On the Performance of a Large-Scale Optical Packet Switch Under Realistic Data Center Traffic

Speaker: Lin Wang

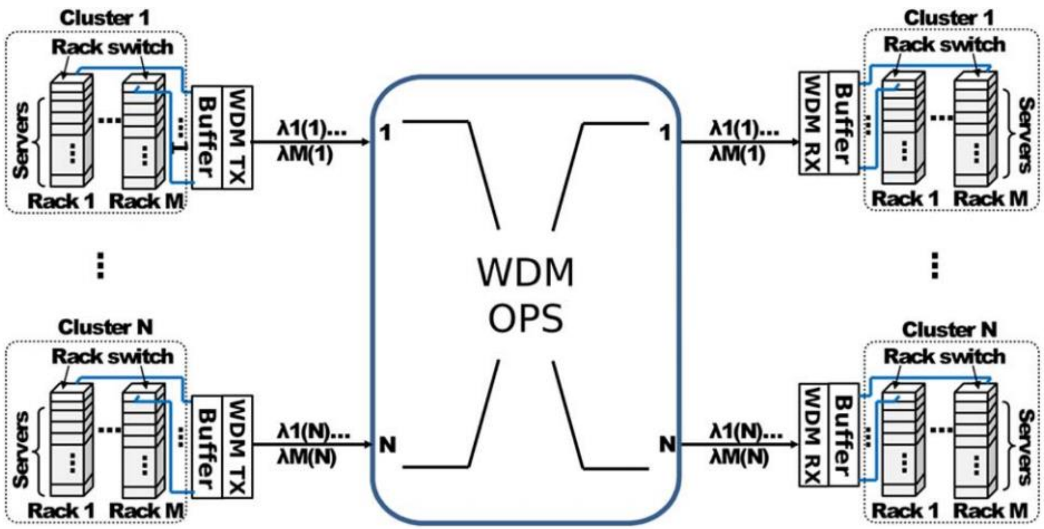
Research Advisor: Biswanath Mukherjee

UC DAVIS

- **Switch architectures and control**
- **Traffic generation**
- **Simulation set up**
- **Results evaluation**
- **PSON architecture**

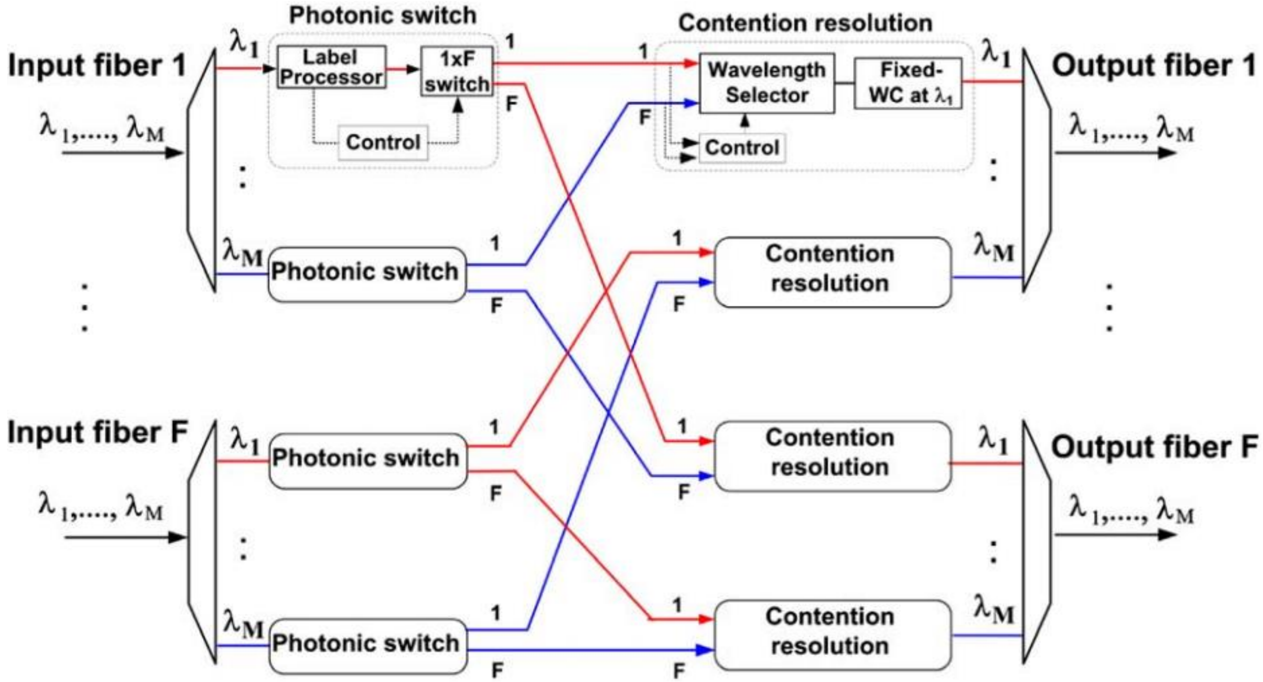
Calabretta, Nicola, et al. "On the performance of a large-scale optical packet switch under realistic data center traffic," *Journal of Optical Communications and Networking* vol. 5, num. 6, pp.:565-573, 2013.

Switch architecture



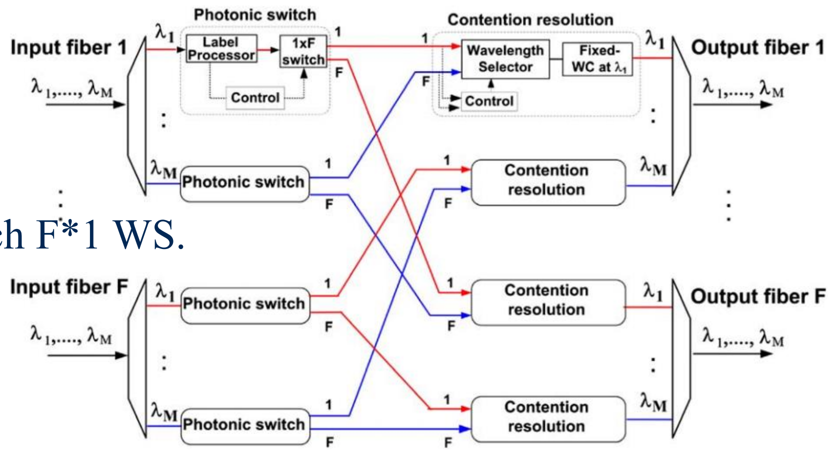
- Wavelength-division multiplexing Optical packet switching (WDM OPS) is based on a strictly nonblocking Spanke-type architecture.
- No centralized control increase the scalability.

Block diagram of the OPS architecture



- WDM OPS architecture with distributed control.
- Input ports is N=F*M.
- Assume the distance between clusters and OPS is 50m.
- Photonic switch has a local control and a paralleled 1*F switch.
- Contention resolution block (CRB) has a F*1 wavelength-selector (WS).

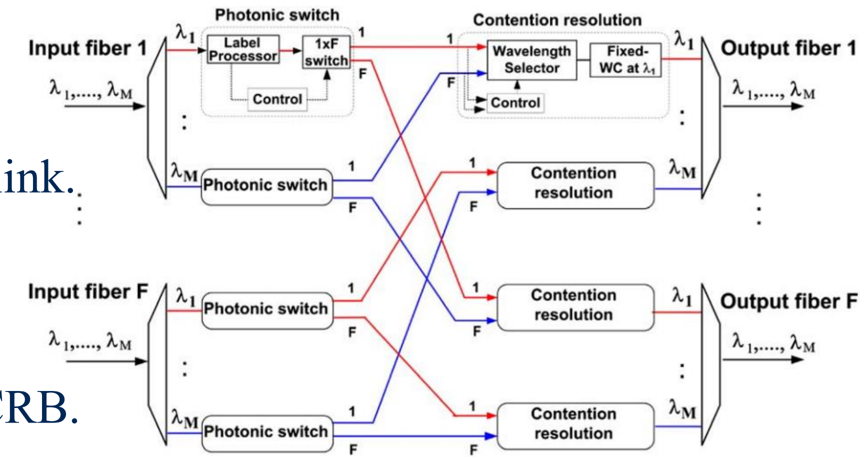
OPS processing



- Contentions occur only between the F input ports of each F*1 WS.
- Optical packet:
  - ❖ Payload carries real data.
  - ❖ Optical label shows the destination.

- Label processor controls the 1*F switch to forward the optical packet to one of F output ports to CRB.
- CRB use M*1 WS and fixed wavelength converters (FWCs) to avoid collisions.
- Outputs of CRB switches reach the destination clusters by optical link.
- M WDM channels are detected by O/E converters at destination clusters.
- Optical packets are converted, buffered and forwarded to M TOR switches.
- Control complexity and configuration time mainly depend on label processing time.

OPS processing

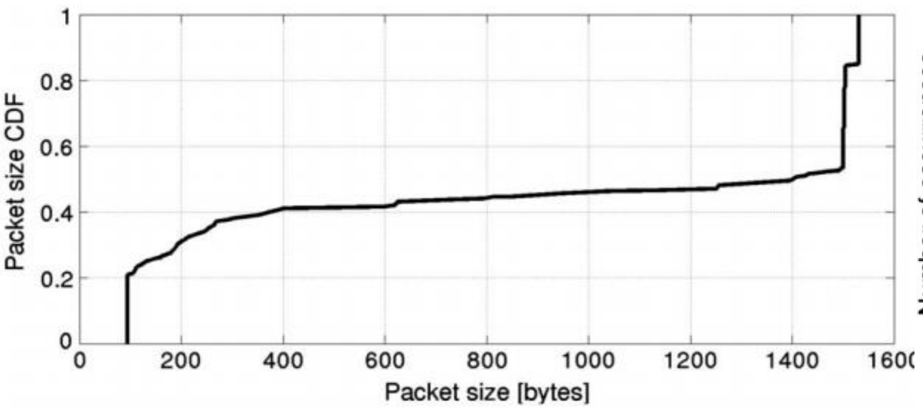


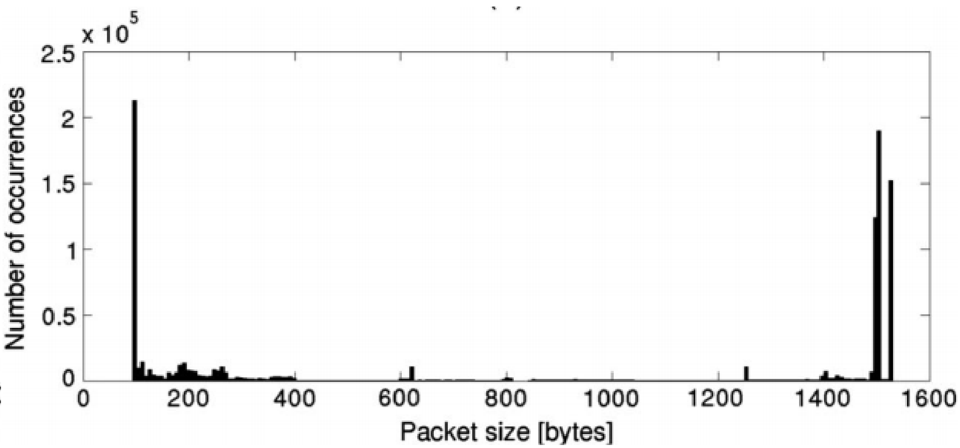
1. Packets are stored in electronic buffer.
2. A copy is sent to the OPS via a 50m optical link.
3. At OPS node
    a. optical label is processed.
    b. Photonic switch is reconfigured.
    c. Forward the packet to the appropriate CRB.
4. At CRB
    a. When packet arrives, check collision.
    b. If no collision, forward it to output port connected with destination cluster.
    c. If two or more packets coming from the same input fiber have the same destination and reach CRB simultaneously, collision happens.
    d. Only one packet is delivered while others are abandoned.
5. At input node, only successfully delivered packets will be acknowledged and erased while others need retransmission.
6. If the input buffer is full, new packets will be dropped which leads to packet loss.

UCDAVIS
UNIVERSITY OF CALIFORNIA

Traffic generation

- Each of the M wavelengths in each cluster receives the input traffic generated by 200 simulated servers.
- The amount of traffic load is normalized and can be scaled from 0 to 1.
- Packet length in real scenarios is mostly found to be a bimodal distribution around 40 bytes and 1500 bytes . [1]-[3]

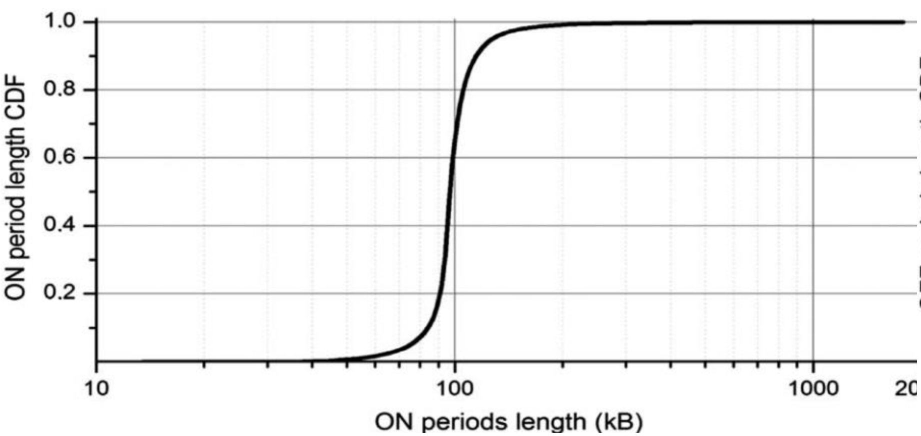


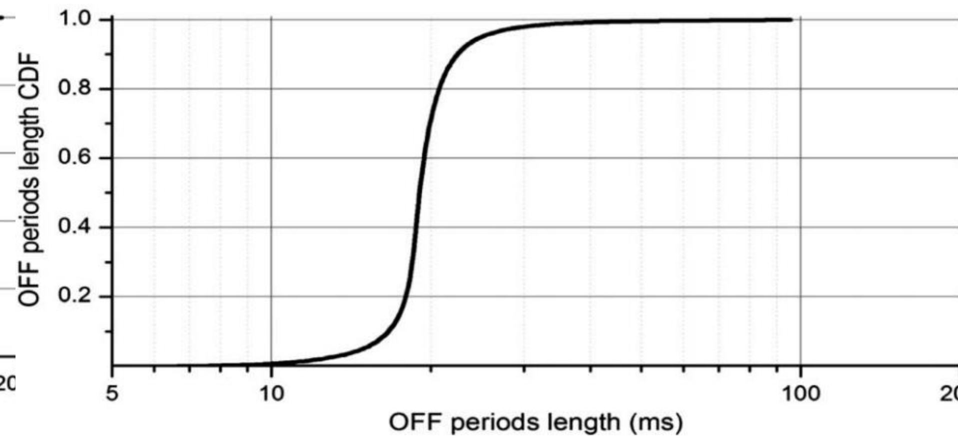


[1] T. Benson, A. Anand, A. Akella, and M. Zhang, “Understanding data center traffic characteristics,” Comput. Commun. Rev., vol. 40, no. 1, pp. 92–99, 2010.
[2] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in Proc. Internet Measurement Conf. (IMC), Melbourne, Australia, Nov. 2010, pp. 267–280.
[3] S. Kandula, S. Sengupta, A. Greenberg, A. Patel, and R. Chaiken, “The nature of datacenter traffic: measurements & analysis,” in Proc. of the 9th ACM SIGCOMM Internet Measurement Conf. (IMC’09), 2009, pp. 202–208.

Traffic generation (Cont.)

- Packet arrival times are modeled matching ON/OFF periods.
- ON/OFF periods follows Pareto distribution.
- ON periods follow the same length distribution regardless of load.
- OFF periods is proportional to the chosen simulation load value.

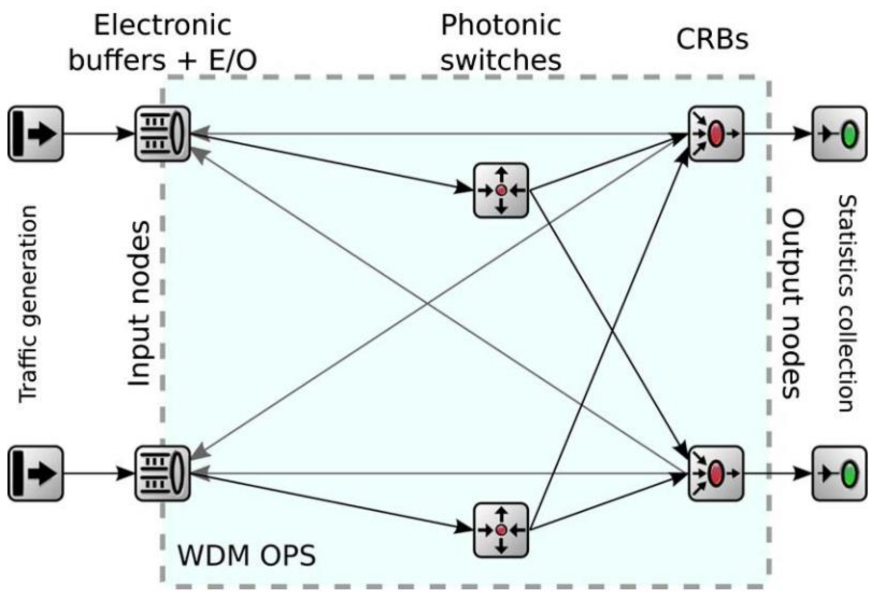




[1] T. Benson, A. Anand, A. Akella, and M. Zhang, “Understanding data center traffic characteristics,” Comput. Commun. Rev., vol. 40, no. 1, pp. 92–99, 2010.
[2] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in Proc. Internet Measurement Conf. (IMC), Melbourne, Australia, Nov. 2010, pp. 267–280.
[3] S. Kandula, S. Sengupta, A. Greenberg, A. Patel, and R. Chaiken, “The nature of datacenter traffic: measurements & analysis,” in Proc. of the 9th ACM SIGCOMM Internet Measurement Conf. (IMC’09), 2009, pp. 202–208.

Simulation Sets

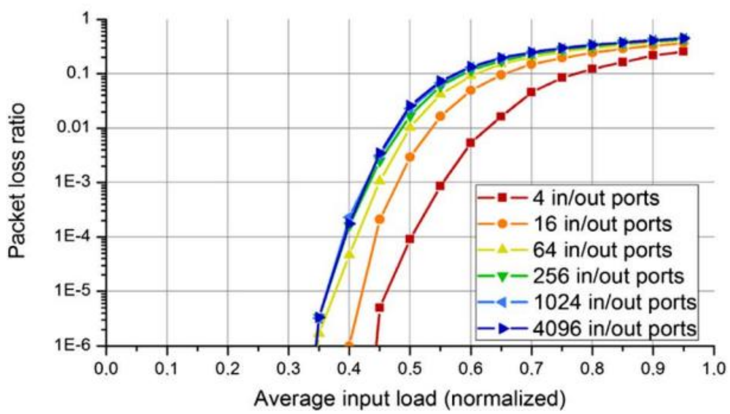
- Use OMNeT++ Network Simulation Framework Software.[4]
- Figure shows the block diagram after the architecture is implemented in the simulation software.
- Data rate is 40Gbit/s.
- M=F=32.
- Delay for optical modules, translates into an RTT is 560ns.

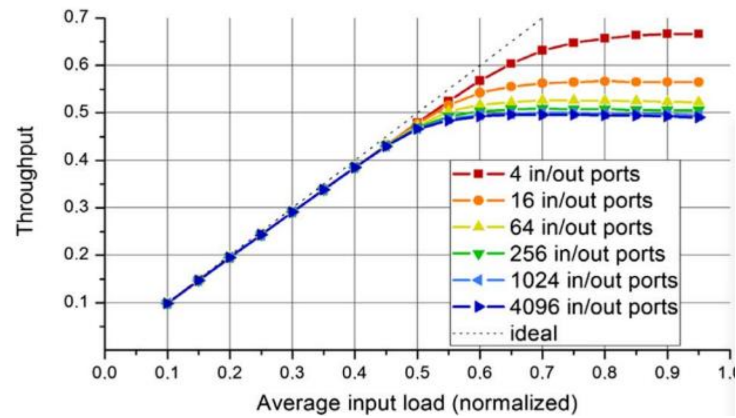


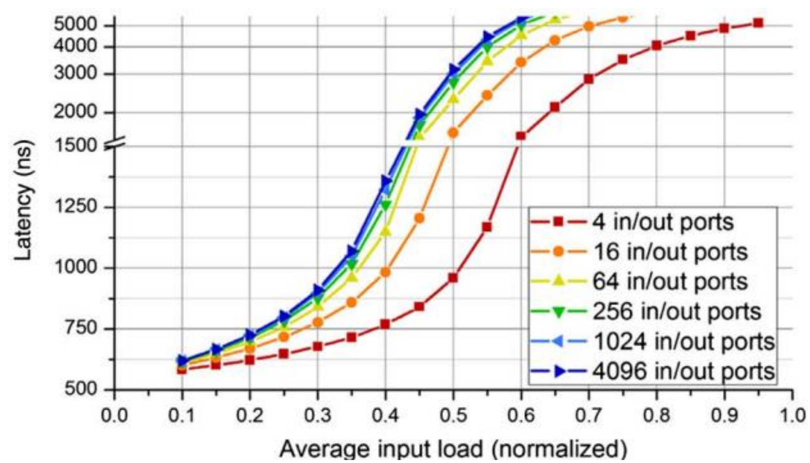
[4] OMNeT++ Network Simulation, http://www.omnetpp.org.

Simulation Results

- Figures below show the effects of increasing number of ports.

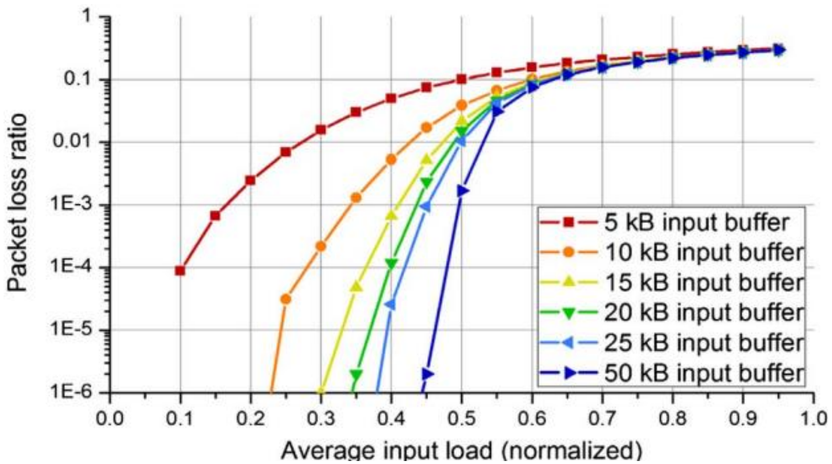


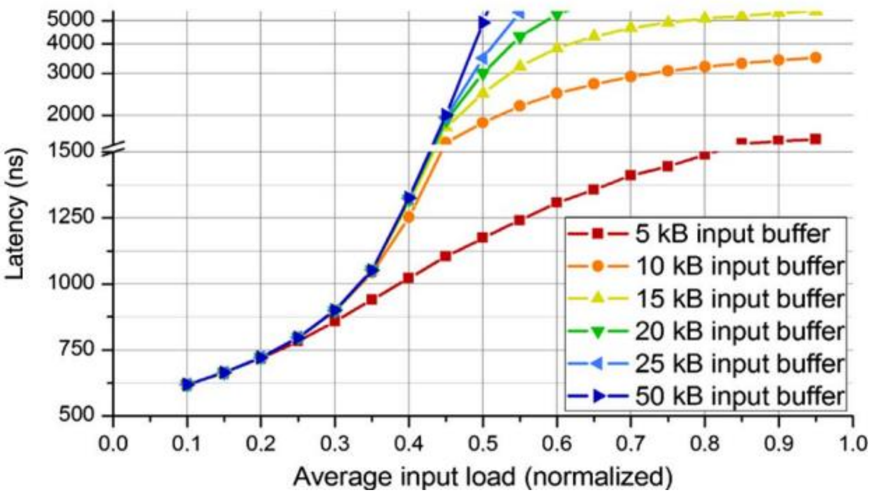




Simulation Results

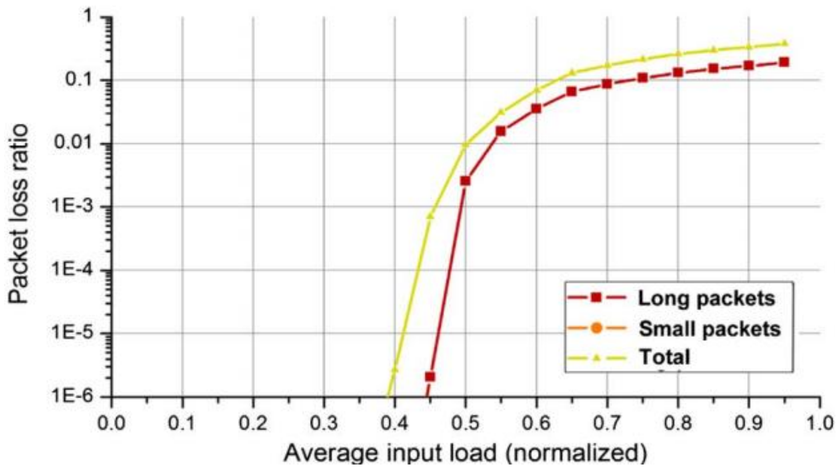
- Figures below show the effects of electrical buffers.

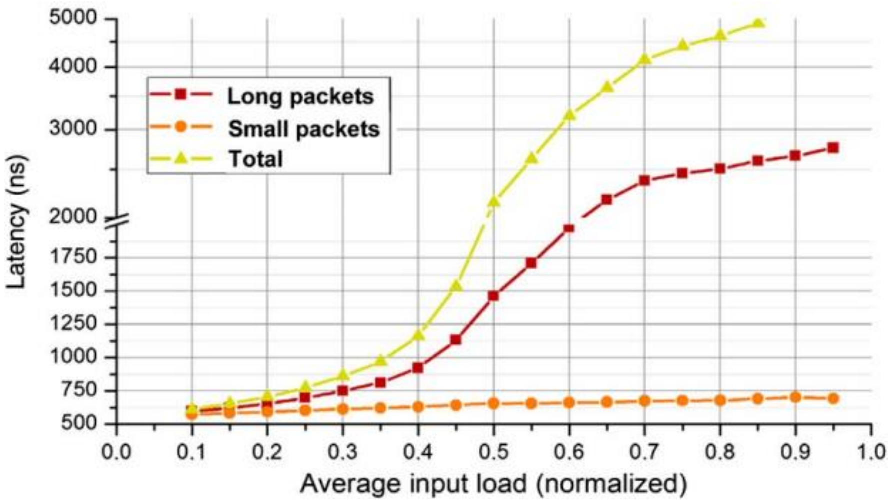




Improved OPS architecture

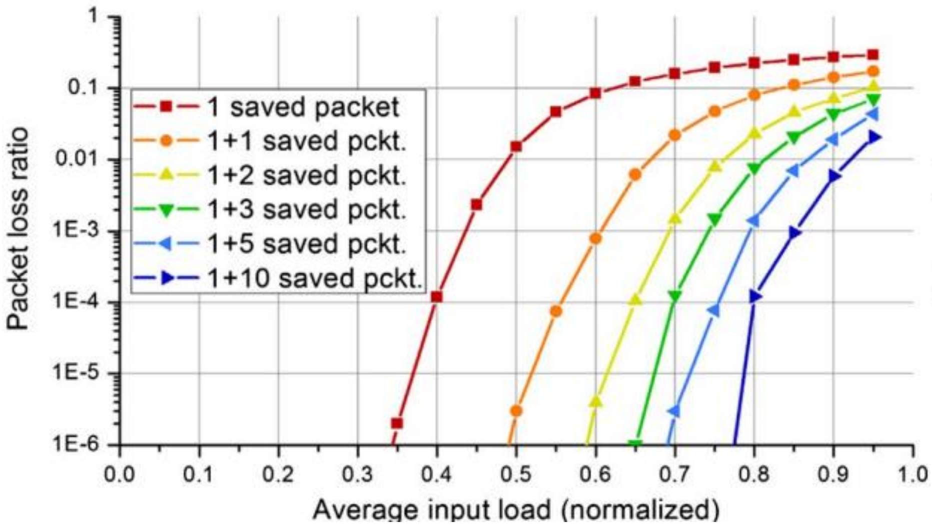
- As the packet length is a bimodal distribution around 40 and 1500 bytes, they propose to use two different OPSs.
- Diverge packets arriving to the clusters to two distinct buffers, one for short packets (5KB) and the other for long packets (15KB).

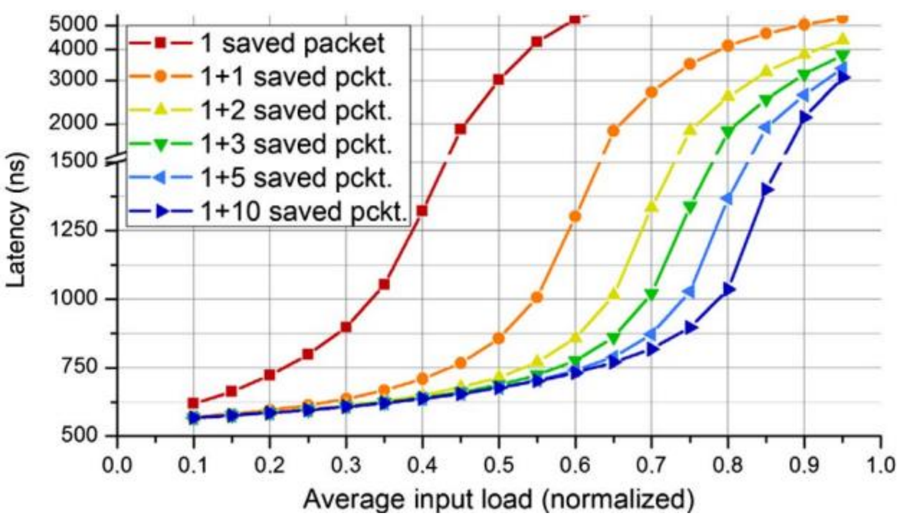




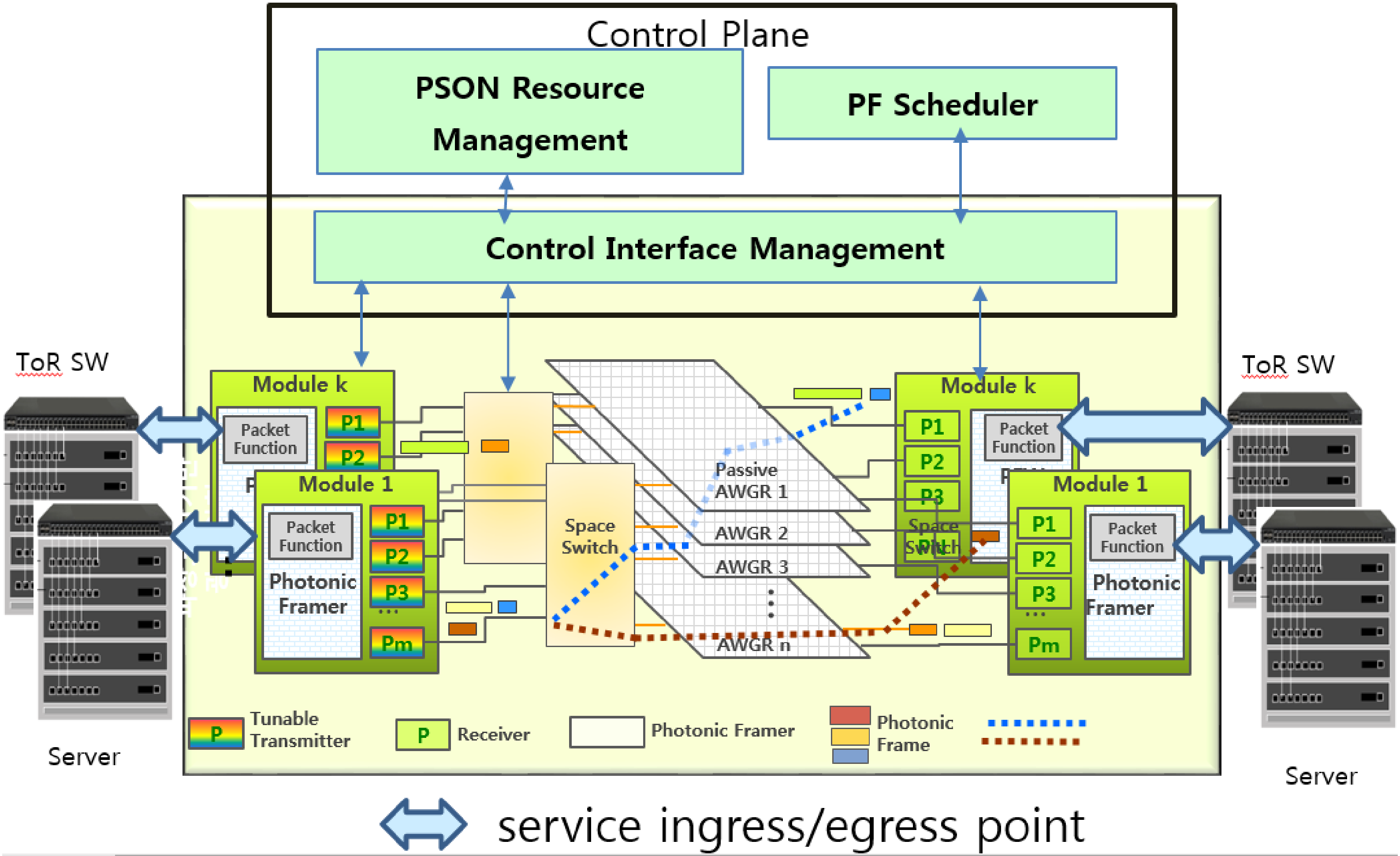
OPS Architecture With Multiple Receivers

- At the CRB, only one packet at a time can be saved and forwarded to the output. The other packets are simply retransmitted at a later time.
- Save more packets in case of contention by using multiple receivers at each output port of the OPS.





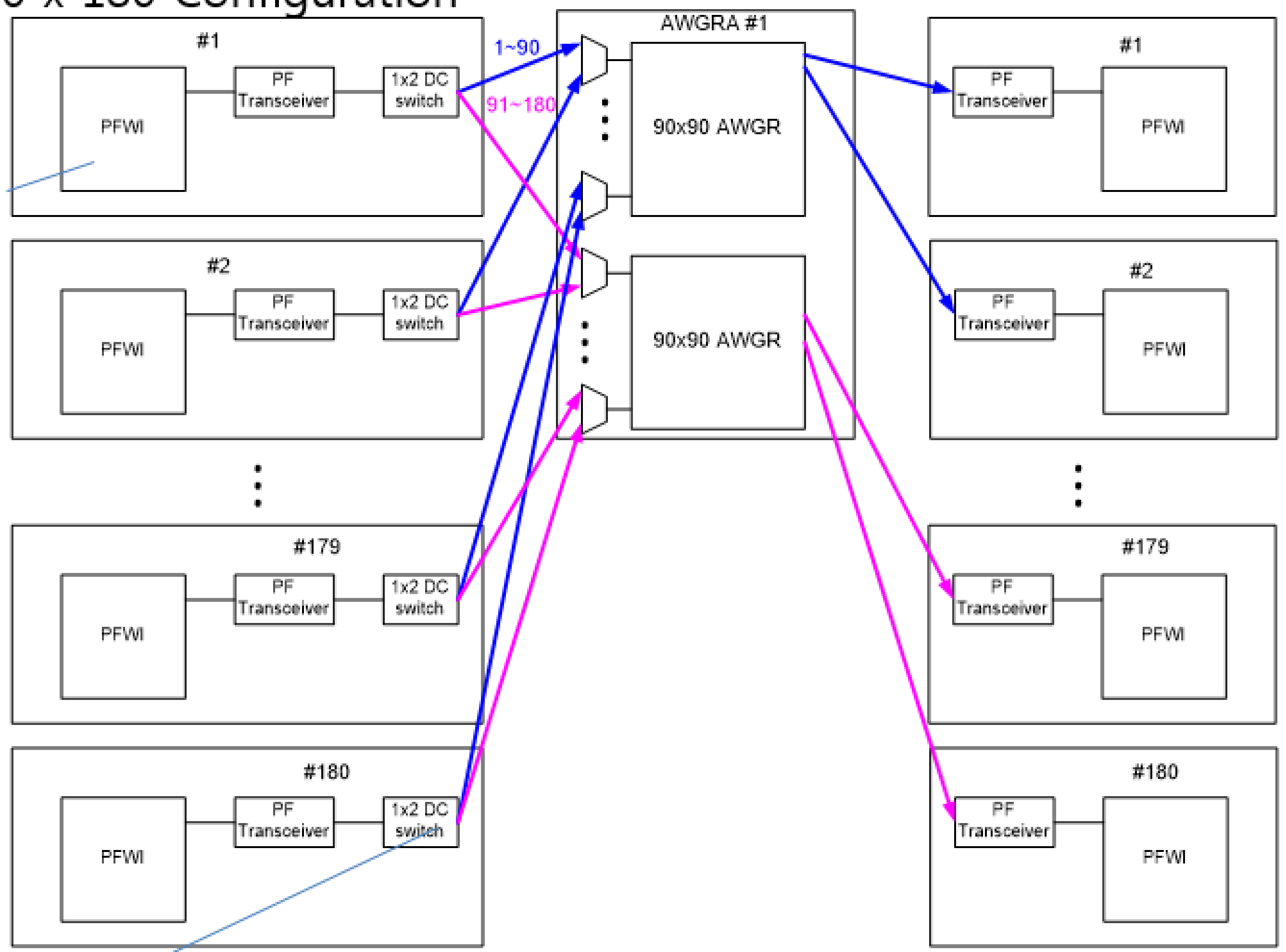
PSON Architecture



A control plane manages tunable transmitters, photonic framers and space switches for data plane with optical switch fabrics (AWGR)

PSON data plane (with optical switch fabric)

- ➢ 180 x 180 Configuration



Photonic Framer

Space switch: Optical path switch to switch optical signal between module and AWGR at sub micro-second speed.

PSON architecture performance evaluation

- **Scheduling algorithm design**
  - ❖ Consider the space switch delay and transmitter tuning delay when we schedule the bandwidth resource to each module.
- **Performance Evaluation**
  - ❖ Effect of frame size
  - ❖ Effect of buffer size
  - ❖ Effect of in/out ports numbers
- **Further work**
  - ❖ Improve the architecture.
  - ❖ Evaluate the performance of modified architecture.

amlwang@ucdavis.edu