A 12-Rack, 180-Server Datacenter Network (DCN) Using Multiwavelength Optical Switching and Full Stack Optimization

Da Wei, Yiran Li, Wei Xu

Institute of Interdisciplinary Information Science (IIIS), Tsinghua University

Lei Xu

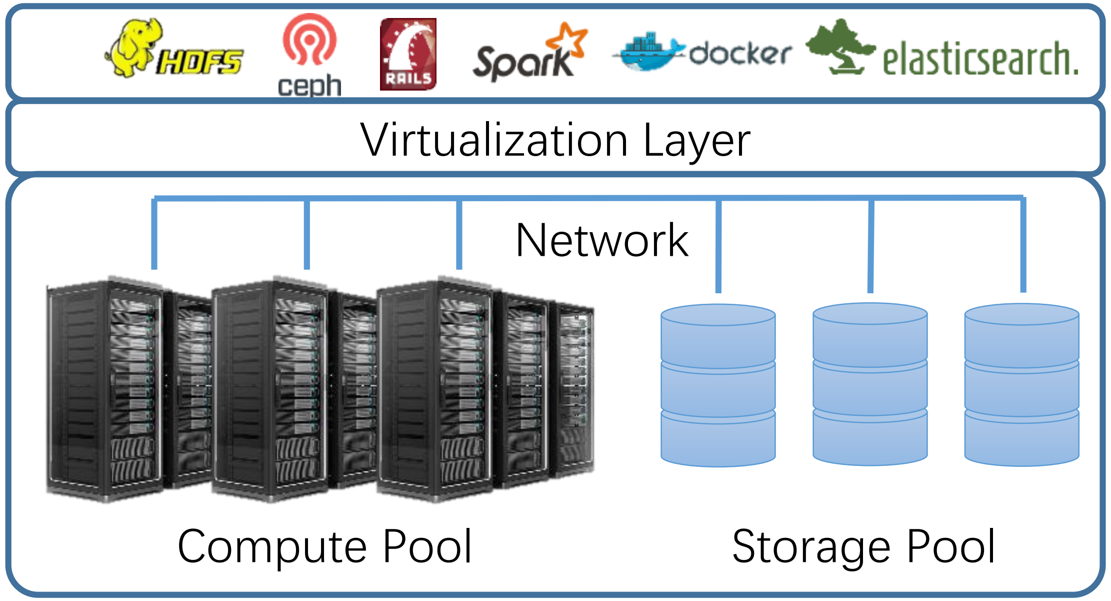
Torray Networks Inc. / Sodero Networks Inc

Xin Jin

Department of Computer Science, Princeton University

Hyper Converged Cloud => More Sophisticated DCNs

- Hyper converged infrastructure
- Different applications running over thousands of servers
- Workloads change fast
- Mix of short and long flows
- Diverse requirements of different applications
  - Search - Latency
  - Hadoop – Throughput
  - …
- We need a **FLEXIBLE** network to cope with the challenges



Hyper converged infrastructure

Previous Work on Optical DCN

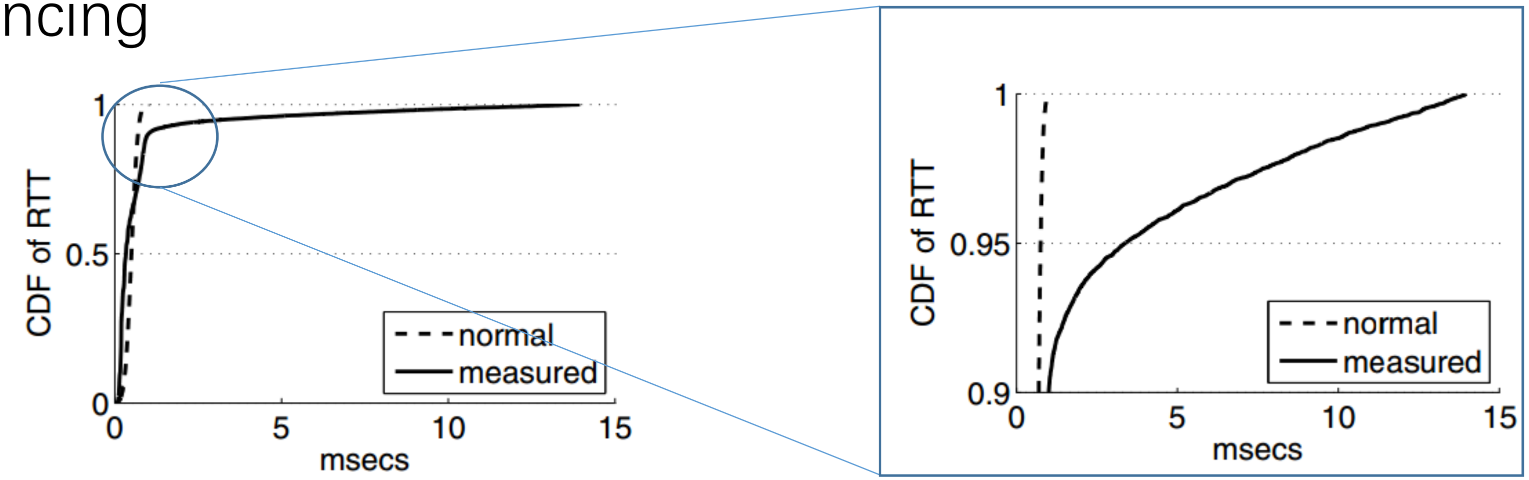
Early demonstrations of optically switched DCN testbed

- K. Chen, A. Singla, A. Singh, L. Xu, Y. Zhang, “**OSA**: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility” , Proc. of USENIX NSDI conference, April 2012.
- G. Wang, D. G. Andersen, M. Kaminsky, M. Kozuch, T. S. E. Ng, K. Papagiannaki, and M. Ryan, “**c-Through**: Part-time Optics in Data Centers'', Proc. ACM SIGCOMM, Aug. 2010.
- N. Farrington, G. Porter, S. Radhakrishnan, H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, “**Helios:** A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers” , Proc. of ACM SIGCOMM, August 2010

Ever since, optical switching for intra- and inter- DCN applications has attracted strong interests in both academia and industry.

Long Tail Latency Issues in DCN

- Tail latency directly impacts the quality of service
- Long tail latency caused by congestions from
  - Traffic bursts
  - Uneven load balancing

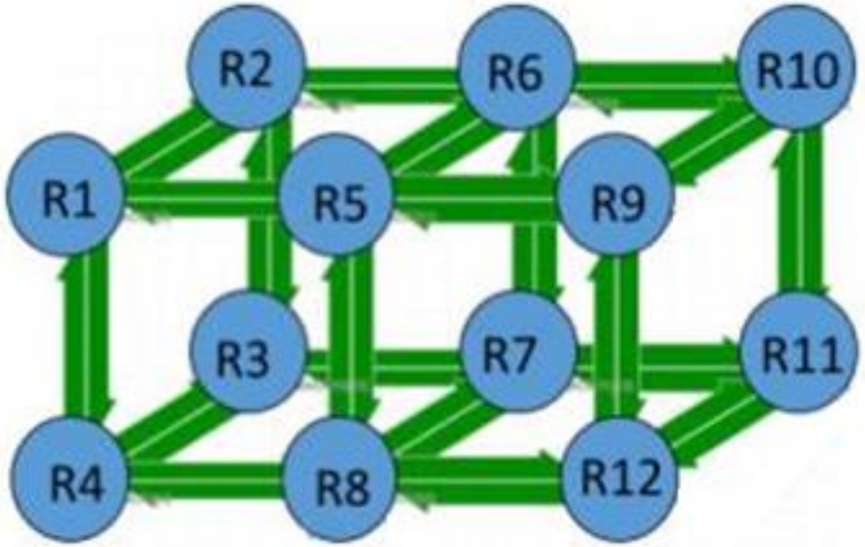


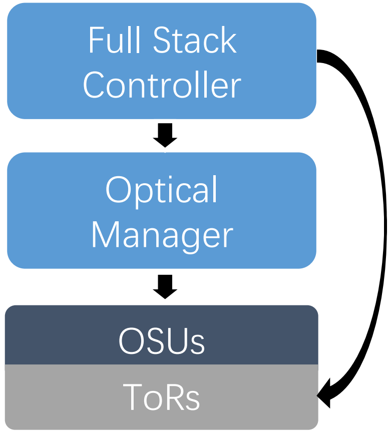
Two orders of magnitude variations in RTT

D. Zats, T. Das, P. Mohan, D. Borthakur, and R. Katz, “DeTail: reducing the flow completion time tail in datacenter networks” , Proc. of ACM SIGCOMM, August 2012

DFabric DCN

- 12 racks, 180 servers
- WSS-based multiwavelength switching and interconnection (without central optical switching matrix)
- Hyper-cube topology
- OpenFlow enabled top-of-rack switches (ToR)
- Full stack controller and optimization





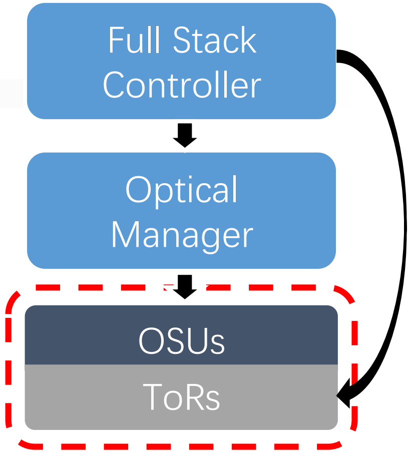
Full Stack Controller

Optical Manager

OSUs

ToRs

Optical Switching Unit (OSU) Design

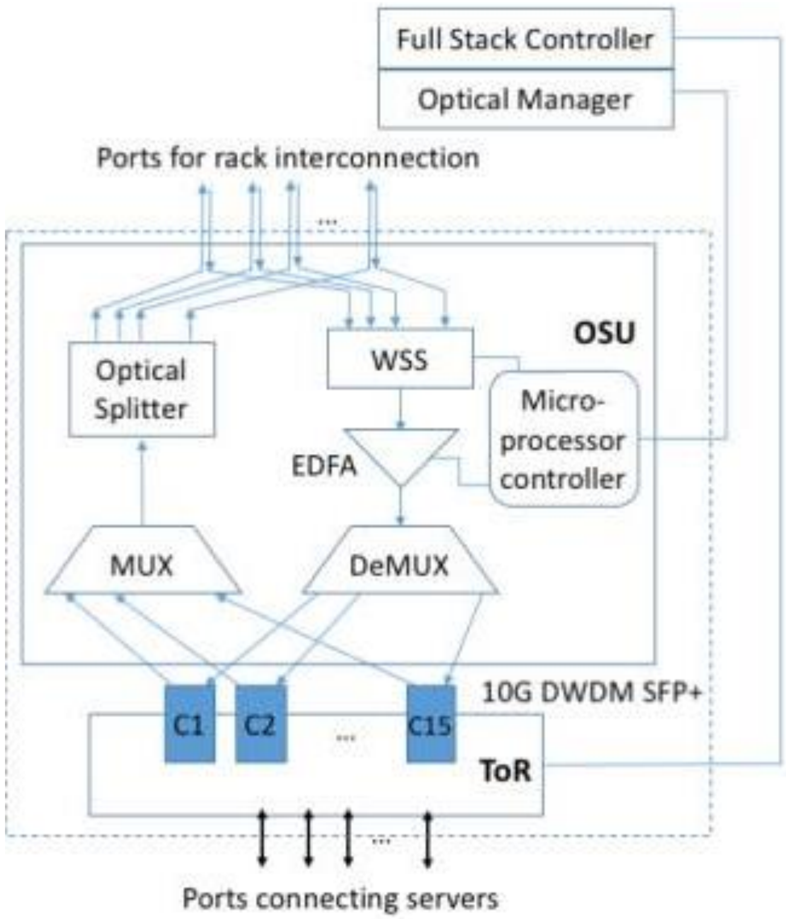


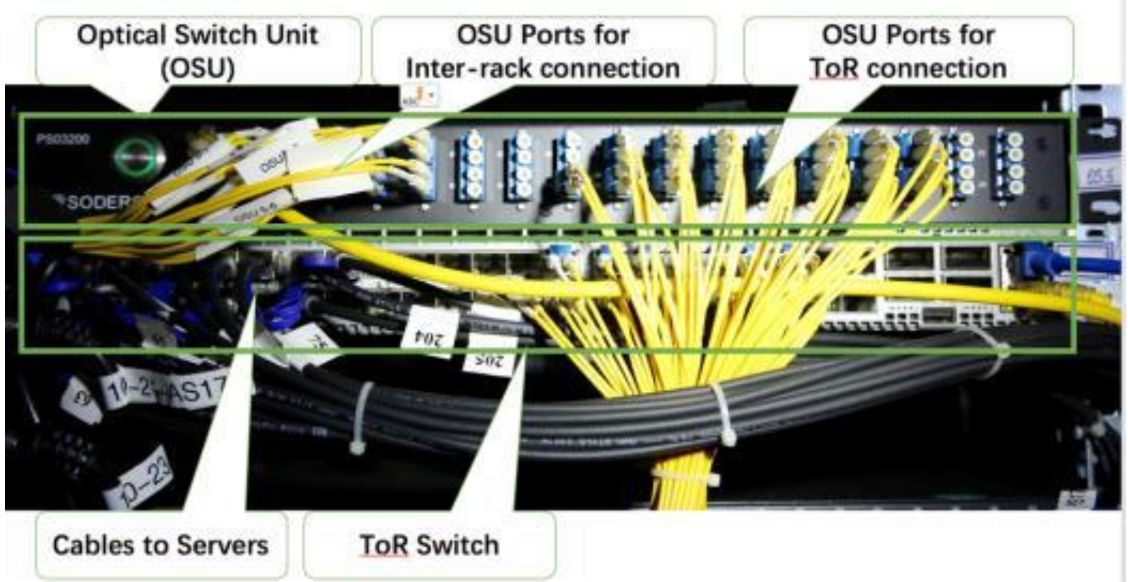
Full Stack Controller

Optical Manager

OSUs

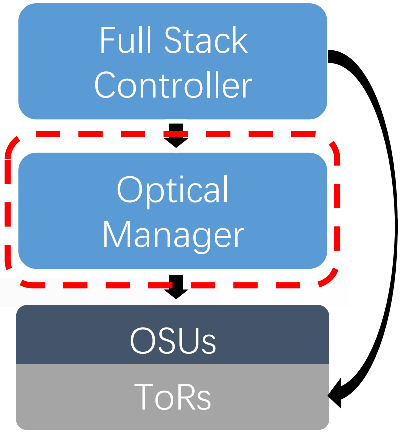
ToRs



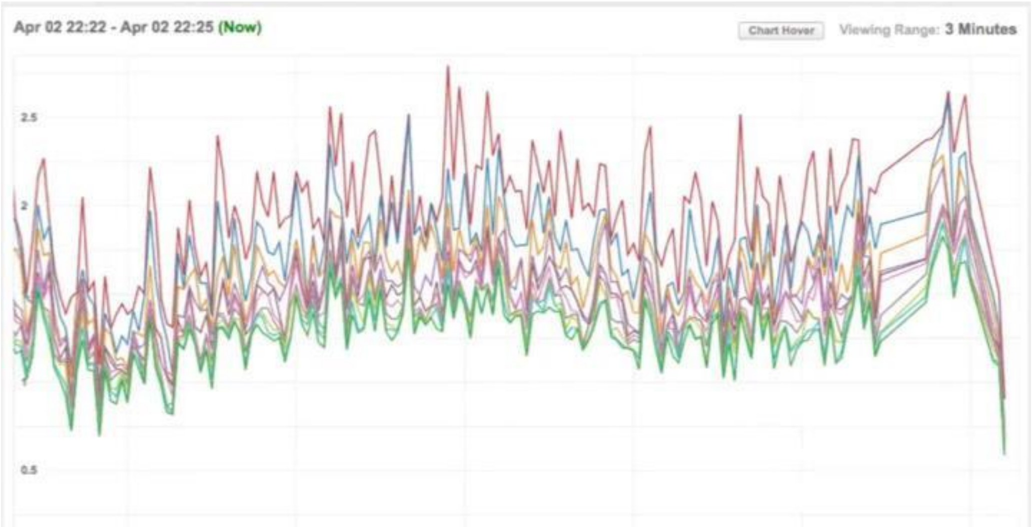

Optical Switch Unit (OSU)
OSU Ports for Inter-rack connection
OSU Ports for ToR connection
Cables to Servers
ToR Switch

Built from off-the-shelf components

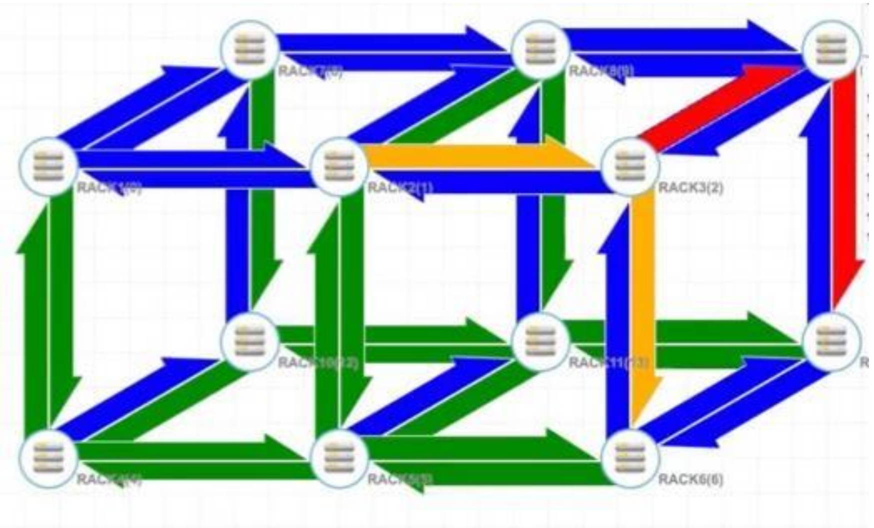
Traffic Monitoring and Visualization



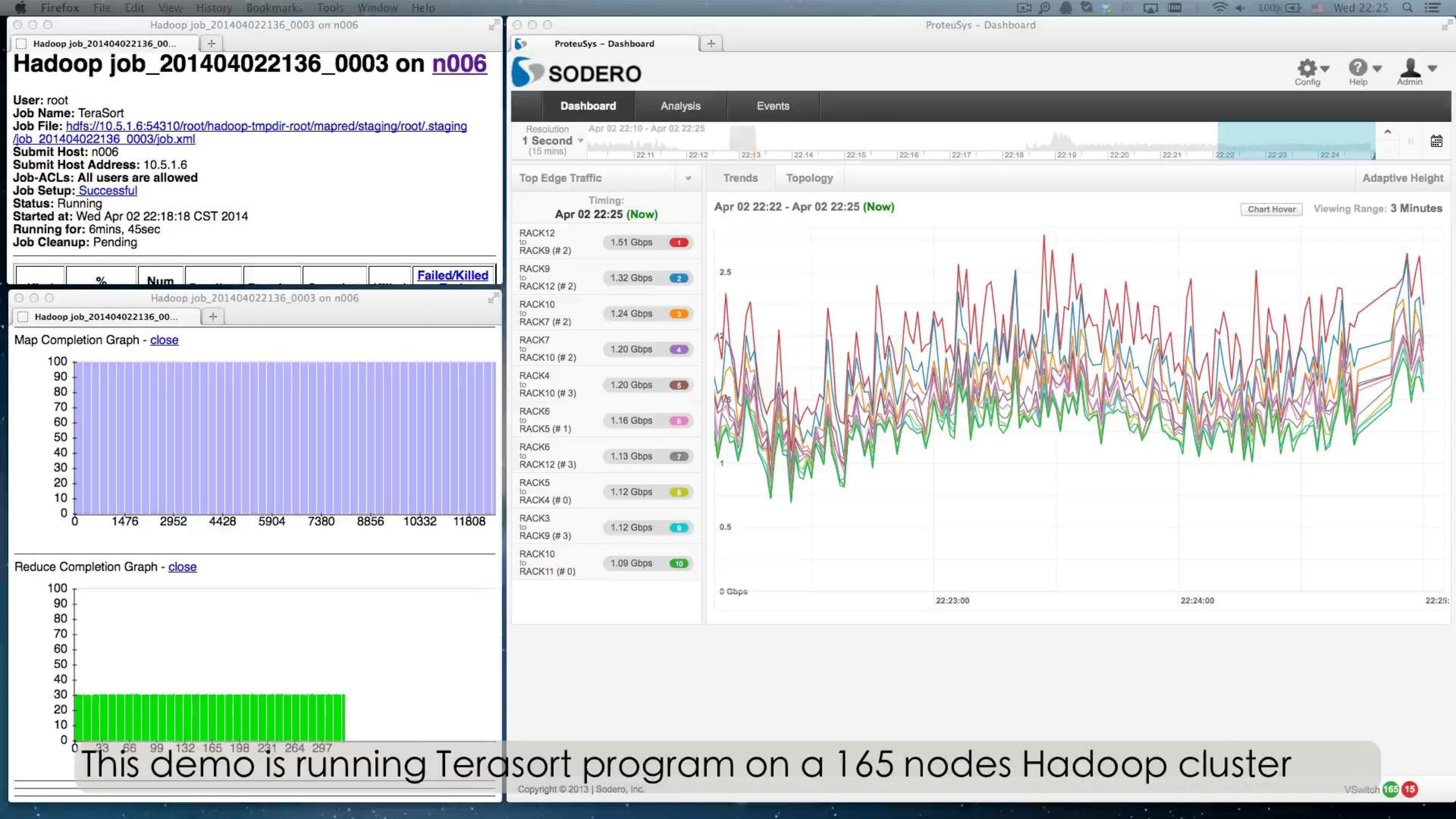
Controlled by the optical manager:



Aggregated real-time network traffic



Real-time per-link utilization



Hadoop job_201404022136_0003 on n006

User: root
Job Name: TeraSort
Job File: hdfs://10.5.1.6:54310/root/hadoop-tmpdir-root/mapred/staging/root/.staging/job_201404022136_0003/job.xml
Submit Host: n006
Submit Host Address: 10.5.1.6
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Running
Started at: Wed Apr 02 22:18:18 CST 2014
Running for: 6mins, 45sec
Job Cleanup: Pending

	%	Num					Failed/Killed

Map Completion Graph - close

Reduce Completion Graph - close

Top Edge Traffic

Timing:
Apr 02 22:25 (Now)

RACK12 to RACK9 (# 2)	1.51 Gbps	1
RACK9 to RACK12 (# 2)	1.32 Gbps	2
RACK10 to RACK7 (# 2)	1.24 Gbps	3
RACK7 to RACK10 (# 2)	1.20 Gbps	4
RACK4 to RACK10 (# 3)	1.20 Gbps	5
RACK6 to RACK5 (# 1)	1.16 Gbps	6
RACK6 to RACK12 (# 3)	1.13 Gbps	7
RACK5 to RACK4 (# 0)	1.12 Gbps	8
RACK3 to RACK9 (# 3)	1.12 Gbps	9
RACK10 to RACK11 (# 0)	1.09 Gbps	10

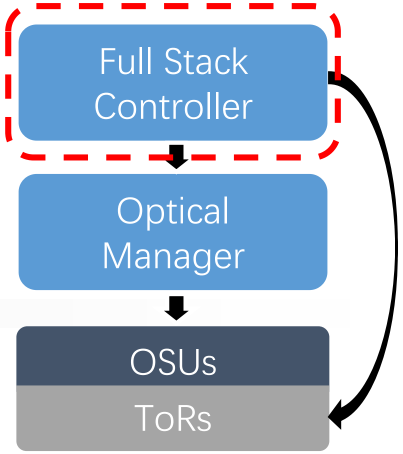
Apr 02 22:22 - Apr 02 22:25 (Now)

Viewing Range: 3 Minutes

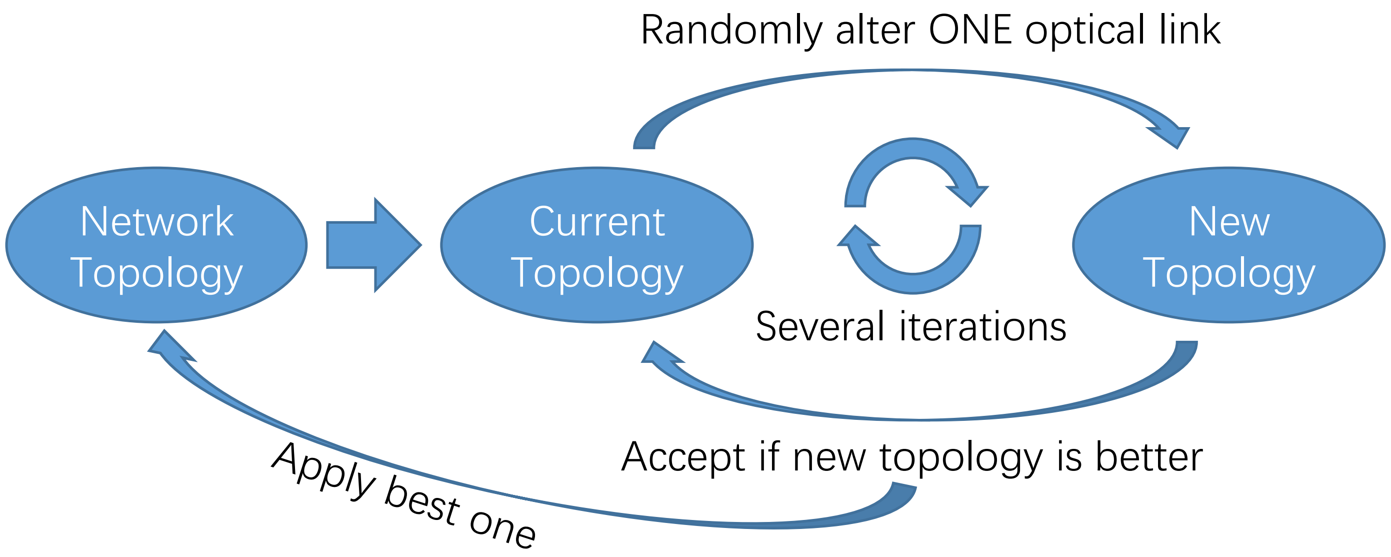
Copyright © 2013 | Sodero, Inc.

This demo is running Terasort program on a 165 nodes Hadoop cluster

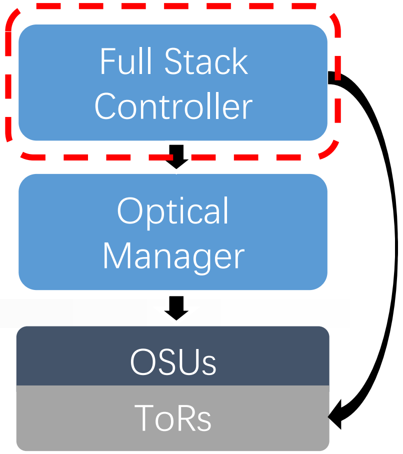
Full-stack optimization



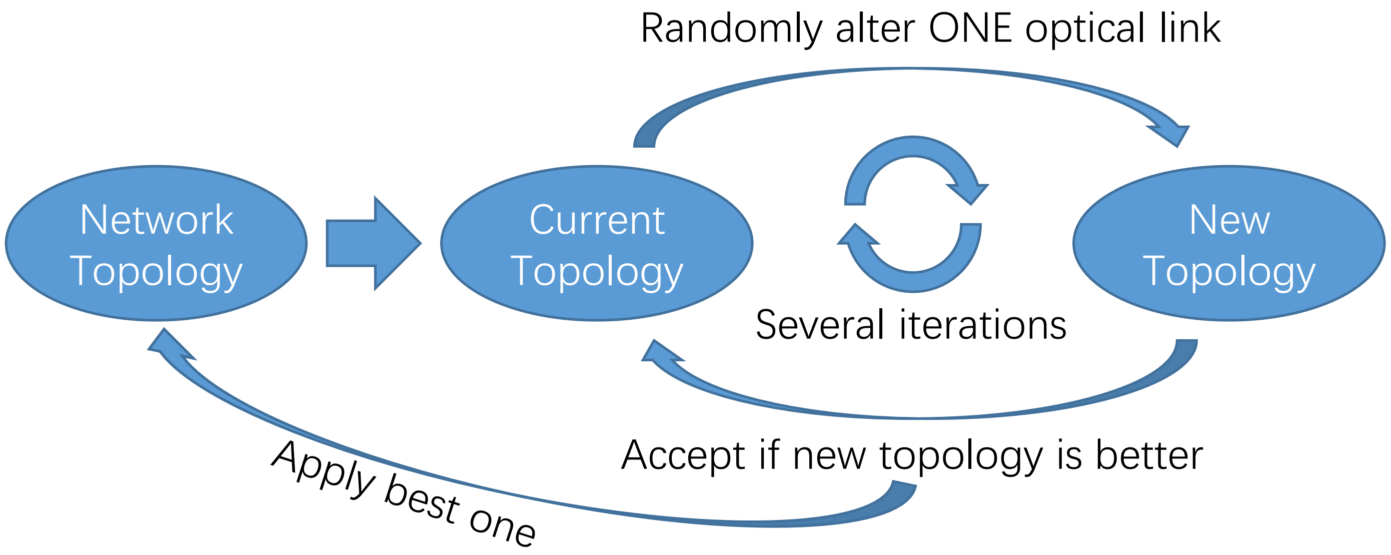
- Balance load on links to avoid congestion
  - Optimization goal: minimize the maximum single link utilization
- Joint optimization of the optical and network layers
  - The problem is NP-hard
  - Randomized approximation algorithm based on simulated annealing



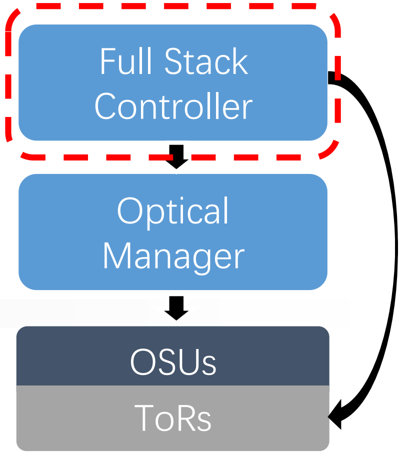
Key Algorithm Ideas



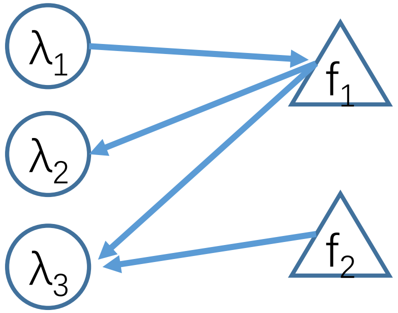
- Reduce search space using network-layer topology as the state
- Starting with topology that is similar to the current one



Consistent Update



- Problem: ensure no packet loss during update process
- Extend the state-of-the-art network update solution Dionysus [3]
- Dionysus uses dependency graph to schedule update operations
- The dependency graph includes two types of nodes:
  - *fNode* - Update operation that moves a flow from an old path to a new path
  - *λNode* – Update operation that moves a wavelength from an old edge to a new edge



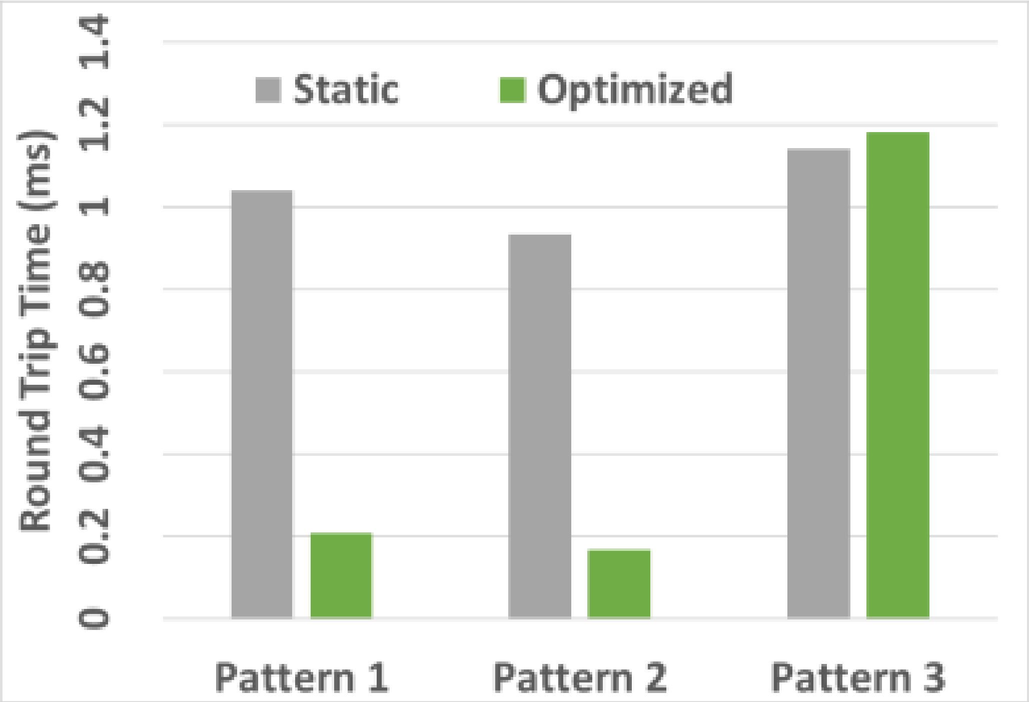
Example of dependency graph

[3] X. Jin, H. Liu, R. Gandhi, S. Kandula, R. Mahajan, M. Zhang, J. Rexford, R. Wattenhofer, "Dynamic scheduling of network updates." Proc. of ACM SIGCOMM, Aug 2014

Results: Long Tail Latency Reduction

- Optimized topology vs. static topology
- Subset of 8 racks with three traffic patterns

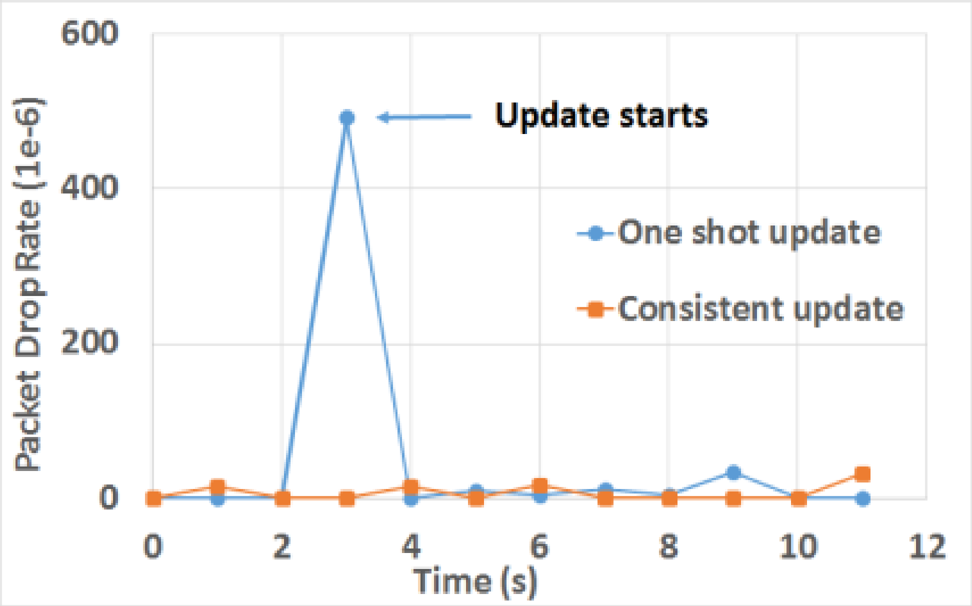
- Pattern 1: Cross-network bulk data transfer
- Pattern 2: Two separate traffic intensive cliques, with limited traffic in between.
- Pattern 3: All-to-all uniformly distributed traffic



99th percentile of round trip time

Results: Effective Consistent Update

- One shot update: move all affected flows onto a default link
- Congestion causes significant packet drop
- No significant change in consistent update



Consistent update vs. one shot update

Conclusion

- We present DFabric: a 12-rack, 180-server DCN using multiwavelength switching and interconnection.
- We implemented real-time network traffic and per-link utilization monitoring, full-stack optimization by jointly optimizing optical switching and network flow routing, and network status consistent update.
- We show benefits in long tail latency reduction and packet loss drop.