Enabling Wide-spread Communications on Optical Fabric with MegaSwitch

SingLab HKUST

Li Chen

Kai Chen, Zhonghua Zhu, Minlan Yu, George Porter, Chunming Qiao, Shan Zhong

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

LUX ET VERITAS

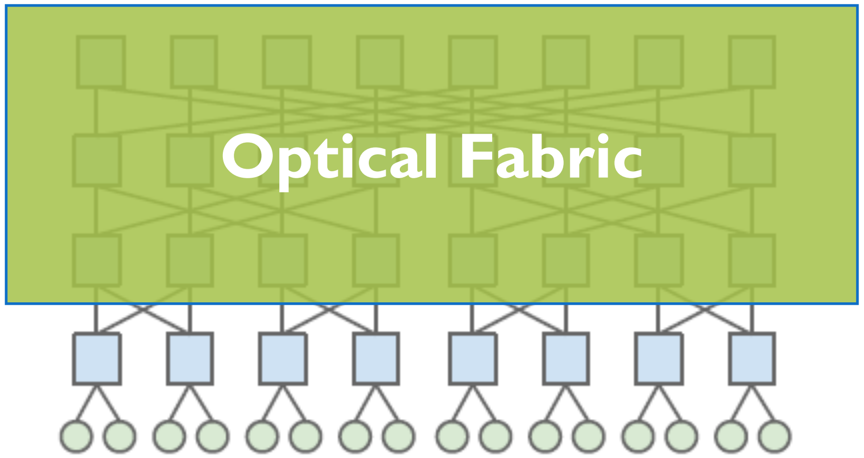
UNIVERSITY OF CALIFORNIA · SAN DIEGO · LET THERE BE LIGHT

UB

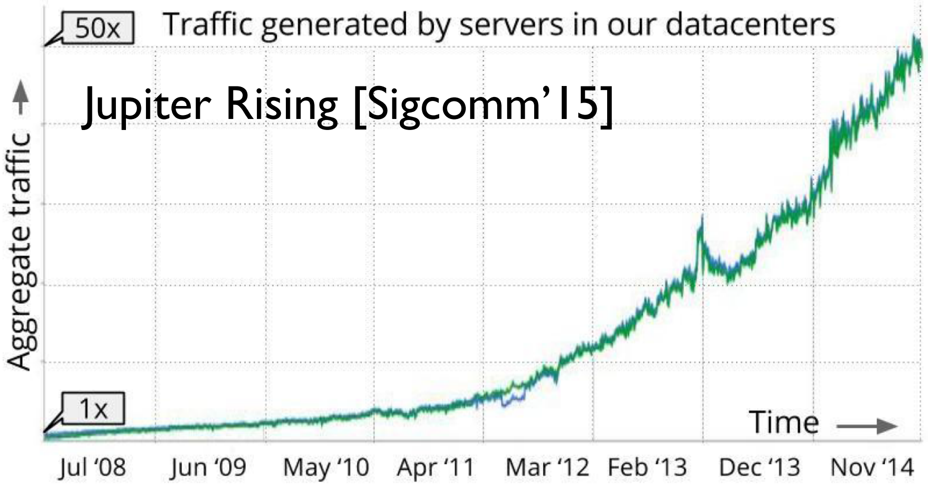
CoAdna

Optical Networking in Data Centers

- Optical networking in data centers
  - Low cost
  - Low power consumption
  - Low wiring complexity
  - High one-to-one bandwidth



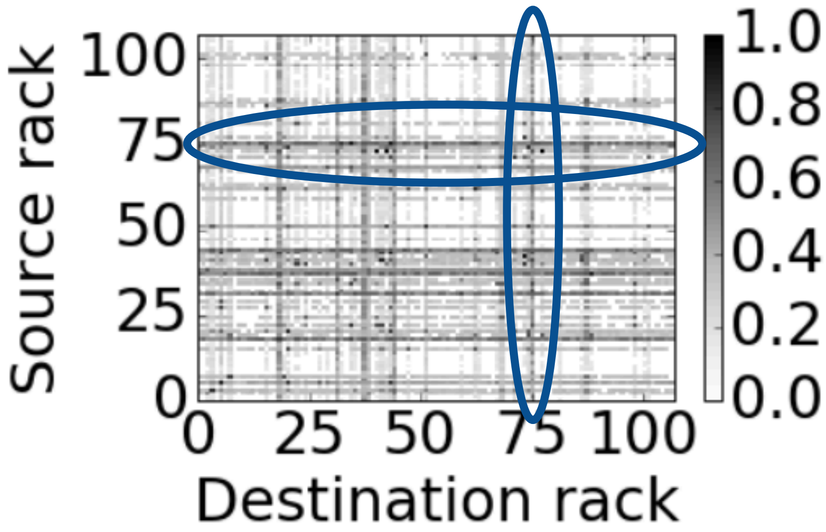
- Data center traffic demand is growing



How to design an optical fabric that enables high bisection bandwidth?

Optical Networking in Data Centers

- Optical fabric usually provides one-to-one high-bandwidth circuits.
- Data Center traffic is wide-spread



Microsoft Data Center Network [ProjecToR, Sigcomm'16]

How to design an optical fabric that supports high-bandwidth & wide-spread traffic?

Prior Works

Prior works reuse wavelengths temporally to meet traffic demand

- 2010: C-Through, Helios
- 2012: OSA
- 2013: Mordia, ReacToR

Dest

Src

	1	2	3
1	-	1	1
2	1	-	1
3	2	1	-

Traffic Demand Matrix

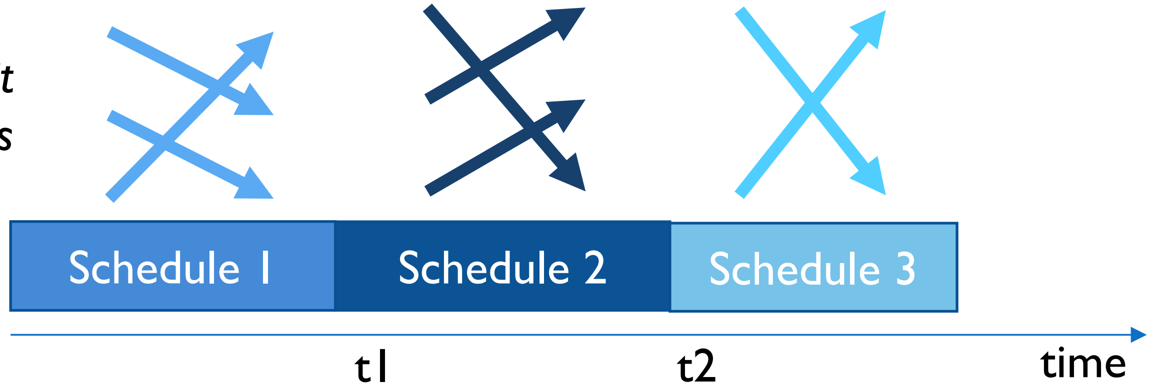
BvN decomp.

	1	2	3
1	-	1	0
2	0	-	1
3	1	0	-

	1	2	3
1	-	0	1
2	1	-	0
3	0	1	-

	1	2	3
1	-	0	1
2	0	-	0
3	1	0	-

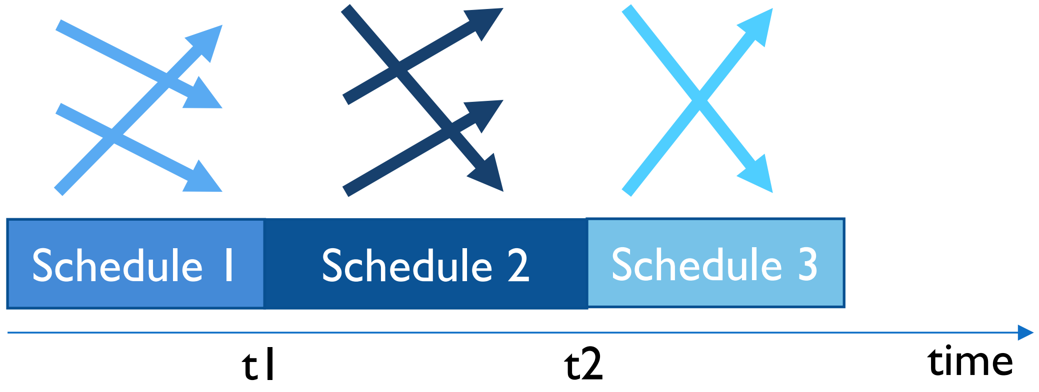
Wavelength/Circuit assignments



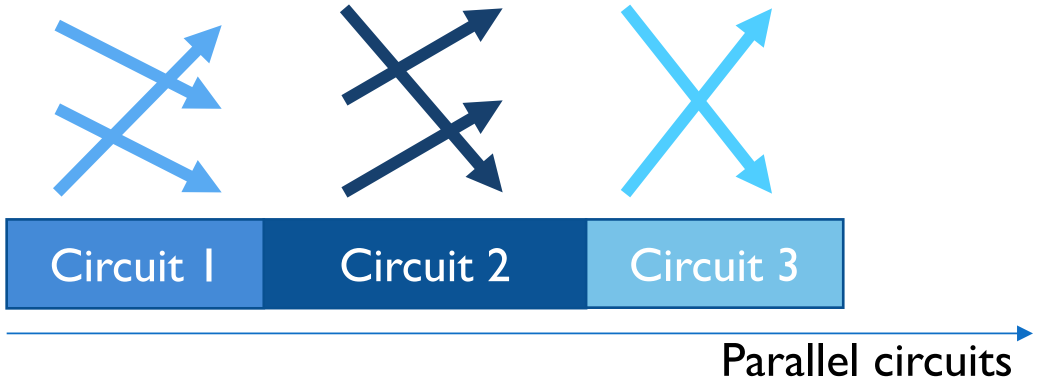
High Bisection Bandwidth + Wide-Spread Connectivity

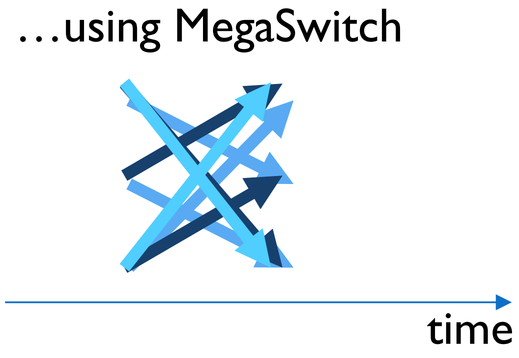
Prior works take several rounds to meet a wide-spread demand

- 2010: C-Through, Helios
- 2012: OSA
- 2013: Mordia, ReacToR



MegaSwitch: Meet a wide-spread demand simultaneously



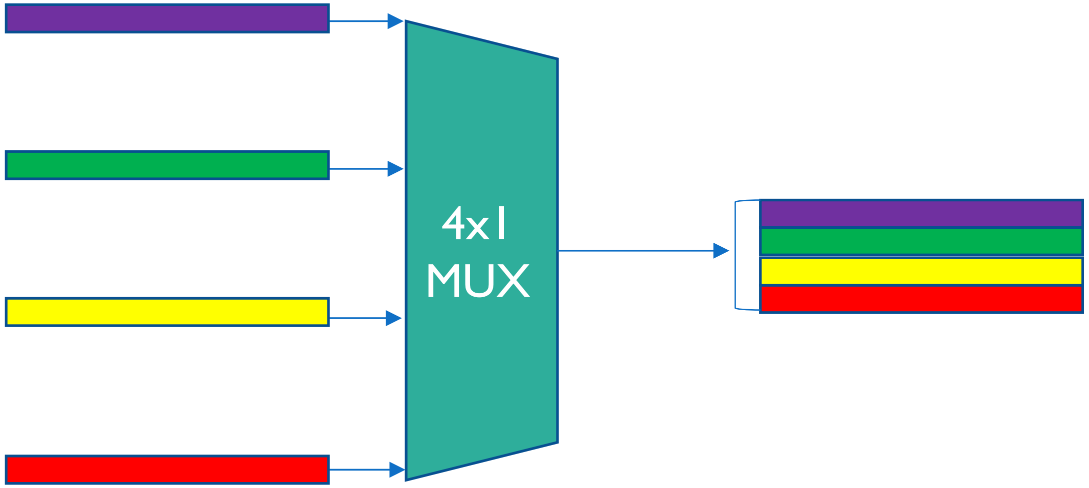


MegaSwitch Data Plane

Enabling Spatial Reuse of Wavelength

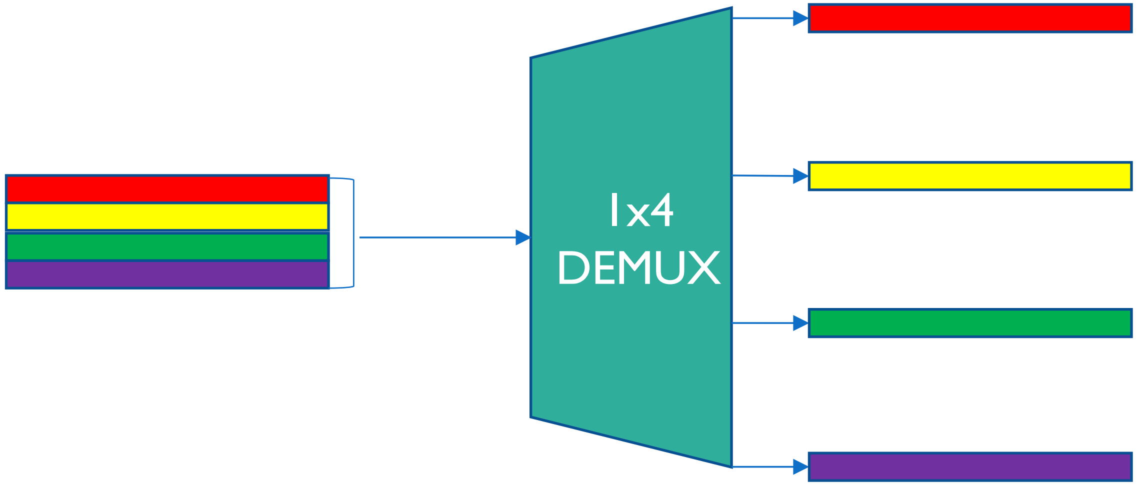
Prototype implementation

Multiplexer



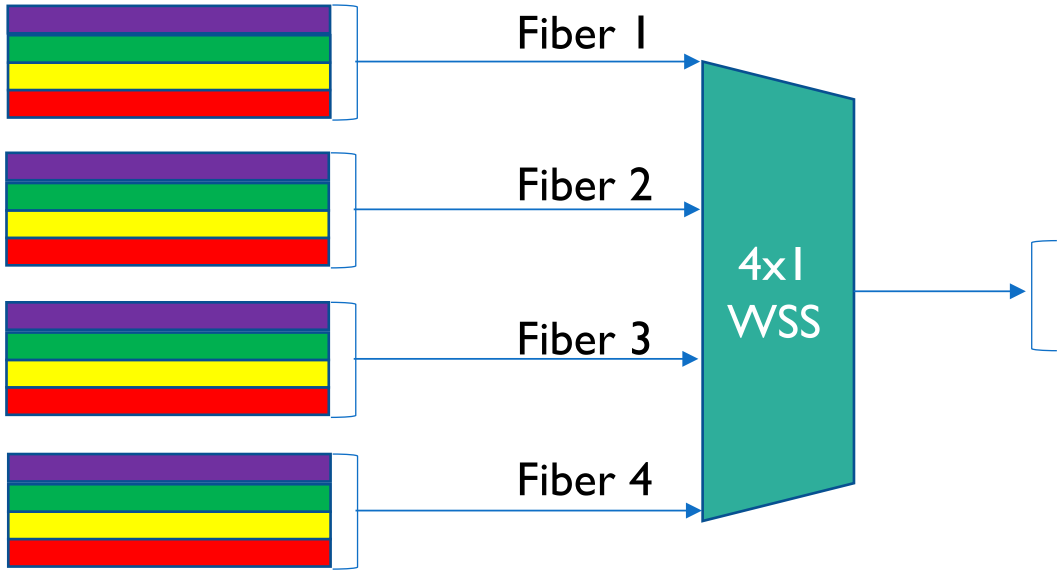
- MUX: k input wavelengths, 1 output fiber

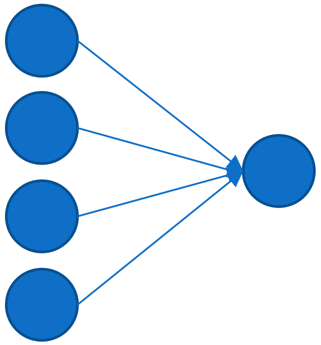
Demultiplexer



- DEMUX: 1 input fiber, k output wavelengths

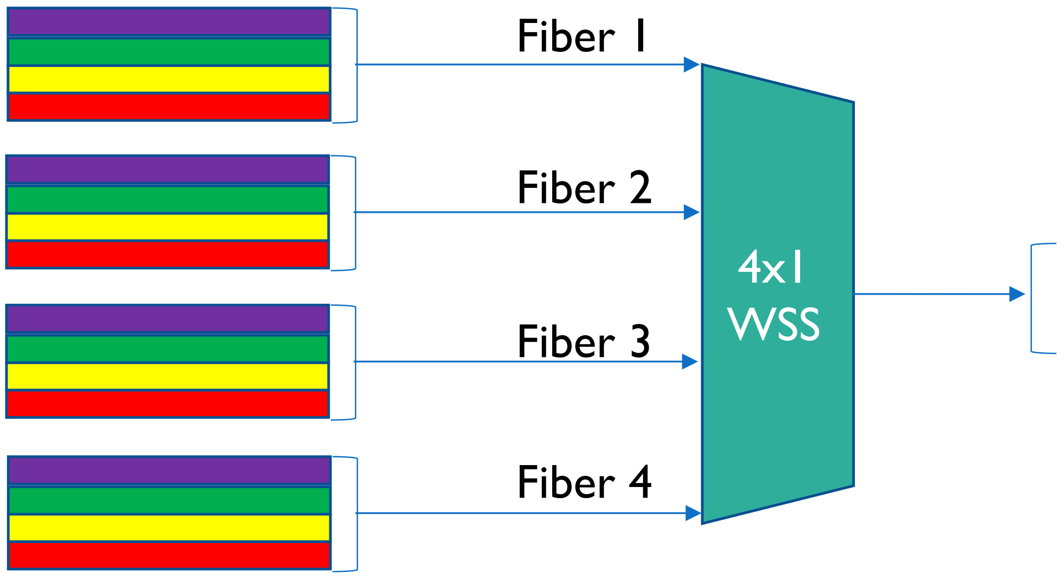
Wavelength Selective Switch

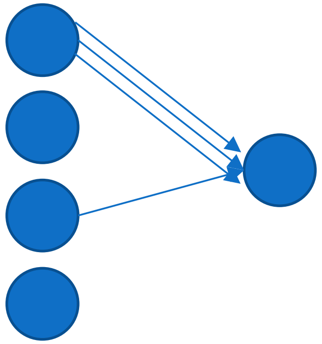




- WSS: w input fibers, 1 output fiber
- Same set of wavelengths can be reused on different fibers.

Wavelength Selective Switch



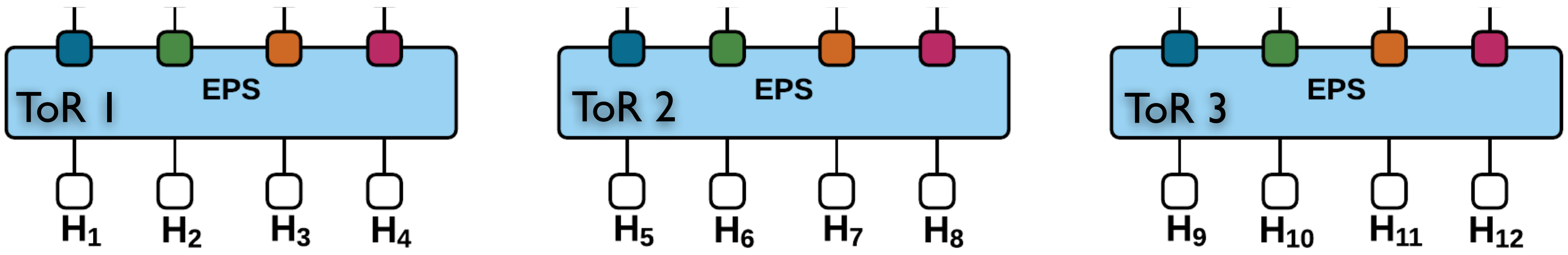


- WSS: w input fibers, 1 output fiber
- Same set of wavelengths can be reused on different fibers.

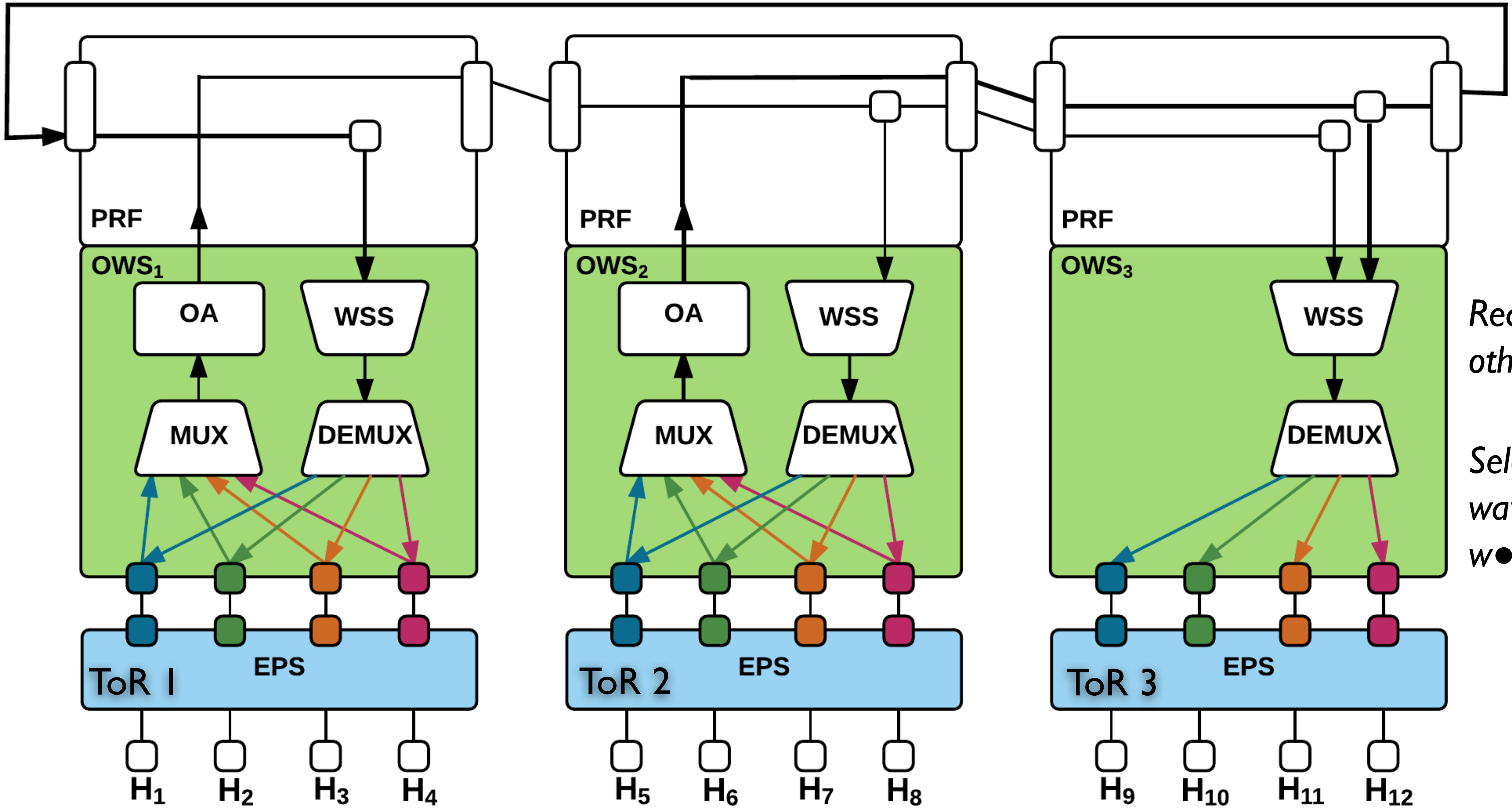
Sending

Each ToR sends on its own fiber

Send using k wavelengths



Receiving



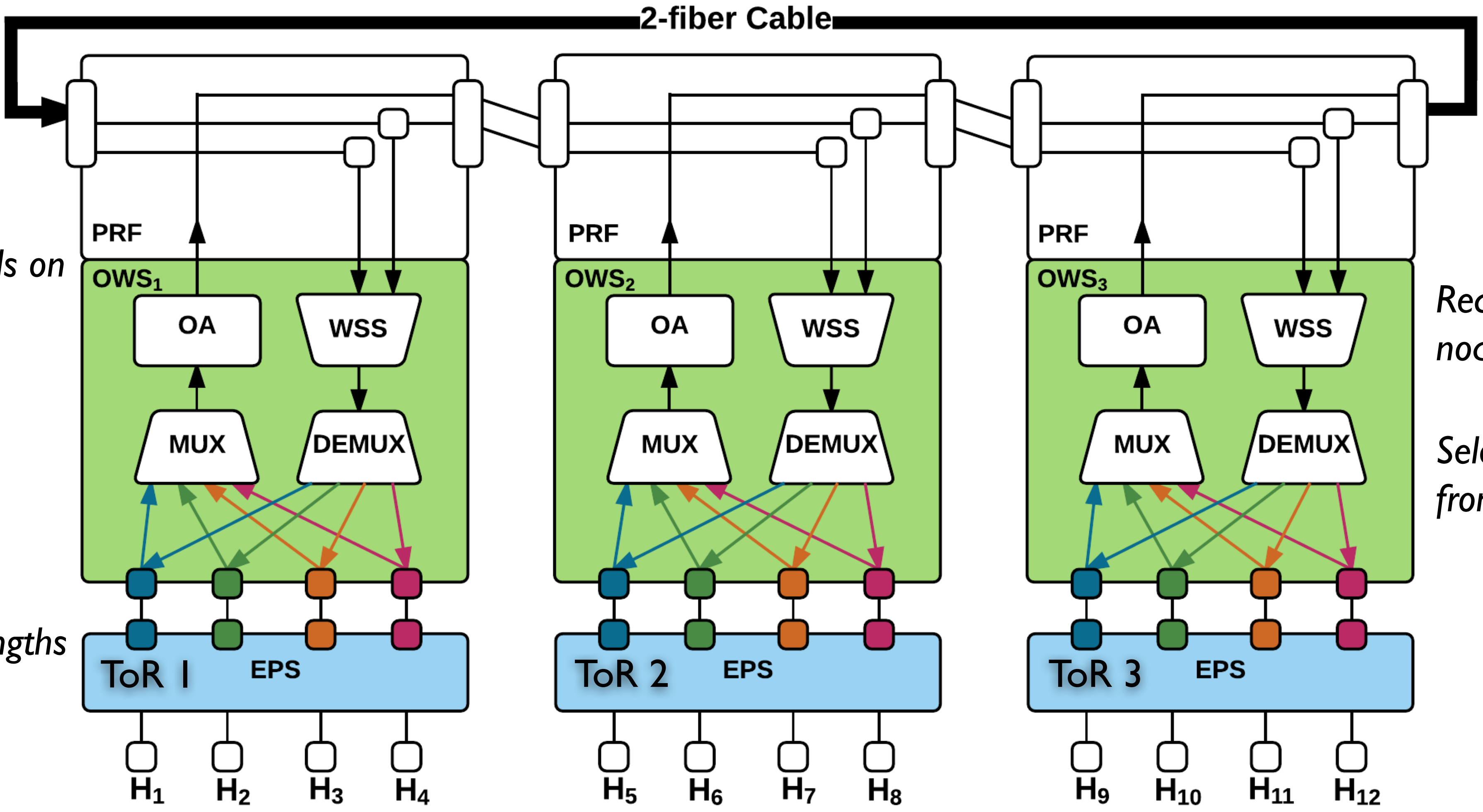
Recv from w other nodes

Select k wavelengths from w●k wavelengths

MegaSwitch: Full 3-Node Example

Each ToR sends on its own fiber

Send using k wavelengths

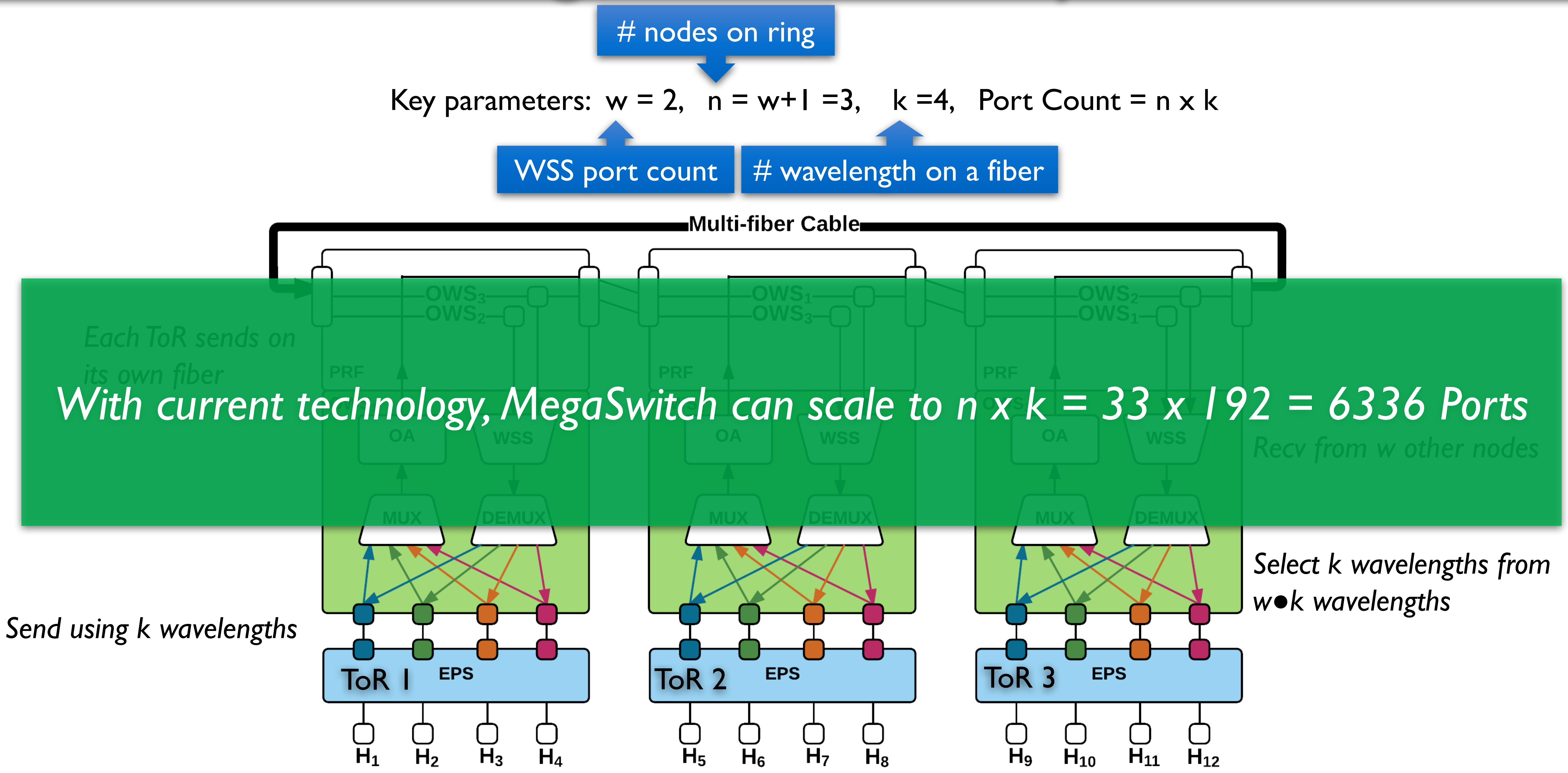


Recv from w other nodes

Select k wavelengths from w●k wavelengths

MegaSwitch: Scalability

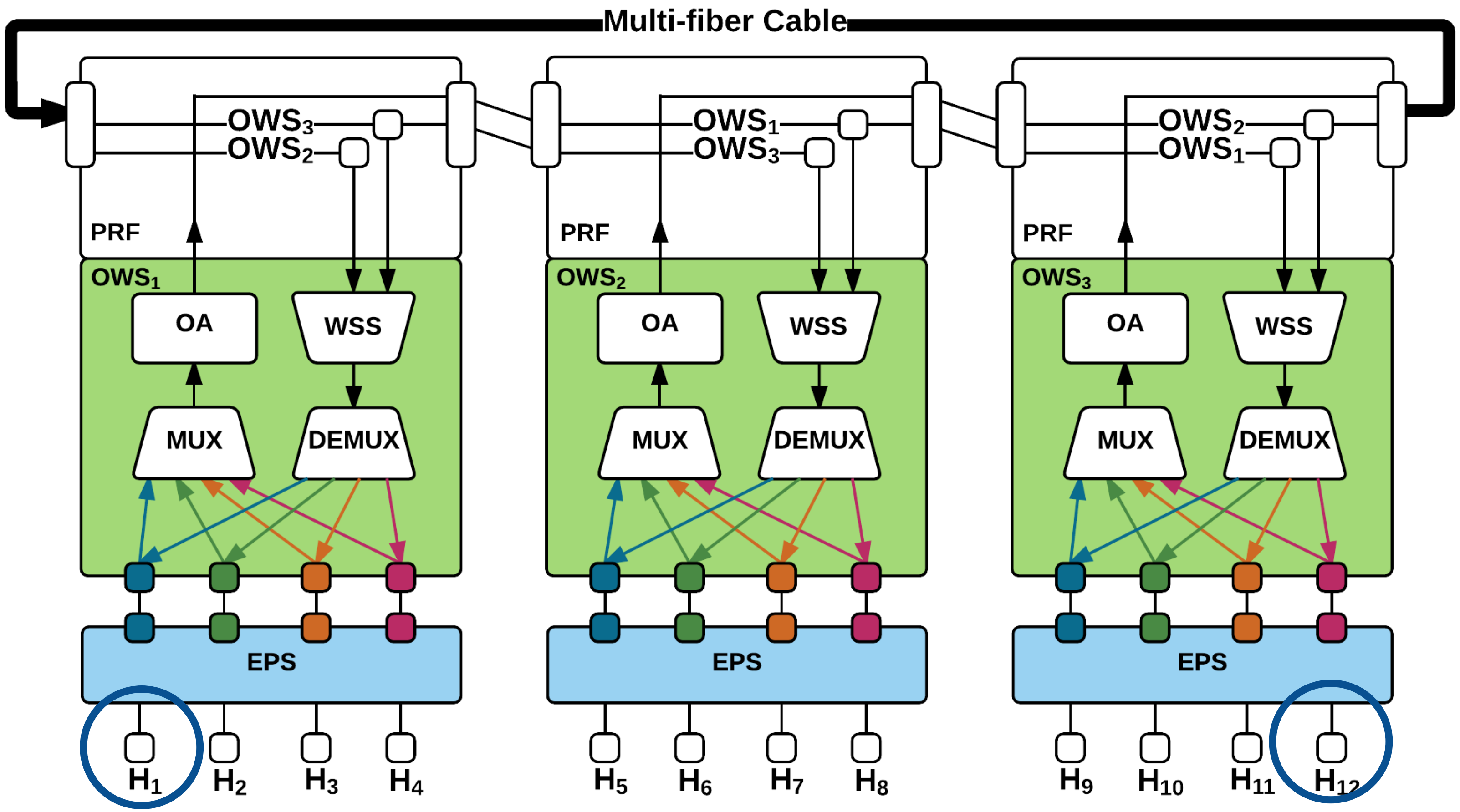
Key parameters: w = 2, n = w+1 =3, k =4, Port Count = n x k



With current technology, MegaSwitch can scale to n x k = 33 x 192 = 6336 Ports

Unicast from H_1 to H_{12}

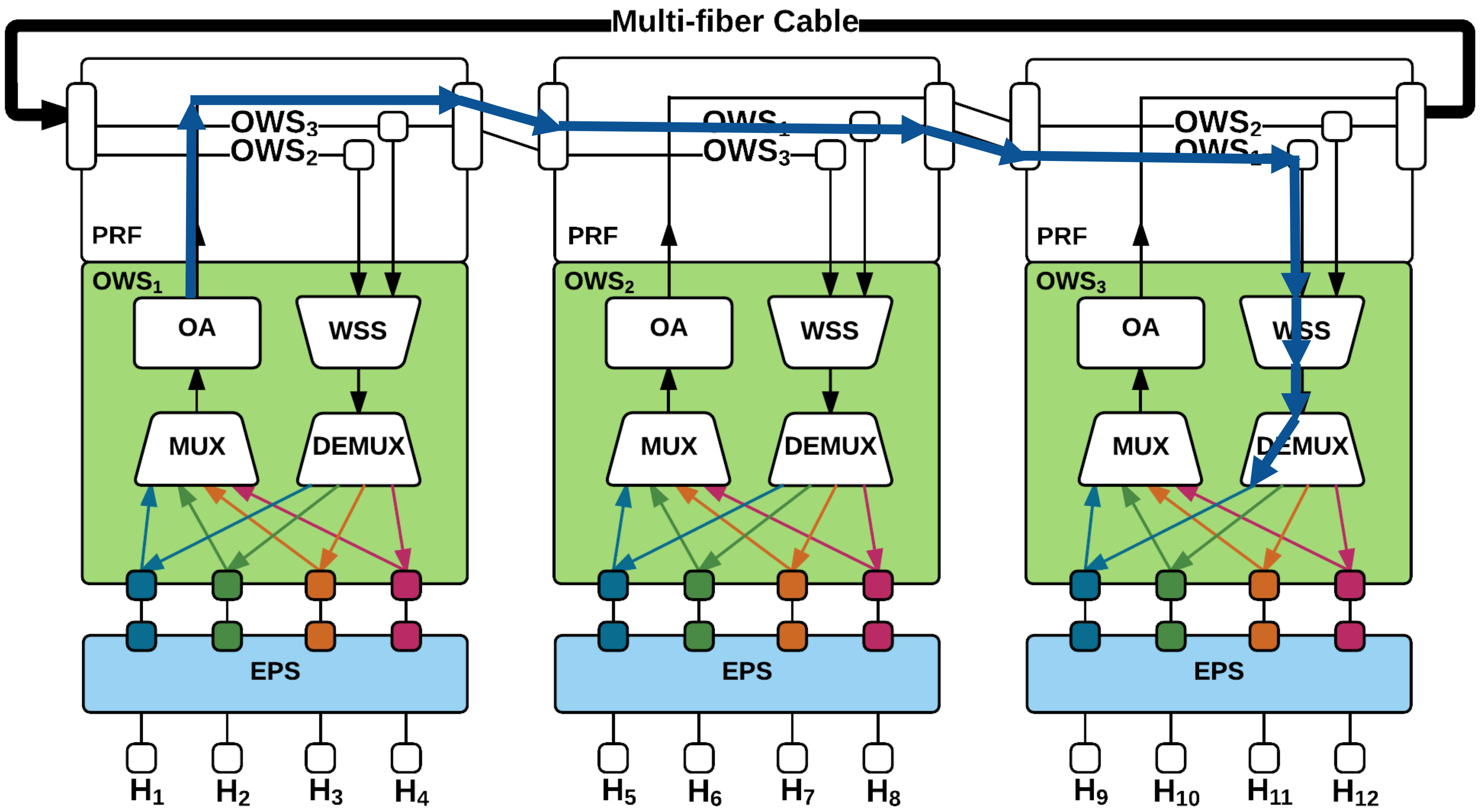
1, Control plane select Blue as the wavelength for the unicast



Unicast from H_1 to H_{12}

1, Control plane select Blue as the wavelength for the unicast

2, Configure WSS in Node 3 to select Blue in Fiber from Node 1

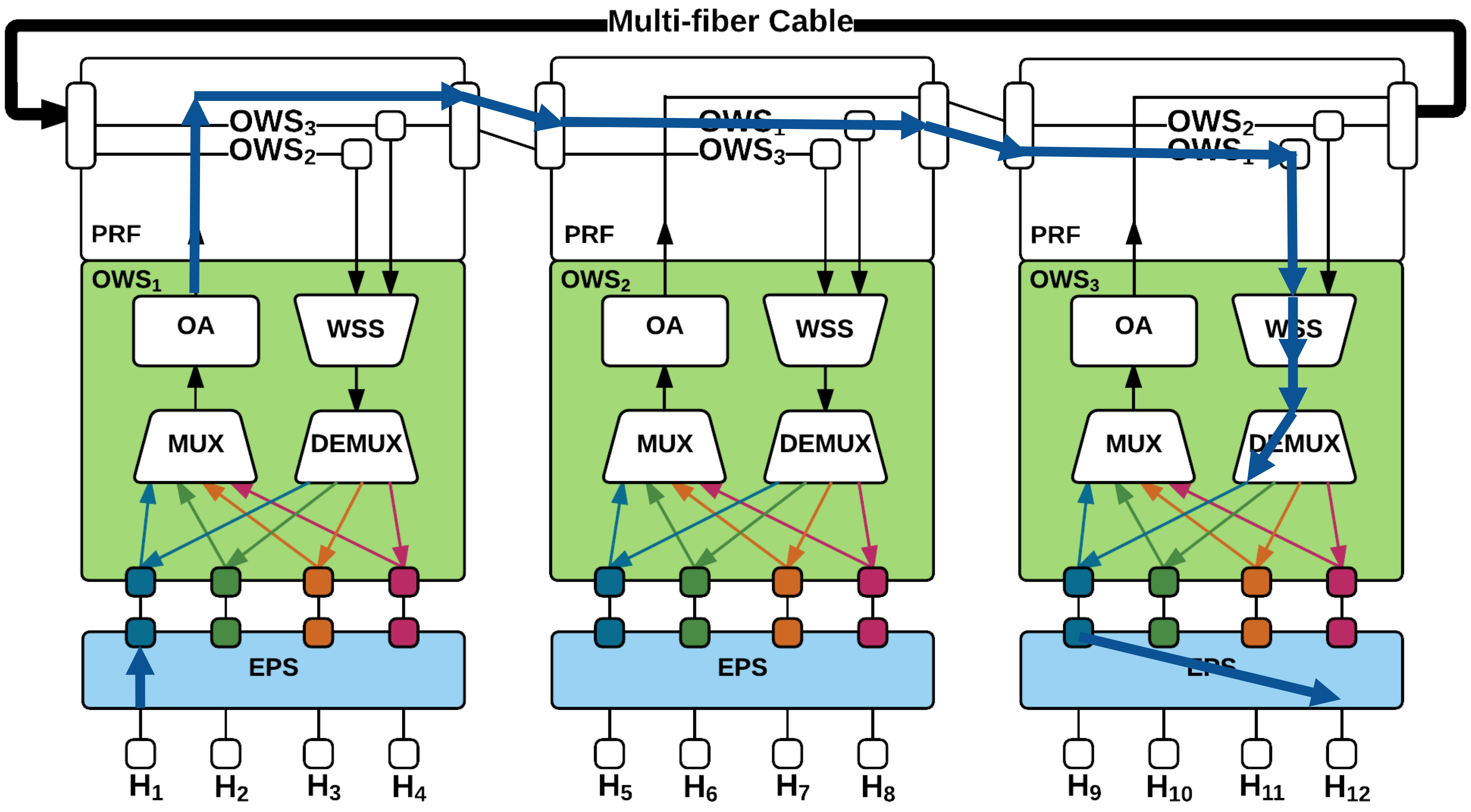


Unicast from H_1 to H_{12}

1, Control plane select Blue as the wavelength for the unicast

2, Configure WSS in Node 3 to select Blue in Fiber from Node 1

3, Setup routing in both EPSes

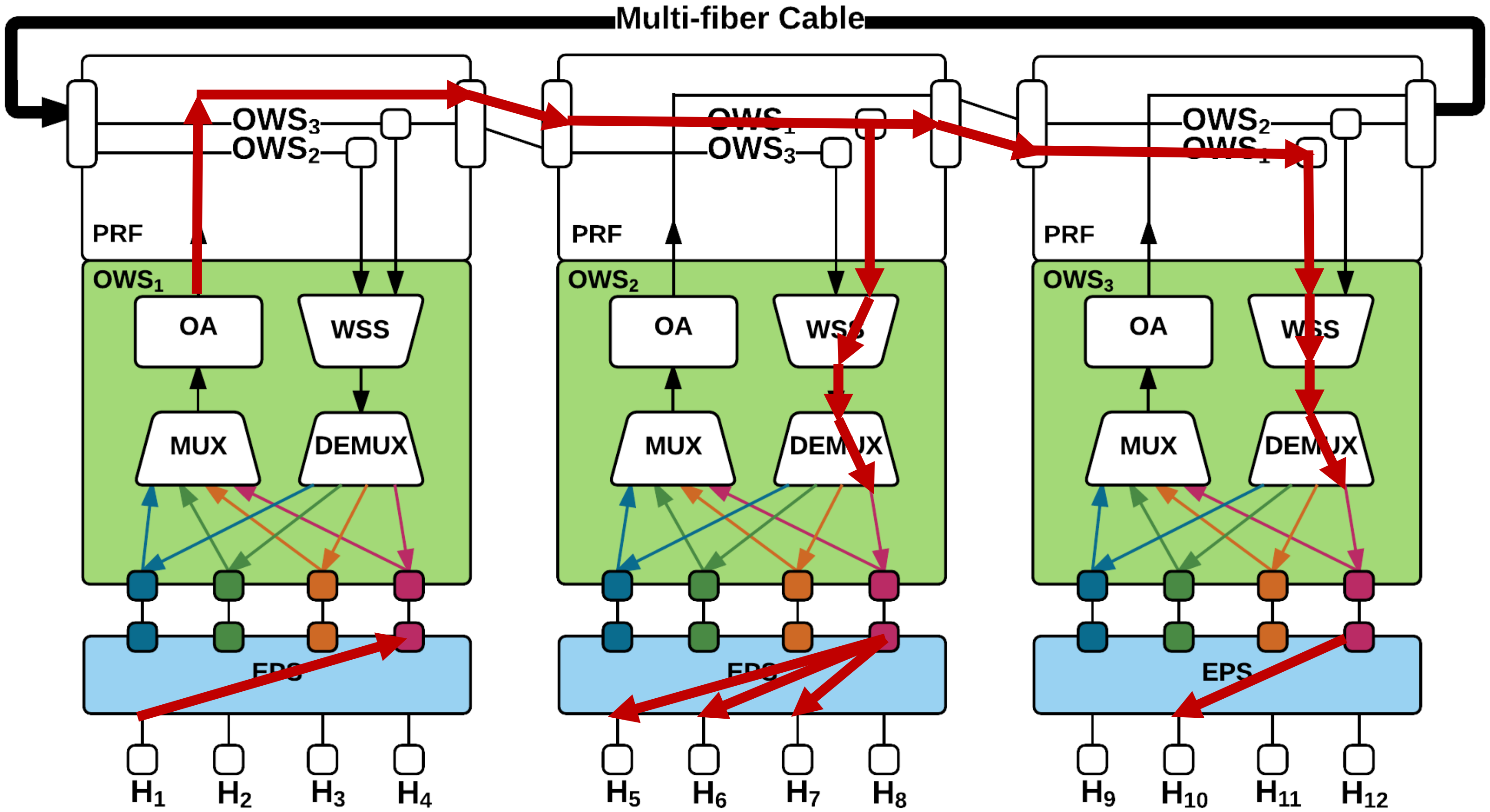


Multicast from H_1 to H_5, H_6, H_7, and H_{10}

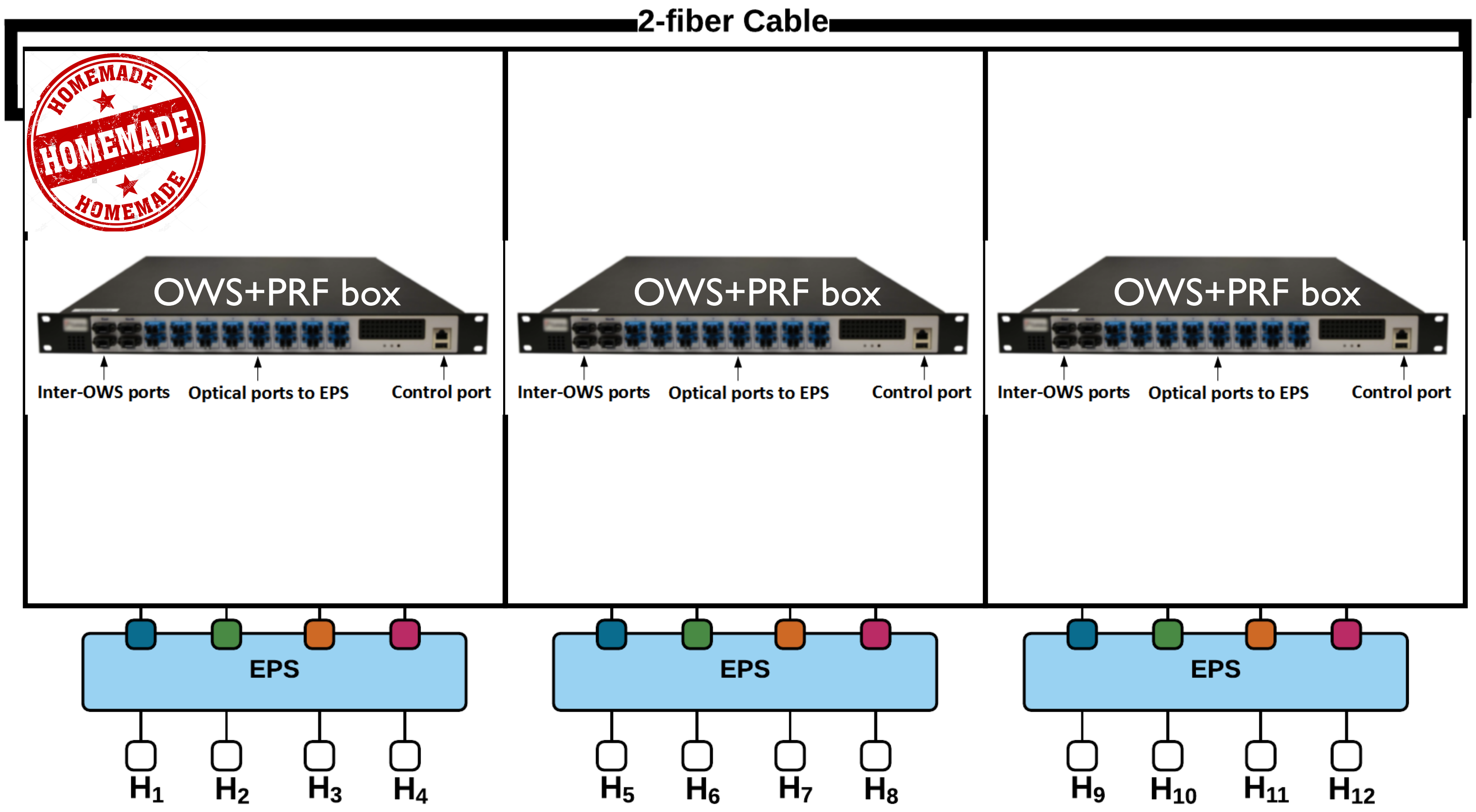
1, Control plane select Red as the wavelength for the multicast

2, Configure WSS in OWS_2 and OWS_3 to select Red in Fiber from OWS_1

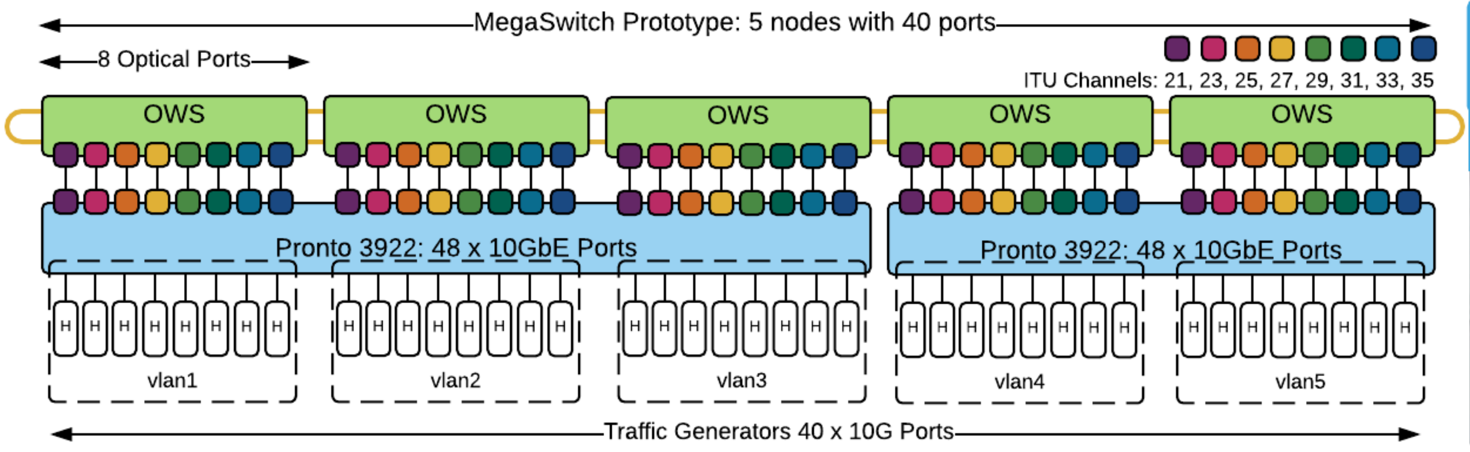
3, Setup routing in both EPSes

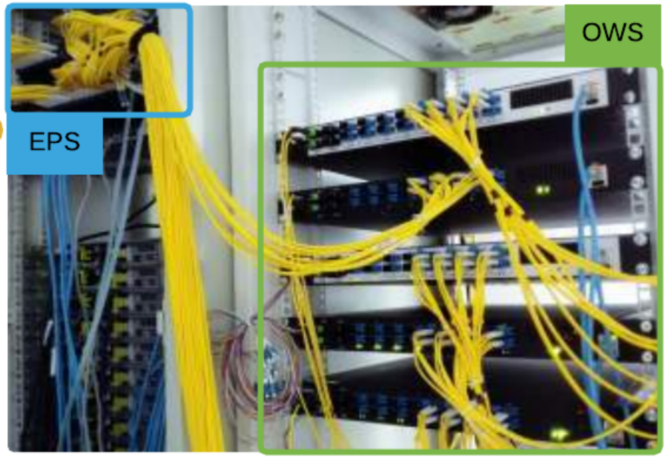


MegaSwitch: Full 3-Node Example

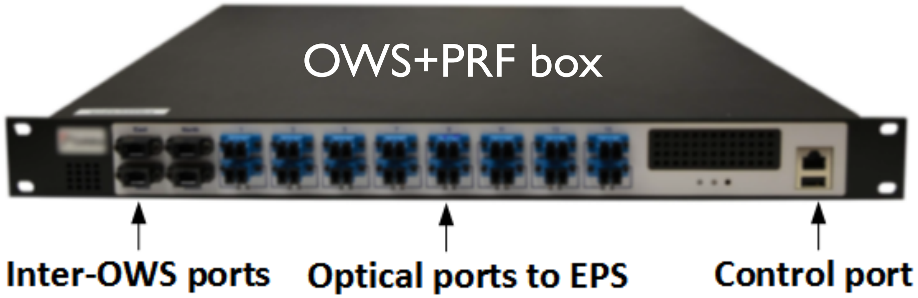


Prototype Implementation





- Implemented OWS+PRF box for practical deployment.
- Implemented prototype with 40 × 10Gbps Ports
  - 5 nodes (OWS-EPS)
  - 8 wavelengths per node



High Port Count & Low Switching Latency

…cannot be achieved at the same time…

- Lowest WSS switching latency reported: 11.5us [Mordia, Sigcomm'13]
  - Digital Light Processing (DLP) technology.
  - Cannot scale beyond 8 ports with 11.5us switching

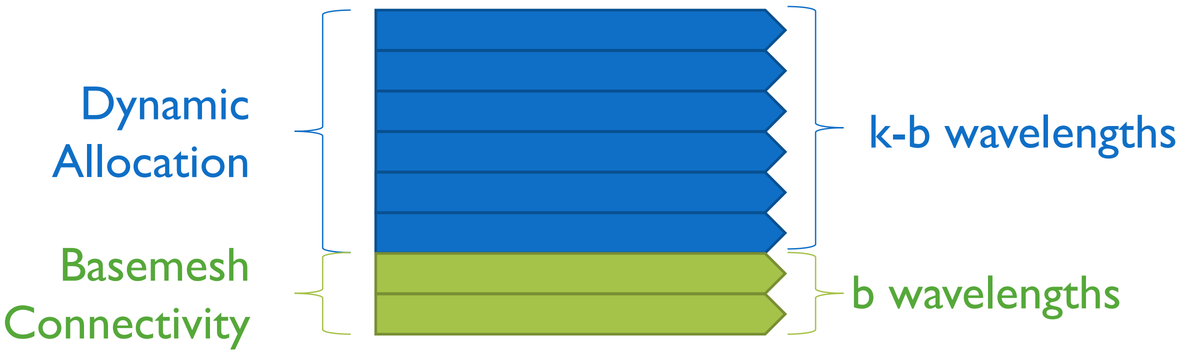
- MegaSwitch need a large WSS port count to scale to more ports
  - Liquid Crystal tech. is a middle-ground in terms of both port count (10~100s) and switching latency (milliseconds).
  - Measured WSS switching latency: ~3ms
    - Milliseconds switching latency is a hard limit for now.
    - Optics community are working on it…
  - How to mitigate impact to short flows?

MegaSwitch Control Plane

Basemesh for latency-sensitive applications

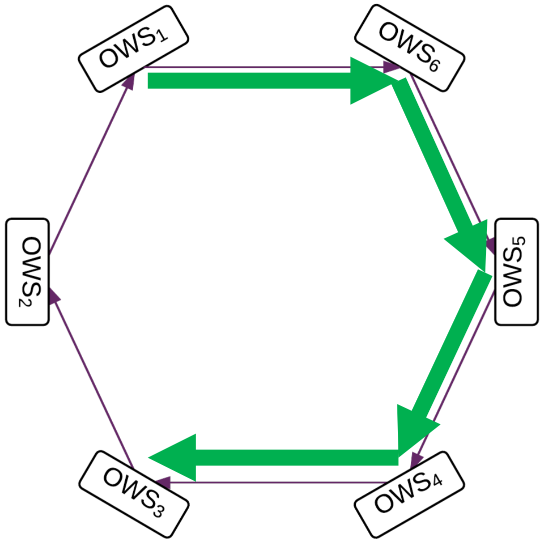
Basemesh

- Problem:
  - …when traffic matrix changes quickly
  - …when traffic matrix is estimated incorrectly
- **Basemesh**: a flexible overlay network on MegaSwitch to provide consistent connectivity for low latency traffic.
  - Each node dedicates b wavelengths to construct an overlay network on fully connected fiber mesh.

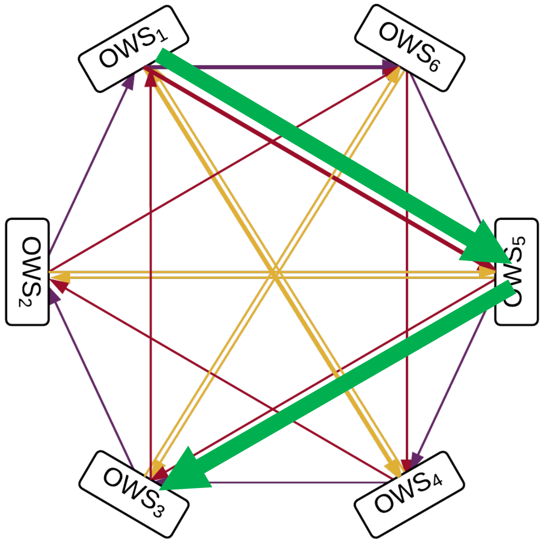


Basemesh: Learning from DHT literature

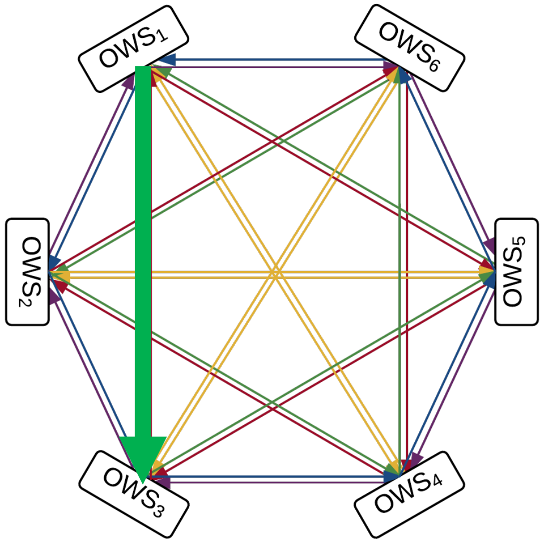
- MegaSwitch uses Symphony [USITS' 13] DHT topology for basemesh construction.
  - b ("routing table size") is adjustable for varying degree of traffic volatility
  - Guaranteed average latency ("average hop count per look-up")



Basemesh b=1
Avg Hops = 2.5



Basemesh b=3
Avg Hops = 1.4



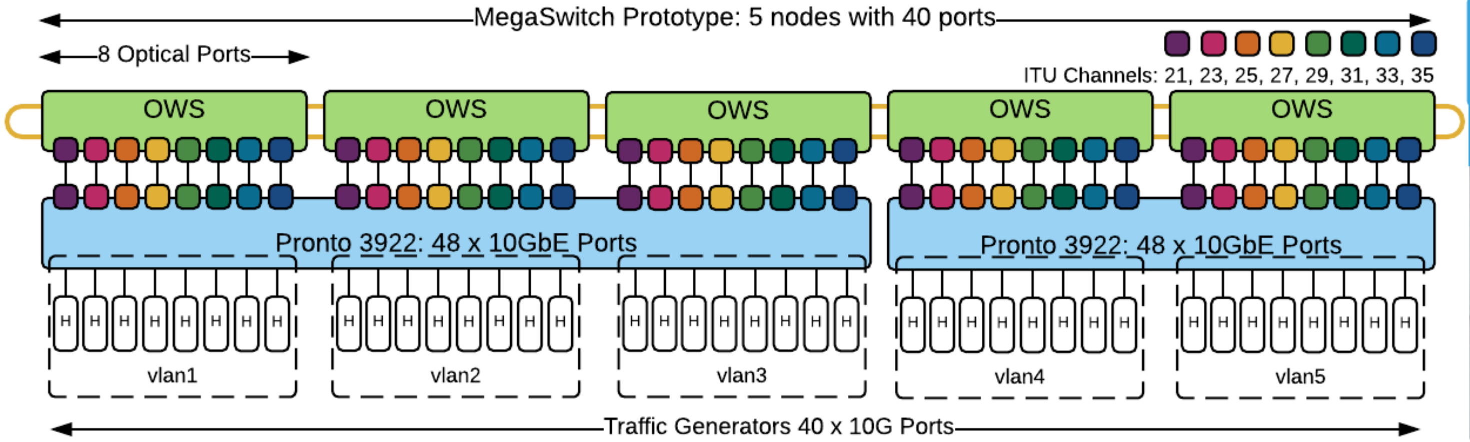
Basemesh b=5, Avg Hops =1
Fully connected mesh network (w=5)
Recommended for low latency apps

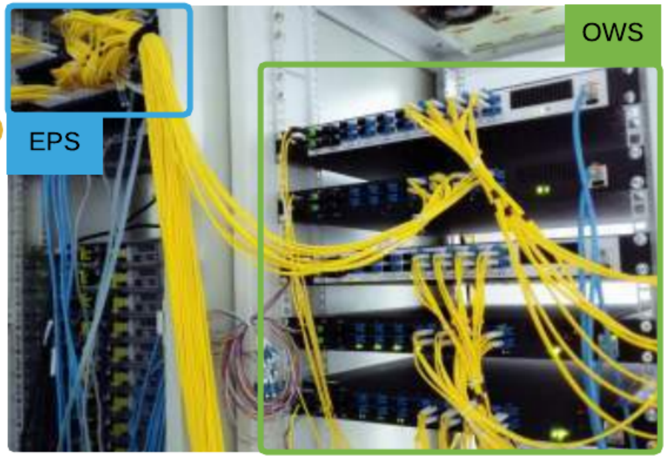
Evaluations

Testbed benchmark

Real application deployments

Prototype Evaluation

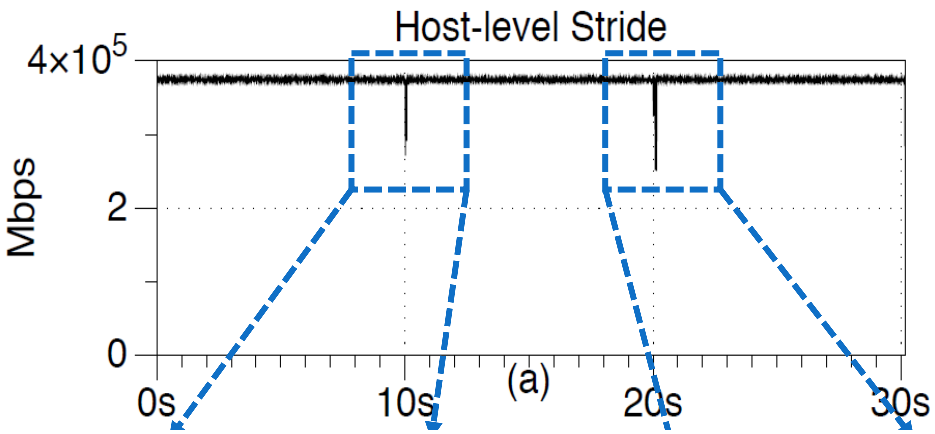




- Setting:
  - 5 nodes (OWS-EPS pair)
  - 8 wavelengths per node
  - 8 servers per node
  - Out-of-band control plane for EPS and OWS
  - Traffic demand matrices are known

Basic measurements

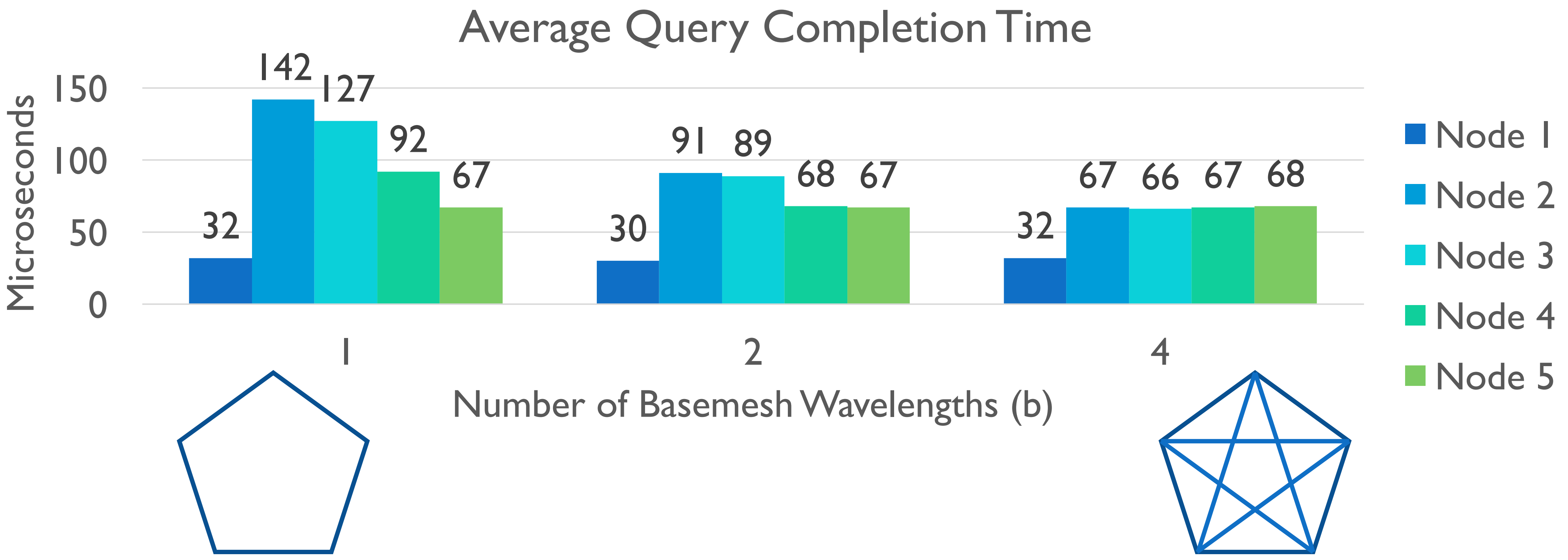
- Host-level stride
  - All-to-all pattern [Helios, Sigcomm'10]
  - Every 10s, one wavelength changes per rack.
- Measured ~20ms total reconfiguration delay.
  - WSS (~3ms), EPS routing (~5ms), transceiver initialization (~10ms)…



MegaSwitch achieves full-bisection bandwidth when circuit is stable

Redis on MegaSwitch

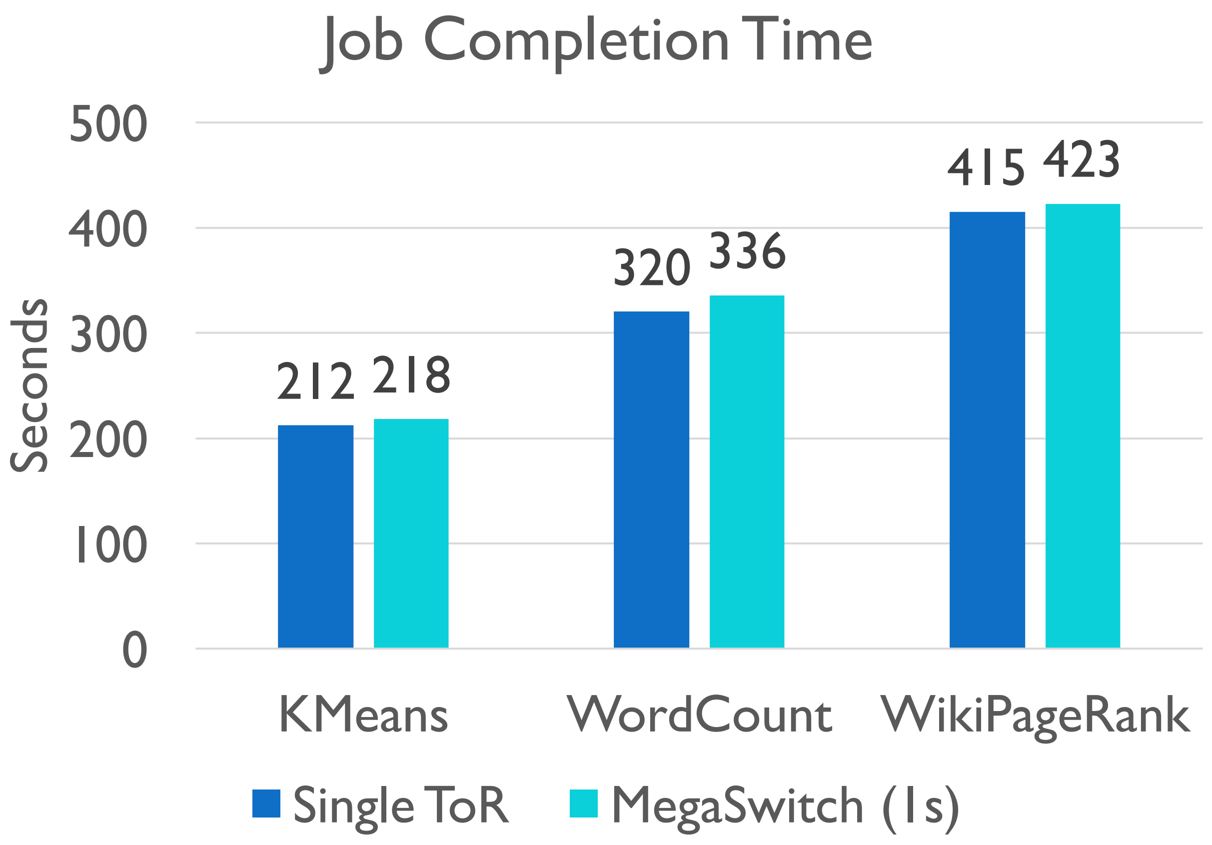
- Latency-sensitive application
  - 1 Million GET/SET requests from all nodes to a server in Node1



Fully connected basemesh → Uniform latency (one-hop)

Apache Spark on MegaSwitch

- Parallel computing applications
- First connect the servers to a single ToR switch, and measure the bandwidth demand
- MegaSwitch updates wavelength assignment every 1 sec.
  - <10 reconfigurations in run-time.



MegaSwitch performs similar to the optimal scenario: all servers in the same rack

Summary

- Spatial reuse of wavelengths to provide non-blocking connectivity for all ports.
- Basemesh to provide consistent connectivity to latency-sensitive flows.
- Practical implementation of a 5-node ring of 40 ports.

MegaSwitch: An optical design that supports wide-spread, high-bandwidth traffic patterns in today's production workloads.

More in our paper: Fault-tolerance, delay measurements, power budget, cost…

Got New Ideas for NSDI'18? Test them in APNet'17!

- The first Asia-Pacific Workshop on Networking
  - Aug. 3-4, 2017 @Hong Kong
- A good venue to test your innovative ideas and get feedback from the community.
- Submit your 6-page paper on/before Apr. 21th, 2017