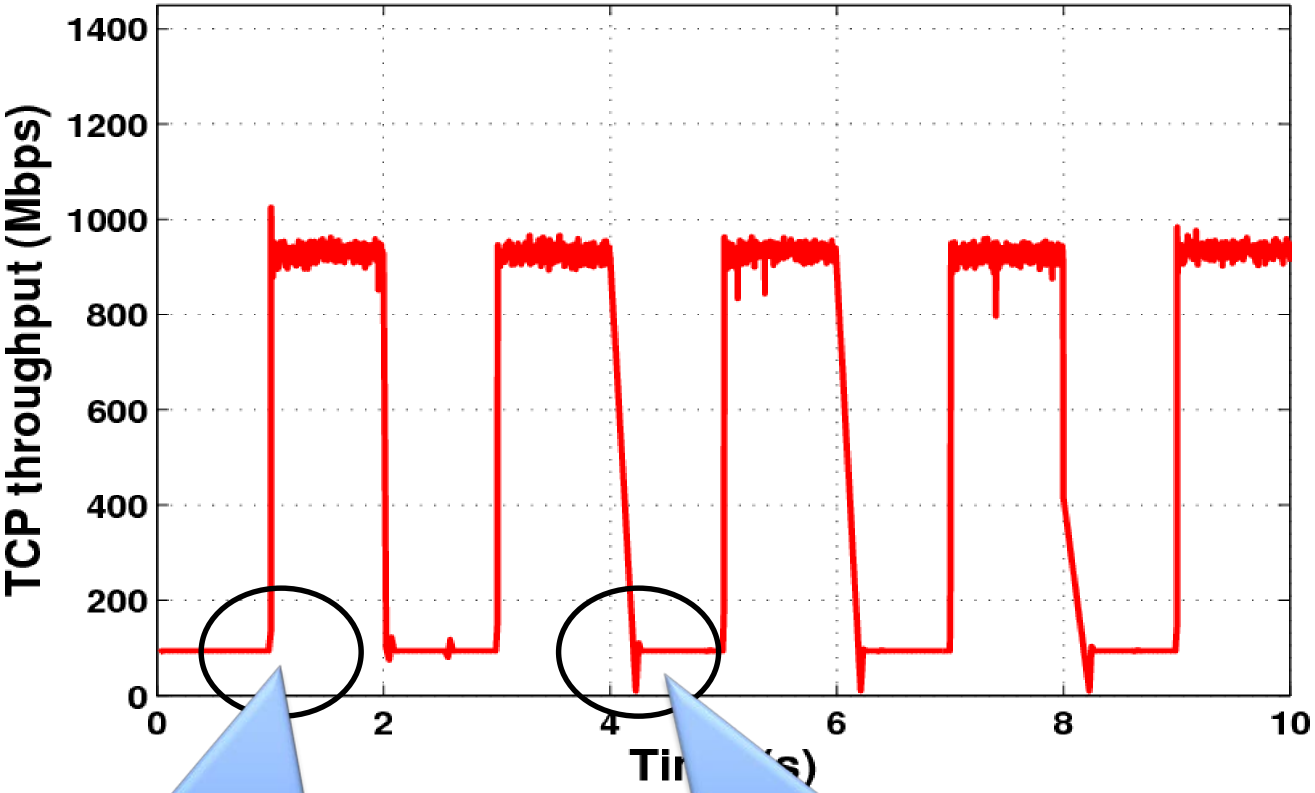


- Optical circuit emulation
  - Optical paths are available only when hosts are notified
  - During reconfiguration, no host can use optical paths
  - 10 ms reconfiguration delay

Evaluation

- Basic system performance:
  - Can TCP exploit dynamic bandwidth quickly?
  - Does traffic control on servers bring significant overhead?
  - Does buffering unfairly increase delay of small flows?
- Application performance:
  - Bulk transfer (VM migration)?
  - Loosely synchronized all-to-all communication (MapReduce)?
  - Tightly synchronized all-to-all communication (MPI-FFT) ?

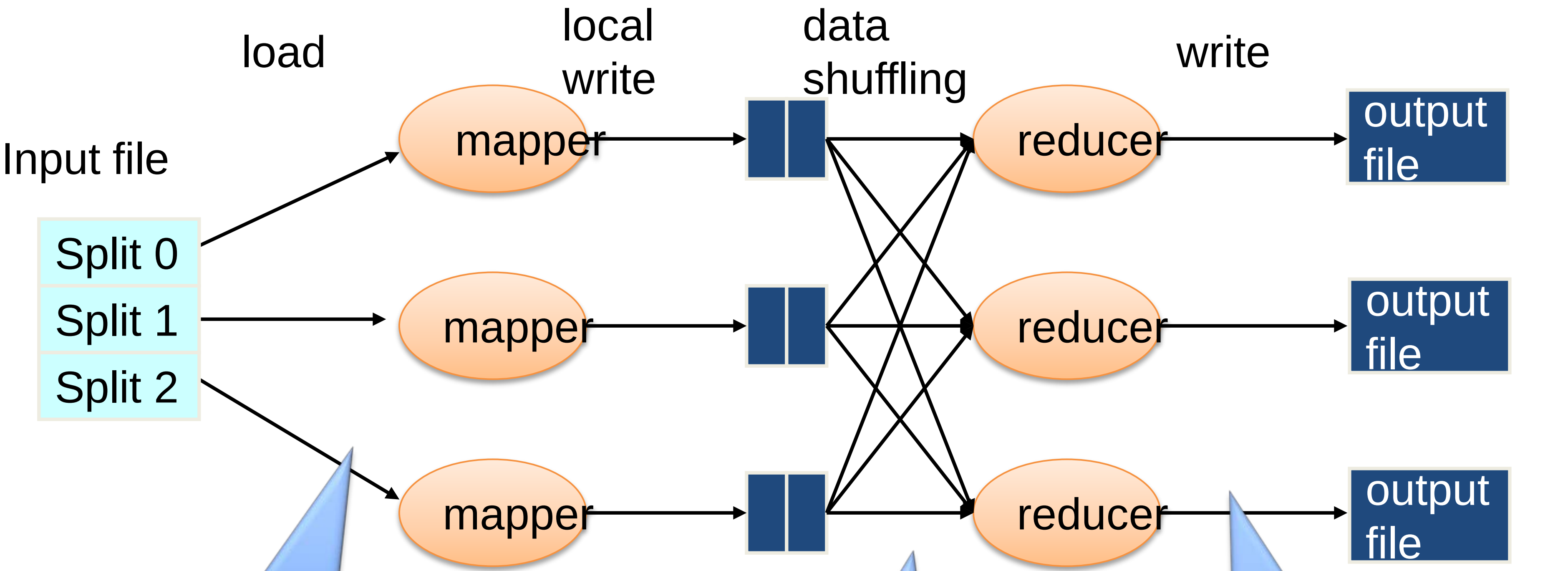
TCP can exploit dynamic bandwidth quickly



Throughput ramps up within 10 ms

Throughput stabilizes within 100ms

MapReduce Overview

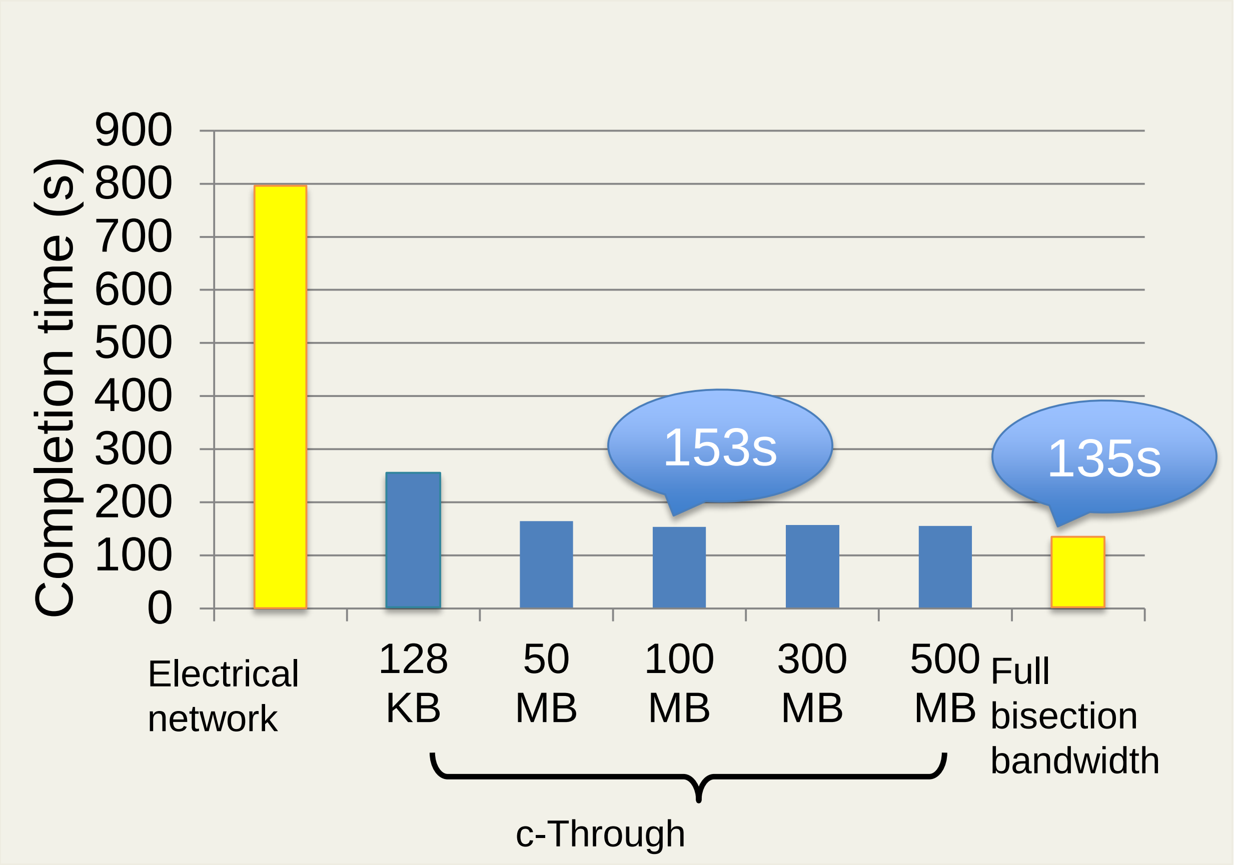


Concentrated traffic in 64MB blocks

Independent transfers: amenable to batching

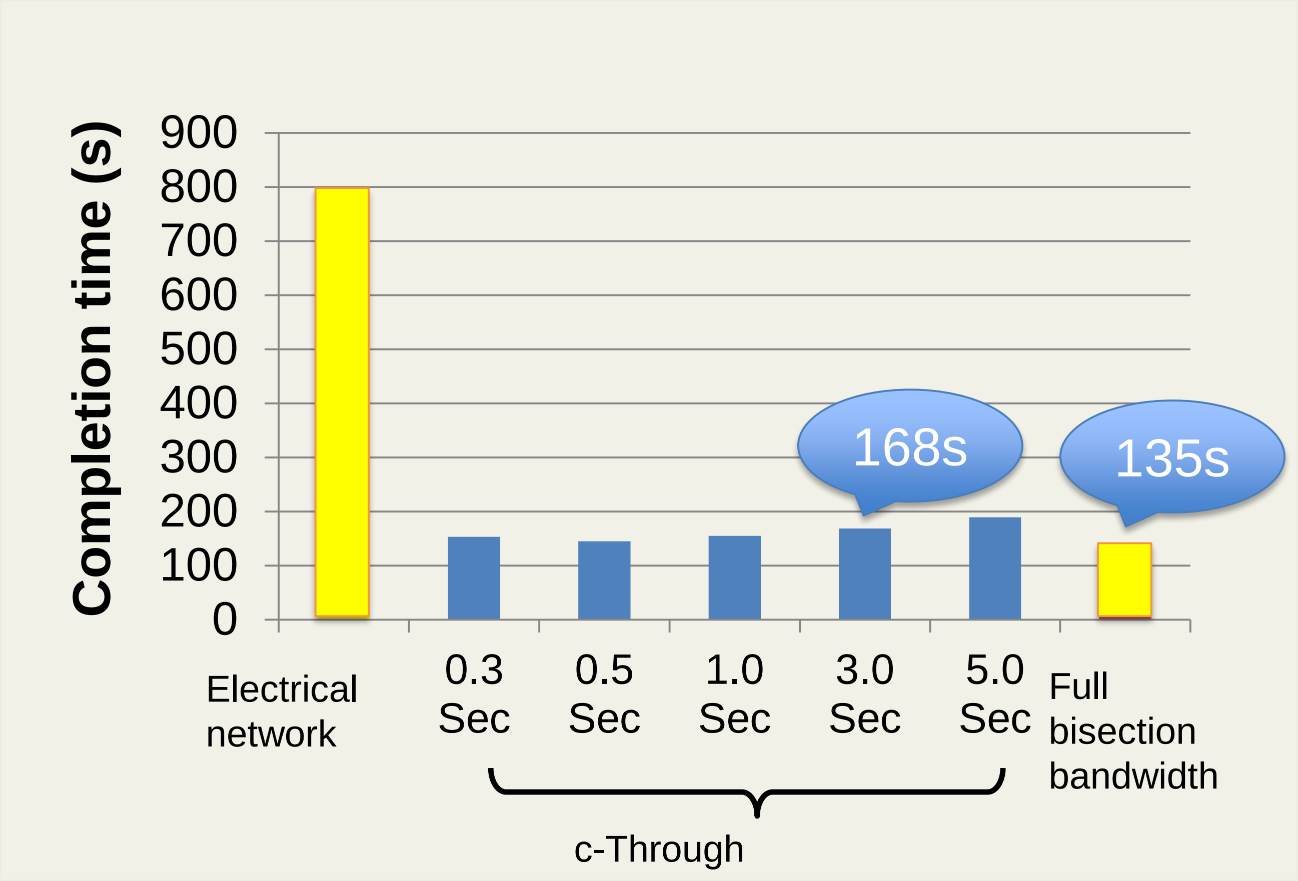
Concentrated traffic in 64MB blocks

MapReduce sort 10GB random data



c-Through varying socket buffer size limit
(reconfiguration interval: 1 sec)

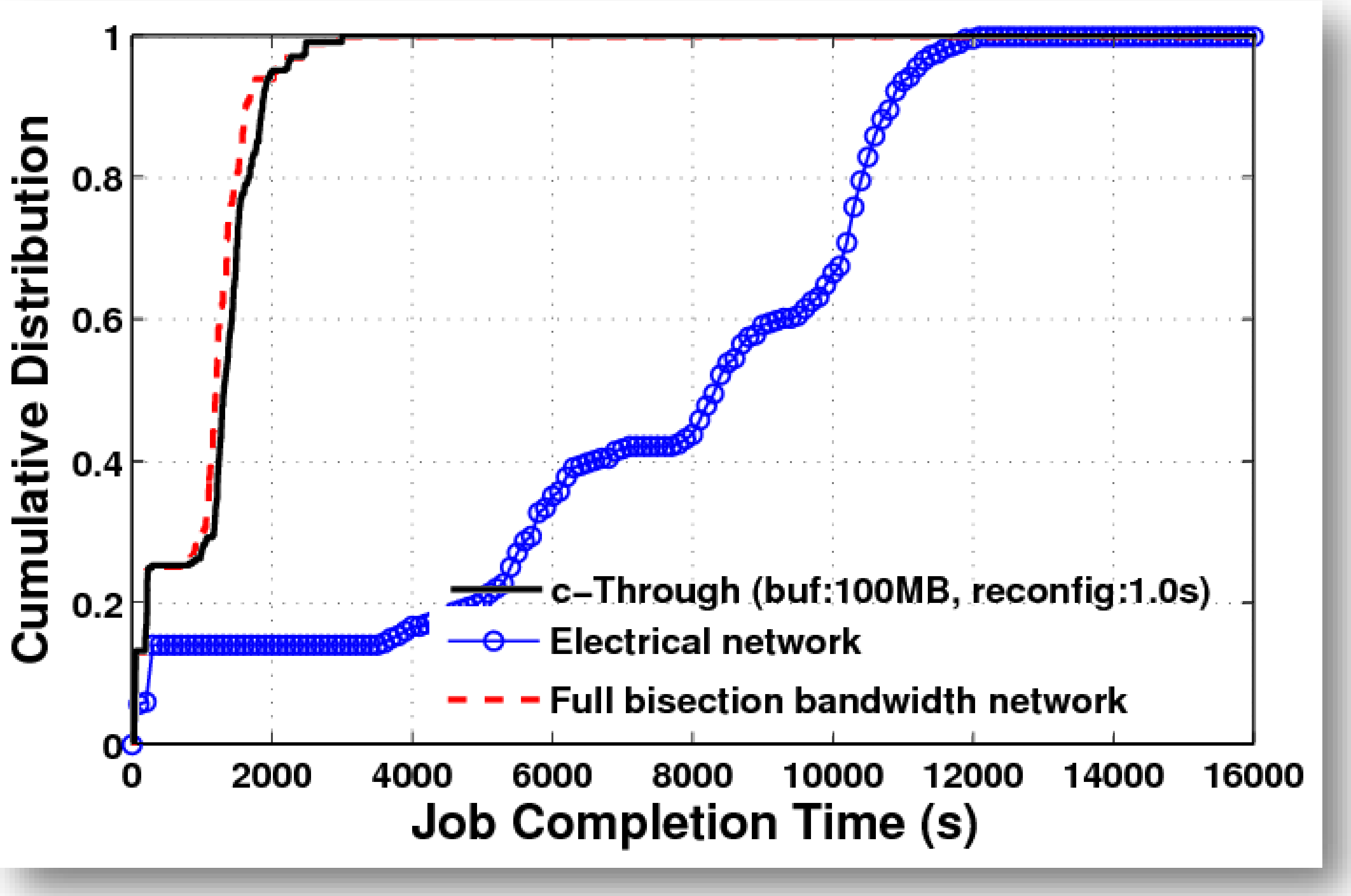
MapReduce sort 10GB random data



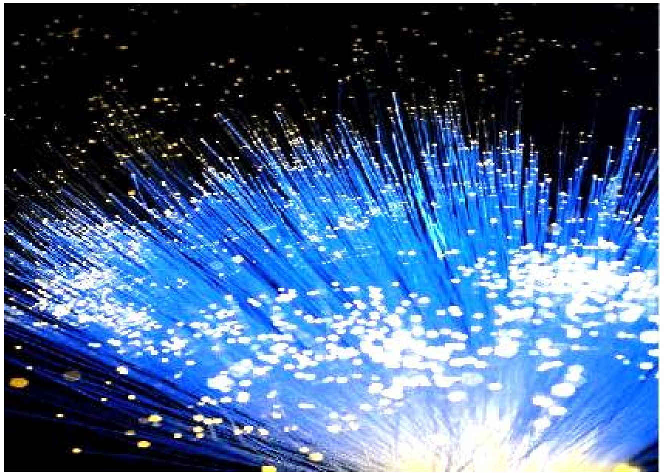
c-Through varying reconfiguration interval
(socket buffer size limit: 100MB)

Yahoo Gridmix benchmark

- 3 runs of 100 mixed jobs such as web query, web scan and sorting
- 200GB of uncompressed data, 50 GB of compressed data



Summary



- Hybrid packet/circuit switched data center network
  - c-Through demonstrates its feasibility
  - Good performance even for applications with all to all traffic
- Future directions to explore:
  - The scaling property of hybrid data center networks
  - Making applications circuit aware
  - Power efficient data centers with optical circuits

Picture from Internet websites.

Related Work

	Link Technology	Modifications Required	Working Prototype
Helios (SIGCOMM ‘10)	Optics w/ WDM 10G-180G (CWDM) 10G-400G (DWDM)	Switch Software	Glimmerglass, Fulcrum
c-Through (SIGCOMM ’10)	Optics (10G)	Host OS	Emulation
Flyways (SIGCOMM ‘11, HotNets ‘09)	Wireless (1G, 10m)	Unspecified	
IBM System-S (GLOBECOM ‘09)	Optics (10G)	Host Application; Specific to Stream Processing	Calient, Nortel
HPC (SC ‘05)	Optics (10G)	Host NIC Hardware	

Before Next time

- Project Interim report
  - **Due Monday, October 27.**
  - And meet with groups, TA, and professor
- Lab3 – Packet filter/sniffer
  - **Due *tomorrow*, Tuesday, October 21.**
- Lab1/2 redux **due Friday, October 24**
- Fractus Upgrade: SAVE ALL YOUR DATA
  - Fractus will be upgraded from October 28th to 30th
  - Can use Red Cloud during upgrade period, then switch back to Fractus

- ***Required review and reading for Wednesday, October 22***
  - "On the Feasibility of Completely Wireless Datacenters," J. Y. Shin, E. G. Sirer, H. Weatherspoon, and D. Kirovski, *IEEE/ACM Transactions on Networking (ToN)*, Volume 21, Issue 5 (October 2013), pages 1666-1680.

- Check piazza: http://piazza.com/cornell/fall2014/cs5413
- Check website for updated schedule