Control Plane Solutions for Scalable and Modular Optically Interconnected Datacenters

Roberto Proietti
University of California, Davis, USA

Opportunities and Challenges for Optical Switching in the Data Center
Sunday, 3 March, 13:00 - 15:30; Room number: 6D
Optical Fiber Communication Conference, San Diego Convention Center

UCDAVIS UNIVERSITY OF CALIFORNIA

1

Outline

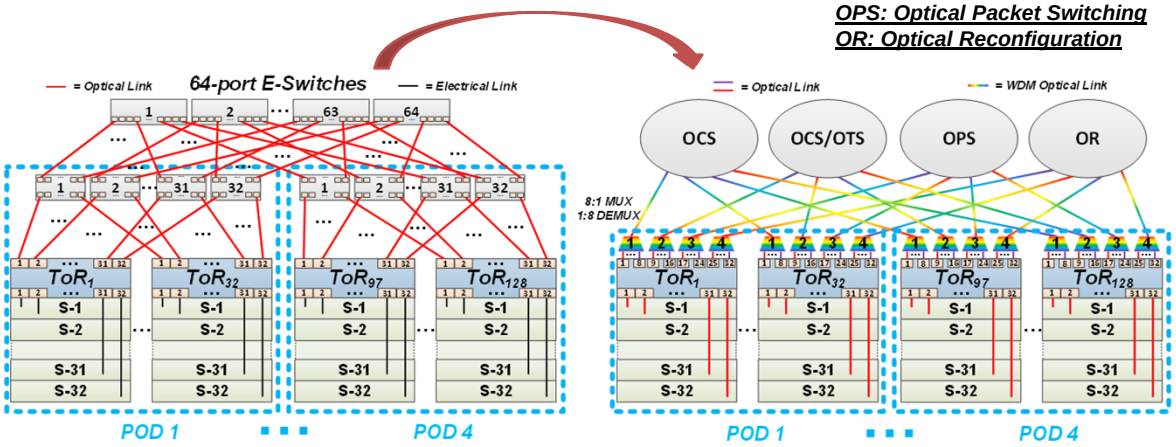
- Optical Switching in Data Centers
  - Motivations
  - Control Plane Requirements
- Control Plane/Data Plane Approaches: Centralized and Distributed
- Scalable & Distributed Control/Data Plane Solutions
- Optical Reconfiguration for Flexible Bandwidth Interconnects
- Conclusions

2

Optical Switching in Data Centers

Replacing Power-Hungry EPS with a FLAT Optical Layer

OCS: Optical Circuit Switching
OTS: Optical Time-Slot Switching
OPS: Optical Packet Switching
OR: Optical Reconfiguration

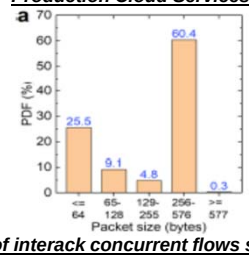


Legacy EPS Architecture ***Hybrid Architecture***

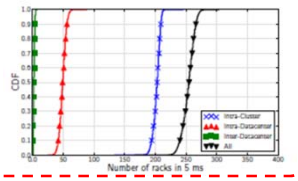
3

Traffic in Data Centers

Packet-Size Distribution from Large Production Cloud Services [1]

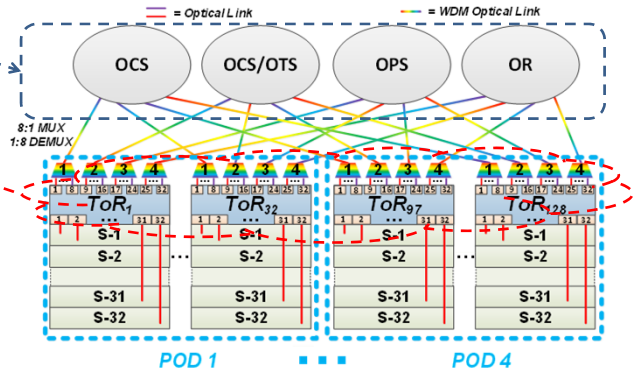


# of interack concurrent flows sent by a single host [2]



Fast switching time
Fast Control Plane Operation

OCS: Optical Circuit Switching
OTS: Optical Time-Slot Switching
OPS: Optical Packet Switching
OR: Optical Reconfiguration



[1] K. Clark et al. "Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," ECOC18 Conference, Rome, Italy
[2] A. Roy et al. "Inside the Social Network's (Datacenter) Network,", SIGCOMM 2015, London, United Kingdom

4

"Ideal" Optical Switches Requirements

- **Fast switching time (nanoseconds)**
- **Fast switch control (nanoseconds)**
- **Buffer-less operation** → (Optical) flow control
- **Scalable** to large input/output ports for implementing a flat DCN → **single stage switch architecture**
- **High Bandwidth per port** → 100 Gb/s or higher
- **Low-loss and Low-power**

5

Centralized or Distributed Control/Data Plane Solutions

	Architecture	CP approach	Switching time	CP latency	Scalability (# of ToR)
OCS	Helios [3]	Centralized	~ 10 ms	O(ms)	256
	⋮	⋮	⋮	⋮	⋮
	OSA [4]	Centralized	~ 10 ms	O(ms)	256
	C-Trough [5]	Centralized	~ 10 ms	O(ms)	256
	Mordia [6]	Centralized	~ 11.5 µs	O(ms)	88
OTS	*RotorNet [7]*	*Distributed*	*~ 20 µs*	*~1 ms*	*2,048*
	HOLST [8]	*Centralized/Distributed*	*(O)ms/~ 10ns*	*O(ms)/O(ns)*	*x*
OPS	*All-Optical TOKEN [9]*	*Distributed*	*O(ns)*	*O(ns)*	*4,096*
	⋮	⋮	⋮	⋮	⋮
	OPS Square [10]	*Distributed*	*~ 10 ns*	*~ 20 ns*	*4,096*
OR	Flex HALL [11]	Centralized	O(µs)	O(ms)	2,592
	⋮	⋮	⋮	⋮	⋮
	[12]	Centralized	O(µs)	O(ms)	x

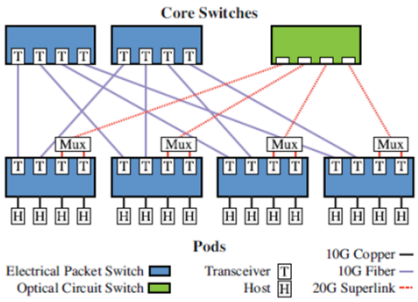
OCS: Optical Circuit Switching; OTS: Optical Time-Slot Switching; OPS: Optical Packet Switching; OR: Optical Reconfiguration

[3] N. Farrington et al., "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," SIGCOMM 2010, New Delhi, India.
[4] K. Chen et al., "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," IEEE Transactions on Networking, 2014
....
[7] W. M. Mellette et al., "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network," SIGCOMM 2017, Los Angeles, CA, USA
...
[9] Proietti et al., "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects" IEEE/OSA JLT, 2013
[10] F. Yan et al., "OPSquare: A Flat DCN Architecture Based on Flow-Controlled Optical Packet Switches," OSA JOCN, 2017
[11] Z. Cao, R. Proietti et al., "Experimental Demonstration of Dynamic Flexible Bandwidth Optical Data Center Network with All-to-All Interconnectivity " ECOC, 2014
[12] A. M. Saleh al., "Elastic WDM Switching for Scalable Data Center and HPC Interconnect Networks" OECC/PS 2016.

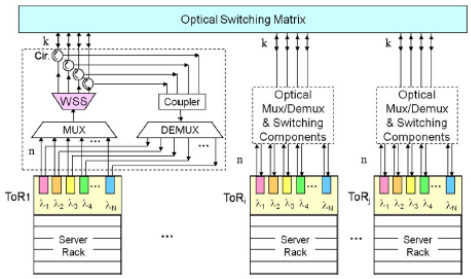
6

OCS Architectures – Centralized CP

Helios [3]



OSA [4]



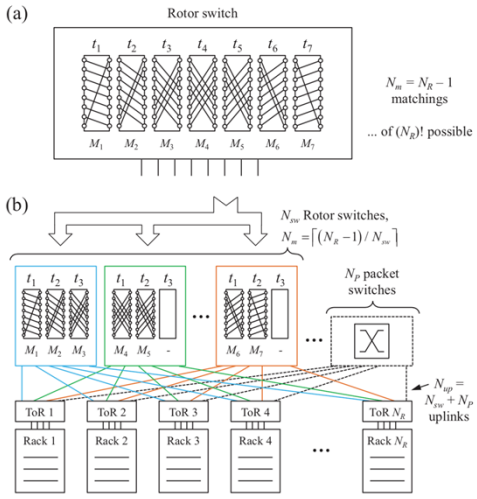
[3] N. Farrington et al., "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," SIGCOMM 2010, New Delhi, India.
[4] K. Chen et al., "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," IEEE Transactions on Networking, 2014

Switching optical circuits according to traffic demand at large scale is very challenging

- **collecting network-wide demand information**
- **traffic classification**
- **determine scheduling of switch configurations**
- **synchronizing the OCSes, the scheduler, and the endpoints**

OTS Architecture with Distributed CP

RotorNet [7]



When scales up,

- High latency: Sequentially step through many matchings
- Fabrication challenge: Monolithic Rotor switch with many matchings

⇩

- Reduced latency: Access matchings in parallel
- Simplifies Rotor switches: More scalable, less expensive

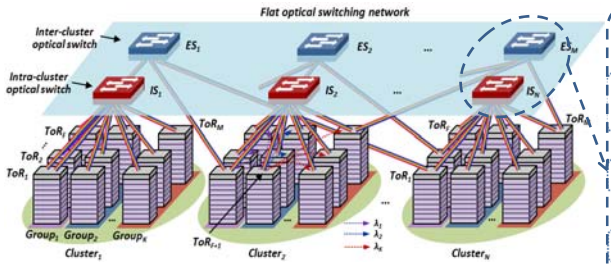
Distributing Rotor matchings = lower latency

OTS: Optical Time-Slot Switching

[6] W. M. Mellette et al., "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network," SIGCOMM 2017, Los Angeles, CA, USA

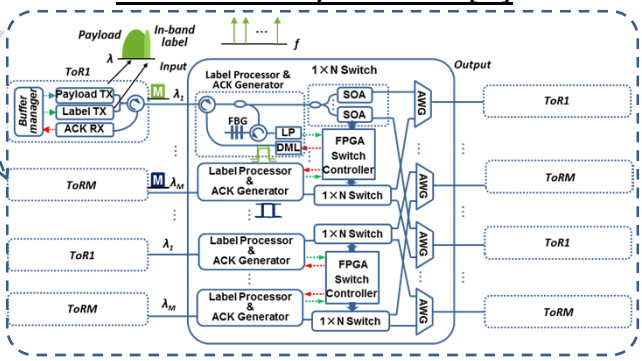
OPS Architectures with Distributed CP

OPSquare [10]



- **Flat connectivity**
  - *High bandwidth*
  - *Low latency*
- **Scalability**
  - *Square of switch radix*
  - *Large interconnectivity*
  - *4096 ToR with 64-port OPS*

Fast Controlled Optical Switch [11]

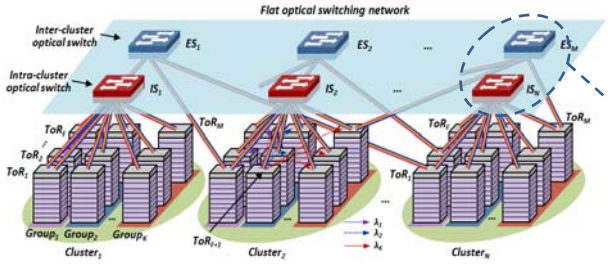


- ***Parallel processing of label bits***
- ***On-the-fly distributed control (ns latency)***
- ***Fast optical flow control and retransmission***

[9] F. Yan et al., "OPSquare: A Flat DCN Architecture Based on Flow-Controlled Optical Packet Switches," OSA JOCN, 2017
[11] W. Miao et al., "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," OSA OpEx, 2014

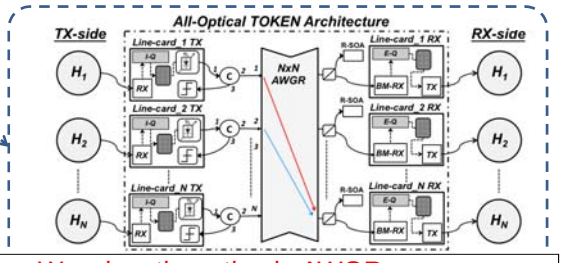
OPS Architectures with Distributed CP

OPSquare [10]



- **Flat connectivity**
  - *High bandwidth*
  - *Low latency*
- **Scalability**
  - *Square of switch radix*
  - *Large interconnectivity*
  - *4096 ToR with 64-port OPS*

All-optical TOKEN for Distributed Control Plane [9]

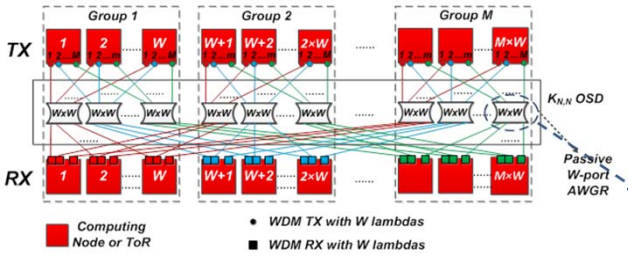


- Wavelength routing in AWGR
- Gain Saturation Effect in Reflective SOAs
- Polarization Diversity TX scheme to send Token requests (control plane) and data on two orthogonal polarizations
- One TOKEN per output port

[9] Proietti et al., "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects" IEEE/OSA JLT, 2013
[10] F. Yan et al., "OPSquare: A Flat DCN Architecture Based on Flow-Controlled Optical Packet Switches," OSA JOCN, 2017

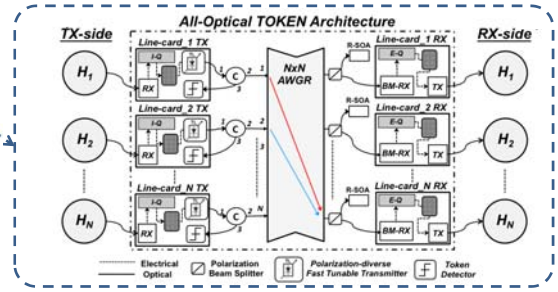
Scalable Distributed Thin-CLOS AO Token Architecture

Thin-CLOS AWGR [13]



N= M×W. There are design tradeoffs based on the value of M and N in terms of complexity, costs, number and performance requirements of the optical components (i.e. AWGs)

All-optical TOKEN for Distributed Control Plane [9]



- ***On-the-fly distributed control***
- ***Fast all-optical flow control and contention resolution***
- ***Nanosecond CP latency***

[9] Proietti et al., "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects" IEEE/OSA JLT, 2013
[13] Proietti et al., "Experimental Demonstration of a 64-port Wavelength Routing Thin-CLOS System for Data Center Switching Architectures," OSA JOCN 2018

11

Distributed Control/Data Plane is Needed

	Architecture	CP approach	Switching time	CP latency	Scalability (# of ToR)
OCS	Helios [3]	Centralized	~ 10 ms	O(ms)	256
	⋮	⋮	⋮	⋮	⋮
	OSA [4]	Centralized	~ 10 ms	O(ms)	256
	C-Trough [5]	Centralized	~ 10 ms	O(ms)	256
	Mordia [6]	Centralized	~ 11.5 µs	O(ms)	88
OTS	*RotorNet [7]*	*Distributed*	*~ 20 µs*	*~1 ms*	*2,048*
	HOLST [8]	*Centralized/Distributed*	*(O)ms/~ 10ns*	*O(ms)/O(ns)*	*x*
OPS	*All-Optical TOKEN [9]*	*Distributed*	*O(ns)*	*O(ns)*	*4,096*
	⋮	⋮	⋮	⋮	⋮
	OPS Square [10]	*Distributed*	*~ 10 ns*	*~ 20 ns*	*4,096*
OR	Flex HALL [11]	Centralized	O(µs)	O(ms)	*2,592*
	⋮	⋮	⋮	⋮	⋮
	[12]	Centralized	O(µs)	O(ms)	X

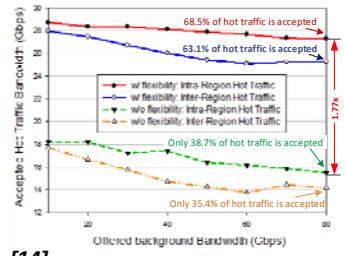
OCS: Optical Circuit Switching; OTS: Optical Time-Slot Switching; OPS: Optical Packet Switching; OR: Optical Reconfiguration

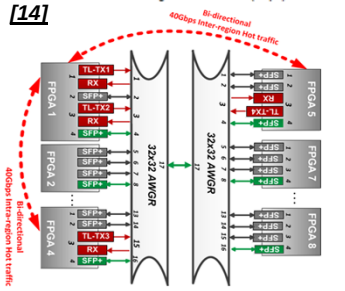
[3] N. Farrington et al., "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," SIGCOMM 2010, New Delhi, India.
[4] K. Chen et al., "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," IEEE Transactions on Networking, 2014
....
[7] W. M. Mellette et al., "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network," SIGCOMM 2017, Los Angeles, CA, USA
...
[9] Proietti et al., "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects" IEEE/OSA JLT, 2013
[10] F. Yan et al., "OPSquare: A Flat DCN Architecture Based on Flow-Controlled Optical Packet Switches," OSA JOCN, 2017
[11] Z. Cao, R. Proietti et al., "Experimental Demonstration of Dynamic Flexible Bandwidth Optical Data Center Network with All-to-All Interconnectivity " ECOC, 2014
[12] A. M. Saleh al., "Elastic WDM Switching for Scalable Data Center and HPC Interconnect Networks" OECC/PS 2016.

12

Optical Reconfiguration For Flexible Bandwidth Interconnects

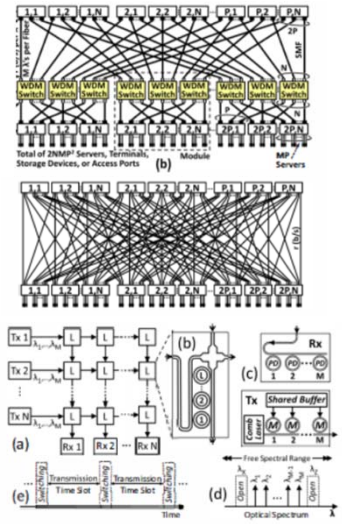
[14]





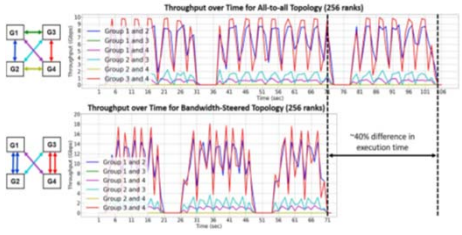
[14] Z. Cao, R. Proietti et al., "Experimental Demonstration of Dynamic Flexible Bandwidth Optical Data Center Network with All-to-All Interconnectivity " ECOC, 2014

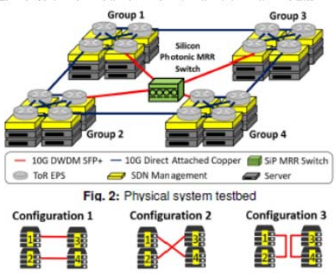
[15]



[15] A. M. Saleh al., "Elastic WDM Switching for Scalable Data Center and HPC Interconnect Networks" OECC/PS 2016

[16]





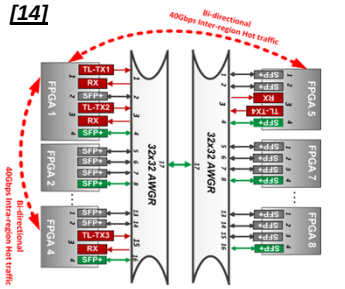
[16] Y. Shen et al., "Accelerating of High Performance Data Centers using Silicon Photonic Switch-enabled Bandwidth Steering " ECOC, 2018

13

Optical Reconfiguration For Flexible Bandwidth Interconnects

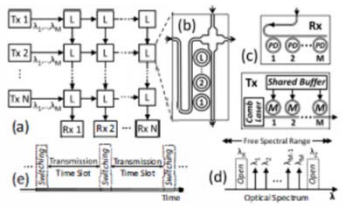
- How to control the reconfiguration operation?
  - *Again, a Centralized Control Plane may be TOO slow unless we can predict when the hotspots will happen or when more bandwidth will be needed*
- We need fast, distributed, hitless and cross-layer solutions to reconfigure the links bandwidth

[14]



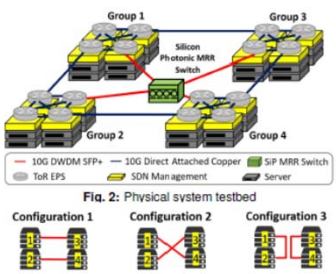
[14] Z. Cao, R. Proietti et al., "Experimental Demonstration of Dynamic Flexible Bandwidth Optical Data Center Network with All-to-All Interconnectivity " ECOC, 2014

[15]



[15] A. M. Saleh al., "Elastic WDM Switching for Scalable Data Center and HPC Interconnect Networks" OECC/PS 2016

[16]



[16] Y. Shen et al., "Accelerating of High Performance Data Centers using Silicon Photonic Switch-enabled Bandwidth Steering " ECOC, 2018

14

Summary

- As a significant portion of the traffic in data centers is composed by small packets or flows, fast switching and fast distributed control/data plane are required
- OPS solutions can achieve fast distributed control plane while overcoming the lack of buffering using optical flow-control techniques
- Challenges & Opportunities:
  - Robust photonic integrated solutions to make OPS approaches commercially viable
  - WDM TRXs with BM-CDR are essential components for enabling any Optical Switching solution BUT will they ever be feasible for datacenter deployment?
  - Can we consider directly connected topologies like Dragon-Fly for Data Centers? If combined with fast optical reconfigurability, they may be a possible solution (no per-packet/flow optical switching required)

M3B.1 • 14:00 - Photonic Switching in Datacenters and Computing Systems, S. J. Ben Yoo; Univ. of California Davis, USA