Congestion Control for Large Scale RoCEv2 Deployments

Hemal V. Shah and Moshe Voloshin

Data Center Solutions Group (DCSG), Broadcom Inc.

Agenda

- **Ethernet for HPC**
- **MPI and Communication Topologies**
- **PFC Challenges**
- **RoCEv2 with Congestion Control**
- **Congestion Control Evaluation**
  - OSU Benchmarks
  - HPCG
  - LAMMPS
  - GPCNeT

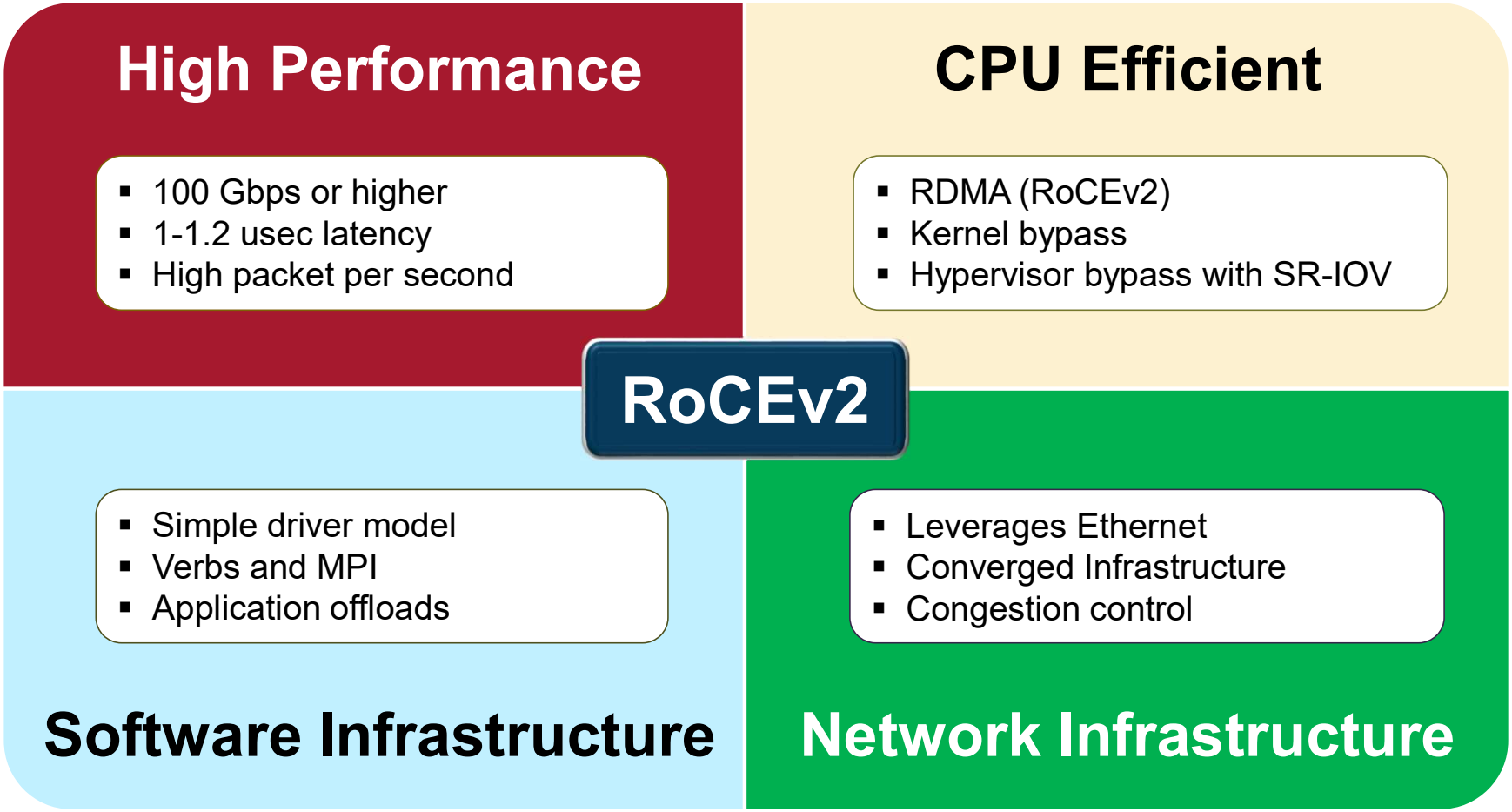
Why Ethernet for HPC? – High Level Analysis

Attribute	InfiniBand	Ethernet	Notes
Scalability (# of Nodes)	1000s	1000s	Ethernet as a fabric scales
Throughput	200G	200G	Ethernet network speeds scale
E2E Latency (B2B) – ½ RT	600 ns	1 – 1.2us	Unloaded RoCE latency < 2x IB latency
Switch Latency	~100ns	200-500ns	ETH switch latency in future ~200-300ns
Transport/Network	IBA Transport	RoCEv2/UDP/IP	Ethernet ecosystem not as limited as IB
Verbs API	OFA OFED	OFA OFED	Common host interface
Congestion Management	FECN/BECN	ECN	Standards based congestion control
Availability	Limited sources	Many sources	Large number of Ethernet suppliers

- **Ethernet has advantages in its ubiquitous deployment → Ethernet is Everywhere**
- **Ethernet ecosystem is mature and standards based → Benefits HPC markets**
- **100/200G Ethernet latencies (w/ RoCE) meets HPC requirements**

BROADCOM

RoCEv2 for HPC



BROADCOM

MPI and Communication Topologies

- **MPI is widely used in HPC/ML clusters as the communication layer**
- **A process group in MPI represents a collection of processes**
- **The number of processes can be 100s per node**
- **The number of nodes can scale to 1000s in a cluster**
- **The communication pattern is represented by a logical topology**
  - Ring, Binary cube, Tree, etc.
- **Selection of logical topologies depends on apps and communication libs**
- **MPI collectives (Gather, Reduce..) can create congestion in the network**

Challenges with PFC without Congestion Control

- Priority Flow Control (PFC) is used for lossless service
- PFC is a point-to-point protocol between two Ethernet endpoints
- PFC can result in congestion spreading
- PFC can create PFC storm due to slow receivers
- PFC may result in transport live-lock

Congestion Control (CC) with RoCEv2

- **ECN based CC schemes do not require any additional infrastructure support**
- **Congestion control without PFC can be sufficient for most of the workloads**
- **CC with PFC addresses PFC storms/live locks & preserves lossless service**
- **Even w/ large number of competing flows switch egress Q peaks are low**
- **Reaction by sender is quick – few 10s of micro-seconds due to low Q level**
  - Even with low marking threshold, network utilization is high
- **Low marking threshold delivers low end-to-end latency with min interference**
- **Low marking threshold leaves majority of switch buffer for incast absorption**
- **Both probabilistic and deterministic marking are possible**

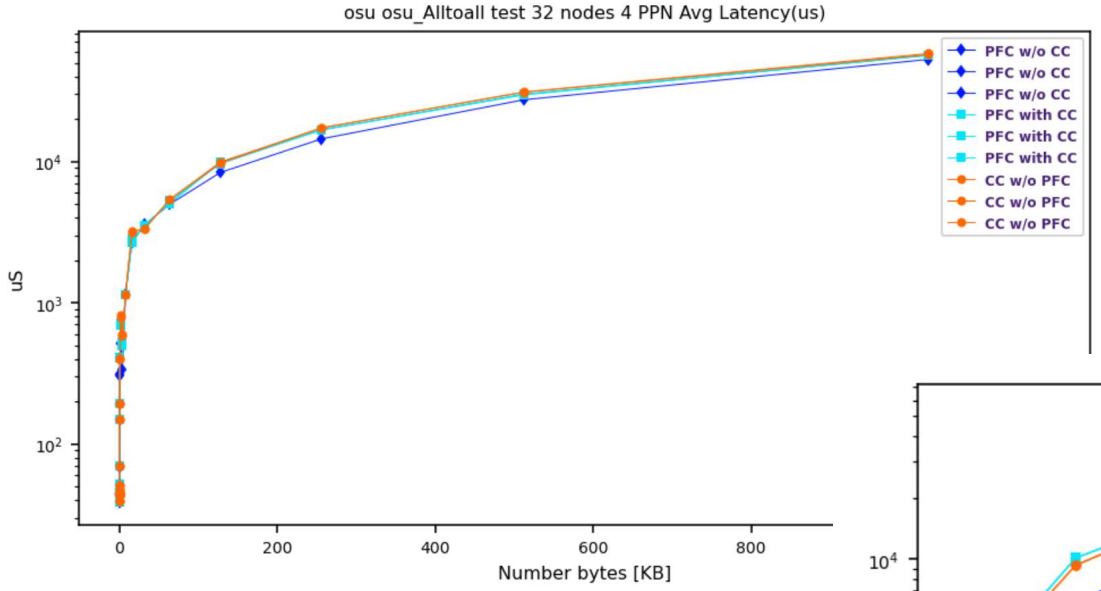
RoCEv2 Application Performance Under Congestion

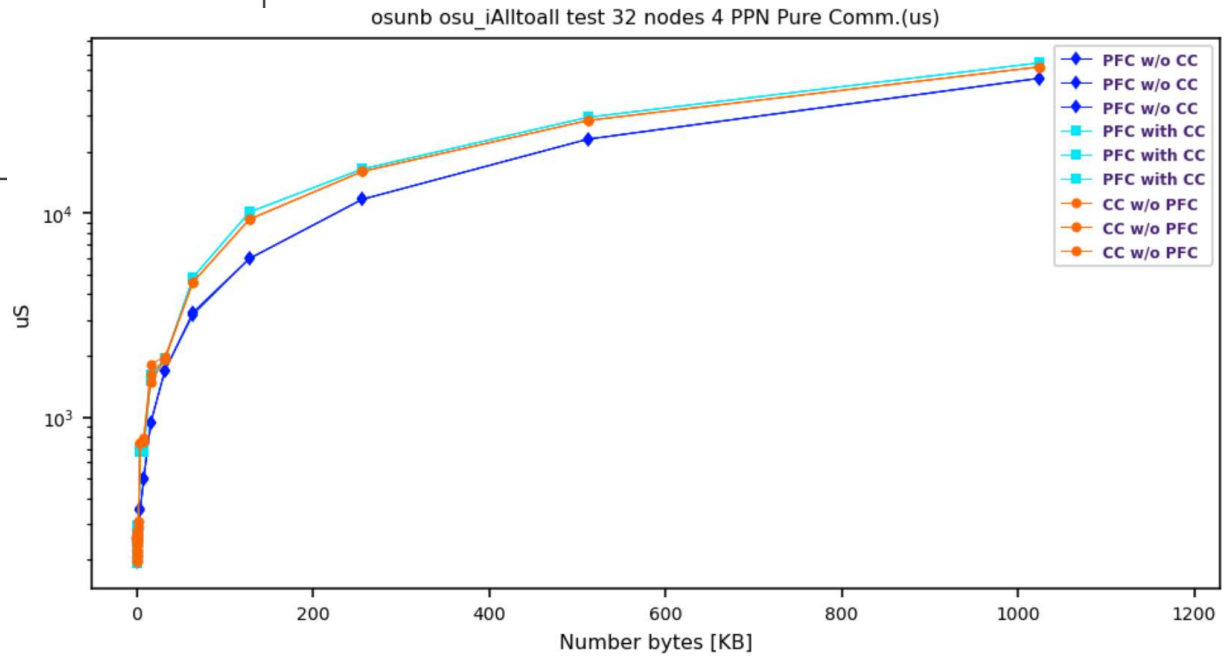
Test Scenario	Overview
OSU Benchmarks	Blocking and non blocking Collective benchmarks for various sizes and with various PPN (Processes Per Node) over 2 to 32 nodes
HPCG	High Performance Conjugate Gradient Benchmark for HPC, with 8, 16, 32 PPN on 8, 16, 32 nodes
LAMMPS	5 benchmarks of Molecular Dynamics with 32,000 atoms per core Scaling efficiency charts relative to CPU time on single core for 32,000 atoms Chart title shows 1 node 1 PPN loop time in seconds for 32,000 atoms

All tests ran with NIC link BW of 100 gbps in 3 configurations: PFC without CC, PFC with CC and CC without PFC

BROADCOM

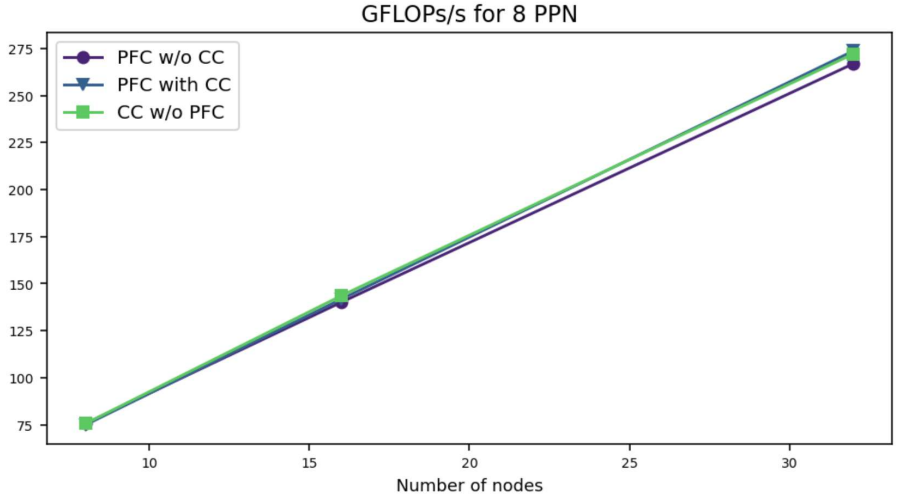
OSU Benchmarks Results – completion time [uS]

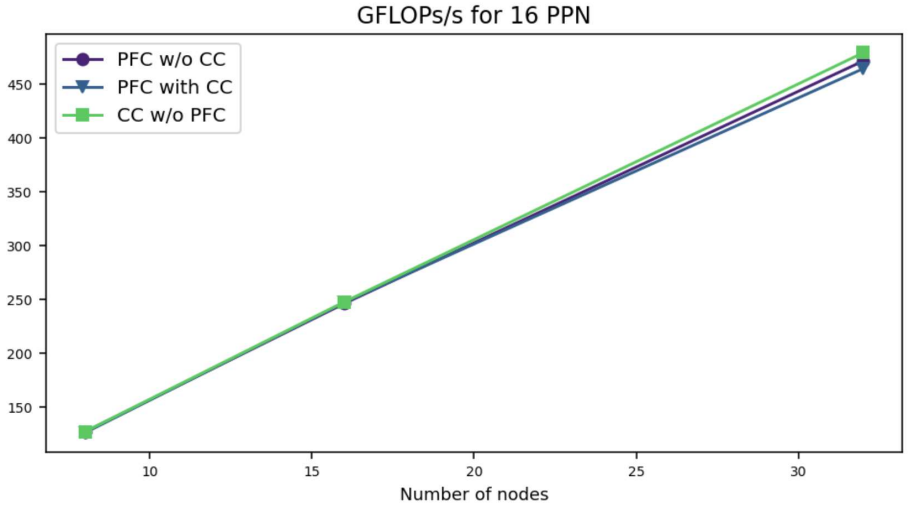


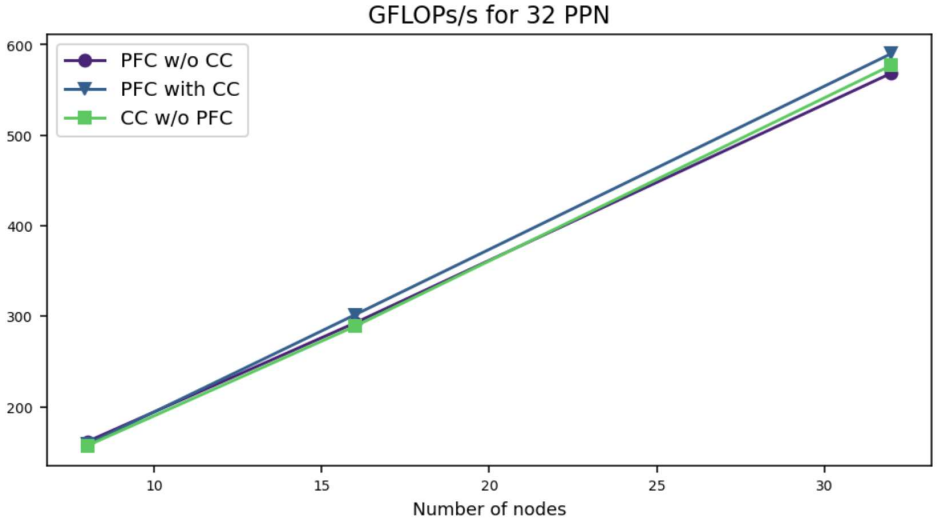


BROADCOM

HPCG Results



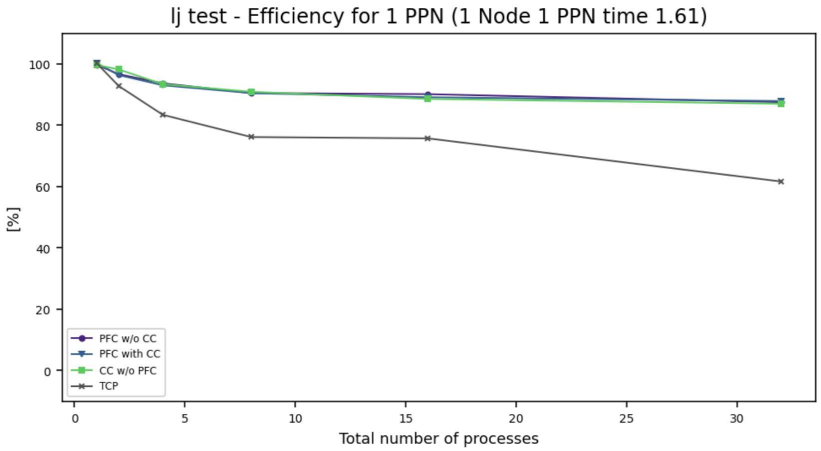


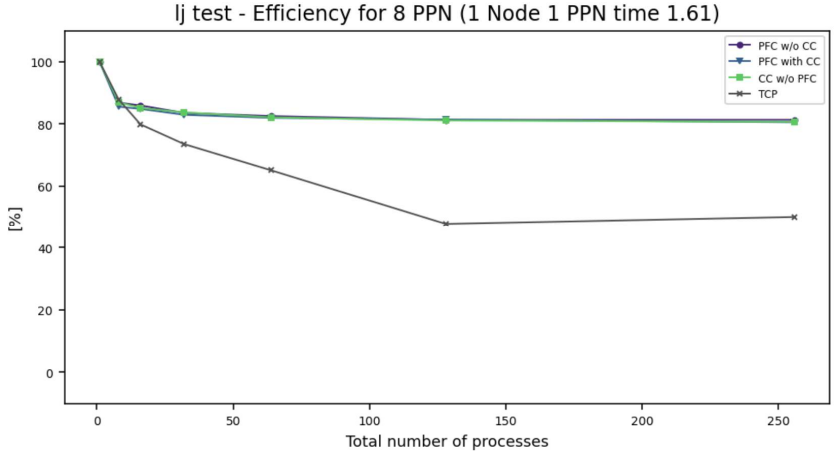


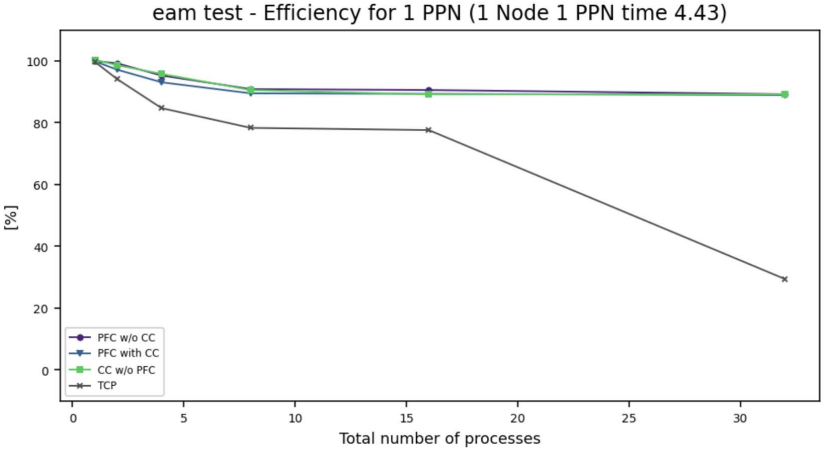
- HPCG application scales well in all configurations with varying PPN

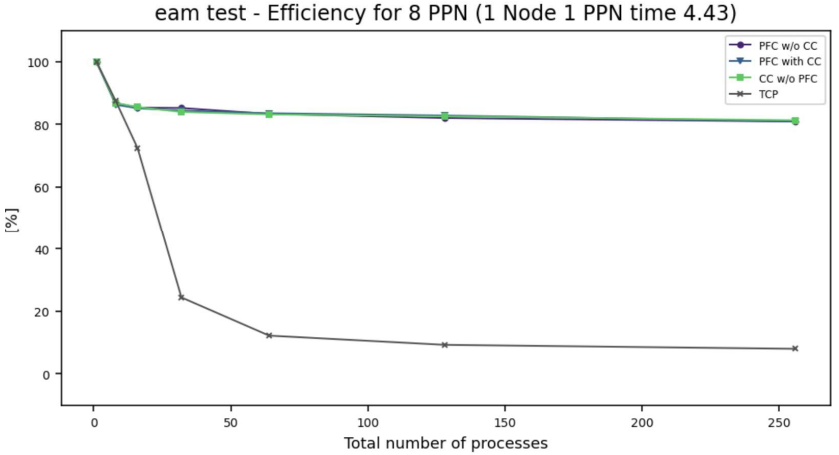
BROADCOM

LAMMPS efficiency vs. # processes



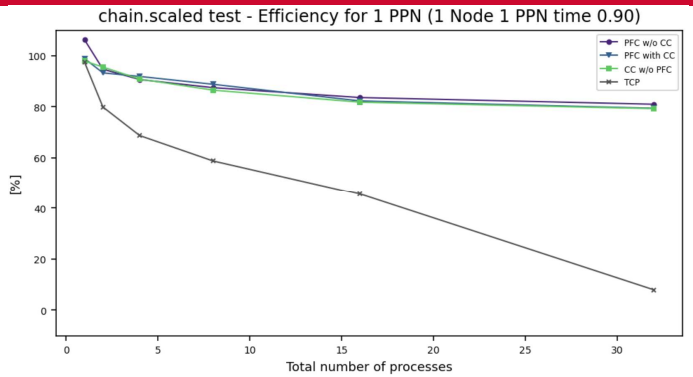


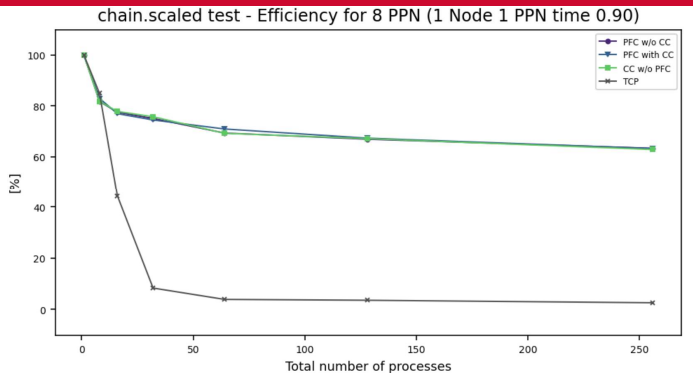


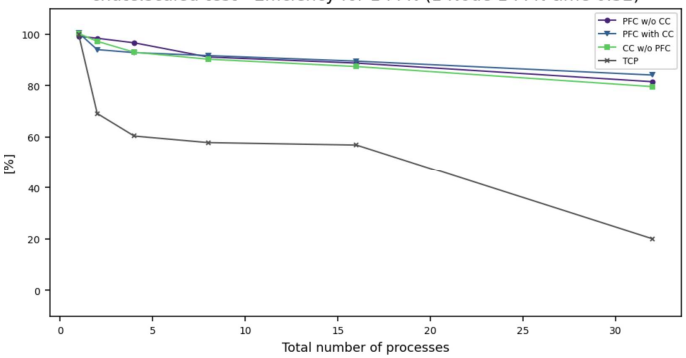


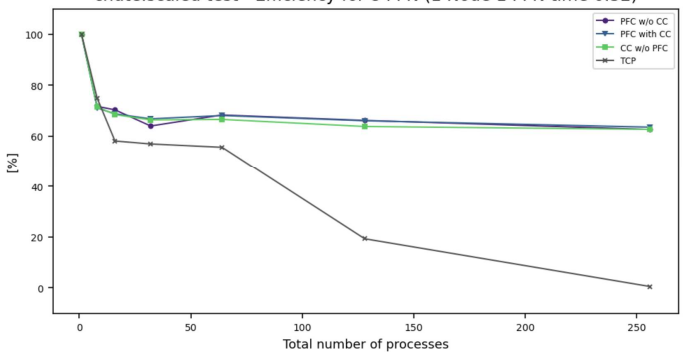
- ROCE significantly outperform TCP in all configurations

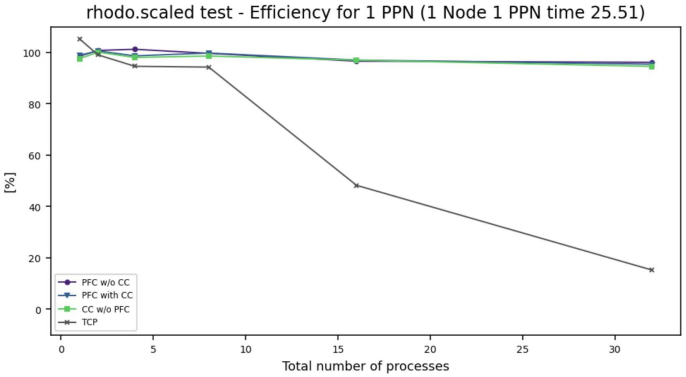
BROADCOM

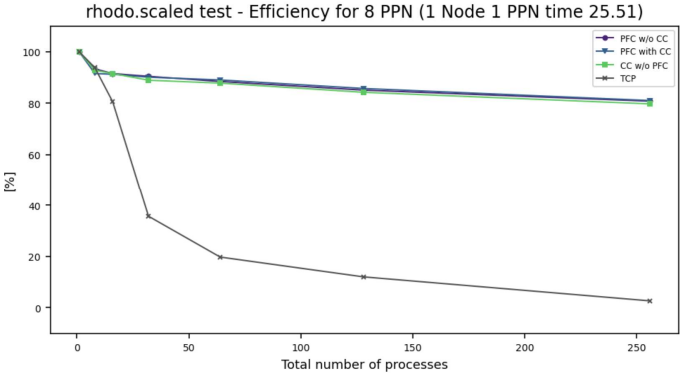

chain.scaled test - Efficiency for 1 PPN (1 Node 1 PPN time 0.90)
PFC w/o CC
PFC with CC
CC w/o PFC
TCP
[%]
Total number of processes


chain.scaled test - Efficiency for 8 PPN (1 Node 1 PPN time 0.90)
PFC w/o CC
PFC with CC
CC w/o PFC
TCP
[%]
Total number of processes


chute.scaled test - Efficiency for 1 PPN (1 Node 1 PPN time 0.32)
PFC w/o CC
PFC with CC
CC w/o PFC
TCP
[%]
Total number of processes


chute.scaled test - Efficiency for 8 PPN (1 Node 1 PPN time 0.32)
PFC w/o CC
PFC with CC
CC w/o PFC
TCP
[%]
Total number of processes


rhodo.scaled test - Efficiency for 1 PPN (1 Node 1 PPN time 25.51)
PFC w/o CC
PFC with CC
CC w/o PFC
TCP
[%]
Total number of processes

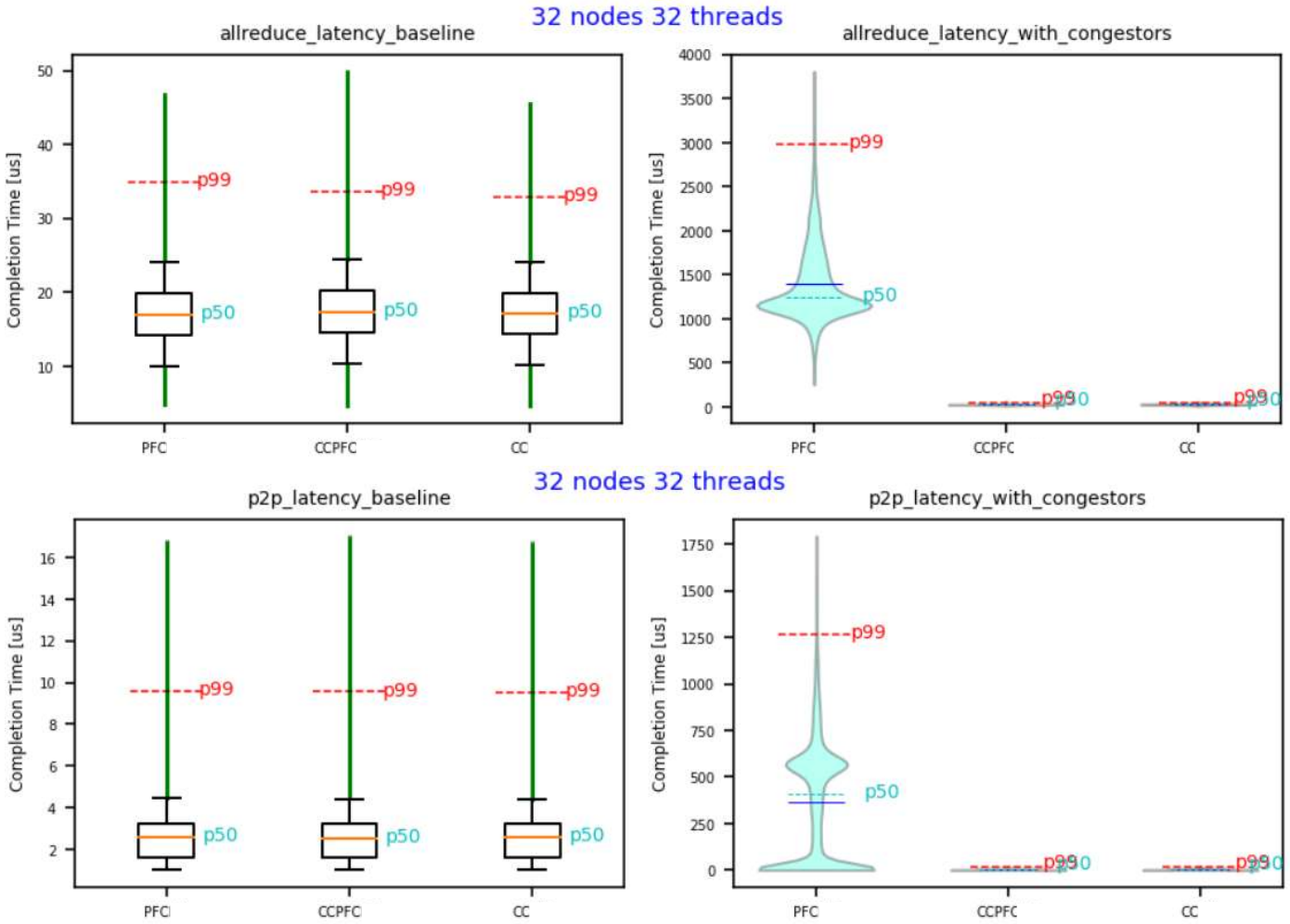

rhodo.scaled test - Efficiency for 8 PPN (1 Node 1 PPN time 25.51)
PFC w/o CC
PFC with CC
CC w/o PFC
TCP
[%]
Total number of processes

BROADCOM®

GPCNeT benchmark

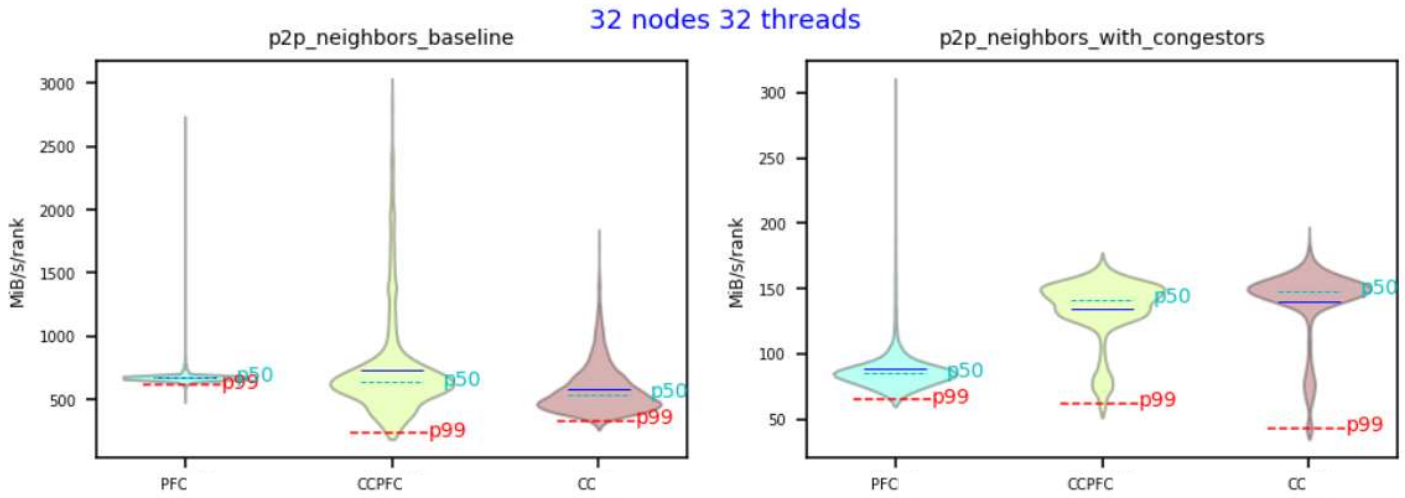
GPCNeT	Global Performance and Congestion Network Test MPI test designed to measure relative performance under load and congestion Designed for large multi-layer switch network • 20% of nodes w/ test tasks: allreduce, p2p latency, random ring neighbor exchange • 80% of nodes assigned with congestor tasks: All2All, incast, RMA put and get • Nodes would share switch buffering resources and will cross path between switches

GPCNeT results – Results on 32 Nodes



BROADCOM

GPCNeT results – Results on 32 Nodes



CCONLY CC+PFC PFC

Network Tests running with Congestion Tests - Key Results			ith Congestion Tests - Key Results		ith Congestion Tests - Key Results	
Name	Congestion		Congestion Impact		Congestion Impact Factor	
	Avg		Avg		Avg	99%
RR Two-sided Lat (8 B)	1.8X		1.8X		120.9X	132.5X
RR Two-sided BW+Sync (131072 B)	3.9X		4.6X		7.9X	9.4X
Multiple Allreduce (8 B)	1.7X		1.5X		79.8X	87.1X

BROADCOM

Congestion Control (CC) for RoCEv2

- **RoCEv2 is designed to scale – no inherent limitation at the protocol level**
- **ROCEv2 demonstrate significant performance advantage over TCP**
- **PFC provides lossless service with a significant interference/blocking**
  - PFC without congestion control should be avoided
- **CC with or without PFC is essential for node and process scaling**
  - ECN marking in switches enable Congestion Notifications to minimize congestion
  - CC algorithms are evolving to provide better congestion avoidance & faster congestion reaction
  - CC algorithm that maintains low switch queue level reduces interference/blocking
- **CC enhancements in NICs further improve performance & scalability**
  - ECN marking/CNP generation
  - Hardware-based congestion control
  - Deterministic marking policy (DCTCP style)

Thank You

BROADCOM®


BROADCOM®
connecting everything®