2022 OFA Virtual Workshop

UPDATE ON PSM3 CAPABILITIES AND ARCHITECTURE

James Erwin, Technical Lead, Software Enabling and Optimization Engineer

Todd Rimmer, Senior Principal Engineer

Intel Corporation

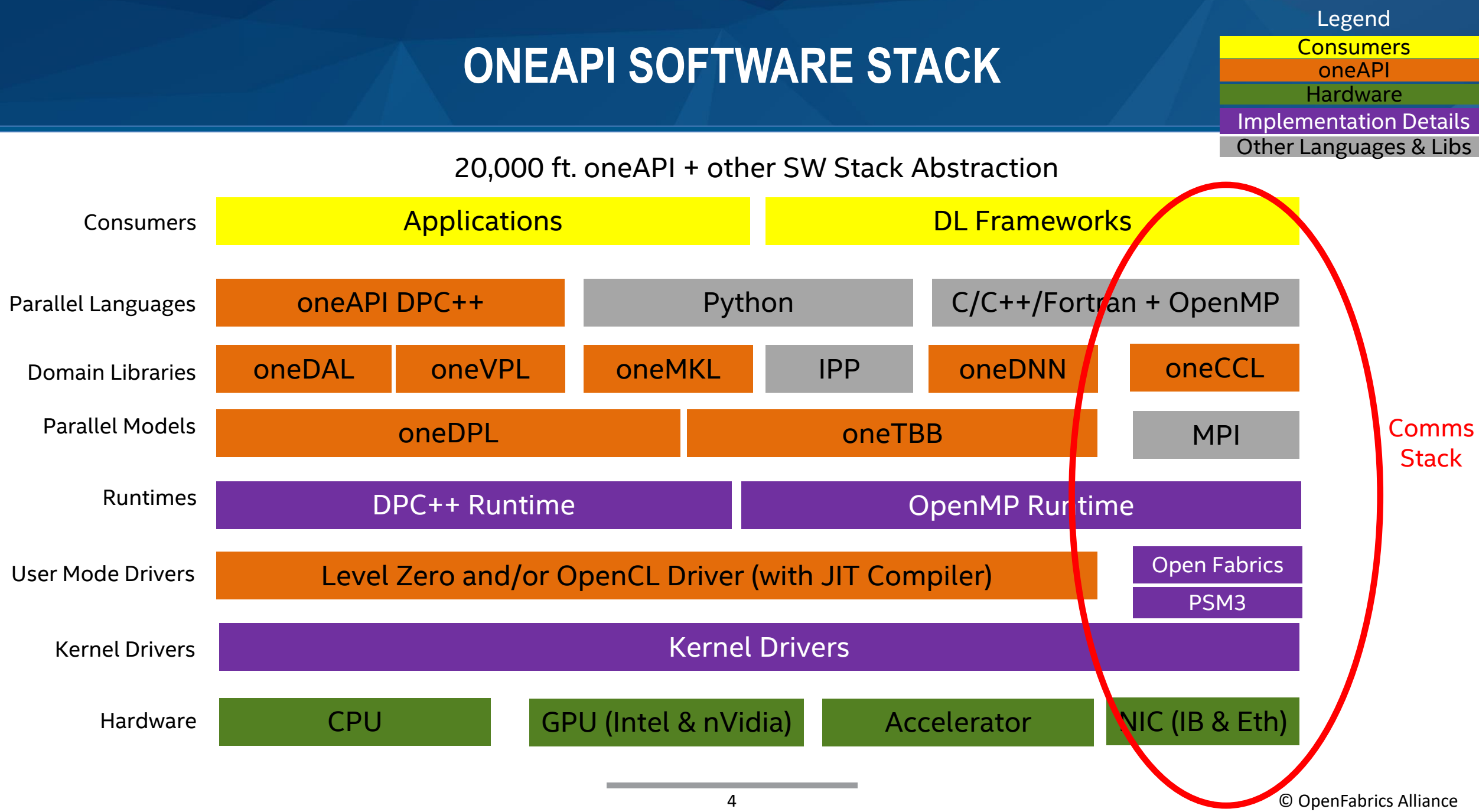
AGENDA

- What is PSM3?
- PSM3 Integration with oneAPI
- Evolution of the PSM3 Architecture
- New Capabilities
- PSM3 Use with NCCL
- Performance

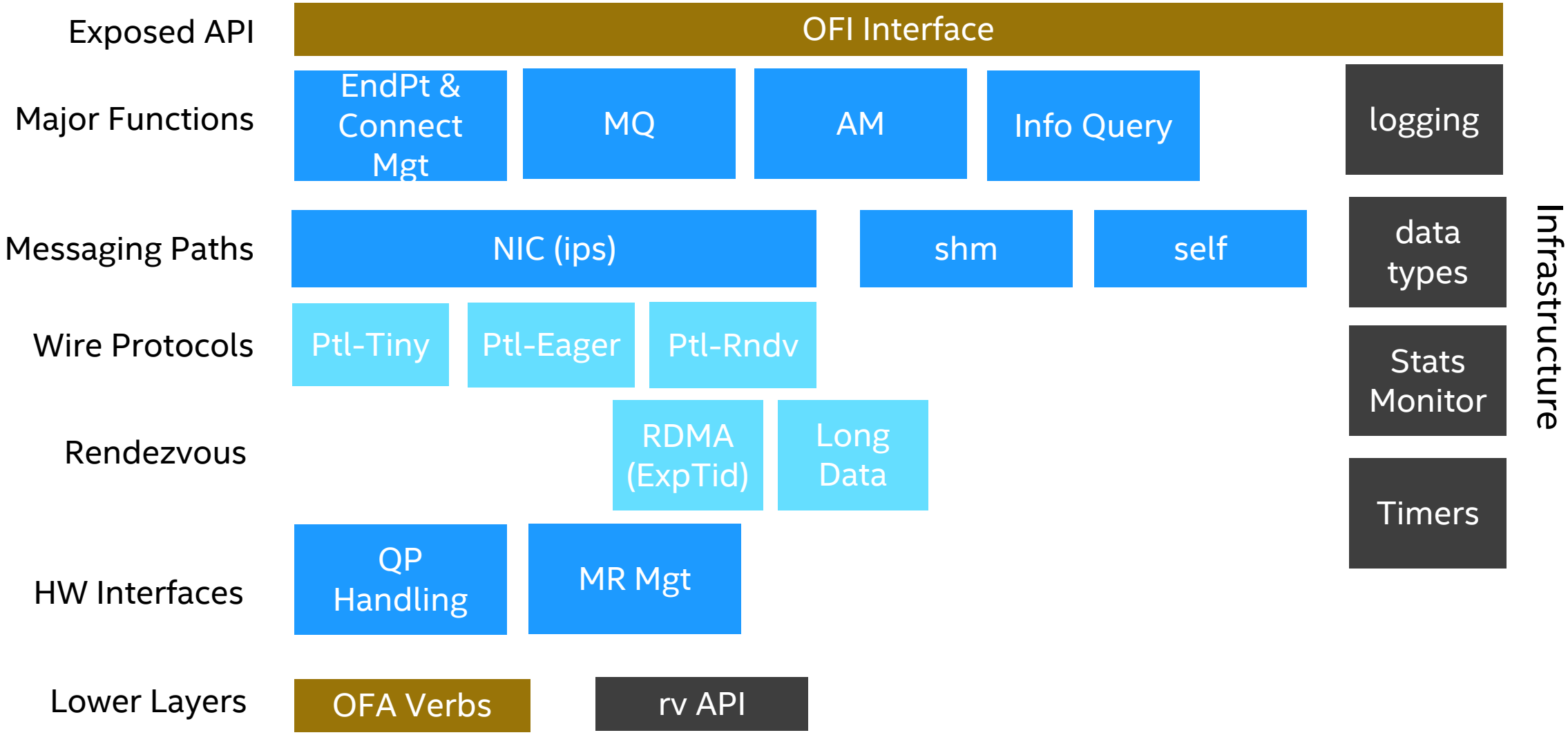
WHAT IS PSM3?

- **PSM3 is an existing libfabric provider**
  - Leverages concepts and code from Intel® Omni-Path Architecture (OPA)
  - Mature and Feature rich
- **PSM3 is designed for Ethernet**
  - Optimizes performance and scalability
  - Uses standard RoCEv2 protocols and APIs
  - Now also supports TCP/IP
- **PSM3 is available upstream now**
  - Latest version (3.2) now available upstream (equates to 11.2 in Intel® Ethernet Fabric Suite)
  - Integrated into libfabric and selected Linux distros
  - Out of Tree code for older distros available on github

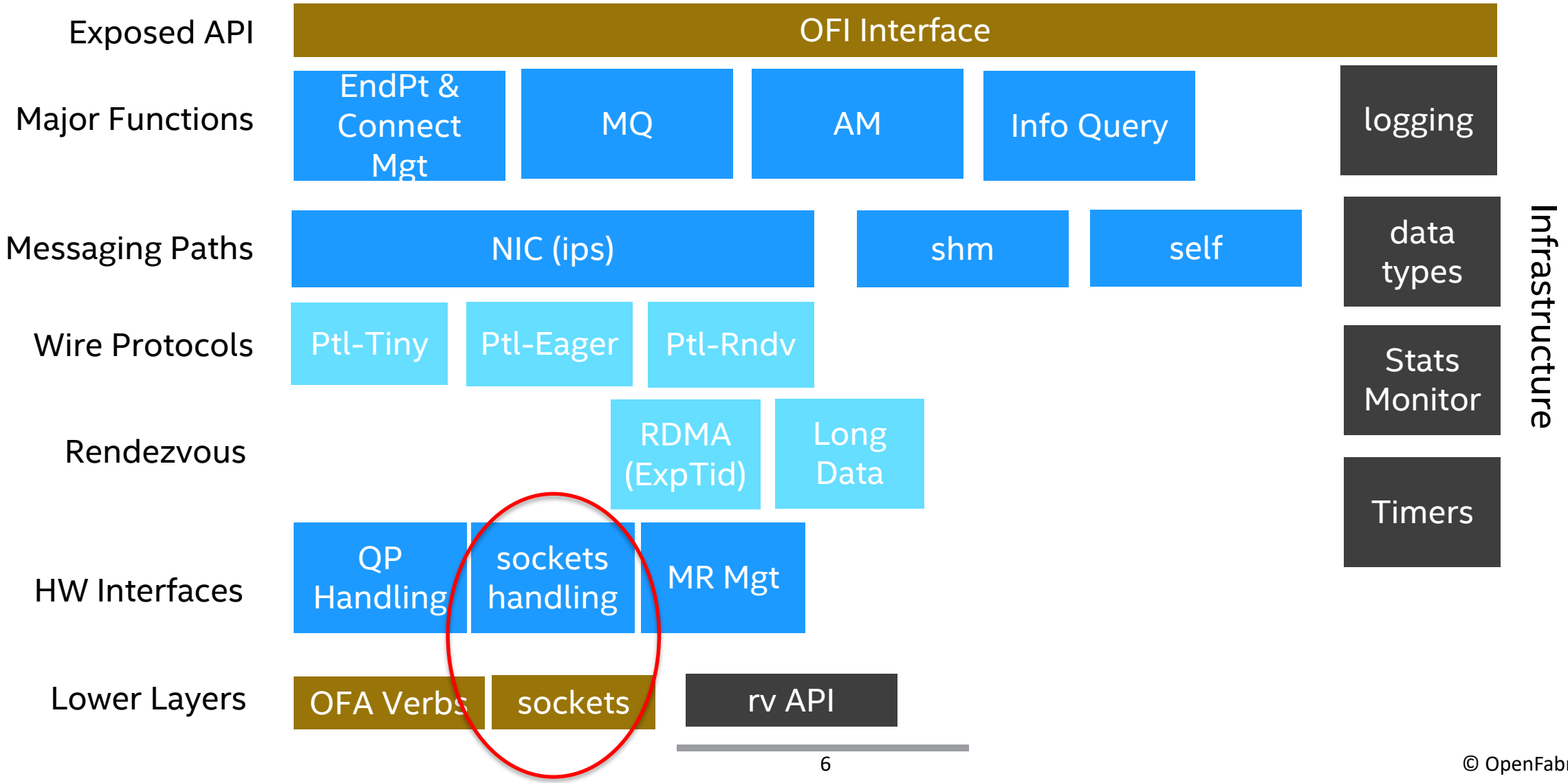
ONEAPI SOFTWARE STACK



PRIOR PSM3 USER SPACE OFI PROVIDER ARCHITECTURE



UPDATED PSM3 USER SPACE OFI PROVIDER ARCHITECTURE

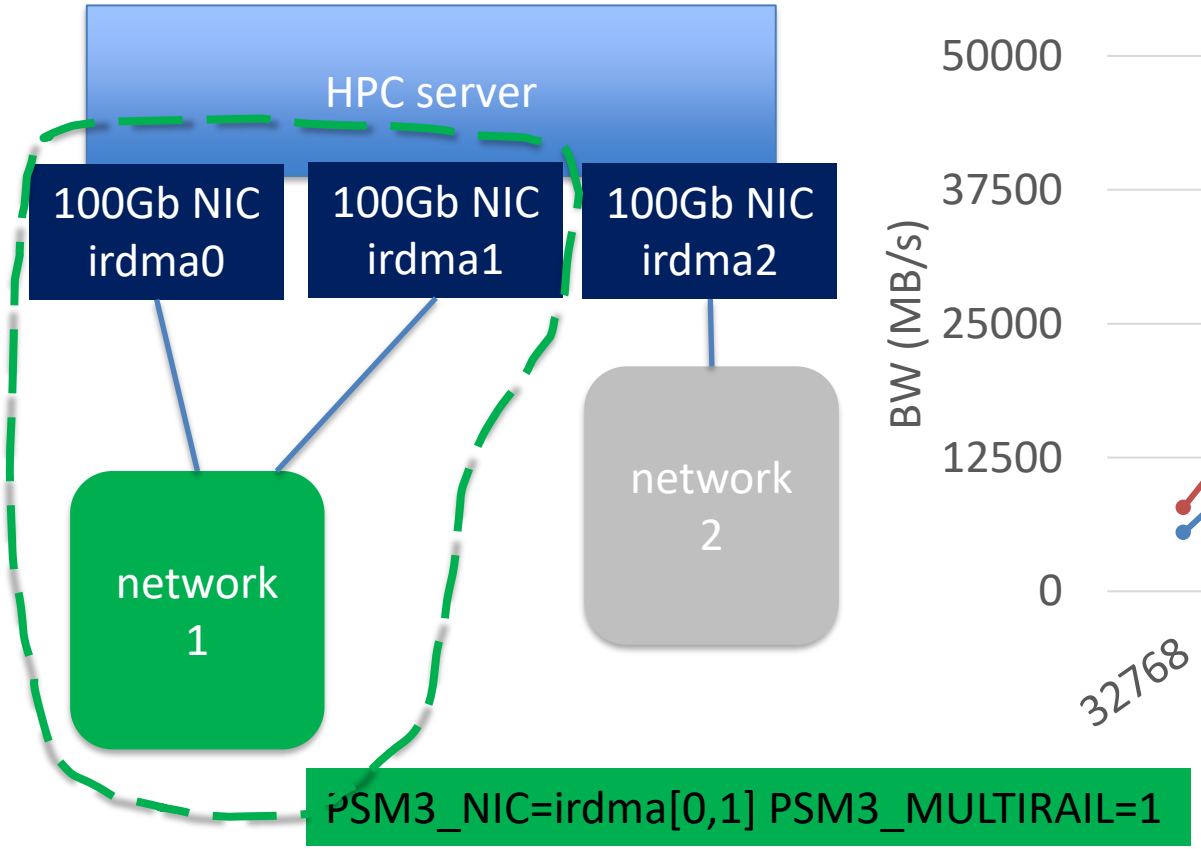


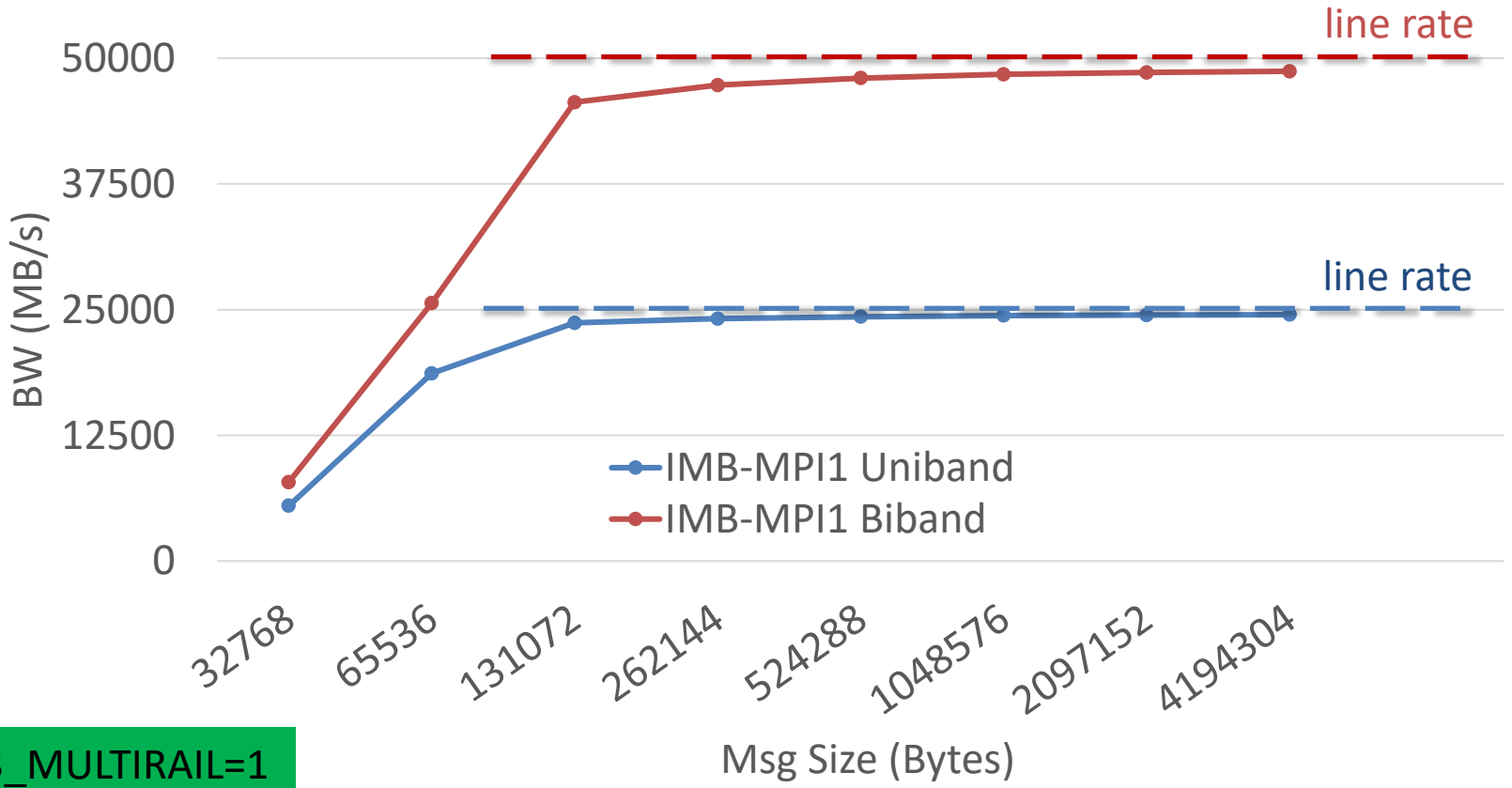
NEW CAPABILITIES

- **Runtime selection of RDMA or TCP/IP**
  - Also build time selectable
- **Support for IPv6 and IPv4 addressing (RoCEv2 and TCP/IP)**
- **Runtime NIC filtering controls**
  - Wildcarded name, Wildcarded IP subnet, Address format, Speed
- **Ongoing tuning and validation with GPUs and CPUs**
  - Both AI and HPC workloads

PSM3 NIC FILTERING AND MULTIRAIL

- **PSM3_NIC wildcards may be used to select specific interfaces/switch planes**
- **Dual rail PSM3 can achieve uni and bi-dir line rate with just 1 process per node**

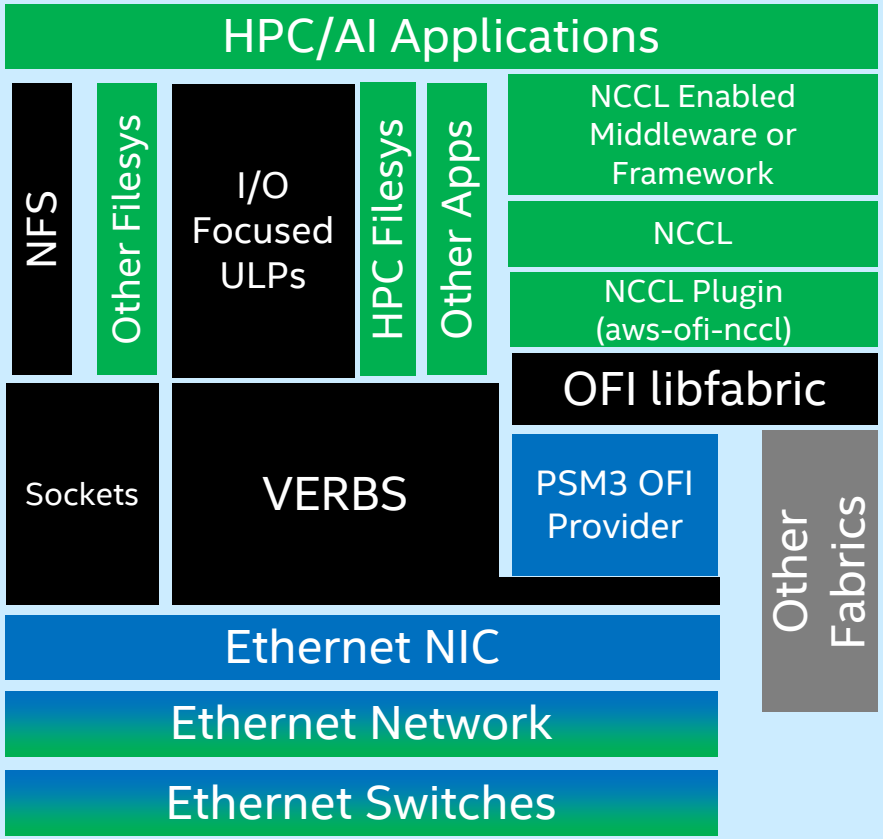




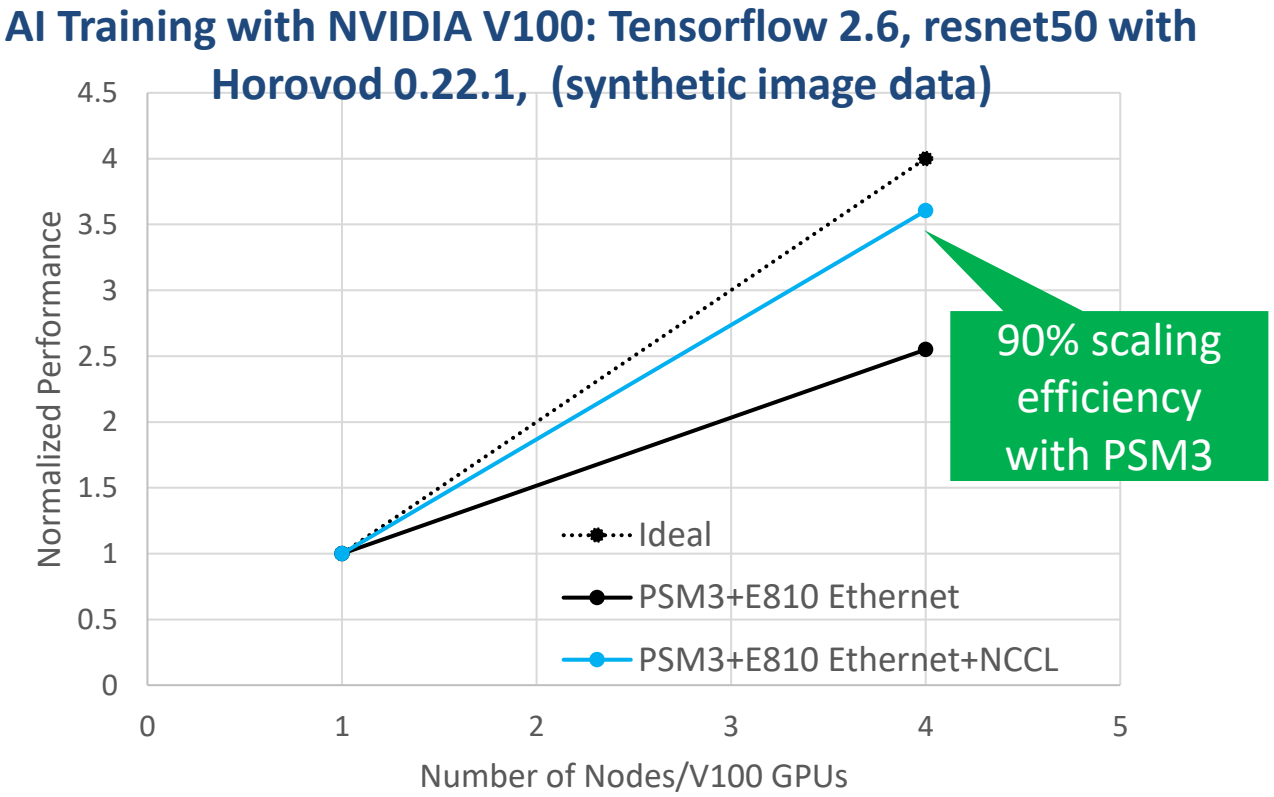
See backup for workloads and configurations. Results may vary.

See configuration #1

USING OFI AND PSM3 WITH NCCL



- **Upstream aws-ofi-nccl plugin**
- **Now supports PSM3 with GPUDirect***

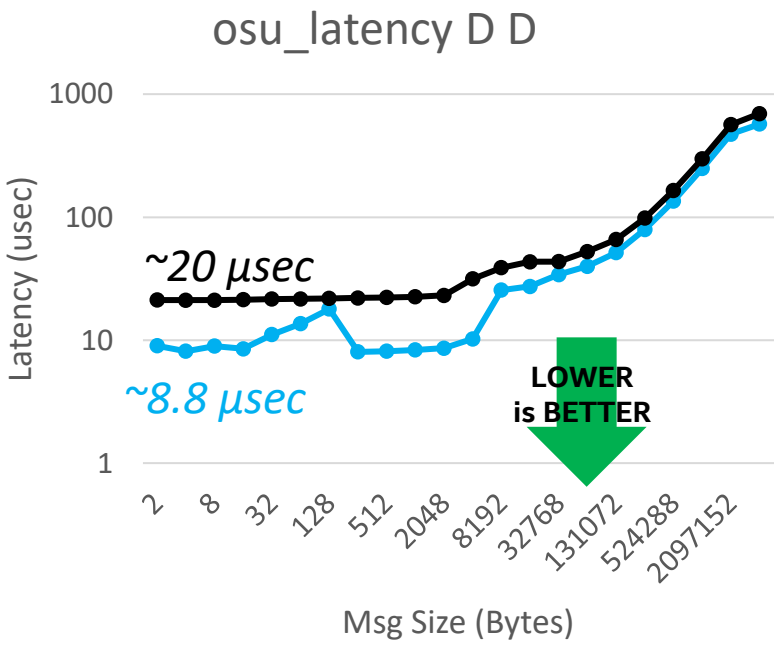


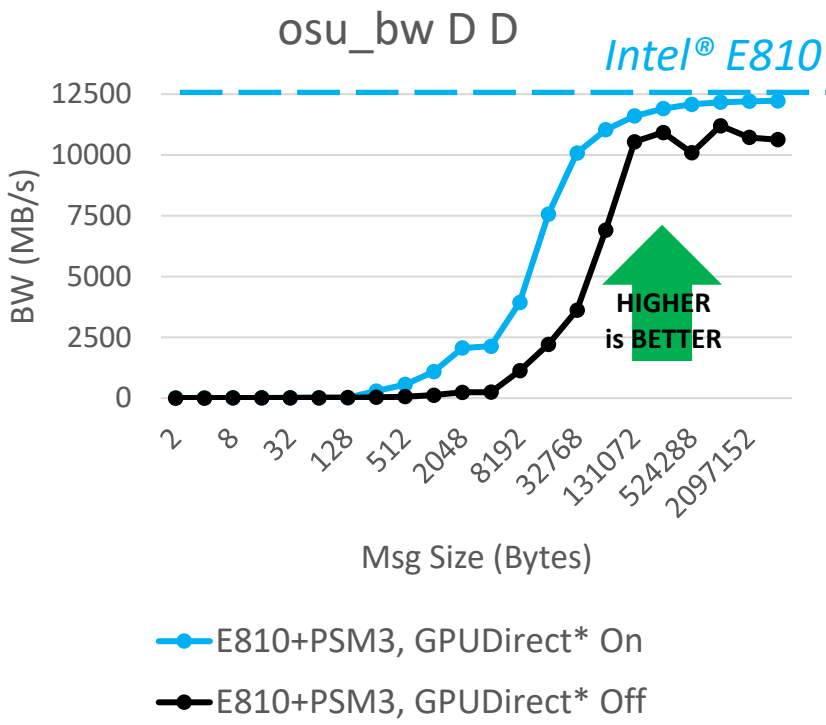
See backup for workloads and configurations. Results may vary.

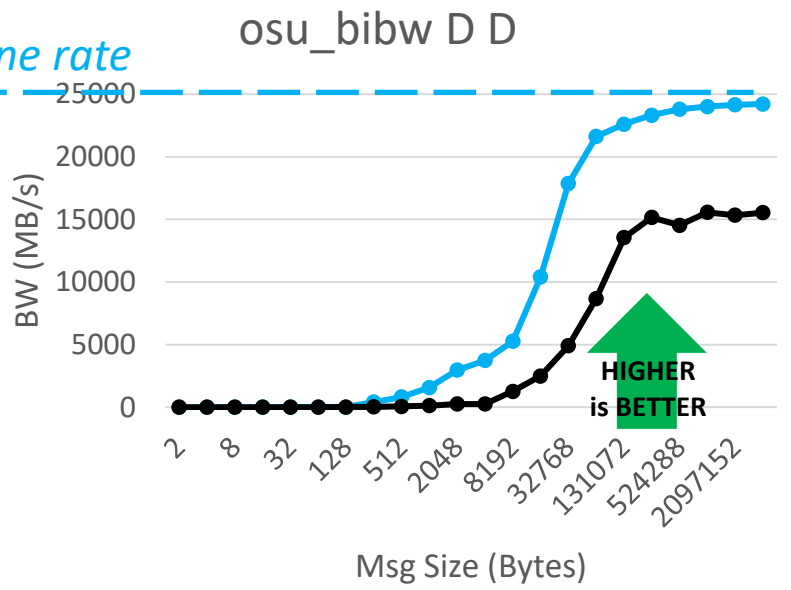
See configuration #2

PSM3 AND GPUDIRECT*

- **Intel® E810 Ethernet and NVIDIA* A100 GPU connected with a PCIe switch**







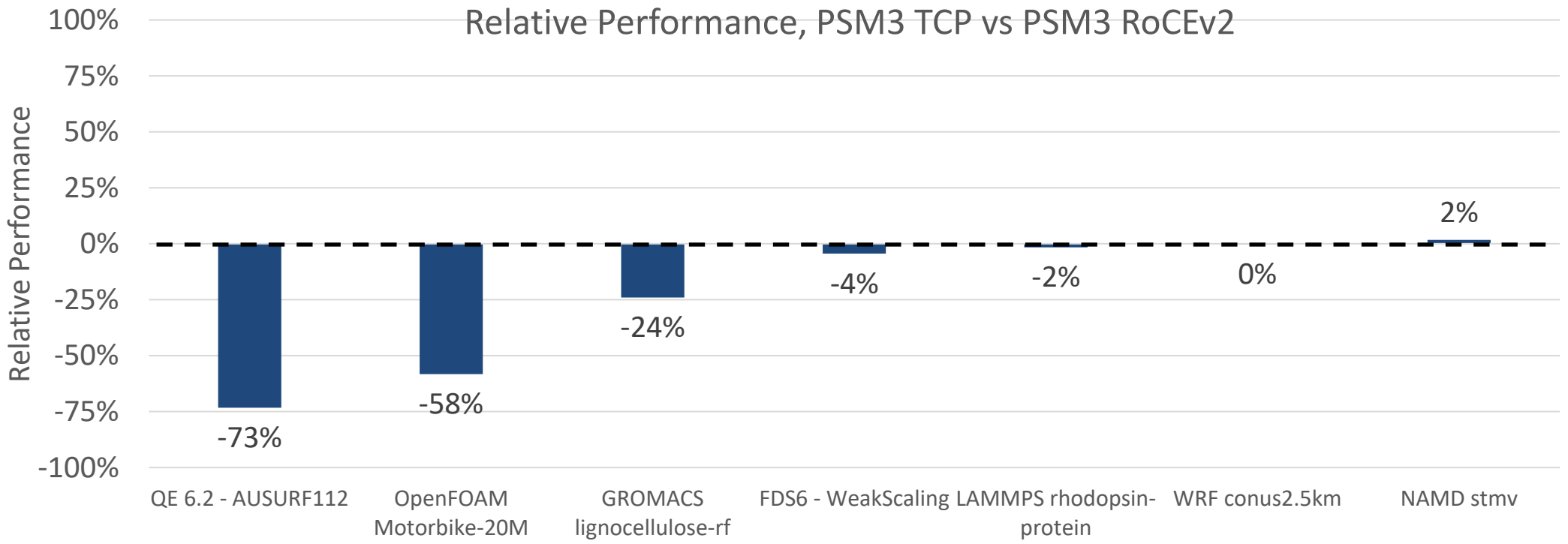
See backup for workloads and configurations. Results may vary.

See configuration #3

OSU Microbenchmarks v5.8

PSM3 TCP VS ROCEV2

- PSM3 TCP performs as well as RoCEv2 for some applications but others greatly benefit from RoCEv2



16 nodes - 832 cores - Dual socket Intel® Xeon® Platinum 8170 processors

See backup for workloads and configurations. Results may vary.

See configuration #4

CONCLUSIONS & FUTURE WORK

- **PSM3**
  - is a robust and proven libfabric provider, purpose-built for HPC on Ethernet
  - works with existing HPC and AI applications and other RDMA/verbs capable networks
  - now supports NCCL and TCP/IP
- **Future work**
  - Tuning for HPC and AI workloads
  - Continue integration with oneAPI and oneCCL for improved performance & portability

See backup for workloads and configurations. Results may vary.

CONFIGURATIONS

1. 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 8.3 (Ootpa). 4.18.0-240.el8.x86_64 kernel. 16xDDR4, 256 GB, 3200 MT/s. irdma version 1.8.45. ice version 1.8.2_2_g4b426405. CVL device firmware-version: 3.10 0x8000acc8 1.3106.0, 144 TxRx queues. pfc_enable: 0x1. Intel Ethernet Fabric Suite 11.2.0.0.259. Intel MPI 2021.5. mpirun -np 2 -ppn 1 -host node1,node2 -genv PSM3_NIC=irdma[0,1] -genv PSM3_MULTIRAIL=1 -genv PSM3_RDMA=1 $I_MPI_ROOT/bin64/IMB-MPI1 Uniband Biband. Mellanox SN2700 Ethernet switch. PFC enabled on priority 0.
2. AI Training with NVIDIA V100: Tensorflow 2.6.2, resnet50 with Horovod 0.22.1, Tests performed on 2 socket Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.4GHz. Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with Intel Pstate driver. Red Hat Enterprise Linux 8.3 (Ootpa). 4.18.0-240.el8.x86_64 kernel. 16xDDR4, 256 GB, 3200 MT/s. irdma version 1.8.45. ice version 1.8.2_2_g4b426405. CVL device firmware-version: 3.10 0x8000aa6d 1.3100.0, 144 TxRx queues.GPU config: slot:b1:00.0 3D controller: NVIDIA Corporation GV100GL [Tesla V100 PCIe 16GB] (rev a1) NUMA node: 1. Driver Version: 495.29.05 CUDA Version: 11.5 openmpi-4.1.1-cuda-ofi as packaged with Intel Ethernet Fabric Suite 11.2.0.0.259. https://github.com/horovod/horovod/blob/master/examples/tensorflow2/tensorflow2_synthetic_benchmark.py. EdgeCore Mavericks (Tofino) switch, PFC enabled on priority 0.
3. 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 disabled with ACPI driver. Red Hat Enterprise Linux 8.4 (Ootpa). 4.18.0-305.el8.x86_64 kernel. 16xDDR4, 256 GB, 3200 MT/s. irdma version 1.7.72. ice version 1.7.16. DDP version 1.3.27.0. CVL device cvl0 firmware-version: 3.10 0x8000ad8e 1.3106.0, 144 TxRx queues. pfc_enable: 0x1.GPU config: slot:4f:00.0 3D controller: NVIDIA Corporation Device 20b5 (rev a1) NUMA node: 0. Driver Version: 510.47.03 CUDA Version: 11.6. Open MPI 4.1.2 as packaged with Intel Ethernet Fabric Suite 11.2.0.0.259. EdgeCore Mavericks (Tofino) switch, PFC enabled on priority 0. Example run command: mpirun -np 2 --map-by ppr:1:node -host node1,node2 -x PSM3_CUDA=1 -x FI_PROVIDER=^psm3;ofi_rxd,shm,sockets,tcp,tcp;ofi_rxm,UDP,UDP;ofi_rxd,verbs,verbs;ofi_rxd,verbs;ofi_rxm -mca mtl ofi -x PSM3_GPUDIRECT=[0 or 1] -x PSM3_RDMA=1 -x PSM3_GPUDIRECT_RDMA_SEND_LIMIT=8388608 ./osu_bw -m 2:4194304 D D
4. PSM3 TCP & RoCE. Tests performed on 16 nodes, 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 1.8.45. ice version 1.8.2_2_g4b426405. CVL device cvl firmware-version: 3.20 0x8000d83e 1.3146.0, 144 TxRx queues. pfc_enable: 0x1.Intel Ethernet Fabric Suite 11.2.0.0.259. Arista 7170 Ethernet switch, PFC enabled on priority 0. FI_PROVIDER=psm3, PSM3_HAL=sockets for PSM3 TCP and PSM3_HAL=verbs for RoCE. All applications are compiled with Intel compilers 2020.2 or newer using default makefiles and run with Intel MPI 2021.5. QuantumESPRESSO Program PWSCF v.6.2 (svn rev. 13899), AUSURF112 benchmark. OpenFOAM v1712, simpleFoam -parallel, motorbike20M benchmark. GROMACS 2020.5 gmx_mpi mdrun -s run.tpr -resethway -noconfout -nsteps 40000, single precision, AVX_512, Intel MKL. FDS6 - Fire Dynamics Simulator FDS6.5.3-2848-gf997a36-master, weak_scaling_test. LAMMPS (10 Feb 2021), rhodopsin protein benchmark, -var x 8 -var y 8 -var z 8 -in in.rhodo.scaled. WRF v3.9.1.1, conus2.5km benchmark. NAMD Git-2020-11-18 for Linux-x86_64-MPI-smp, Based on Charm++/Converse 61002 for mpi-linux-x86_64-smp-mpicxx. Running in SMP mode: 128 processes, 5 worker threads (PEs) + 1 comm threads per process, 640 PEs total

NOTICES & 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.

2022 OFA Virtual Workshop

THANK YOU

James Erwin, Technical Lead, Software Enabling and Optimization Engineer

Todd Rimmer, Senior Principal Engineer

Intel Corporation