2021 OFA Virtual Workshop

PERFORMANCE SCALED MESSAGING 3
PERFORMANCE STUDIES

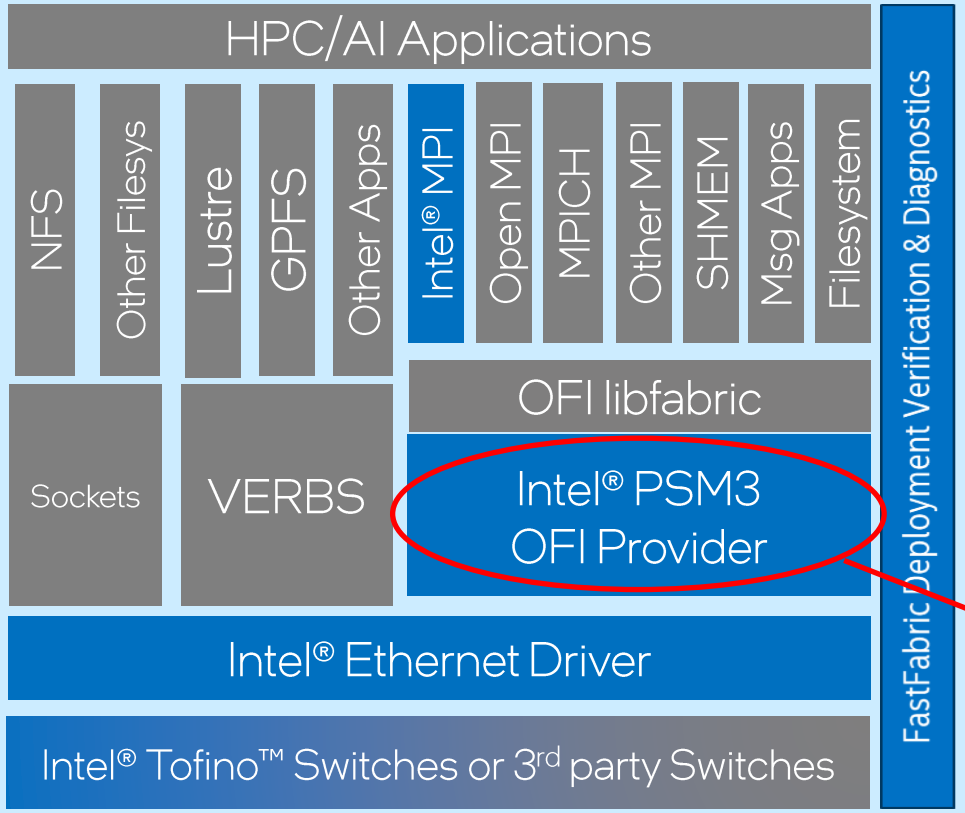
James Erwin, Software and Performance Engineer

Intel Corporation

OUTLINE

- PSM3 Overview
- PSM3 Performance Tuning Basics
- MPI Latency beyond 2 nodes - at scale
- MPI Bandwidth vs RDMA mode - UD and RC
- MPI Bandwidth Saturation
- Multi-rail with PSM3
- QP footprint - PSM3 vs RxM
- Application performance at scale
- Priority Flow Control - setup, validation, debug

HPC/AI COMMS ARCHITECTURE WITH PSM3



3rd generation Performance Scaled Messaging (PSM3)

- Evolution of PSM (TrueScale) & PSM2 (OPA)
- Enhanced for Ethernet and RoCE v2

Compatible with existing MPI applications

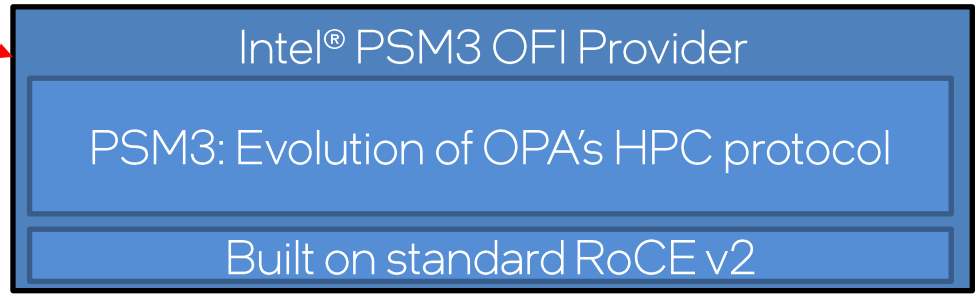
- No code changes necessary

Leverages the OpenFabrics Alliance*

- Standards based software

Provides communications for OneAPI

- Open common environ for CPU/GPU/Accel



PSM3 - PERFORMANCE TUNING BASICS

User Libraries

- **Use the latest available PSM library**
- **Intel MPI 2019.10 and newer preferred (but works with older versions too)**

System/OS

- **Latest ice (Intel® Ethernet) and irdma (Intel RDMA) drivers**
- **Rendezvous (RV) kernel module**
- **Enable Intel® Turbo Boost Technology and performance CPU governor**

Network

- **Priority Flow Control configuration (important for high stress traffic)**

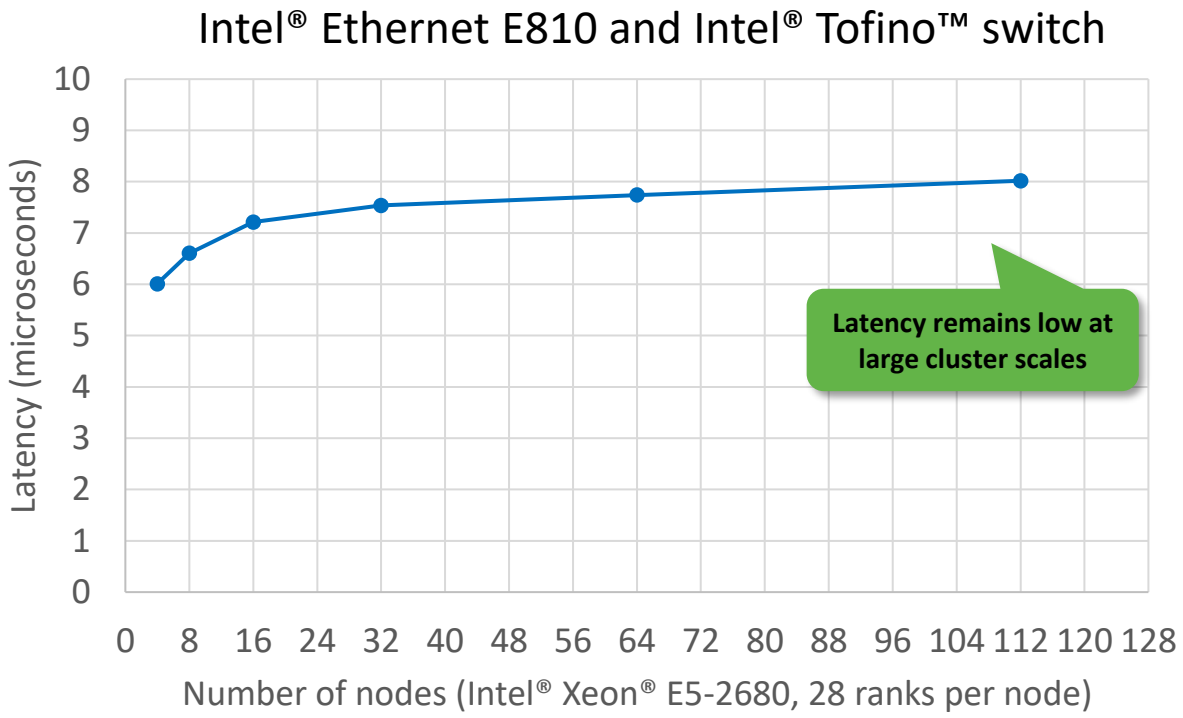
*Intel® Ethernet Fabric Suite

Consult the ***Intel® Ethernet Fabric Suite Performance Tuning User Guide*** for more detail

PSM3 - LATENCY BEYOND 2 NODES

HPCC 1.5.0, Randomly Ordered Ring Latency

- **Point to point latency is not the only latency that matters**
- **HPC app performance is sensitive to low latency at larger cluster scales**
- **PSM3 latency remains ~ flat**
  - Tested up to 112 nodes, 3,136 cores



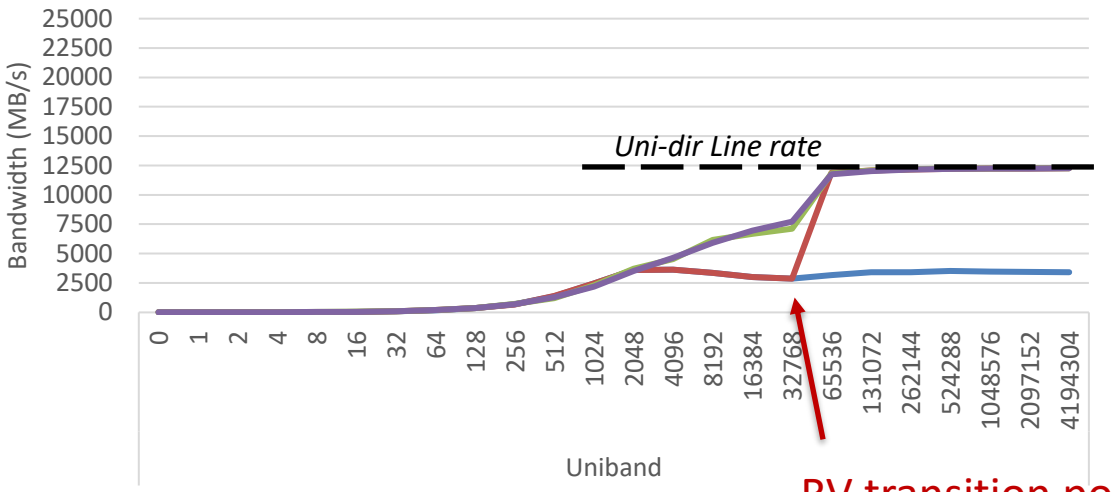
Intel MPI 2019.10, Intel Ethernet Fabric Suite PV Release

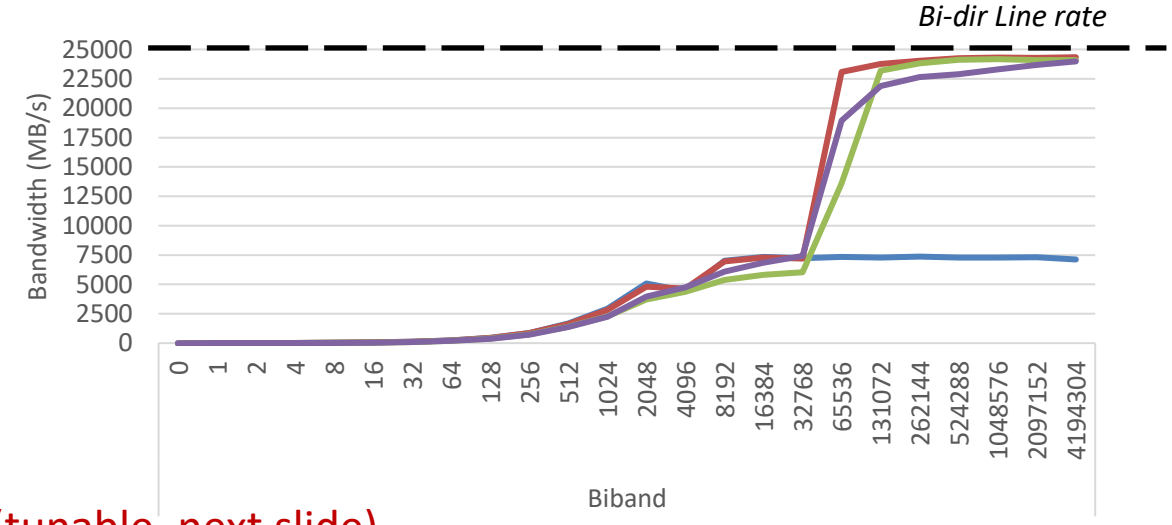
PSM3 - BANDWIDTH - UD AND RC

Intel MPI Benchmarks, 2 nodes, 1 rank per node

- **UD mode has lowest memory footprint and highest scalability**
- **RC modes deliver line rate performance using a single MPI rank**

•PSM3_RDMA=0 – UD only
•PSM3_RDMA=1 – Kernel RC QP for Rendezvous RDMA
•PSM3_RDMA=2 – User space RC QP for Rendezvous RDMA, UD for everything else.
•PSM3_RDMA=3 – User space RC QP for Rendezvous RDMA and eager data, control via UD





RV transition point (tunable, next slide)

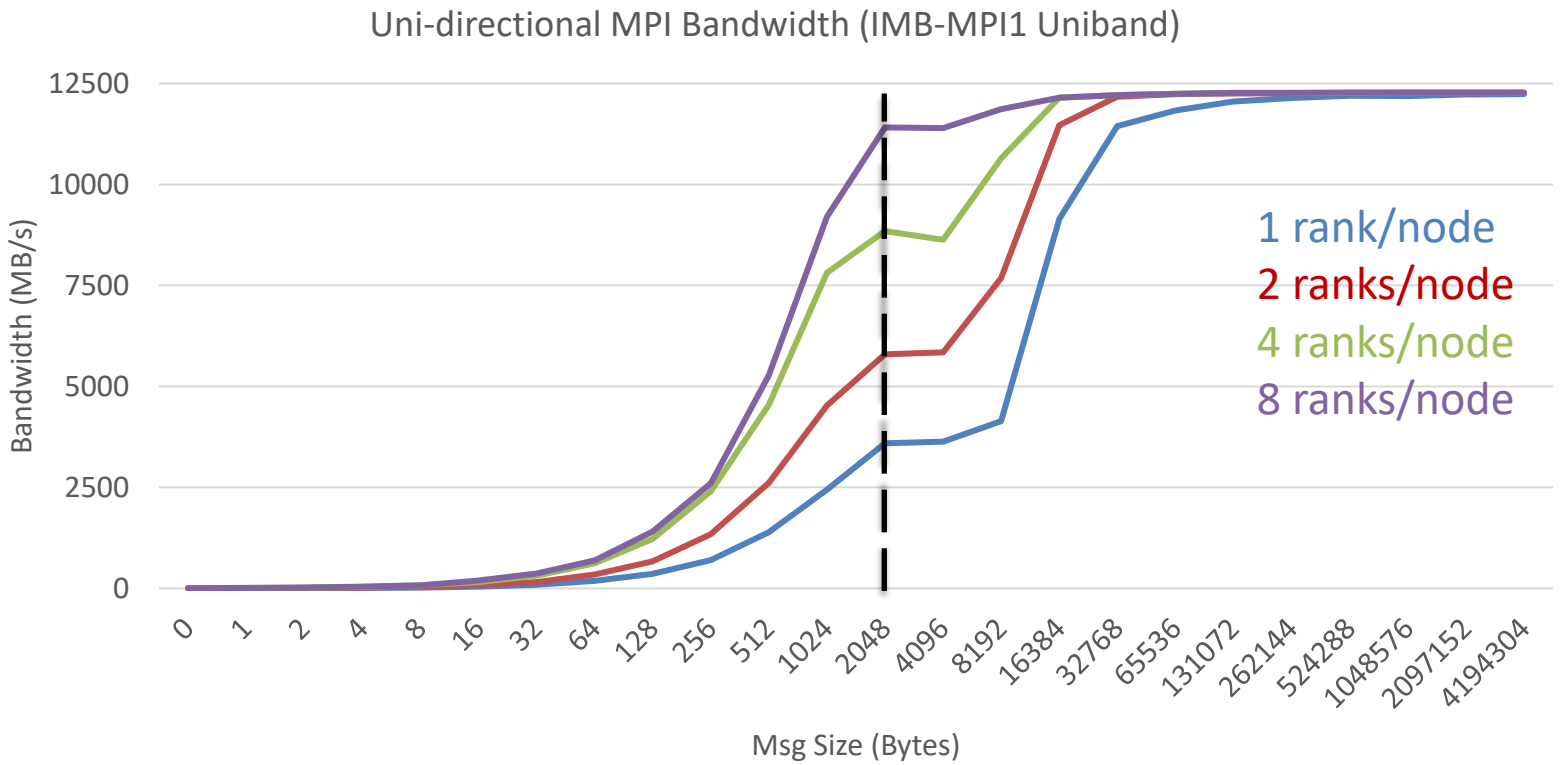
average of 10 runs shown

Intel MPI 2019.10, Intel Ethernet Fabric Suite 11.0.0.0.162

PSM3 - BANDWIDTH SATURATION

Intel MPI Benchmarks, 2 nodes, various rank per node

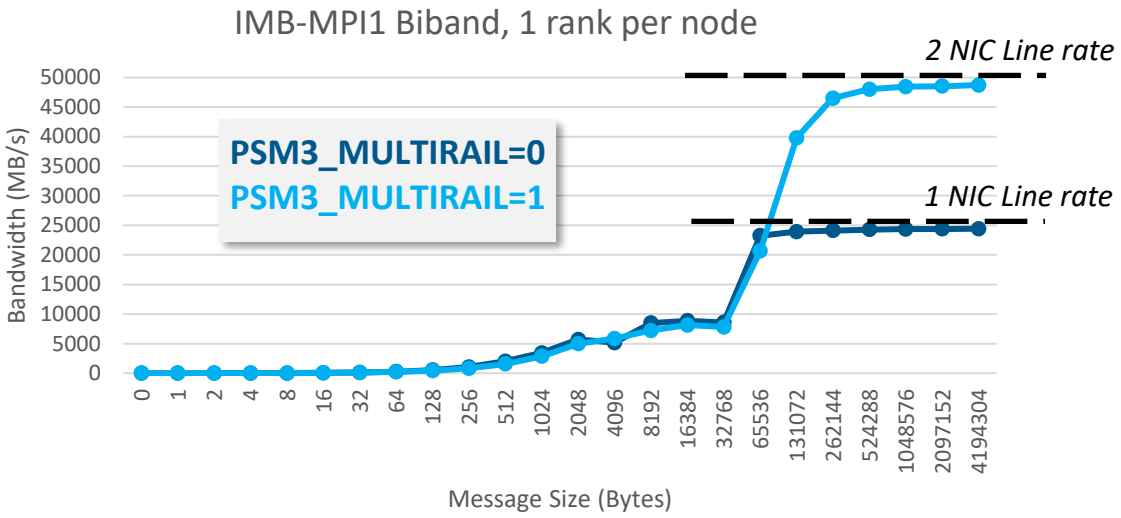
- Kernel RC mode (PSM3_RDMA=1)
- Bandwidth is saturated around **2K msg size with 8 ranks**
- Additional PSM3 tuning:
  - PSM3_MQ_RNDV_NIC_THRESH=8000 (default 64000)
  - Transition to RV at smaller msg sizes
  - May become default in a future release



Intel MPI 2019.10, Intel Ethernet Fabric Suite 11.0.0.0.162

PSM3 - MULTIRAIL PERFORMANCE

Intel MPI Benchmarks, 2 nodes, 1 rank per node



- PSM3_RDMA=1 (Kernel RC)
- PSM3_MULTIRAIL=1 forces an MPI rank to stripe messages across all available RDMA ports
- PSM3_MULTIRAIL_MAP can be more prescriptive about which ports to use

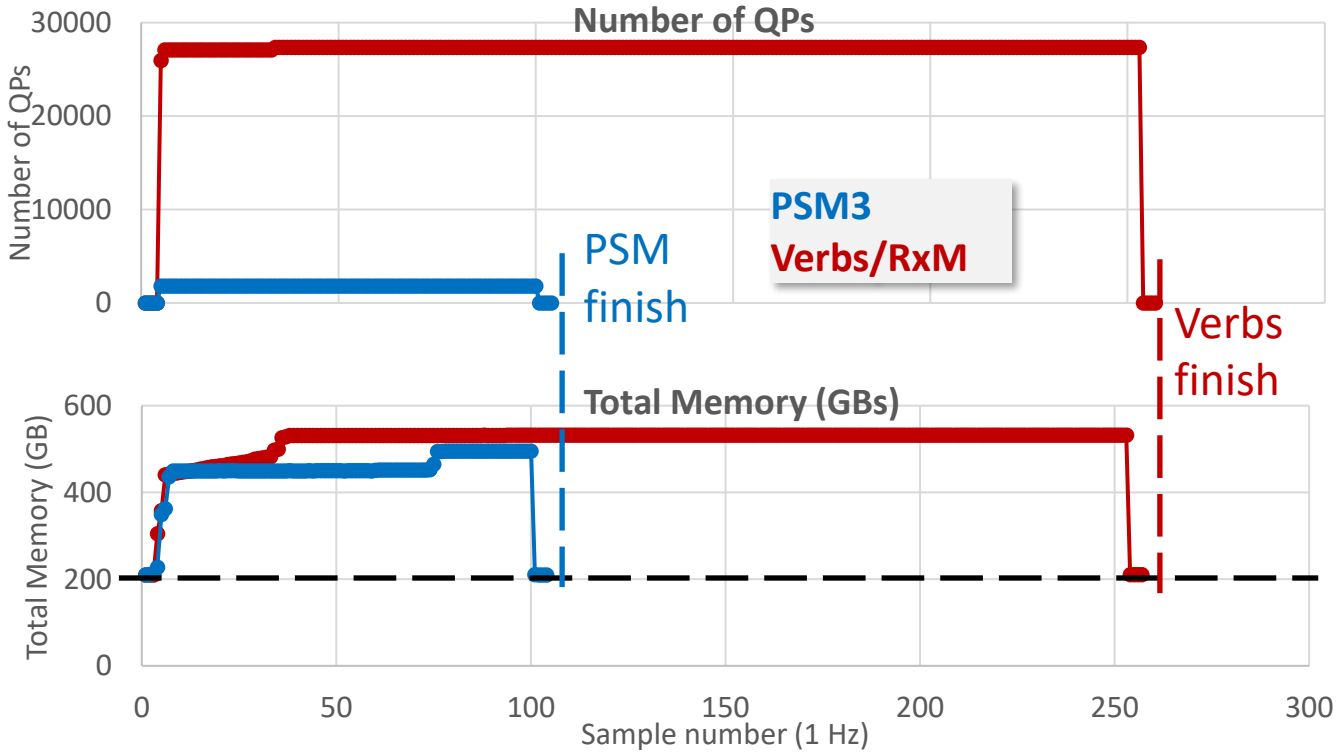
Intel® Xeon® Platinum 8360Y Processor

Intel MPI 2019.10, Intel Ethernet Fabric Suite 11.0.0.0.162

QP FOOTPRINT - PSM3 VS VERBS/RXM

WRF 3.9.1.1 - CONUS 2.5KM Benchmark

https://www2.mmm.ucar.edu/wrf/WG2/benchv3/



- 64 nodes: Intel® Xeon® CPU E5-2680 v4
- PSM3 in UD mode (PSM3_RDMA=0) has low QP and memory footprint relative to Verbs/RxM

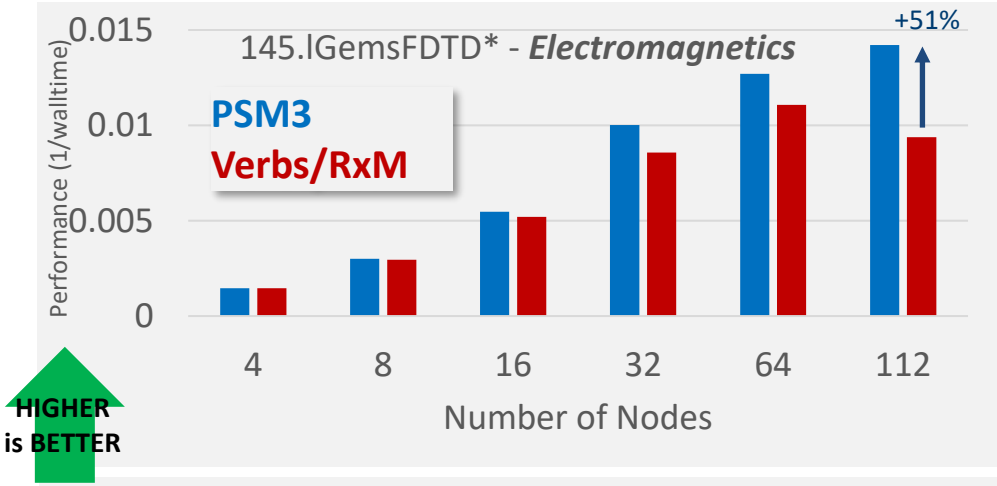
PSM3_RDMA=1 ~ 495GB, 9696 QPs

Intel® Ethernet E810 and Intel® Tofino™ switches

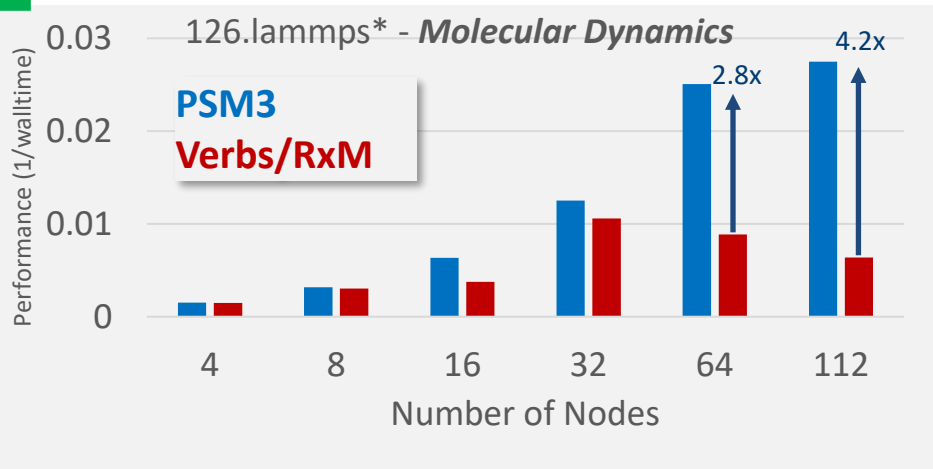
Intel MPI 2019.9, Intel Ethernet Fabric Suite 11.0.0.0.162

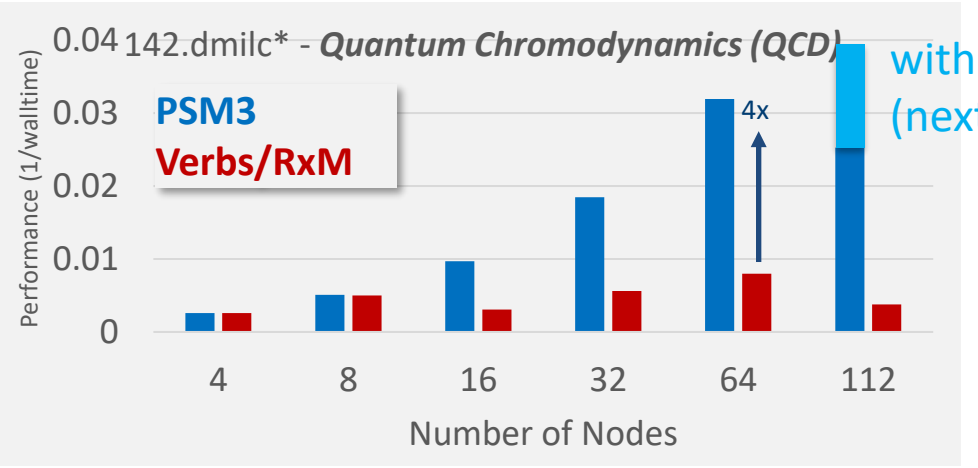
HPC APPLICATION PERFORMANCE AT SCALE

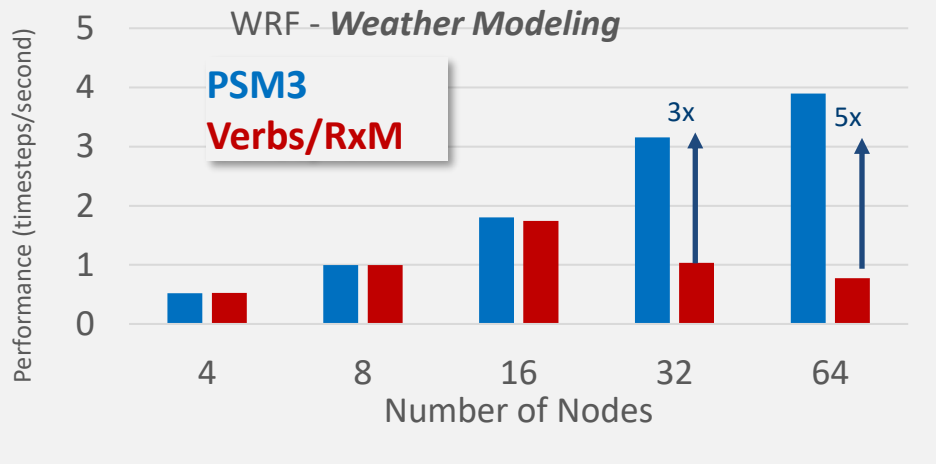
up to 112 nodes - 3,136 cores



HIGHER is BETTER







- PSM3 enables applications to scale on Ethernet
- Verbs does not scale as well
- All performance shown is UD (PSM3_RDMA=0)
- PFC is functional but not fully tuned on this cluster (next slide...)

Intel MPI 2019.9, Intel Ethernet Fabric Suite PV Release

*SPEC MPI2007 results have not been submitted to or reviewed by SPEC and are therefore marked as estimates

PRIORITY FLOW CONTROL

- Various PFC recipes available
  - Common: DCB configured on switch, willing mode on NICs
- Switch documentation for PFC is not always clear or complete
  - PFC configuration must be validated
- A symptom that PFC is not configured/tuned correctly: pacing the senders (reducing PSM3_FLOW_CREDITS) improves performance
  - PSM3_PRINT_STATS is a light-weight profiling tool that highlights drops/retries
- PFC is configured properly IFF:
  - No Tx/Rx drops (Lossless) - you can check the hosts (ethtool) and switch counters
  - Xon/Xoff counters (both Tx and Rx) are present
- PFC may need to be "tuned"
  - Example: Arista 7170 PFC "Headroom" needs to be increased from 18,560 bytes to at least 48,800 bytes (32 nodes)
  - Results on previous slide are without headroom increase. PSM3 throttling improves performance as a work-around
  - Detailed guidance from Intel is available

```
[user@node01 ~]$ ethtool -S cvl0 | grep priority_0
     tx_priority_0_xon.nic: 20287
     tx_priority_0_xoff.nic: 244655362
     rx_priority_0_xon.nic: 42871
     rx_priority_0_xoff.nic: 43117
```

host tx to switch

host rx from switch

SUMMARY AND NEXT STEPS

- Summary
  - PSM3 is a new libfabric provider that allows for scalable application performance
  - Two-node microbenchmarks are not indicative of application performance at scale
  - PSM3 latency, memory, and QP counts scale well
  - Proper configuration of PFC is important for RoCE performance, especially at larger scale & high core counts
- Next Steps
  - Explore applications that benefit from single-threaded MPI bandwidth (RDMA mode 1)
  - Expand range of application testing
    - PSM3 statistics, PSM3_PRINT_STATS, can help identify sub-optimal behavior and potential tunings
  - Develop PFC recommendations for larger clusters

CONFIGURATIONS

- PSM3 - Latency Beyond 2 nodes: Tests performed on 2 socket Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with Intel Pstate driver. CentOS Linux 8 (Core). 4.18.0-147.8.1.el8_1.x86_64 kernel. 8xDDR4, 256 GB, 2400 MT/s. irdma version 1.3.19. ice version 1.3.2. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0. HPCC 1.5.0 Randomly Ordered Ring Latency, best of 3 runs. Intel MPI 2019.10. IEFS 11.0.0.0.69, PSM3_RDMA=0 (default). Tofino: 2-tier switch fabric, DCS-7170-32CD TOR/edge, DCS-7170-64C spines. PFC enabled on priority 0. PFC enabled on priority 0 on TOR switches only, not spines.
- PSM3 - Bandwidth - UD and RC and PSM3 - Bandwidth Saturation: Tests performed on 2 socket Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with Intel Pstate driver. Red Hat Enterprise Linux 8.1 (Ootpa). 4.18.0-147.el8.x86_64 kernel. 12xDDR4, 196608 MB, 2666 MT/s. irdma version 1.3.19. ice version 1.3.2. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0. Arista 7170 (Intel® Tofino™) Ethernet switch, PFC enabled on priority 0. Tofino: Arista DCS-7170-32CD-F, 4.22.1FX-CLI.
- PSM3- Multirail performance: Tests performed on 2 socket Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40GHz. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with Intel Pstate driver. Red Hat Enterprise Linux 8.2 (Ootpa). 4.18.0-193.el8.x86_64 kernel. 16xDDR4, 256 GB, 3200 MT/s. irdma version 1.3.19. ice version 1.4.8. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0. Intel Ethernet Fabric Suite. Arista 7170 (Intel® Tofino™) Ethernet switch, PFC enabled on priority 0. Tofino: Arista DCS-7170-32CD-F, 4.22.1FX-CLI. One Ethernet NIC on socket 0, one Ethernet NIC on socket 1. 1 port active per NIC. PSM3_MULTIRAIL=1 PSM3_RDMA=1
- QP Footprint - PSM3 vs Verbs/RxM: Tests performed on 2 socket Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with Intel Pstate driver. CentOS Linux 8 (Core). 4.18.0-147.8.1.el8_1.x86_64 kernel. 8xDDR4, 256 GB, 2400 MT/s. irdma version 1.3.19. ice version 1.3.2. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0.Intel MPI 2019.10, FI_PROVIDER=psm3 and verbs. 2-tier switch fabric, Tofino: 2-tier switch fabric, DCS-7170-32CD TOR/edge, DCS-7170-64C spines. PFC enabled on priority 0. PFC enabled on priority 0 on TOR switches only, not spines.
- HPC Application performance at scale: Tests performed on 2 socket Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with Intel Pstate driver. CentOS Linux 8 (Core). 4.18.0-147.8.1.el8_1.x86_64 kernel. 8xDDR4, 256 GB, 2400 MT/s. irdma version 1.2.22. ice version 1.2.0_rc36. Intel® Ethernet E810 firmware-version: 2.15 0x800049c3 1.2789.0. Intel MPI 2019.9, FI_PROVIDER=psm3 and verbs. 2-tier switch fabric, Arista DCS-7060CX-32S TOR/edge, Arista DCS-7260CX-64 spines. PFC enabled on priority 0. 145.lGemsFDTD, 142.dmilc, 126.lammps stand-alone runs from SPEC MPI2007 results have not been submitted to or reviewed by SPEC and are therefore marked as estimates. WRF v3.9.1.1, conus2.5km.

2021 OFA Virtual Workshop

THANK YOU

James Erwin, Software and Performance Engineer

Intel Corporation