Remote Priority Flow Control (Remote PFC)

OCP Summit 2021

Contact: jeremias.blendin@intel.com
Team: Jeremias Blendin, Yanfang Le, JK Lee, Grzegorz Jereczek

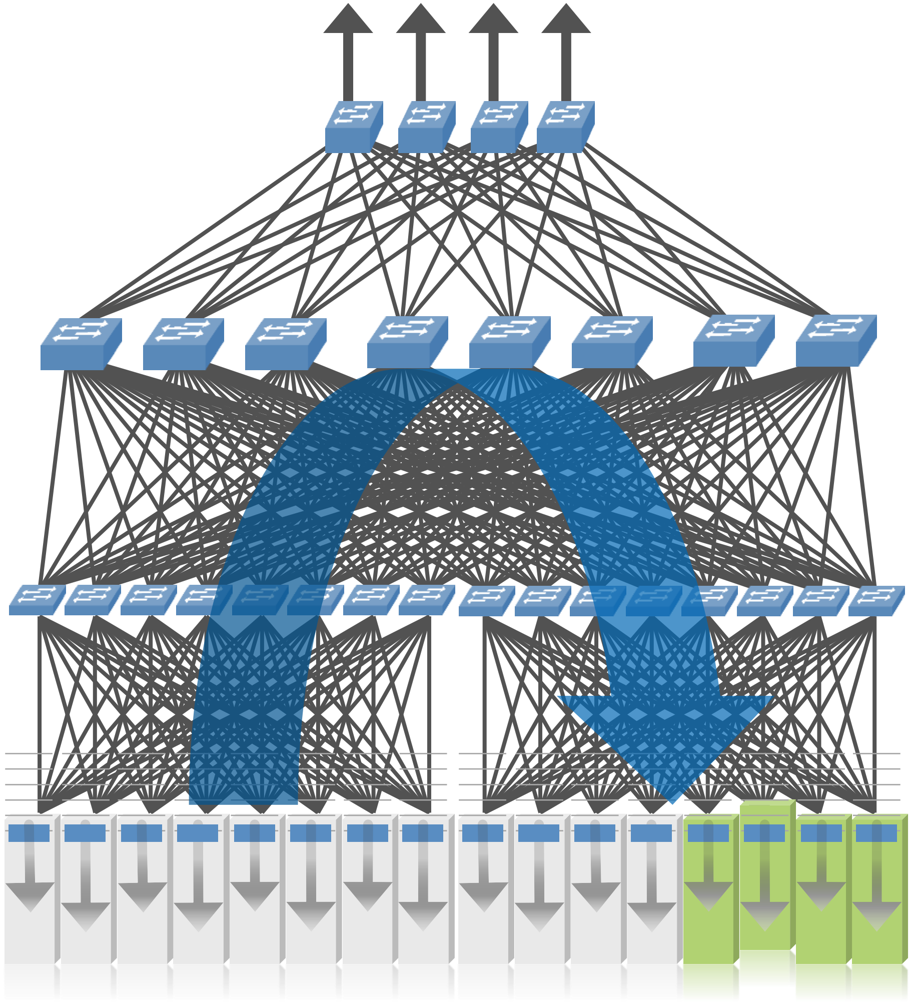
Intel, Barefoot Switch Division

Remote PFC at a Glance

Remote Priority Flow Control (Remote PFC) specifically improves the performance of incast (many senders, one receiver) heavy workloads such as AI deep learning clusters. It does so by "flattening the curve" of incast traffic and achieves a significant reduction of the data center switch queue utilization and flow completion time (FCT) compared to the state of the art. Remote PFC uses Intel® Tofino™ 2 Programmable Ethernet Switch ASIC’s & Intel® Tofino™ 3 Intelligent Fabric Processor’s unique programmability features and SONiC PINS’ flexibility to achieve sub round trip time (RTT) edge-to-edge signaling of congestion in data centers.

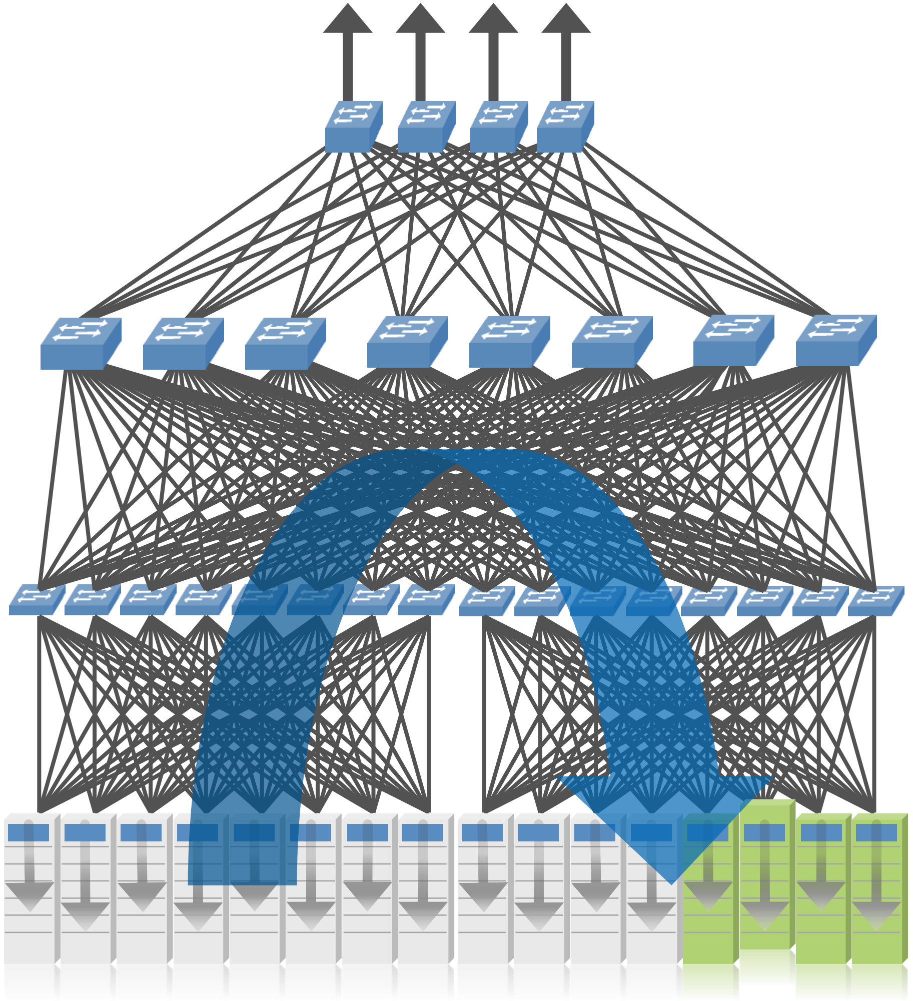
Incast Congestion in Data Centers

- **Incast**
  - Cause: many-to-one traffic pattern
  - Mostly at the last-hop
  - Governs max/tail latency
  - Tail latency can have a big performance impact on RDMA-style workloads
  - High incast ratios require reaction at congestion-free base network RTT scale



Solution space

- Edge-to-edge (e2e) congestion control
  - Detect congestion in e2e path and adjust TX rates
  - Requires multiple RTTs to react
  - Part of e2e transport such as TCP, RoCEv2
- Hop-by-hop flow control by example of IEEE 802.1Qbb PFC
  - Low-latency xon/xoff signal to previous hop queue
  - Designed to prevent packet loss
  - Complex configuration and operational side-effects
    - Incurs head-of-line blocking (HoL)
    - PFC storm, deadlocks



Need for a new, low-latency edge-to-edge flow control mechanism!

Remote PFC's Approach to Flow Control

Remote PFC is an in-network flow control mechanism. Remote PFC leverages Intel® Tofino™ 2 Programmable Ethernet Switch ASIC's advanced programmability to detect queue build and to signal congestion across the data center to stop the contributing sender NICs directly. PFC is used for flow control enforcement between top-of-rack (ToR) switches and NICs for backwards compatibility and low latency. Thereby, Remote PFC combines the strength of edge-to-edge signaling with the strength of low-latency flow control to implement sub-RTT remote PFC signaling.

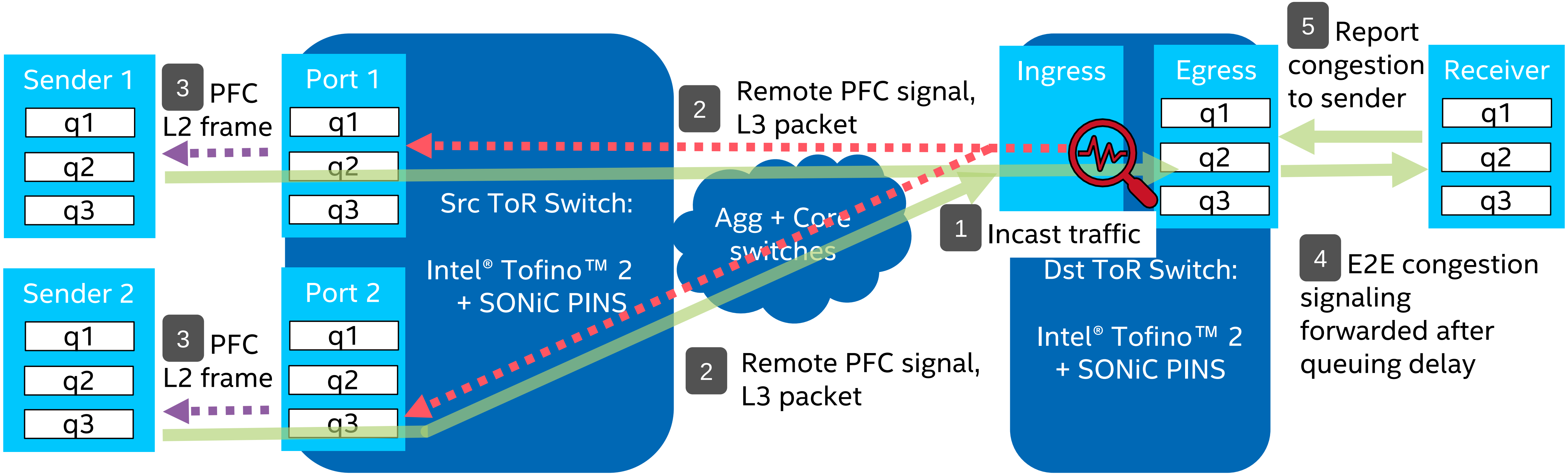
Remote PFC Edge-to-Edge View

What is Remote PFC?

- Edge-to-Edge signaling of congestion
- Flow control that instantly 'flattens the curve'
- Signaling + 'source' flow ctrl all in sub-RTT

Remote PFC does not target/does target

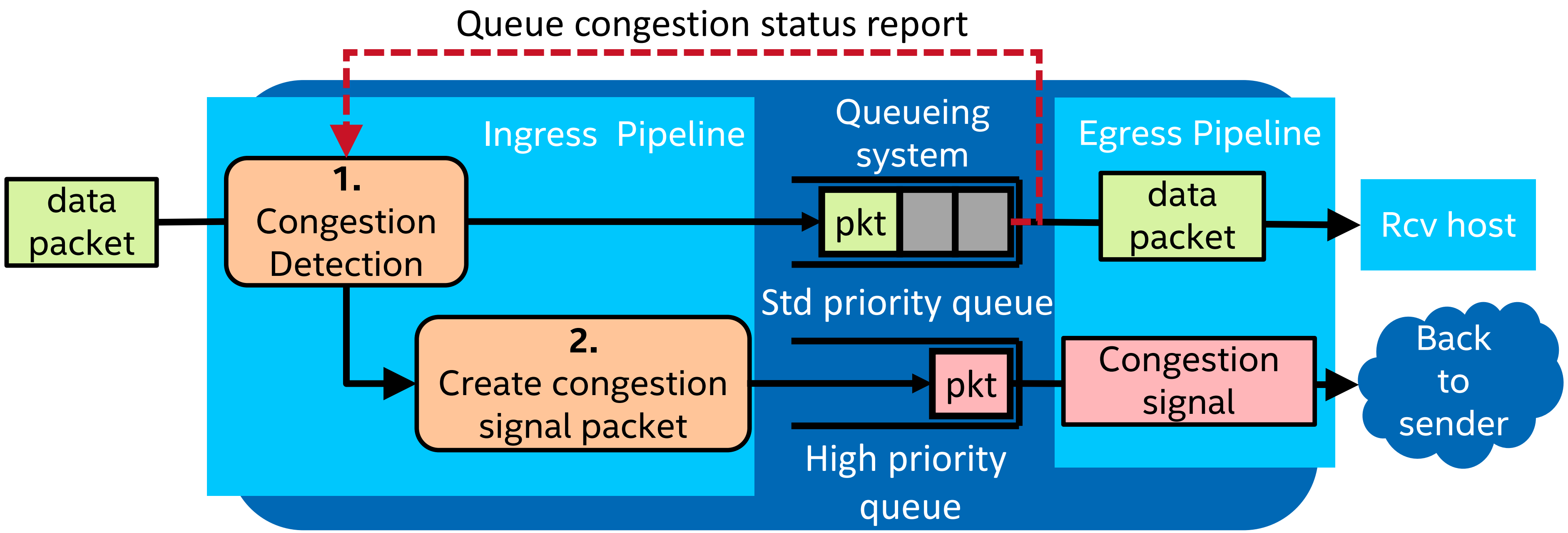
- ~~aim 100% lossless~~ vs min switch buffering
- ~~e2e congestion ctrl~~ vs NIC flow ctrl
- ~~Pause Agg/Core switches~~ → no PFC side effects
- ~~Need greenfield deployment~~ → ToR-only upgrade



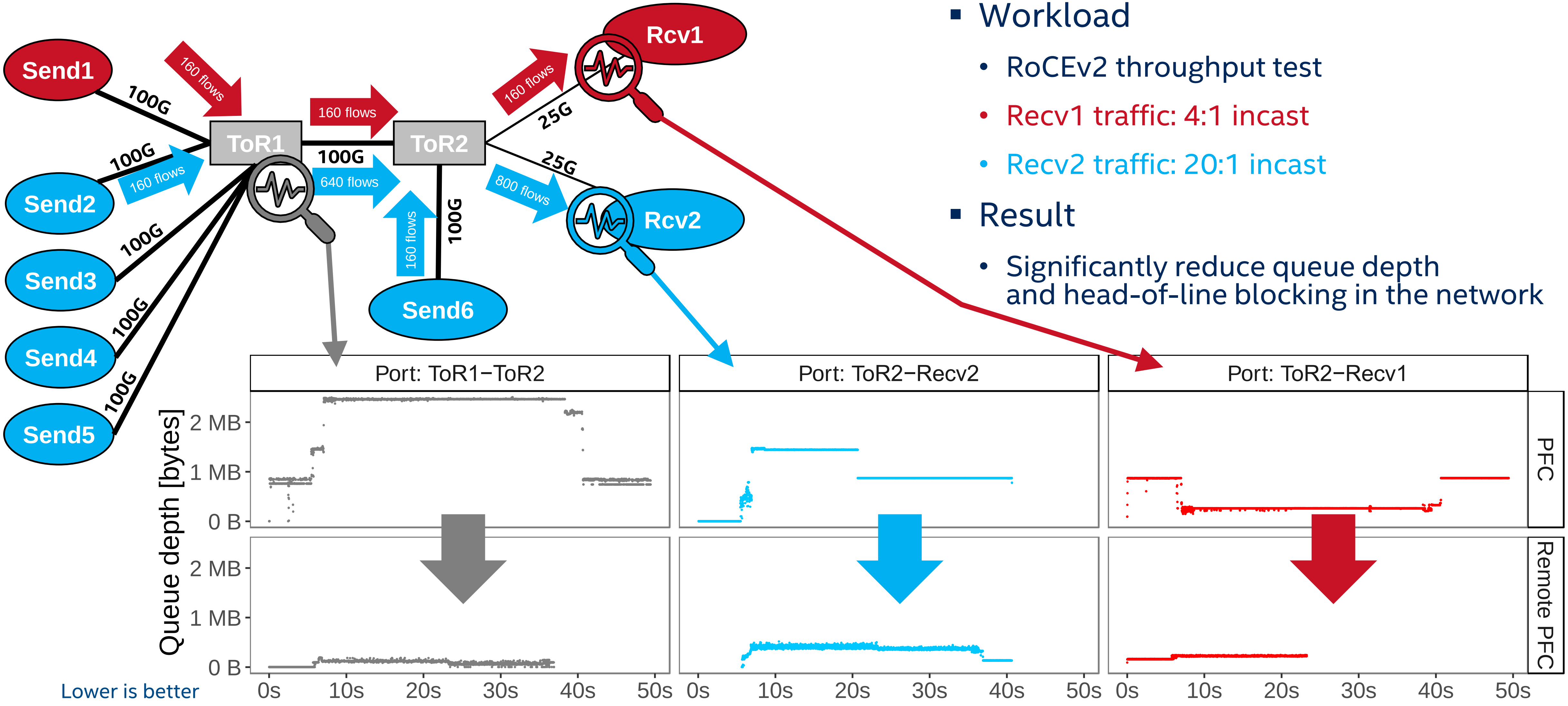
Sender 1
q1
q2
q3
Sender 2
q1
q2
q3
3 PFC L2 frame
3 PFC L2 frame
Port 1
q1
q2
q3
Port 2
q1
q2
q3
Src ToR Switch:
Intel® Tofino™ 2 + SONiC PINS
2 Remote PFC signal, L3 packet
2 Remote PFC signal, L3 packet
Agg + Core switches
1 Incast traffic
Ingress
Egress
q1
q2
q3
Dst ToR Switch:
Intel® Tofino™ 2 + SONiC PINS
5 Report congestion to sender
Receiver
q1
q2
q3
4 E2E congestion signaling forwarded after queuing delay

Intelligent Congestion Detection

1. The programmable logic checks the congestion status of an outgoing queue before enqueuing a packet
2. If congestion is detected, a notification packet is created that skips the congestion and is sent directly back to the sender



Remote PFC’s Effect on Queue Depth

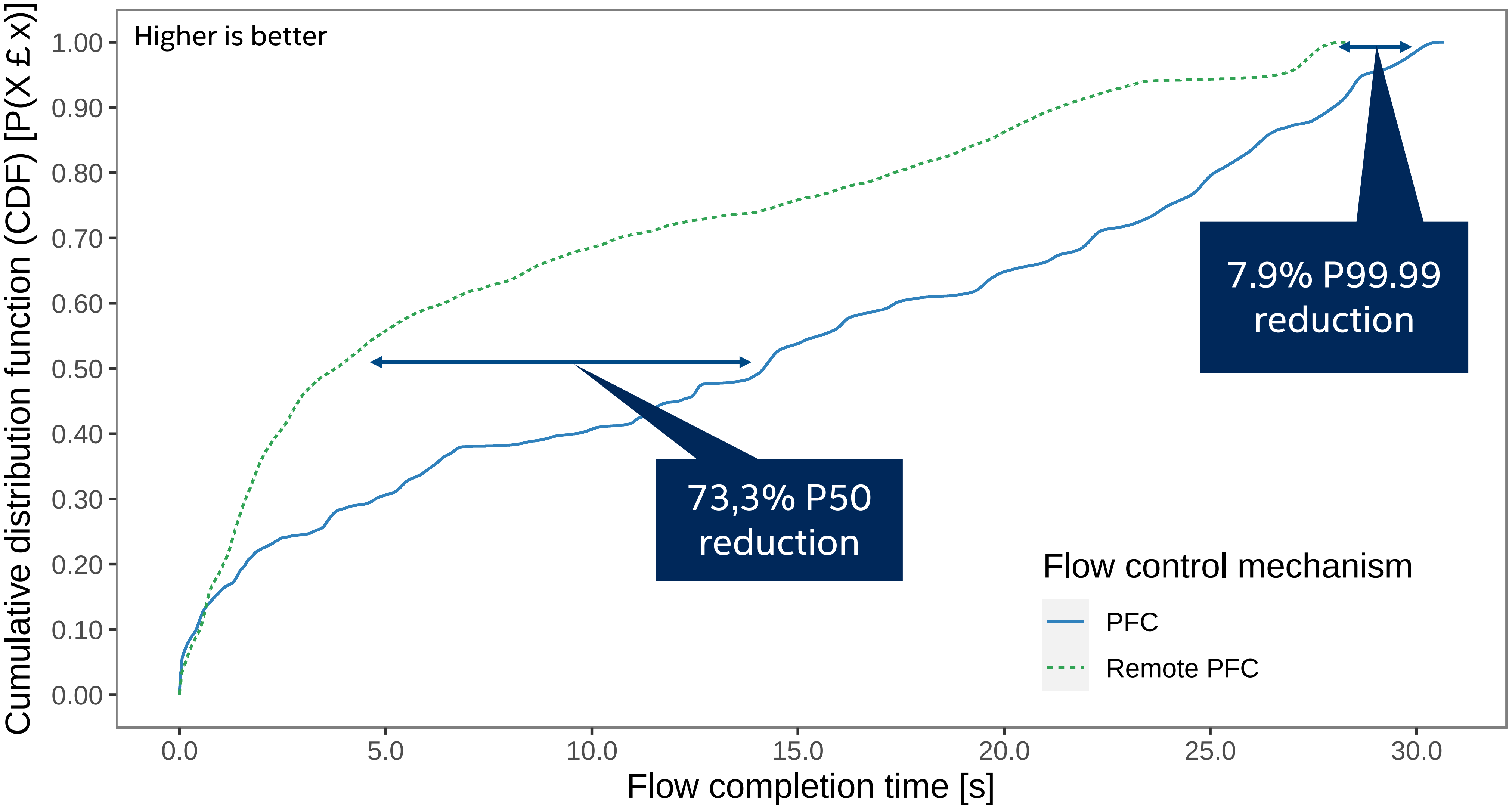


Lower is better

- **Workload**
  - RoCEv2 throughput test
  - Recv1 traffic: 4:1 incast
  - Recv2 traffic: 20:1 incast
- **Result**
  - Significantly reduce queue depth and head-of-line blocking in the network

See backup for workloads and configurations. Results may vary.

Remote PFC’s Effect on Flow Completion Time



73,3% P50 reduction

7.9% P99.99 reduction

See backup for workloads and configurations. Results may vary.

Summary

- Remote PFC
  - Flattens the buffer utilization curve for incast workloads in data centers
  - Leverages the programmability of Intel® Tofino™ 2/Tofino™ 3-based ToR switches for sub-RTT edge-to-edge congestion signaling
  - Compatible with standard NICs that support IEEE 802.1Qbb PFC
  - SONiC PINS enables Remote PFC's rapid deployment in production environments
- Future
  - Upstream to SAI
  - Ongoing efforts to standardize Remote PFC at IEEE 802.1
  - Generalize the Remote PFC approach to providing flow control directly in the protocol engine in the sender as Source Flow Control (SFC)

Notices and Disclaimers

- Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
- Your costs and results may vary.
- Intel technologies may require enabled hardware, software or service activation.
-

intel®

Switch Config

	Switch Config1 (Remote PFC “off”, PFC “on”)	**Switch Config2 (Remote PFC “on”, PFC “off”)**
Test by	Intel	
Test date	04/08/2021	
SUT Setup		
Platform	Accton AS9516 32d-r0	
# Switches	2 (ToR1, ToR2)	
HWSKU	Newport	
Ethernet switch ASIC	Intel® Tofino™ 2 Programmable Ethernet Switch ASIC	
SDE version	9.5.0-9388-pr	
OS	SONiC.master.111-dirty-20210201.022355	
Buffer Pool allocation	Ingress Lossless pool size is 7.6MB and lossy pool size is 7.6MB. Egress lossless pool size is 16.7MB, and lossy pool size is 6.4MB.	
Remote PFC threshold	N/A	100KB
PFC threshold	Headroom size is 184KB, dynamic threshold is 4.	N/A

Server Config

	Two server models (A and B) are used at the same time in the testbed	
Server model	Model A	Model B
Test by	Intel	Intel
Test date	04/08/2021	04/08/2021
Server Setup		
Platform	Intel S2600WFT	Supermicro X10DRW-i
# Nodes	3 (Send 6, Recv 1, 2)	5 (Send 1, 2, 3, 4, 5)
# Sockets	2	2
CPU	Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
Cores/socket, Threads/socket	18/36	8/16
Microcode	0x5003003	0xb000038
HT	On	On
Turbo	On	On
Power management (disabled/enabled)	enabled	enabled
# NUMA nodes per socket (1, 2, 4…)	2	2
Prefetcher’e enabled (svr_info)	Yes	Yes
BIOS version	SE5C620.86B.02.01.0008.031920191559	3.0a
System DDR Mem Config: slots / cap / speed	6 slots / 16GB / 2934 (*)	8 slots / 32 GB / 2133
Total Memory/Node (DDR, DCPMM)	96, 0	256, 0
NIC	1x 2x100GbE Mellanox ConnectX-6 NIC	1x 2x100GbE Mellanox ConnectX-6 NIC
PCH	Intel C620	Intel C610/X99
Other HW (Accelerator)	RoCEv2 protocol engine in Mellanox ConnectX-6 NIC	RoCEv2 protocol engine in Mellanox ConnectX-6 NIC
OS	Ubuntu 20.04.2 LTS	Ubuntu 20.04.2 LTS
Kernel	5.4.0-66-generic	5.4.0-66-generic
Workload	Custom trace based on Homa (Sigcomm 2018) “Facebook Hadoop” dataset	Custom trace based on Homa (Sigcomm 2018) “Facebook Hadoop” dataset
Compiler	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Libraries	MLNX_OFED_LINUX-5.1-2.5.8.0 (OFED-5.1-2.5.8)	MLNX_OFED_LINUX-5.1-2.5.8.0 (OFED-5.1-2.5.8)
NIC driver	mlx5_core	mlx5_core
NIC driver version	5.1-2.5.8	5.1-2.5.8
NIC Firmware version	20.28.2006 (MT_0000000224)	20.28.2006 (MT_0000000224)

*The memory population is per system. For server Model A only half of the memory channels are used per socket. This is a sub-optimal memory configuration compared to the best-known configuration where all memory channels are populated but is not a performance-critical issue. The performance-critical path for the workload runs in the RoCEv2 hardware engine of the RDMA NIC and accesses the memory controllers of the CPUs directly. The maximum network throughput on the NIC is limited to the port speed of 100Gbps. The maximum load on the memory controller is limited to 12.5GB/s and hence the memory controller is not a performance limiter.