RDMA over Commodity Ethernet at Scale

Chuanxiong Guo, Haitao Wu, Zhong Deng, Gaurav Soni,
Jianxi Ye, Jitendra Padhye, Marina Lipshteyn

ACM SIGCOMM 2016

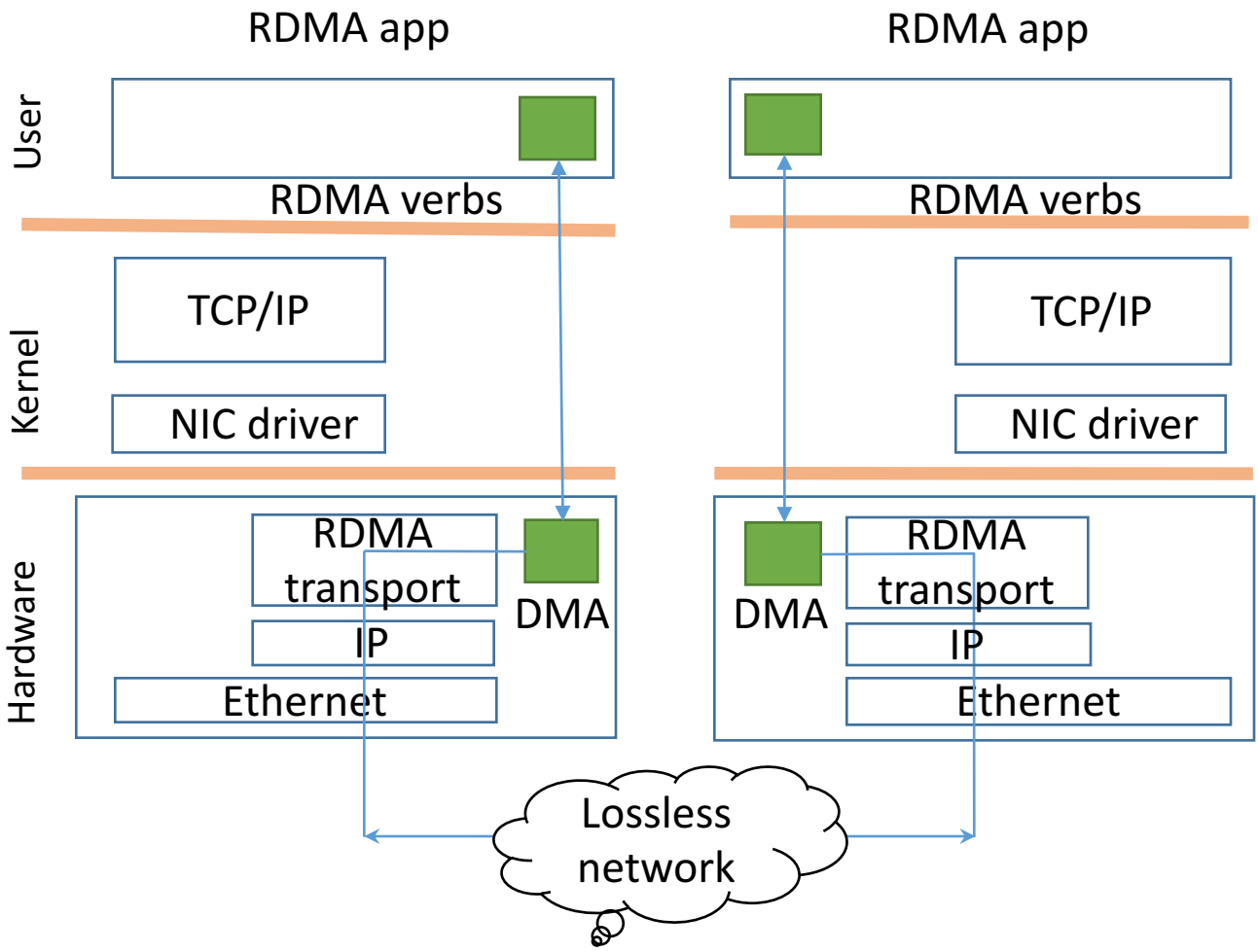
August 24 2016

Outline

- RDMA/RoCEv2 background
- DSCP-based PFC
- Safety challenges
  - RDMA transport livelock
  - PFC deadlock
  - PFC pause frame storm
  - Slow-receiver symptom
- Experiences and lessons learned
- Related work
- Conclusion

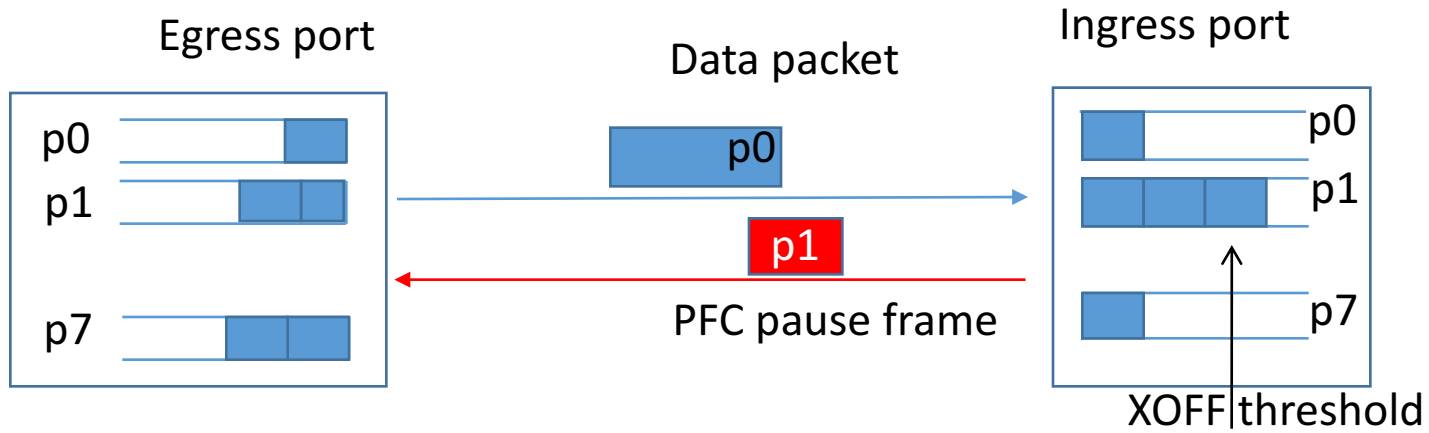
RDMA/RoCEv2 background

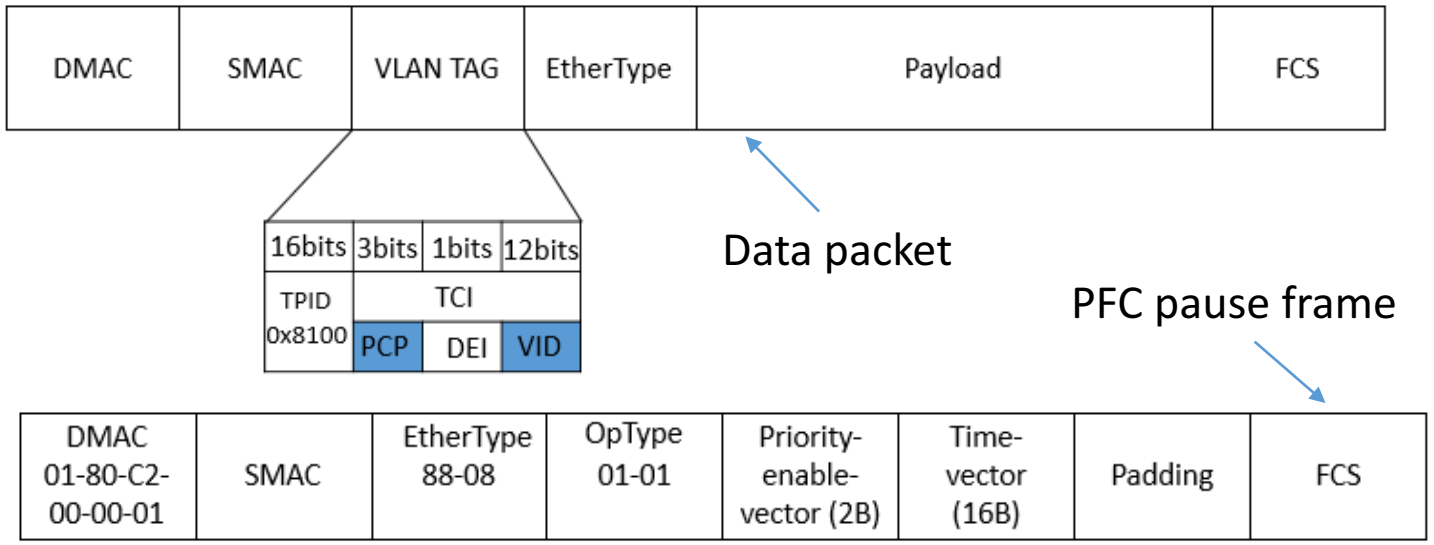
- RDMA addresses TCP's latency and CPU overhead problems
- RDMA: Remote Direct Memory Access
  - RDMA offloads the transport layer to the NIC
  - RDMA needs a lossless network
- RoCEv2: RDMA over commodity Ethernet
  - DCQCN for connection-level congestion control
  - PFC for hop-by-hop flow control



Priority-based flow control (PFC)

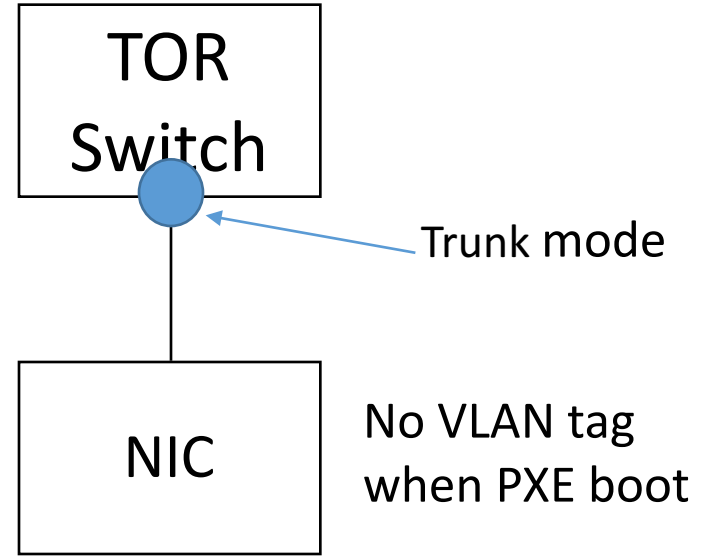
- Hop-by-hop flow control, with eight priorities for HOL blocking mitigation
- The priority in data packets is carried in the VLAN tag
- PFC pause frame to inform the upstream to stop





DSCP-based PFC

- Issues of VLAN-based PFC
  - It breaks PXE boot
  - No standard way for carrying VLAN tag in L3 networks
- DSCP-based PFC
  - DSCP field for carrying the priority value
  - No change needed for the PFC pause frame
  - Supported by major switch/NIC venders



Data packet

DMAC	SMAC	EtherType 0800	IP header	DSCP		IP payload	FCS

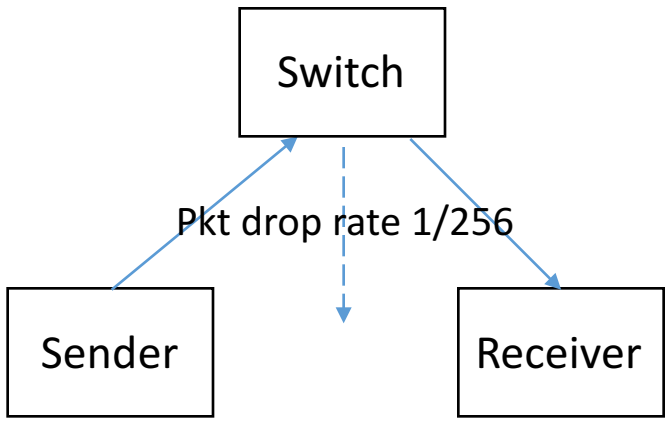
PFC pause frame

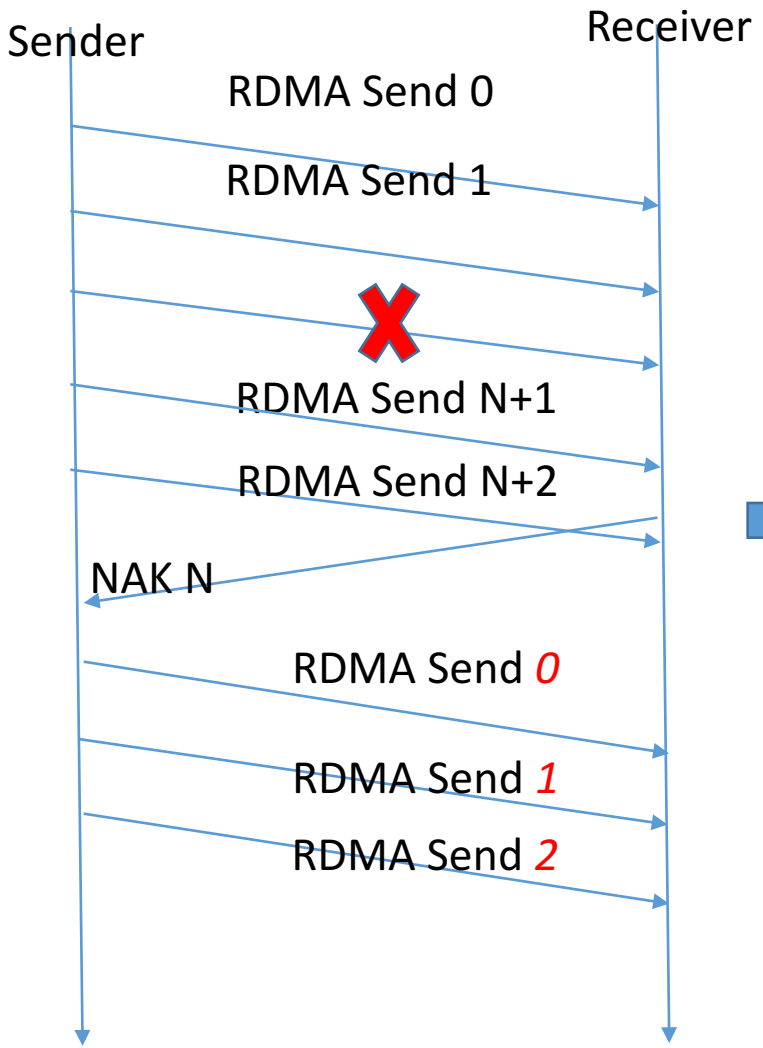
DMAC 01-80-C2-00-00-01	SMAC	EtherType 88-08	OpType 01-01	Priority-enable-vector (2B)	Time-vector (16B)	Padding	FCS

Outline

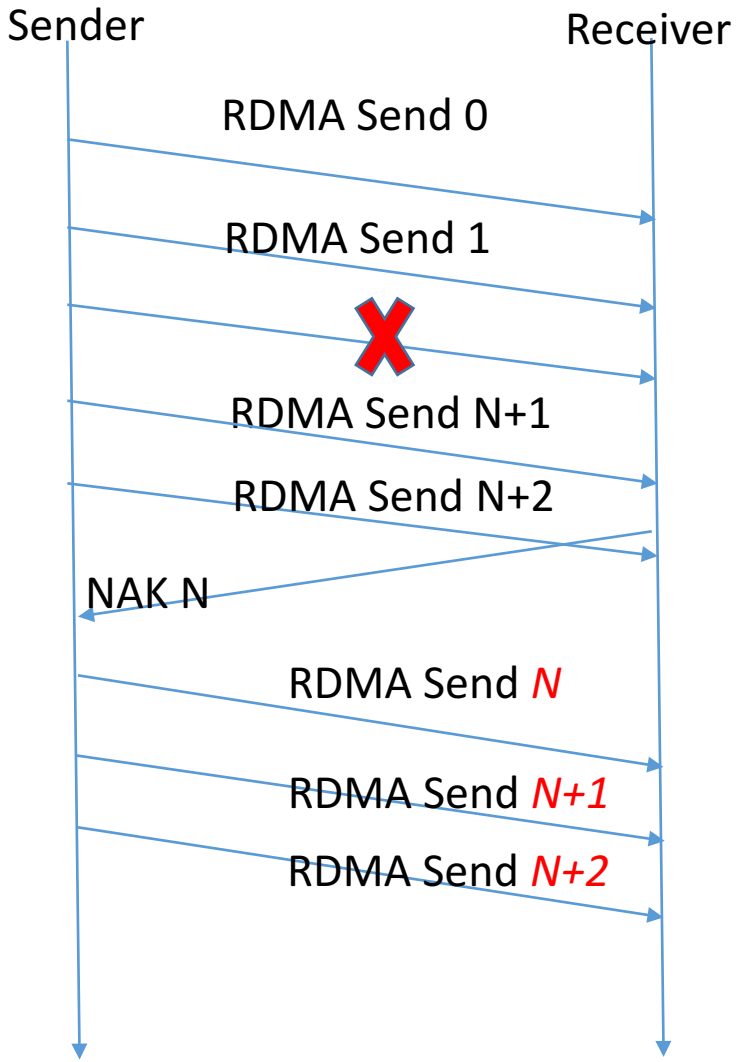
- RDMA/RoCEv2 background
- DSCP-based PFC
- *Safety challenges*
  - RDMA transport livelock
  - PFC deadlock
  - PFC pause frame storm
  - Slow-receiver symptom
- Experiences and lessons learned
- Related work
- Conclusion

RDMA transport livelock





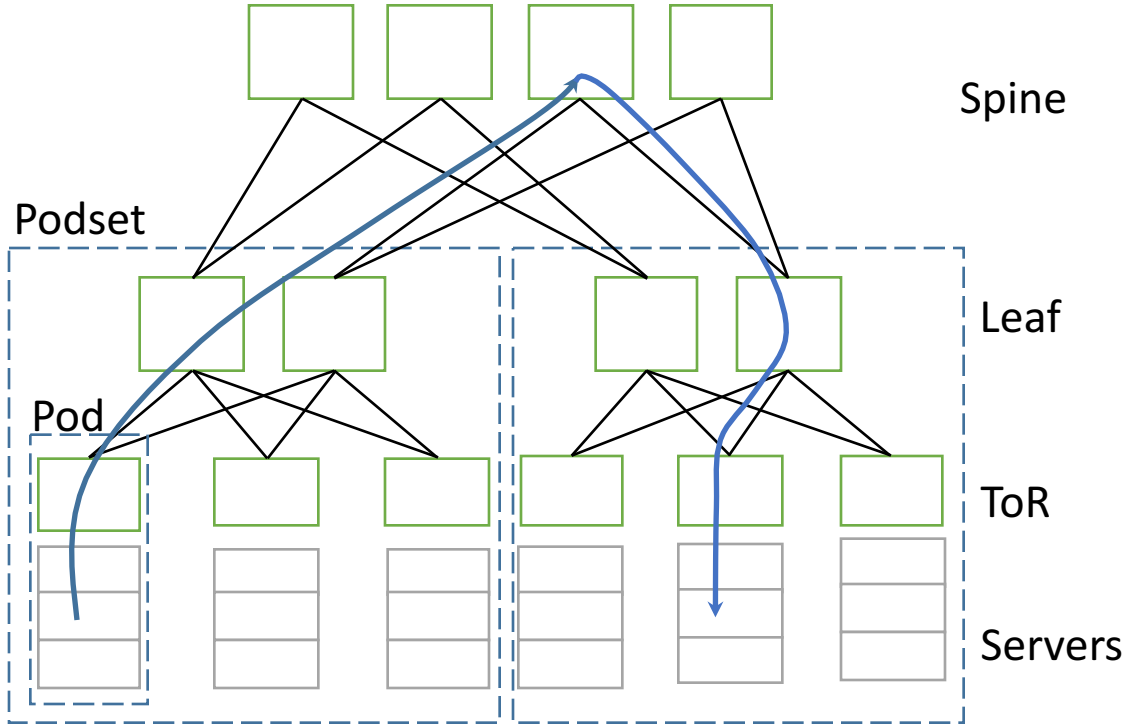
Go-back-0



Go-back-N

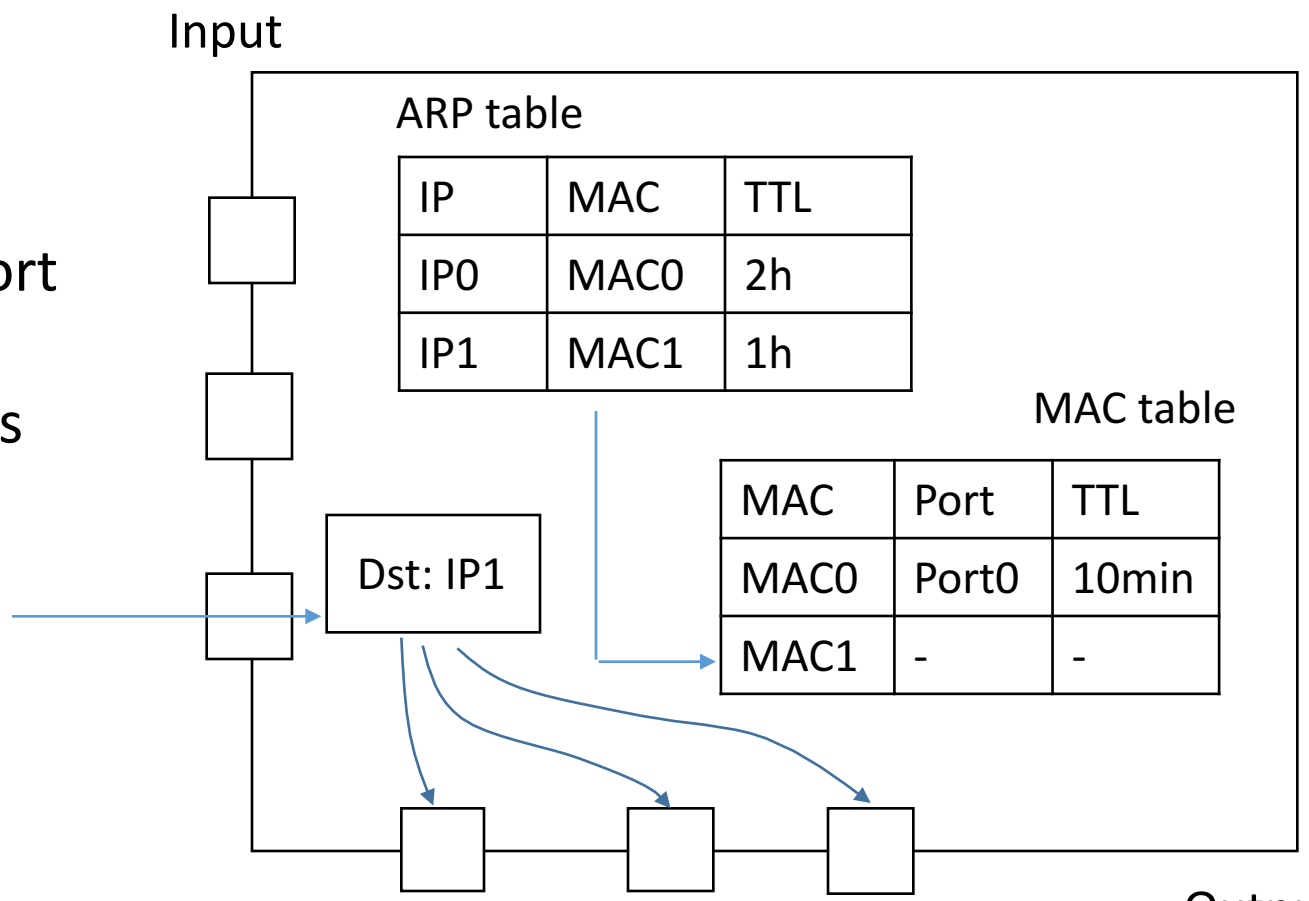
PFC deadlock

- Our data centers use Clos network
- Packets first travel up then go down
- No cyclic buffer dependency for up-down routing -> no deadlock
- But we did experience deadlock!

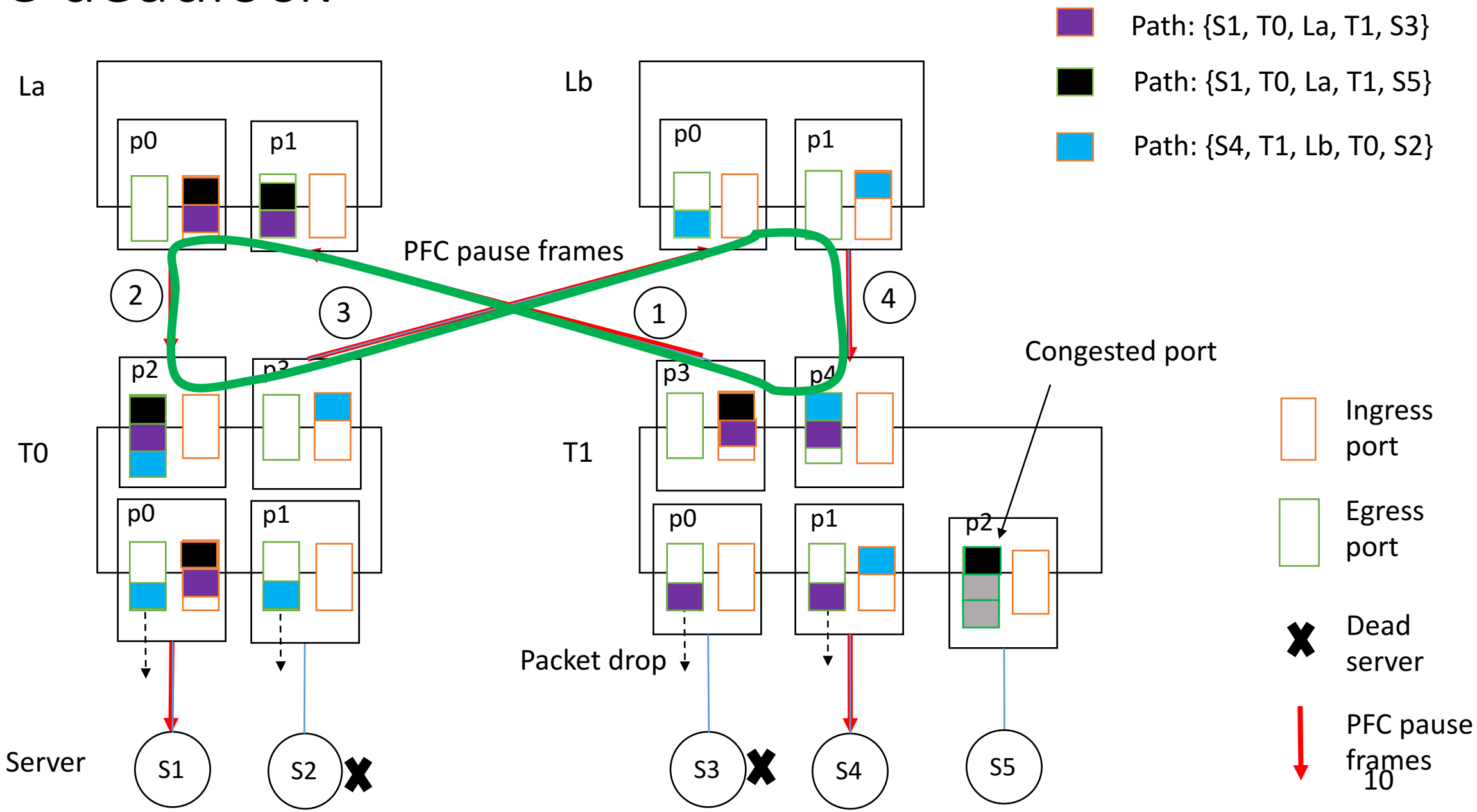


PFC deadlock

- Preliminaries
  - ARP table: IP address to MAC address mapping
  - MAC table: MAC address to port mapping
  - If MAC entry is missing, packets are flooded to all ports



PFC deadlock

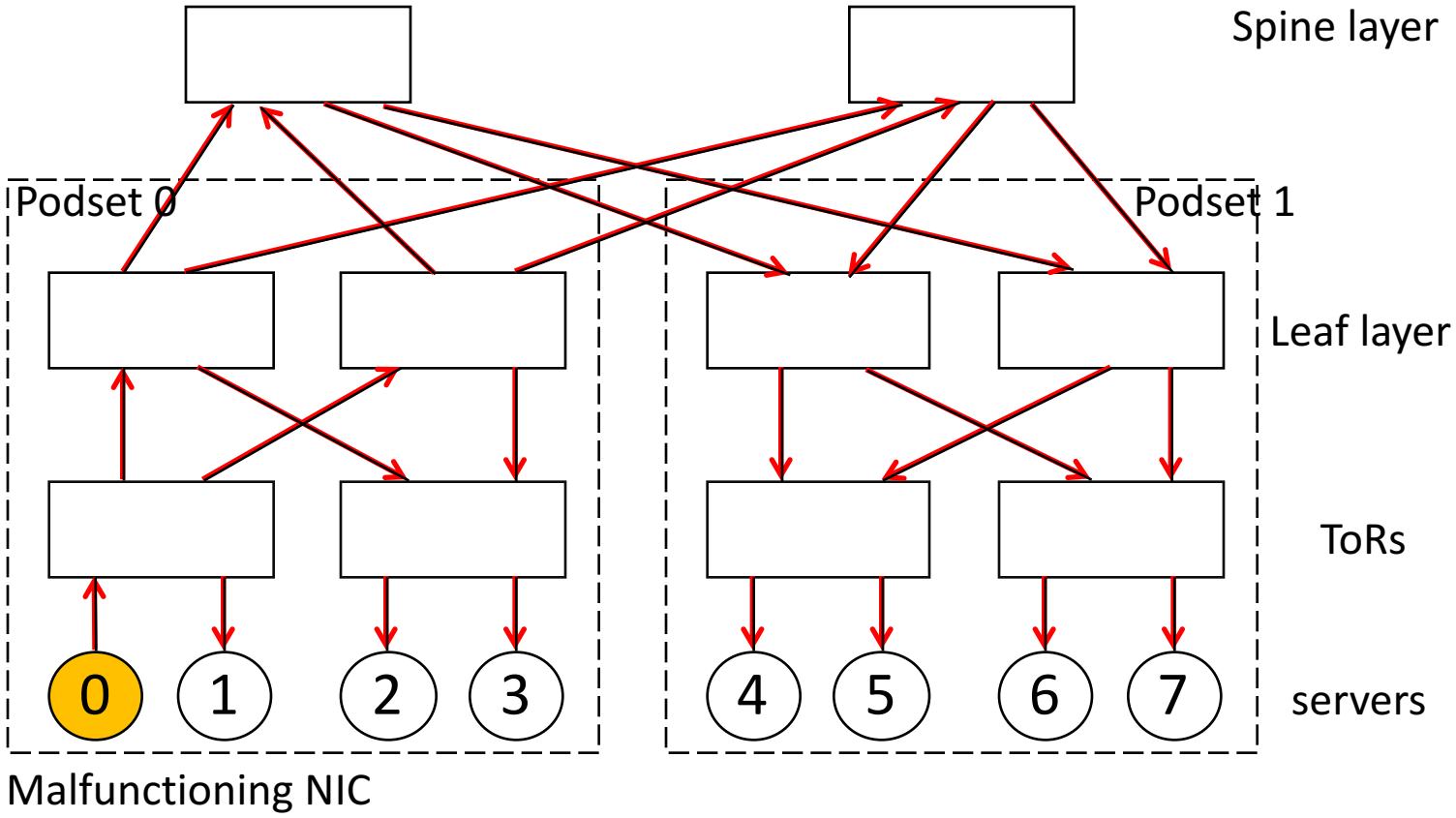


PFC deadlock

- The PFC deadlock root cause: the interaction between the PFC flow control and the Ethernet packet flooding
- Solution: drop the lossless packets if the ARP entry is incomplete
- Recommendation: do not flood or multicast for lossless traffic
- Call for action: more research on deadlocks

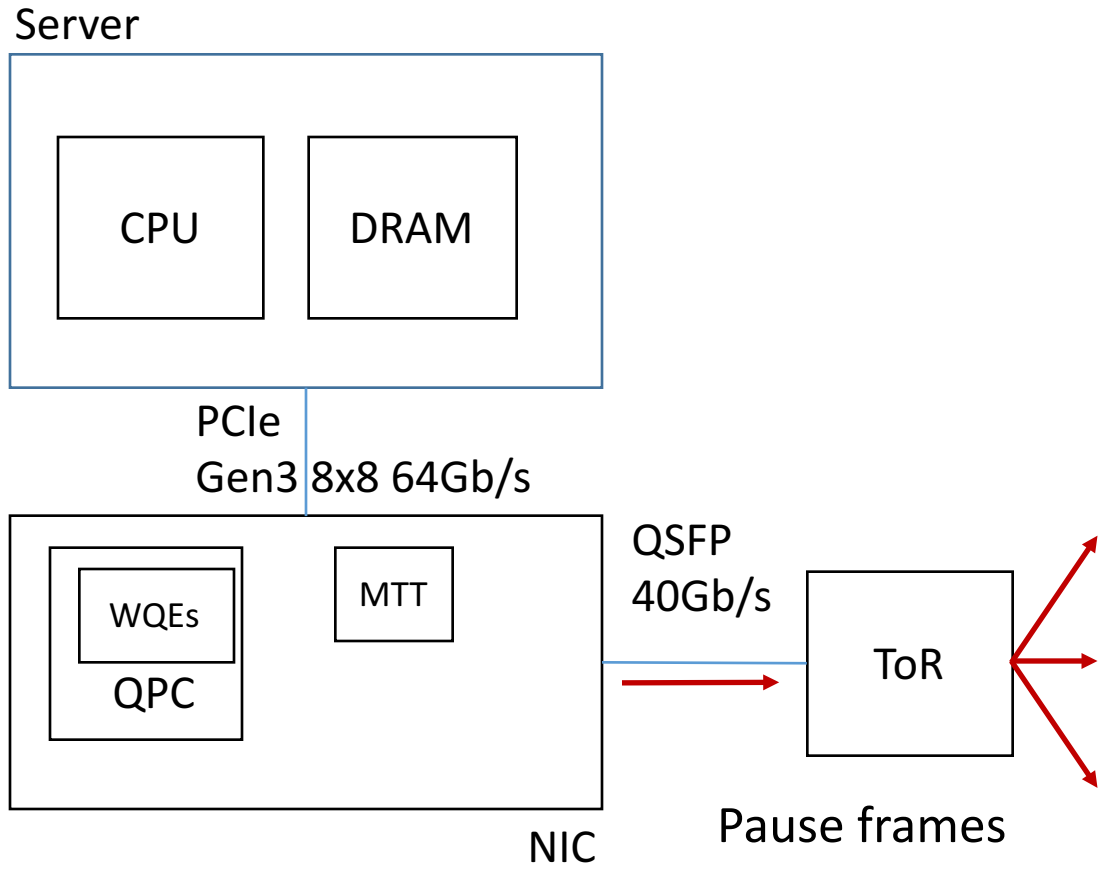
NIC PFC pause frame storm

- A malfunctioning NIC may block the whole network
- PFC pause frame storms caused several incidents
- Solution: watchdogs at both NIC and switch sides to stop the storm



The slow-receiver symptom

- ToR to NIC is 40Gb/s, NIC to server is 64Gb/s
- But NICs may generate large number of PFC pause frames
- Root cause: NIC is resource constrained
- Mitigation
  - Large page size for the MTT (memory translation table) entry
  - Dynamic buffer sharing at the ToR

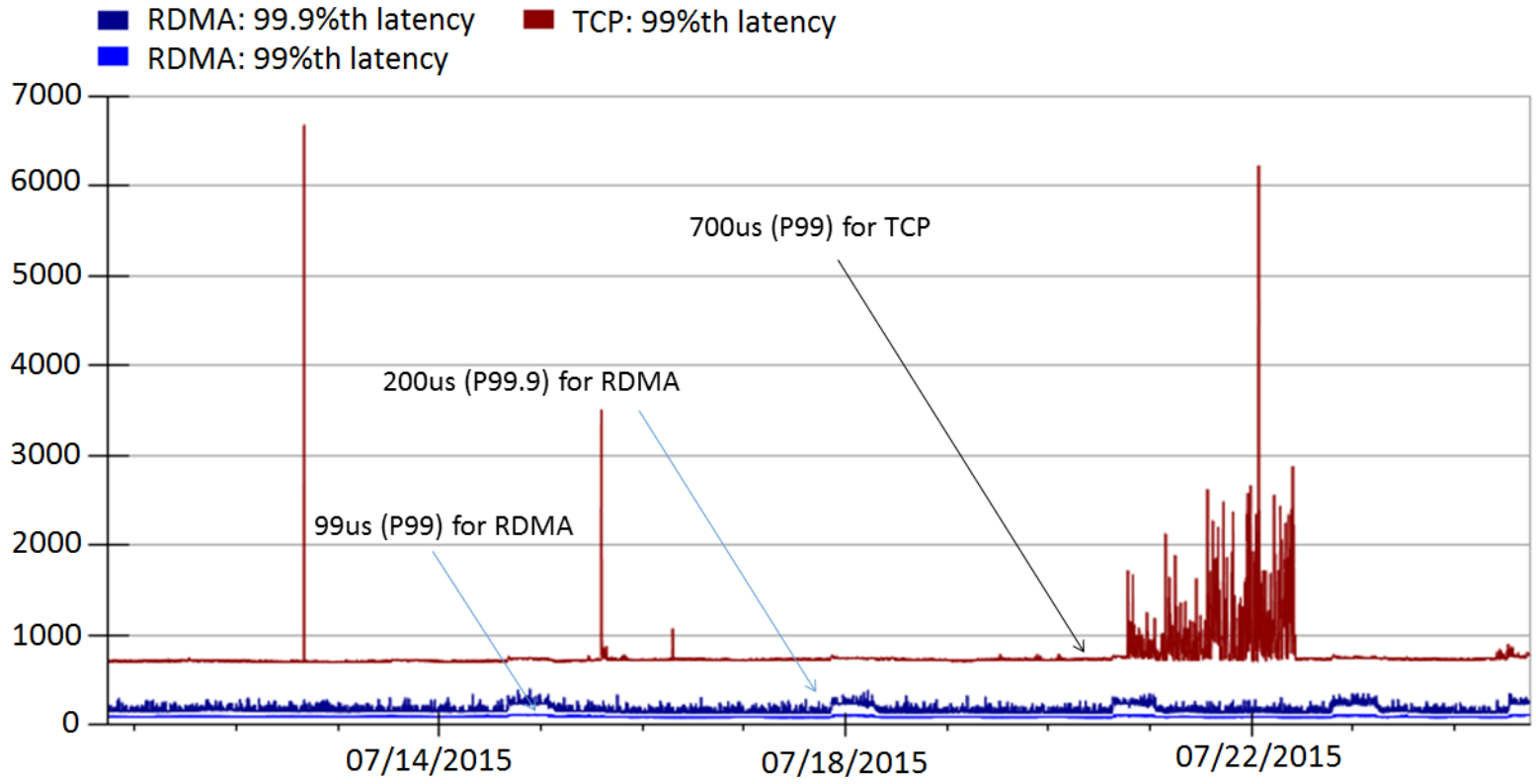


Outline

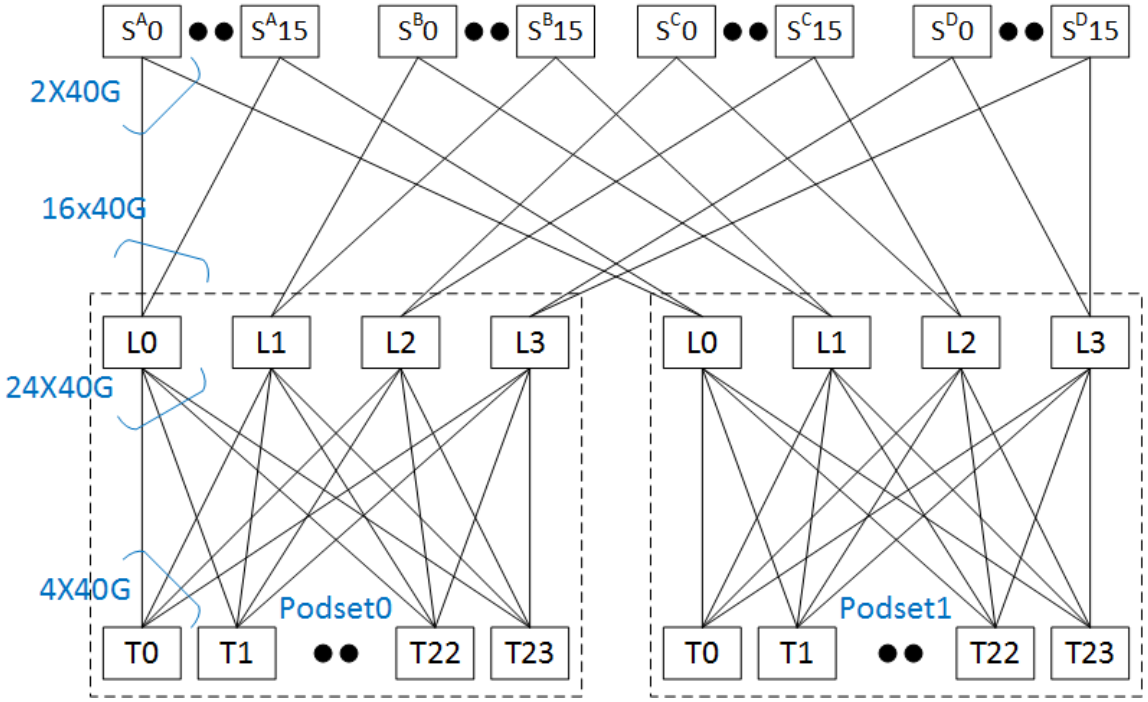
- RDMA/RoCEv2 background
- DSCP-based PFC
- Safety challenges
  - RDMA transport livelock
  - PFC deadlock
  - PFC pause frame storm
  - Slow-receiver symptom
- *Experiences and lessons learned*
- Related work
- Conclusion

Latency reduction

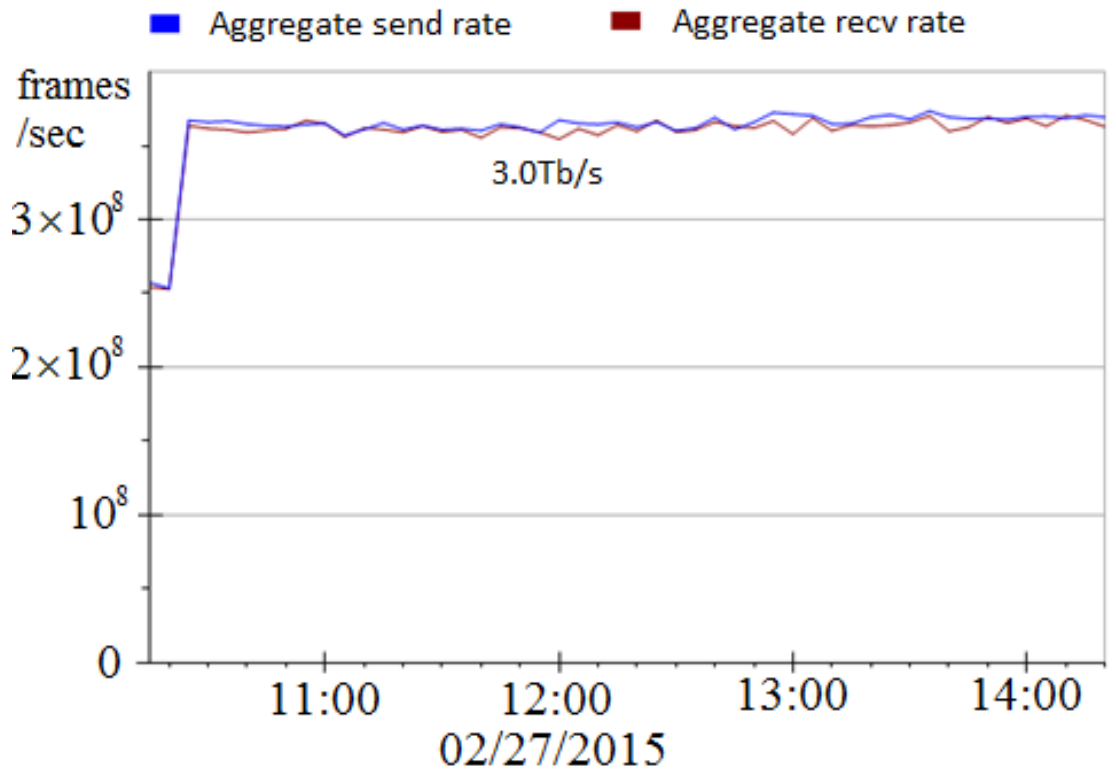
- RoCEv2 deployed in Bing world-wide for one and half years
- Significant latency reduction
- Incast problem solved as no packet drops



RDMA throughput

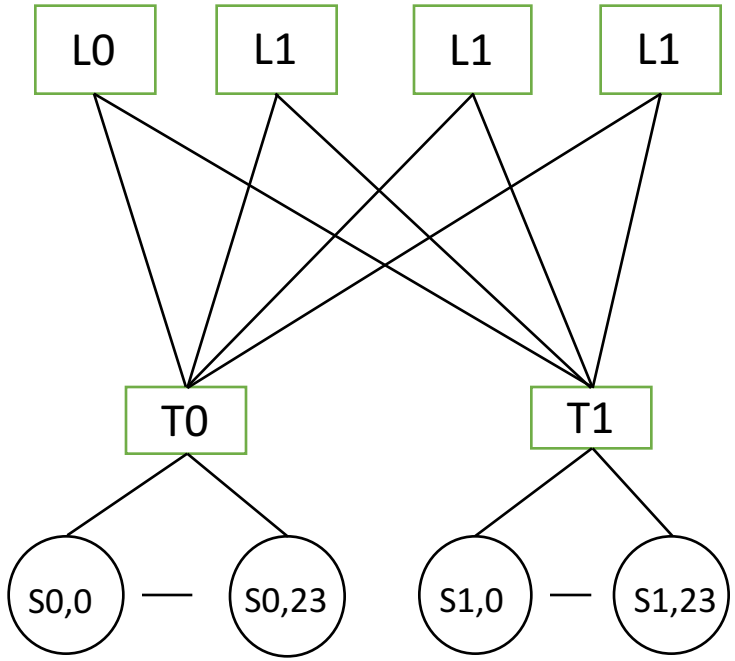


- Using two podsets each with 500+ servers
- 5Tb/s capacity between the two podsets

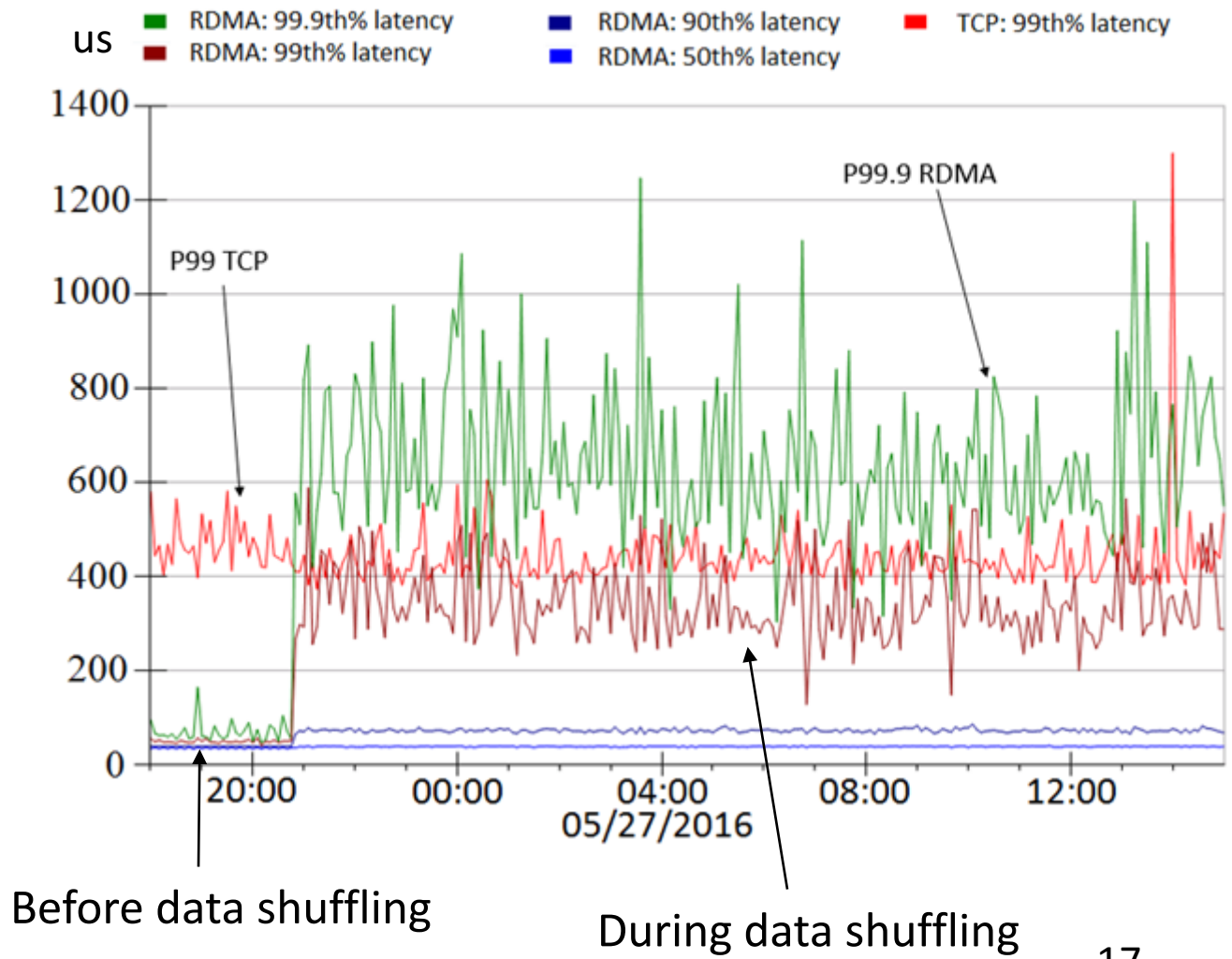


- Achieved 3Tb/s inter-podset throughput
- Bottlenecked by ECMP routing
- Close to 0 CPU overhead

Latency and throughput tradeoff



- RDMA latencies increase as data shuffling started
- Low latency vs high throughput



Lessons learned

- Deadlock, livelock, PFC pause frames propagation and storm did happen
- Be prepared for the unexpected
  - Configuration management, latency/availability, PFC pause frame, RDMA traffic monitoring
- NICs are the key to make RoCEv2 work
- Loss vs lossless: Is lossless needed?

Related work

- Infiniband
- iWarp
- Deadlock in lossless networks
- TCP perf tuning vs. RDMA

Conclusion

- RoCEv2 has been running safely in Microsoft data centers for one and half years
  - DSCP-based PFC which scales RoCEv2 from L2 to L3
  - Various safety issues/bugs (livelock, deadlock, PFC pause storm, PFC pause propagation) can all be addressed
- Future work
  - RDMA for inter-DC communications
  - Understanding of deadlocks in data centers
  - Lossless, low-latency and high-throughput networking
  - Applications adoption