Dos-A Scalable Optical Switch for Datacenters

Speaker: Lin Wang

Research Advisor: Biswanath Mukherjee

Ye, X. et al., "DOS: A scalable optical switch for datacenters," *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*. ACM, 2015.

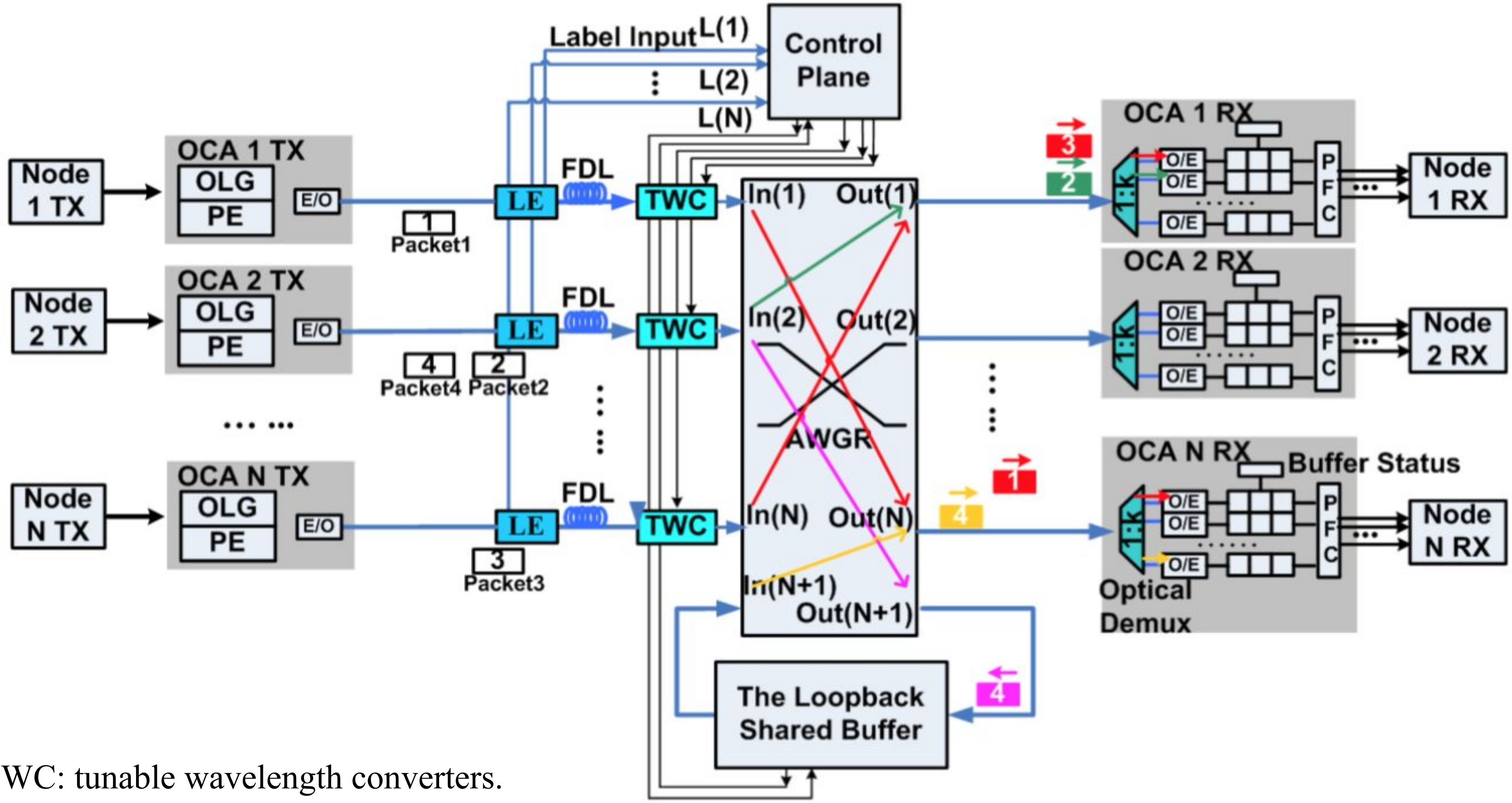
UCDAVIS

Major Differences compare (Telecom vs Datacenters)

- **More latency reduction is required for data center applications (100’s of nanoseconds as opposed to 10’s or 100’s of microseconds).**

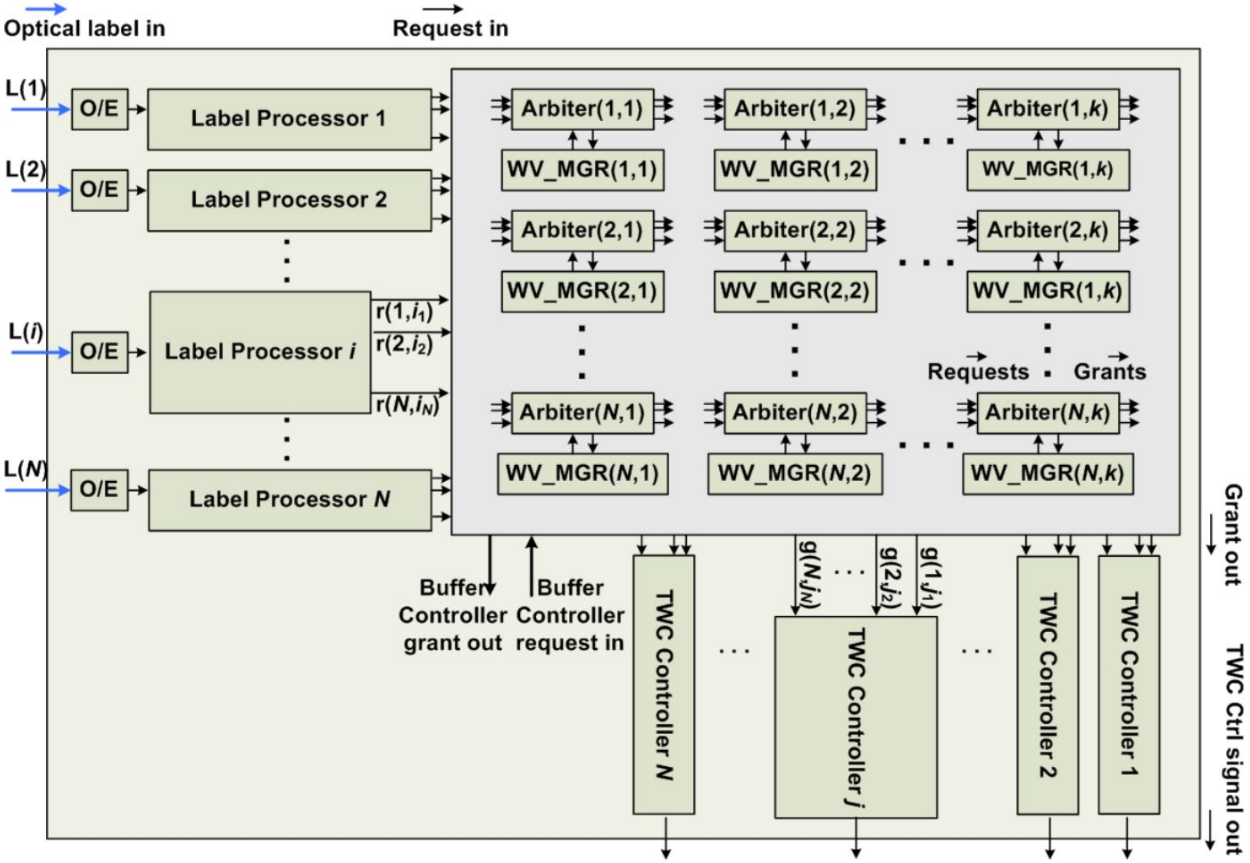
- **Data center switches need to connect many more nodes (e.g. hundreds or thousands in large data centers).**

Datacenter optical switch (DOS) architecture



TWC: tunable wavelength converters.
ULCF: uniform loss and cyclic frequency AWGR.
Loopback shared buffer.
Control block: process label and arbitration by checking resource availability.
OCA: optical channel adapter.

DOS Control Plane



Problem: If each RX only has k receivers, then no more than k packets on different wavelengths can be received successfully.

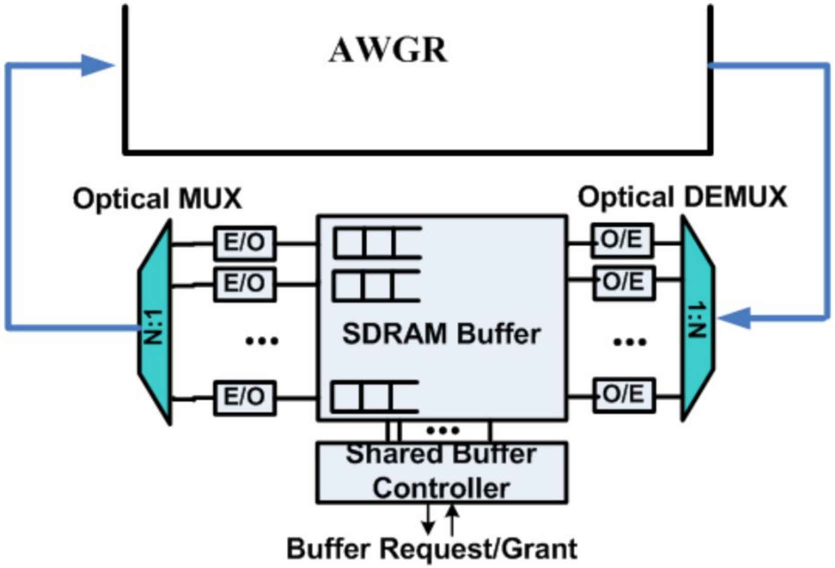
Solution: Define a wavegroup as a set of wavelengths that will come out from the same output port of the optical RX. Therefore, arbitration is necessary to guarantee that at most k packets will arrive at the AWRG output port in one cycle.

Shared SDRAM Buffer

Why need buffer?

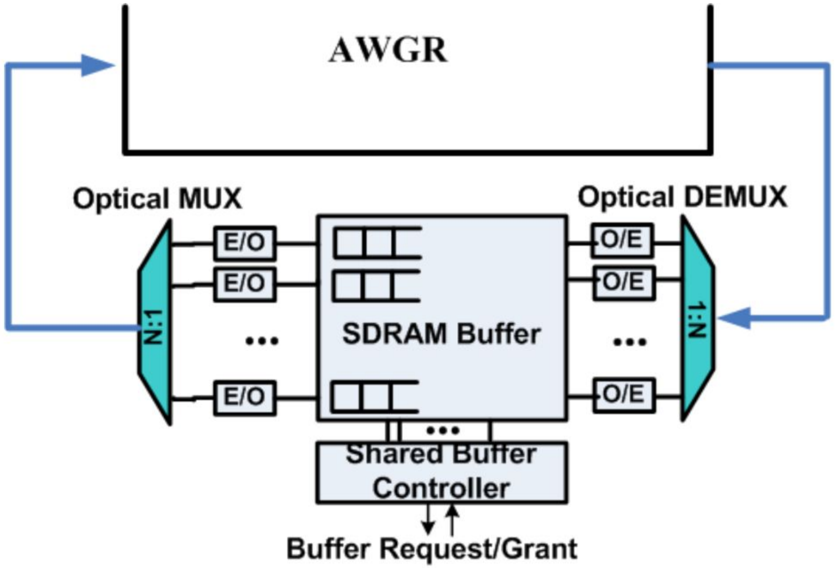
- In data center application, packet drop is more critical (unlike telecom applications).
- Timeout and retransmission could result in an unacceptable latency for a computing application.

Solution:
Put delayed or unsuccessful packets into SDRAM buffer and proceed them immediately after the corresponding wavegroup is available.



Loopback buffer system

Shared SDRAM Buffer



Loopback buffer system

1. Shared buffer receives failed packets in a arbitration cycle.
2. Packets on different wavelengths are separated by optical DEMUX.
3. Packets are converted from optical to electrical domain and stored in SDRAM.
4. SDRAM sends requests to buffer controller.
5. In next arbitration cycle, buffer requests have highest priority and will be approved if wavelength is idle.
6. SDRAM buffer sends delayed packets to AWGR outports.

Shared SDRAM Buffer

In-band Flow Control

- Why need flow control?
  SDRAM buffer size is limited.

- Solution:
  Introduce in-band ON-OFF flow control using little overhead.

- Steps:

1. When occupied SDRAM buffer exceeds a threshold, the certain bits in a delayed packet header is set.
2. End nodes receive delayed packet and check the certain bits.
3. If bits are set, end node temporarily suspend transmission.
4. When occupied SDRAM buffer becomes small, certain bits are reset back.
5. Then end node receives new packets indicating buffer is not much occupied now, they will restart transmission.

Arbitration In DOS

- **Compared with traditional N*N electronic switch.**

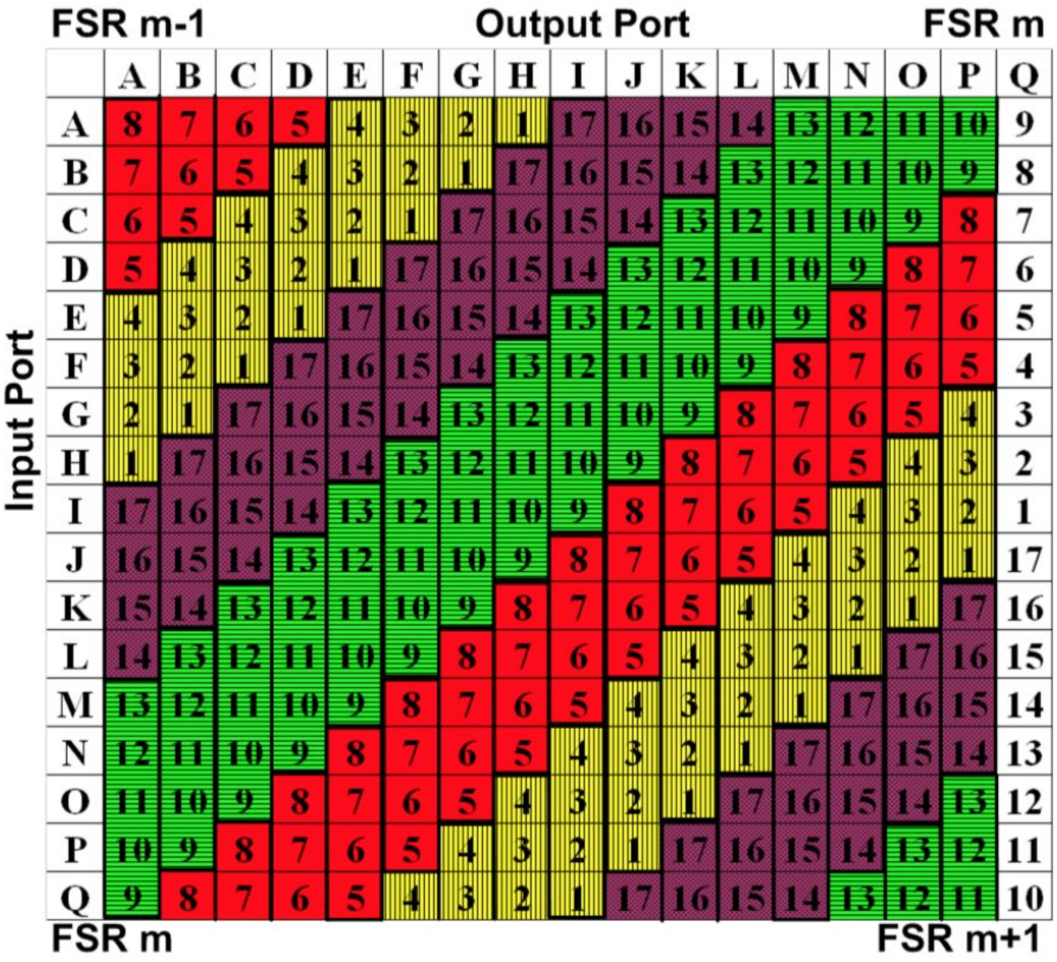
1. No packet is buffered at input.
2. All labels are processed in time.
3. No input will generate repeated requests except SDRAM buffer.
4. As VOQs are not used, input needs only one request and accept grant when notified by controller plane.
5. Only 2-phase arbiter is enough and O(log2N) iterations are not necessary.

- Optimization

1. AWGR provides wavelength parallelism and cyclic operation;
2. Reduce the inputs contending for the same output by increasing k number of wavelengths allowed per AWGR output.

Arbitration In DOS

- Example of 16-way optical switch with 1:4 optical DEMUX for each output.
- To accommodate the loopback packets from SDRAM, a 17*17 AWGR is necessary.

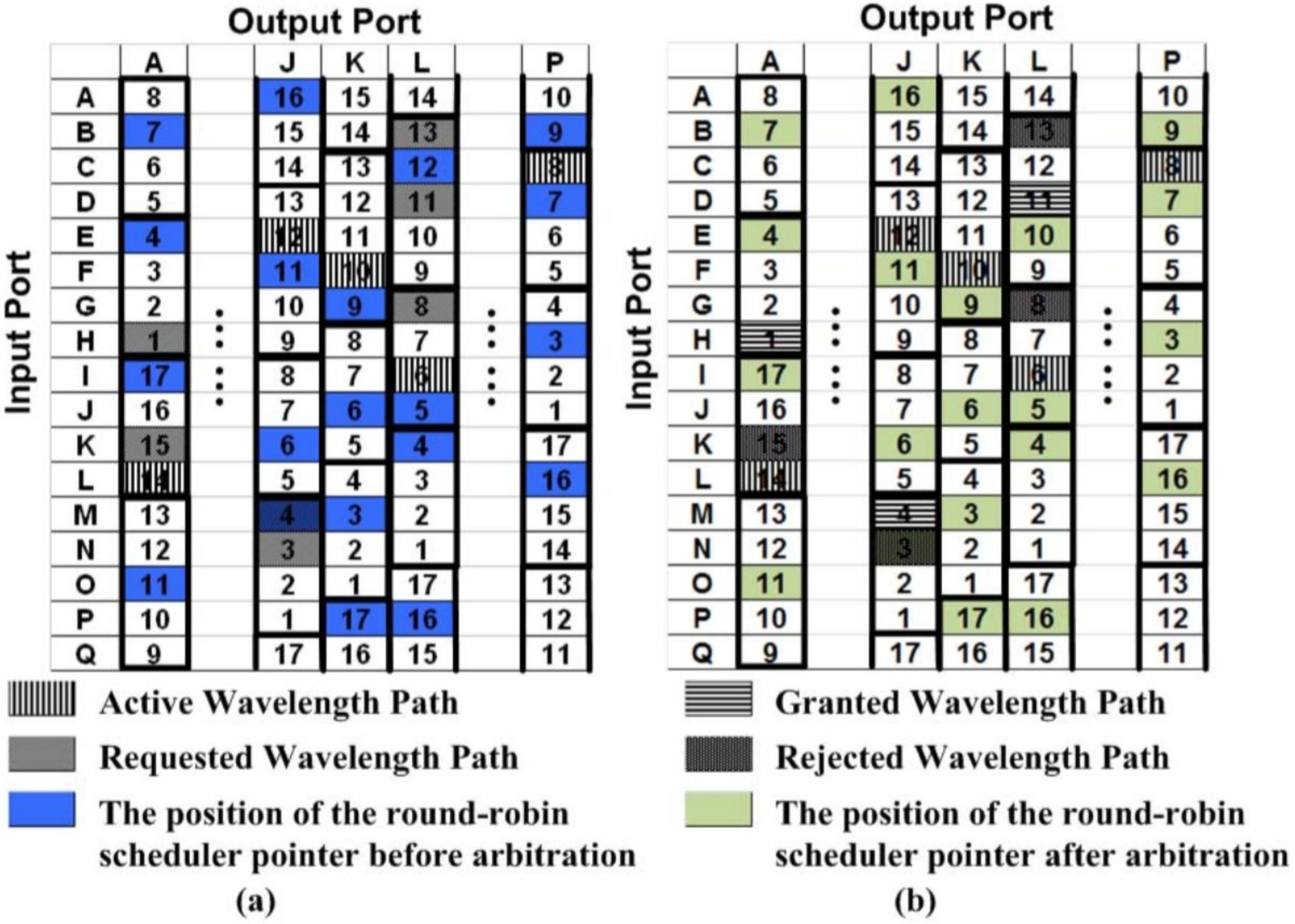

FSR m-1
Output Port
FSR m
Input Port
FSR m
FSR m+1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	8	7	6	5	4	3	2	1	17	16	15	14	13	12	11	10	9
B	7	6	5	4	3	2	1	17	16	15	14	13	12	11	10	9	8
C	6	5	4	3	2	1	17	16	15	14	13	12	11	10	9	8	7
D	5	4	3	2	1	17	16	15	14	13	12	11	10	9	8	7	6
E	4	3	2	1	17	16	15	14	13	12	11	10	9	8	7	6	5
F	3	2	1	17	16	15	14	13	12	11	10	9	8	7	6	5	4
G	2	1	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
H	1	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
I	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
J	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	17
K	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	17	16
L	14	13	12	11	10	9	8	7	6	5	4	3	2	1	17	16	15
M	13	12	11	10	9	8	7	6	5	4	3	2	1	17	16	15	14
N	12	11	10	9	8	7	6	5	4	3	2	1	17	16	15	14	13
O	11	10	9	8	7	6	5	4	3	2	1	17	16	15	14	13	12
P	10	9	8	7	6	5	4	3	2	1	17	16	15	14	13	12	11
Q	9	8	7	6	5	4	3	2	1	17	16	15	14	13	12	11	10

Arbitration In DOS

- **Example of how arbitration works**

1. Check whether wavelength is idle;
2. Collect requests;
3. Use round-robin pointer to decide which input should be granted.



(a)

	A		J	K	L		P
A	8		16	15	14		10
B	7		15	14	13		9
C	6		14	13	12		8
D	5		13	12	11		7
E	4		12	11	10		6
F	3		11	10	9		5
G	2	⋮	10	9	8	⋮	4
H	1	⋮	9	8	7	⋮	3
I	17	⋮	8	7	6	⋮	2
J	16	⋮	7	6	5	⋮	1
K	15		6	5	4		17
L	14		5	4	3		16
M	13		4	3	2		15
N	12		3	2	1		14
O	11		2	1	17		13
P	10		1	17	16		12
Q	9		17	16	15		11

▥ Active Wavelength Path

◼ Requested Wavelength Path

🟦 The position of the round-robin scheduler pointer before arbitration

(b)

	A		J	K	L		P
A	8		16	15	14		10
B	7		15	14	13		9
C	6		14	13	12		8
D	5		13	12	11		7
E	4		12	11	10		6
F	3		11	10	9		5
G	2	⋮	10	9	8	⋮	4
H	1	⋮	9	8	7	⋮	3
I	17	⋮	8	7	6	⋮	2
J	16	⋮	7	6	5	⋮	1
K	15		6	5	4		17
L	14		5	4	3		16
M	13		4	3	2		15
N	12		3	2	1		14
O	11		2	1	17		13
P	10		1	17	16		12
Q	9		17	16	15		11

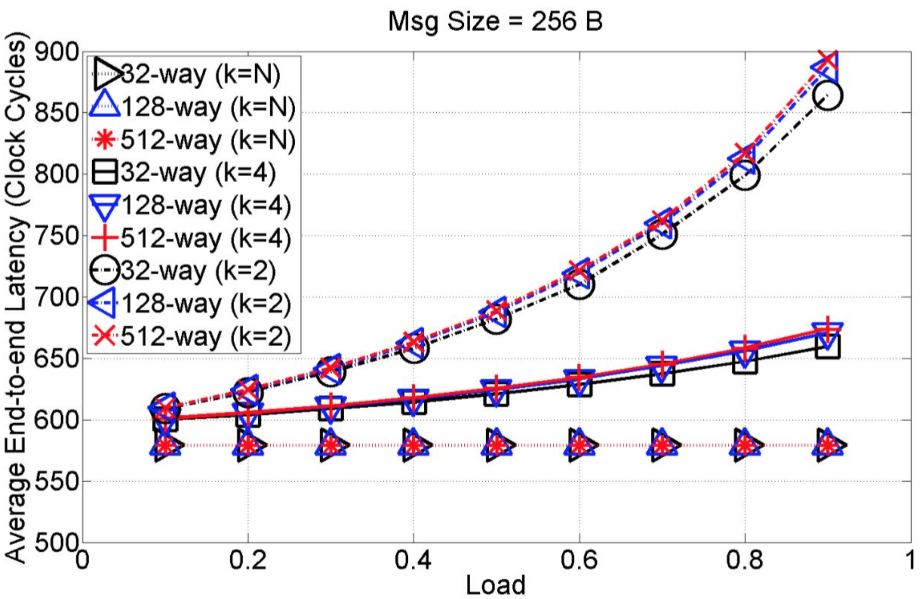
▤ Granted Wavelength Path

▩ Rejected Wavelength Path

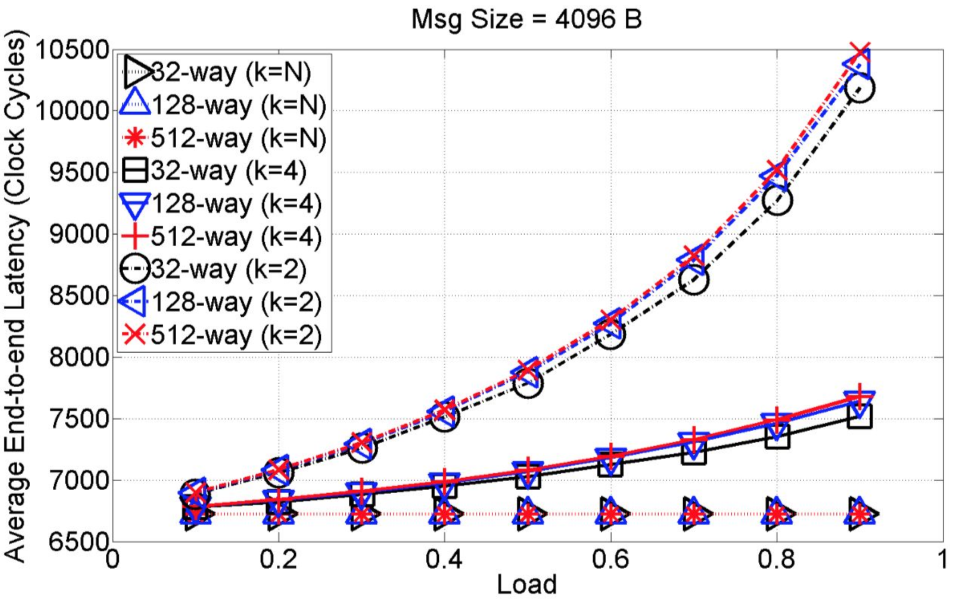
🟩 The position of the round-robin scheduler pointer after arbitration

UC DAVIS
UNIVERSITY OF CALIFORNIA

Simulation Results

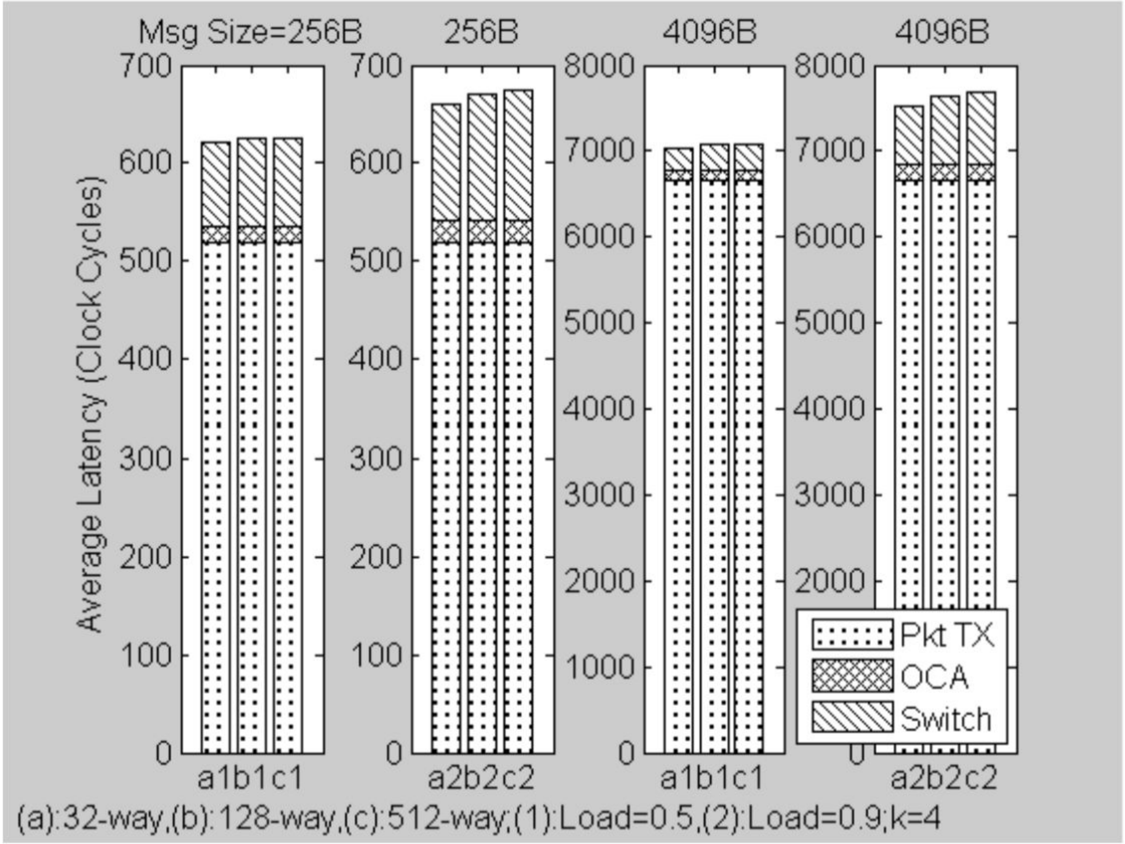


End-to End latency for DOS with message size 256 bytes



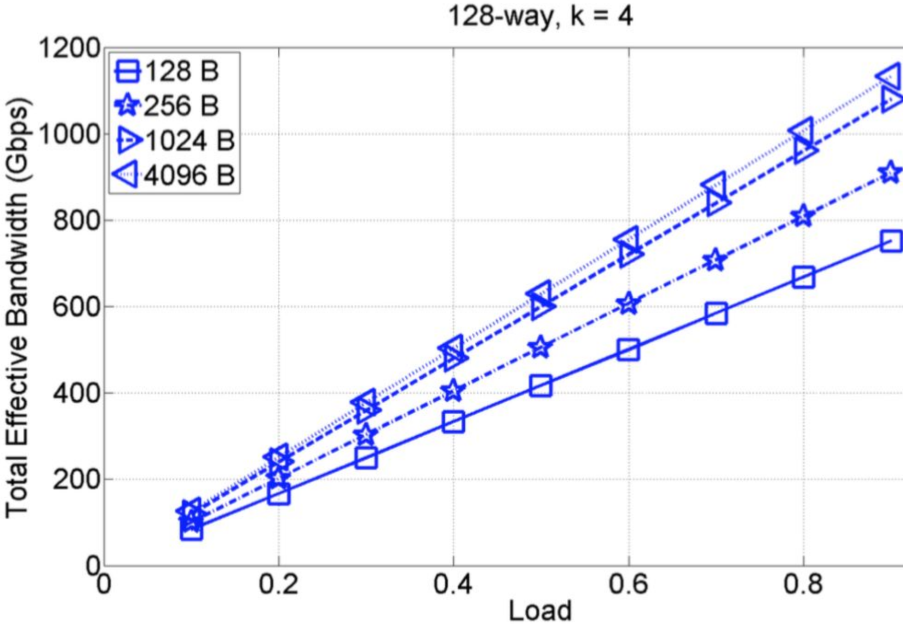
End-to End latency for DOS with message size 4096 bytes

Simulation Results

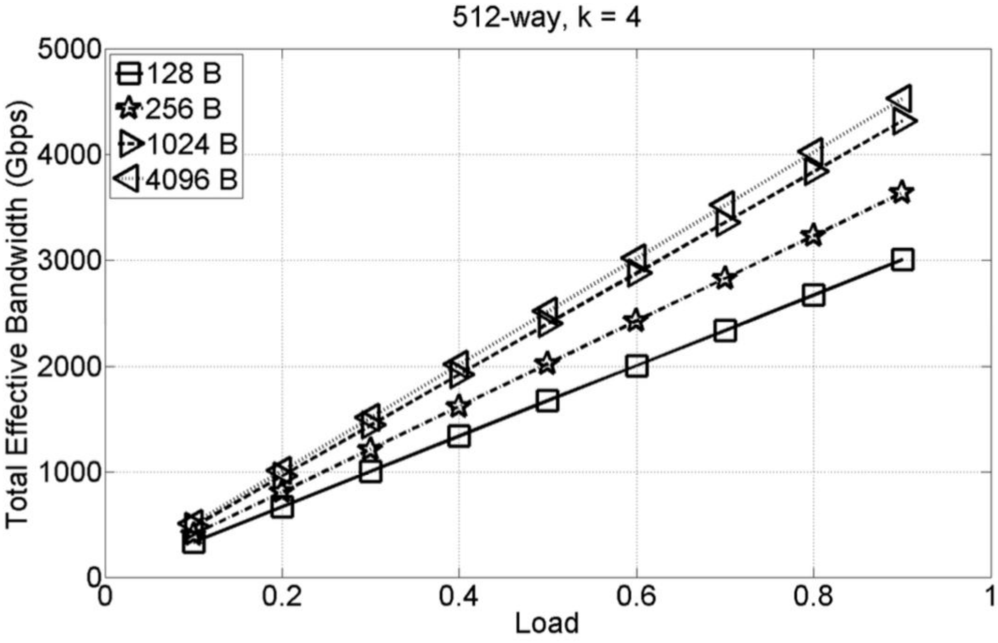


The breakdown of the end-to-end latency.

Simulation Results

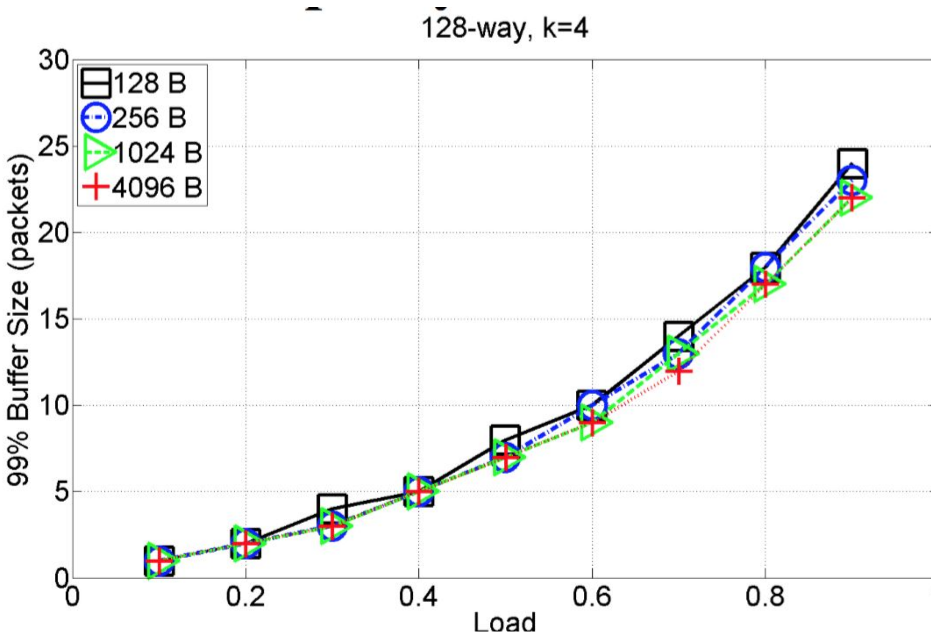


Effective bandwidth versus load for DOS system with 128 ways

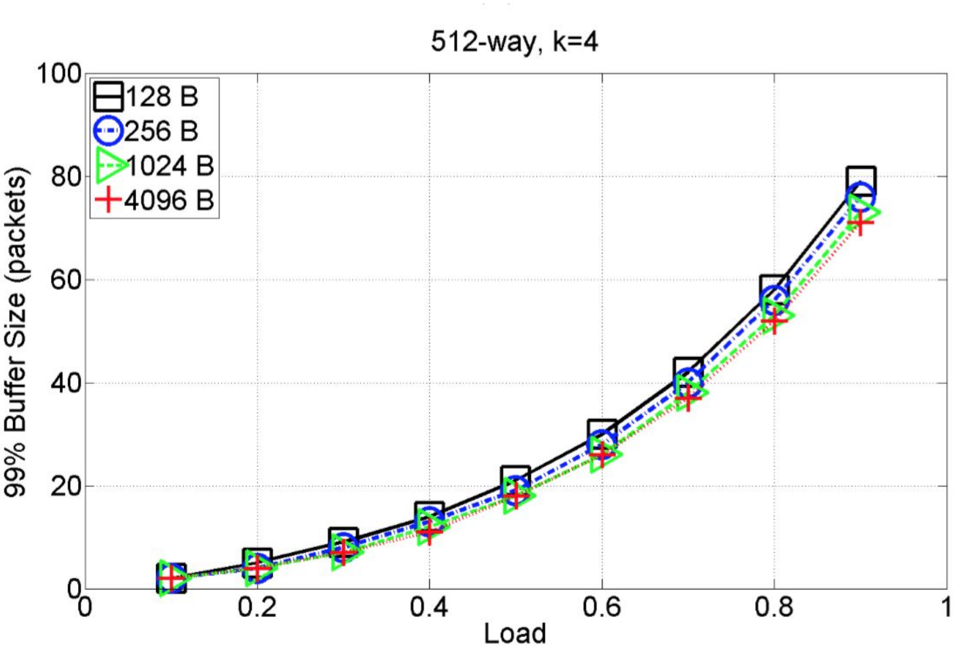


Effective bandwidth versus load for DOS system with 512 ways

Simulation Results



Buffer occupancy measurement
for 128way DOS



Buffer occupancy measurement
for 512-way DOS

UC DAVIS
UNIVERSITY OF CALIFORNIA

amlwang@ucdavis.edu