MVAPICH4 over C-DAC's Trinetra Network

Rakesh Kumar Yadav
C-DAC, India
August 19, 2025

13th Annual MVAPICH User Group (MUG)
2025 Conference

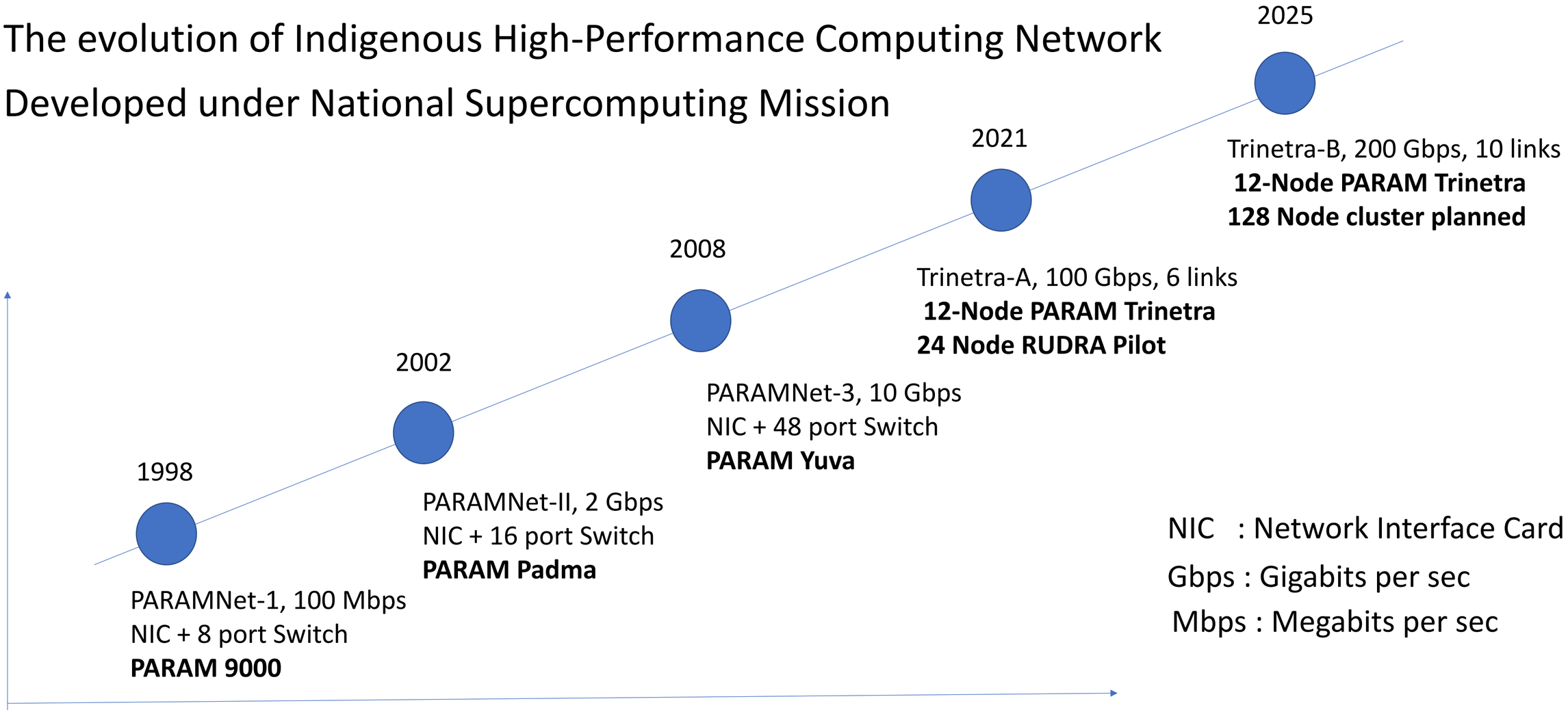
Talk Outline

- Introduction of Trinetra HPC Network
- Trinetra Software Stack
  - Features
- Adaptation of MVAPICH4
- Benchmarking and Performance tuning
- HPC Applications on Trinetra
- Experiments with AI Frameworks on Trinetra
- MVAPICH4 on RISC-V Architecture
- Future Work
- Concluding Remarks

Introduction of Trinetra HPC Network

C-DAC's HPC Networks

- The evolution of Indigenous High-Performance Computing Network
- Developed under National Supercomputing Mission



Timeline : PARAMNet Development over the years

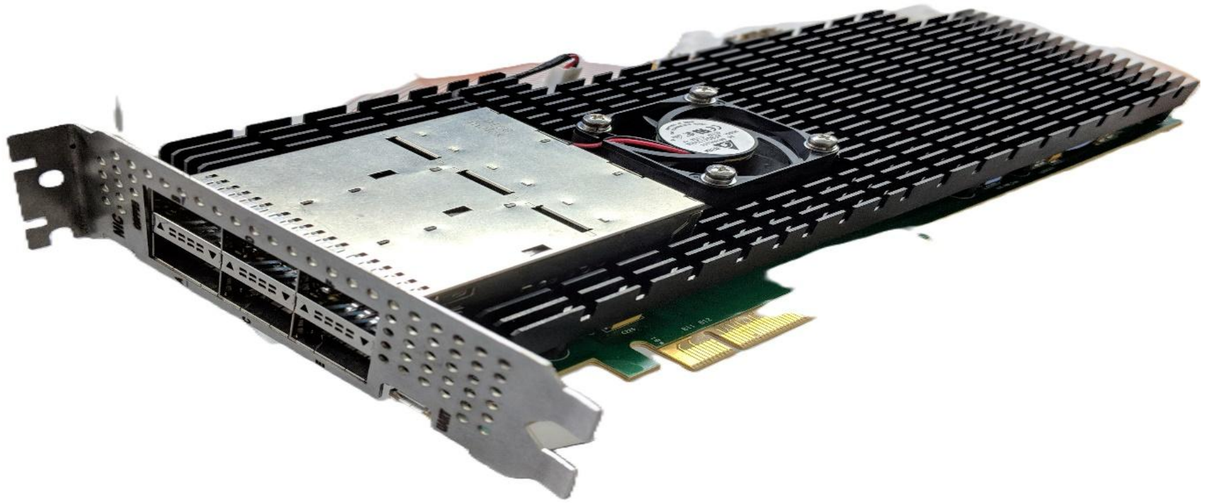
NIC : Network Interface Card

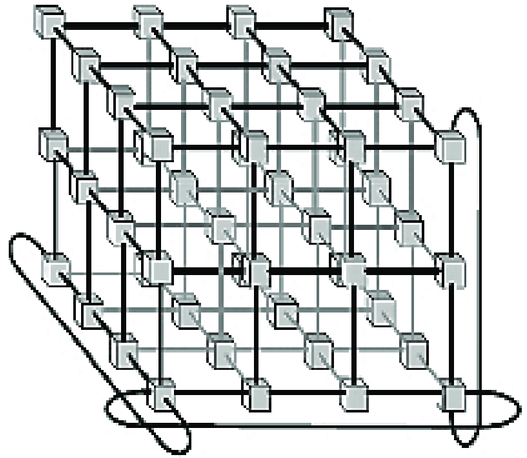
Gbps : Gigabits per sec

Mbps : Megabits per sec

Trinetra-A Network

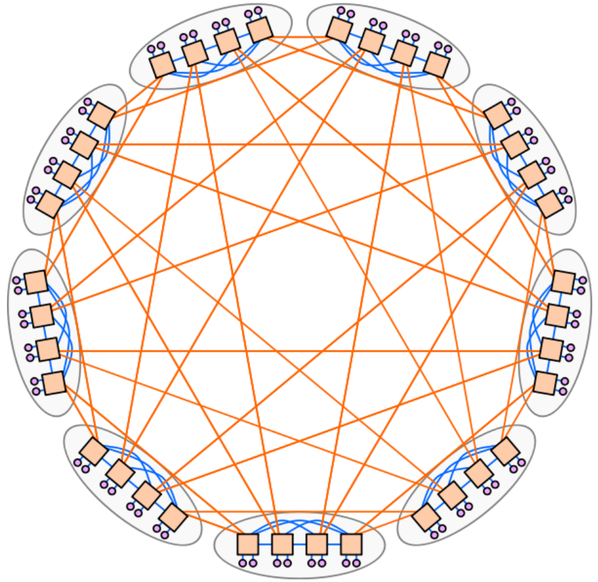
- Physical link Speed : 100 Gigabits per sec
- Host Interface : PCIe Gen3 x8
- No. of links : 6 links
- FPGA : Xilinx Virtex Ultrascale VU095
- Topology : Switchless 3D Torus





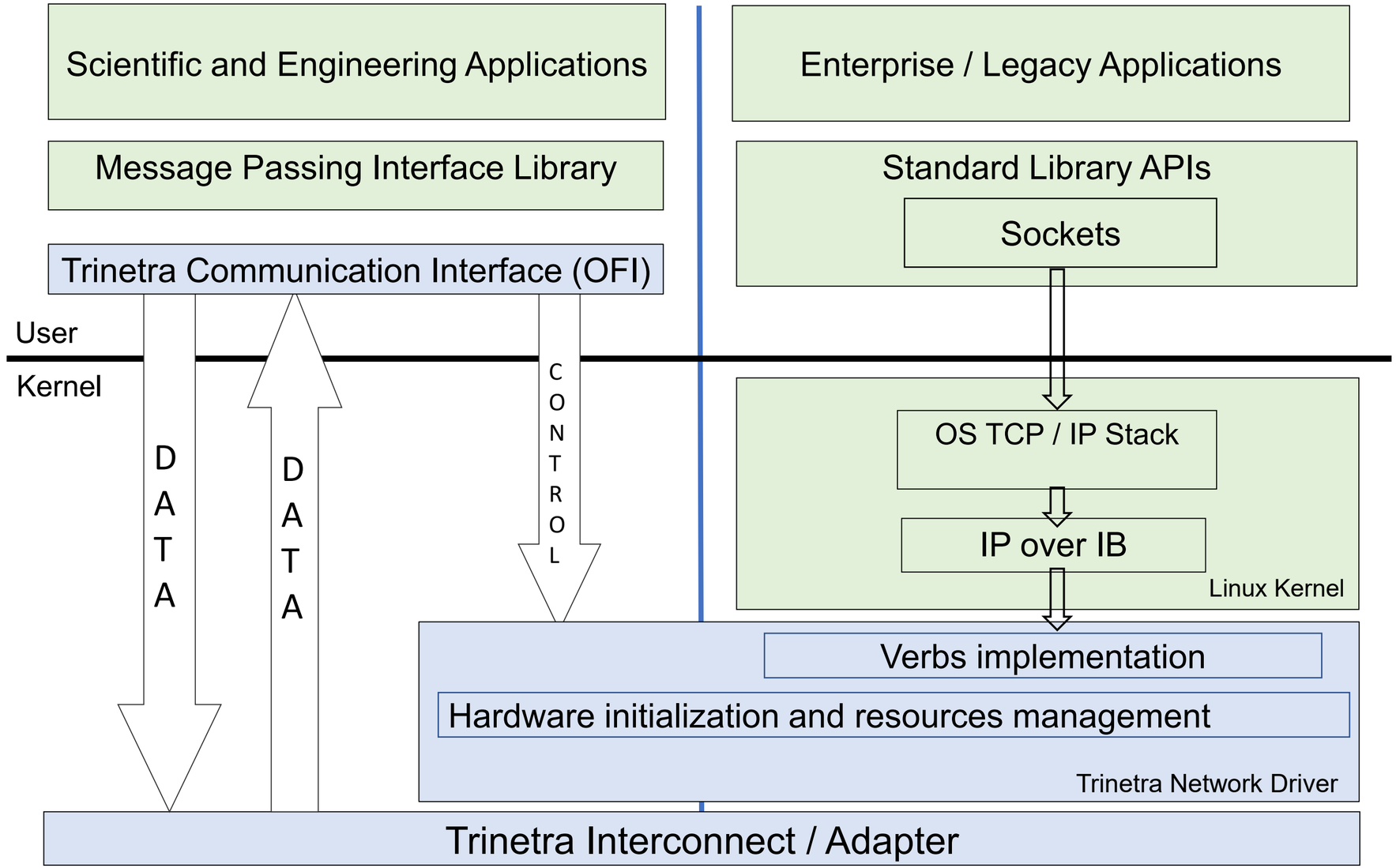
Trinetra-B Network

- Physical link Speed : 200 Gigabits per sec
- No. of links : 10 links
- Host Interface : PCIe Gen3 x16
- FPGA : Xilinx Virtex Ultrascale+ VU27P
- Topology : Switchless Hybrid (Hypercube, Mesh, Torus, Dragonfly)



Trinetra Software Stack

Software Stack for Trinetra-A & Trinetra-B



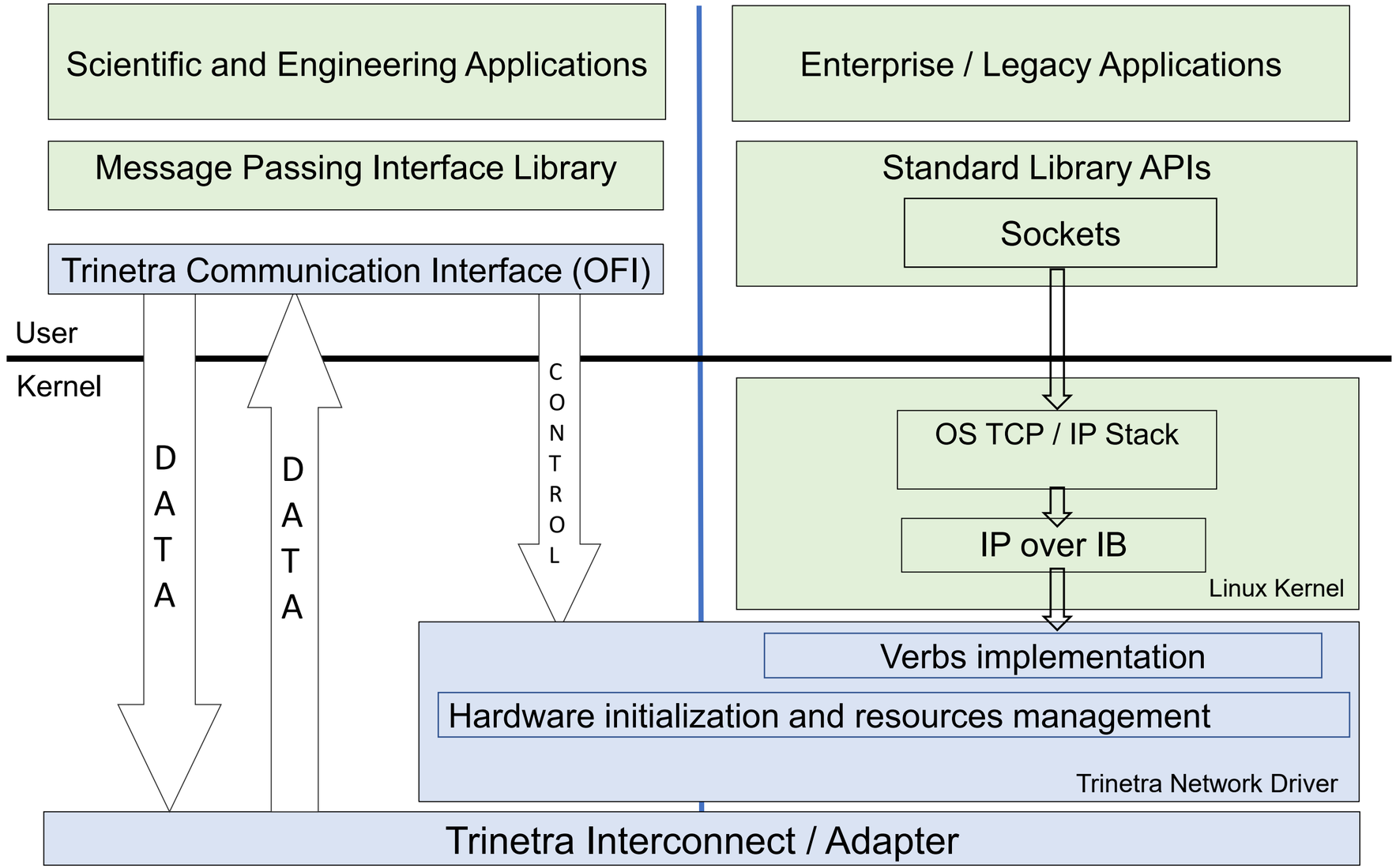
Software Stack Features

- Hybrid (Onload + Offload) approach for transport protocol processing
- Onload transport protocol processing
  - Reliable data transfer using re-transmissions and Acknowledgements
  - Adaptive flow control based on credits
  - Packetization and assembly of packets
  - Out-of-order packet handling
  - In-order message delivery
- Tag matching
- Kernel bypass mechanism

Software Stack Features

- Configurable software packet size (1KB, 2KB, 4KB, 8KB, 16KB, 32KB)
- Two Protocols for data transfer based on message length
  - Single Fragment Queue – SFQ
    - Used for small messages
    - Uses Programmed IO
  - EAGER
    - Used for medium/large messages
    - Uses Programmed IO and DMA
- Configurable protocol switch threshold
- Latency optimized path for tiny and small length messages
- Open Fabric Interface (OFI) adopted for communication with Trinetra network.

Software Stack for Trinetra-A & Trinetra-B



IPoIBoT: Enabling Sockets over Trinetra

- Enables Legacy (sockets/IP based) applications/services.
- Uses IPoIB component of OFED/upstream Linux.
- Validation
  - Iperf 3.9
  - NetPerf 2.7
  - FIO 3.35 over NFS
  - OMB 7.5 over mvapich-4.0

Adaptation of MVAPICH4

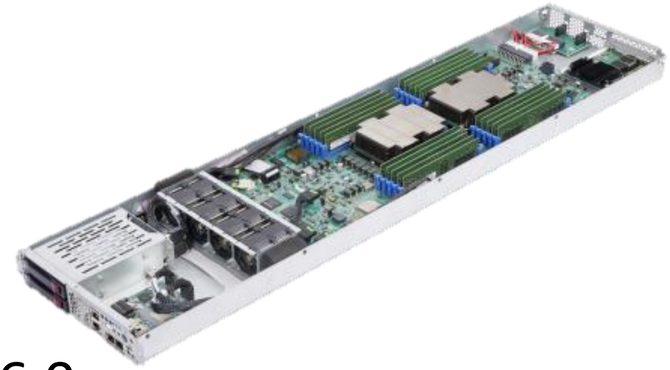
MPI Support

- MVAPICH4 is available as a primary MPI implementation on Trinetra.
  - Open Fabric Interface (OFI) support
- Adaptation of MVAPICH-PLUS
  - Work in progress
- Long association with MVAPICH team
  - Excellent interaction, help from the team
- Other MPI implementations supported
  - MPICH
  - OpenMPI
  - IntelMPI

Benchmarking and Performance Tuning

Trinetra-A Testbed at C-DAC, Pune

- Cluster – 24 nodes based on C-DAC's indigenously designed **Rudra servers**, Intel Xeon Gold 6240R @ 2.4GHz, 48 cores
- Operation System – AlmaLinux 8.9
- RAM – 196 GB per node
- Primary network – **Trinetra-A 100Gbps (2 x 2 x 6)**
- InfiniBand – Nvidia HDR 100 with Mellanox OFED-24.04-0.6.6.0
- MPI library – mvapich4.0, mvapich3.0, mvapich2-2.3.7, Libfabric – 2.1.0
- MPI Benchmark – IMB-2021.10, OMB (as part of MVAPICH tar)
- HPL – 2.3, OpenBLAS – 0.3.3
- Compiler – gcc-8.5, icx-2024.2



Trinetra-B Testbed at C-DAC, Pune

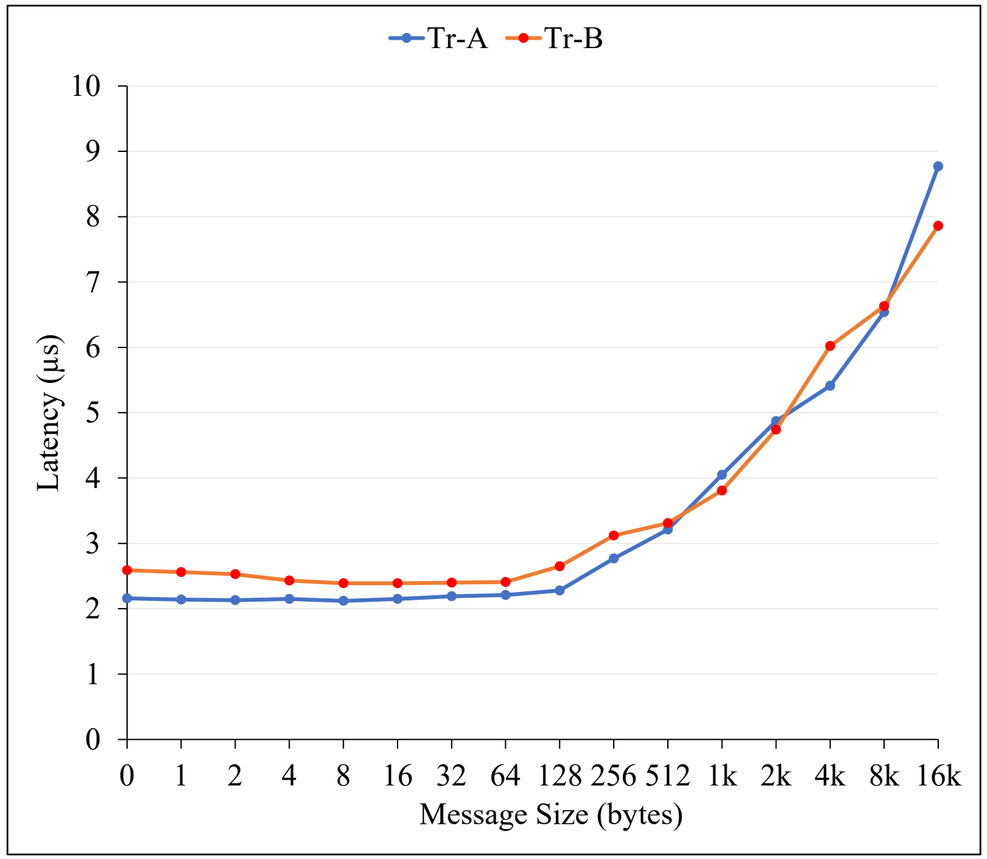
- Cluster – 12 nodes, Intel Xeon Gold 5118 @ 2.3 GHz, 24 cores
- Operation System – AlmaLinux 8.9
- RAM – 92 GB per node
- Primary network – **Trinetra-B 200Gbps (4 x 3)**
- InfiniBand stack – OFED-4.17-rc1
- MPI library – mvapich4.0, mvapich3.0, mvapich2-2.3.7, Libfabric – 2.1.0
- MPI Benchmark – IMB-2021.10, OMB (as part of MVAPICH tar)
- HPL – 2.3, OpenBLAS – 0.3.3
- Compiler – gcc-8.5, icx-2024.2

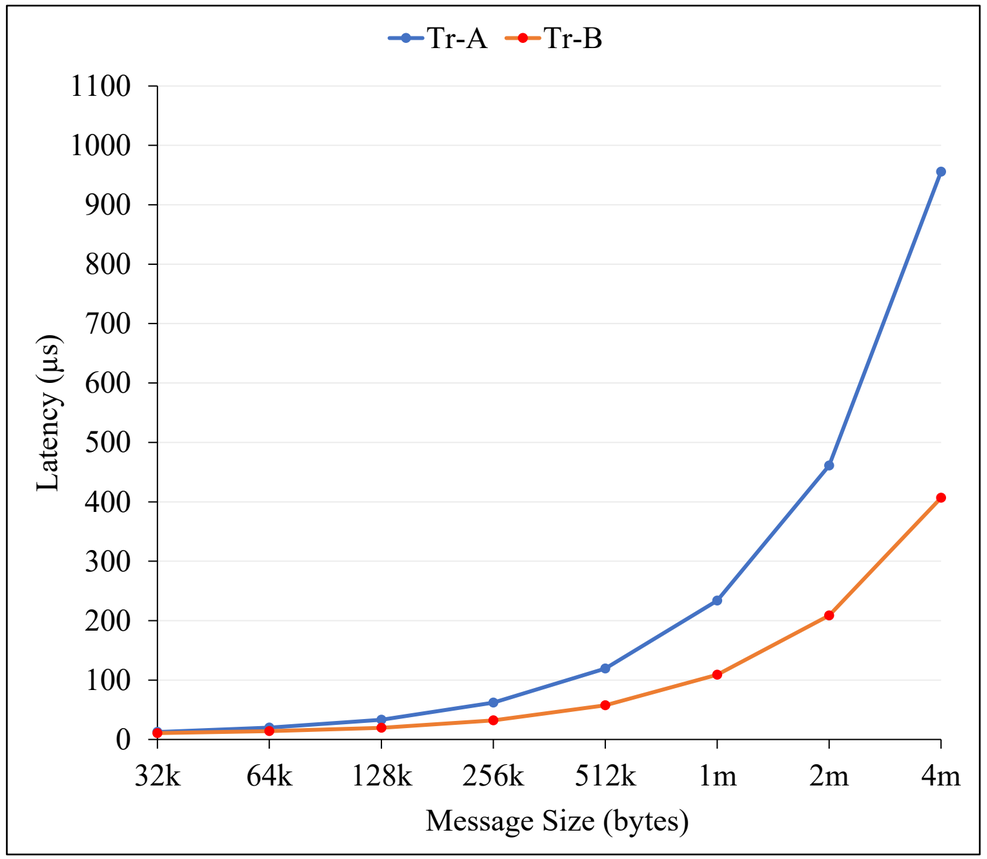
Benchmarking and Validation

- Validated with
  - OSU MPI benchmark (OMB)
  - Intel MPI benchmark (IMB) suite
  - NAS Parallel Benchmarks
  - High Performance Linpack (HPL)
  - High Performance Conjugate Gradients (HPCG) Benchmark

- HPC applications executed on Trinetra
  - WRF, LAMMPS, Quantum Espresso, SeisAccomo2D, NAMD

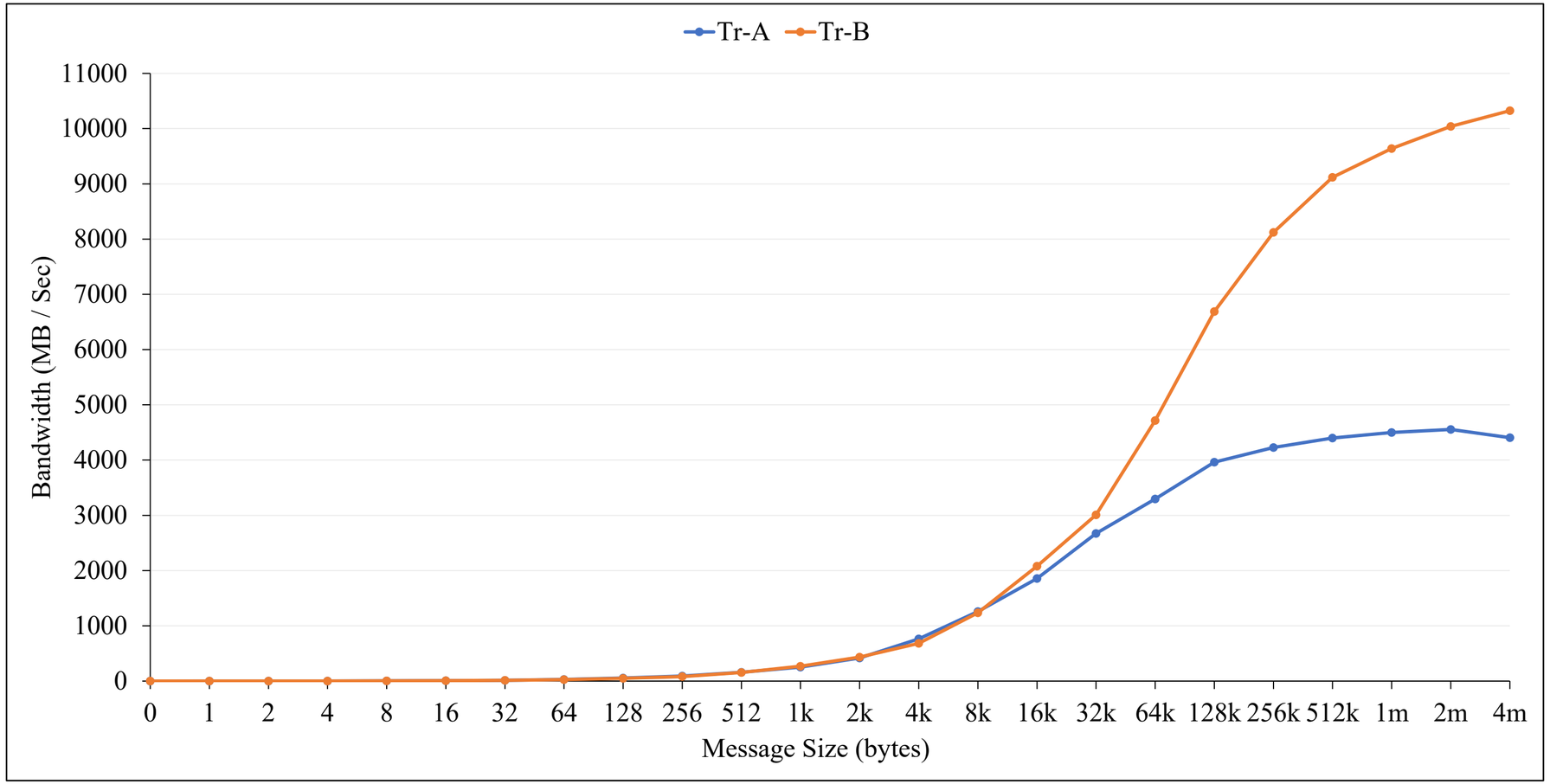
Micro-benchmarks: Trinetra-A and Trinetra-B Latency





MPI - MVAPICH4, Latency - Tr-A 2.1 (us), Tr-B 2.4 (us)

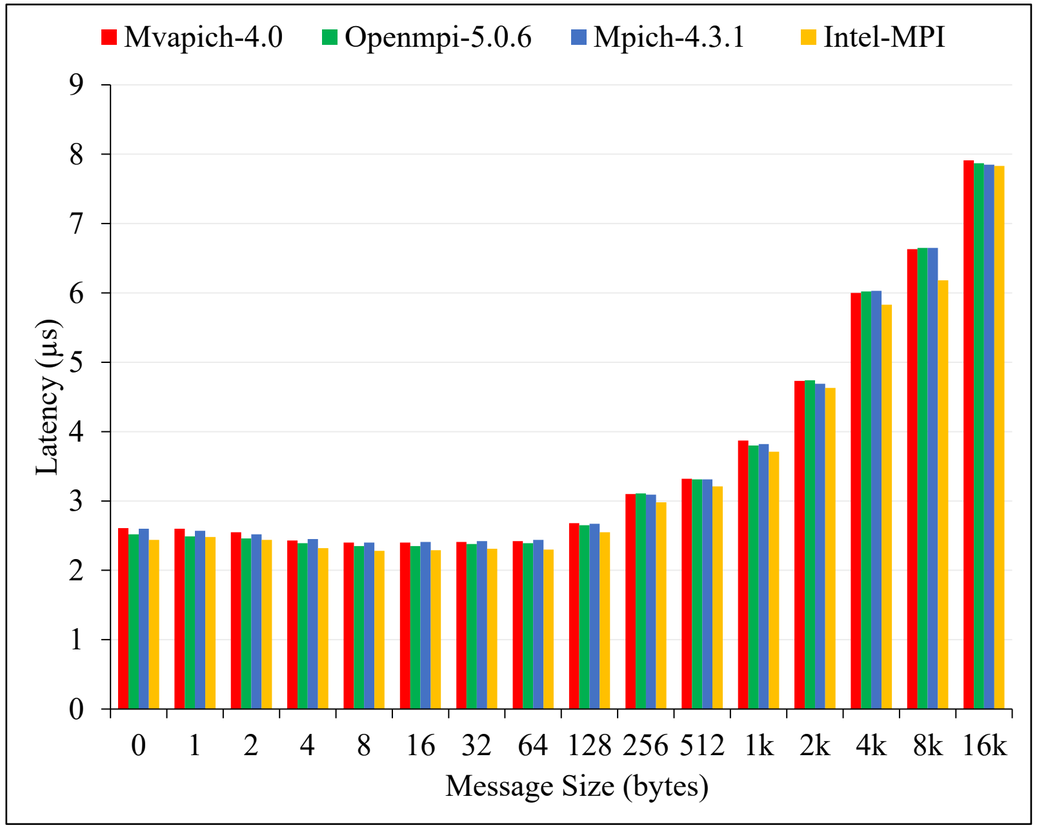
Micro-benchmarks: Trinetra-A and Trinetra-B Bandwidth

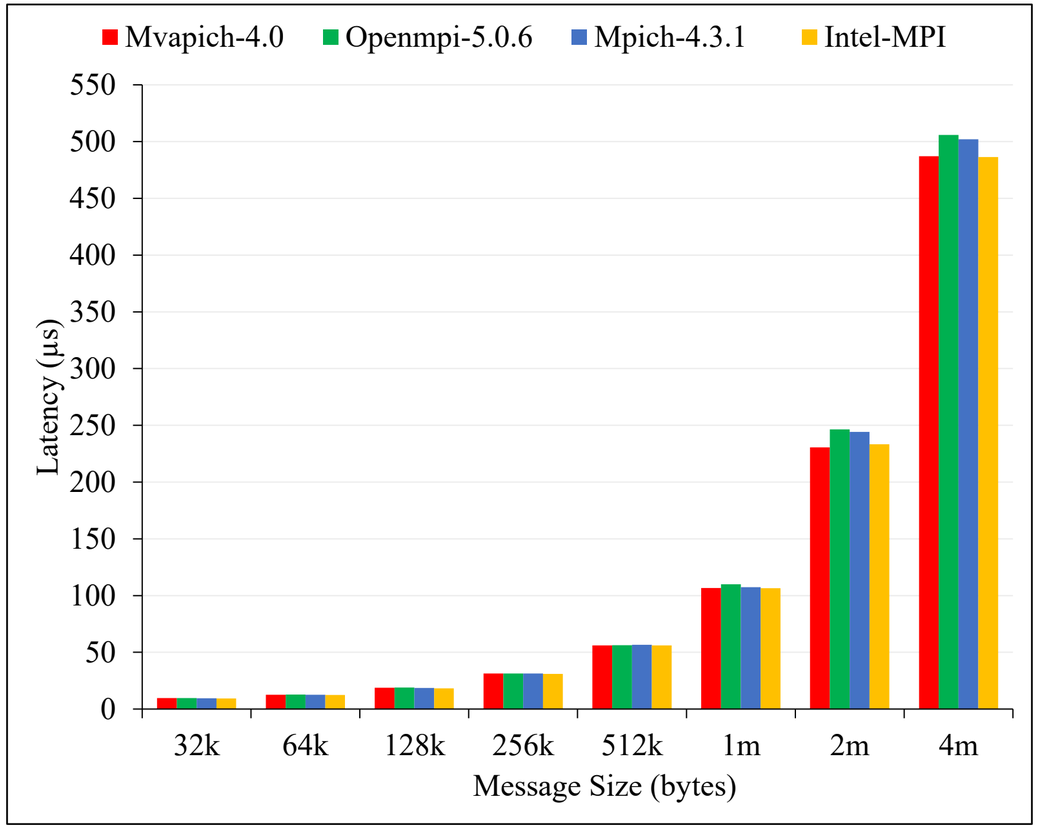


MPI - MVAPICH4

Trinetra-A BW = 4554 MB/s, Trinetra-B BW = 10324 MB/s

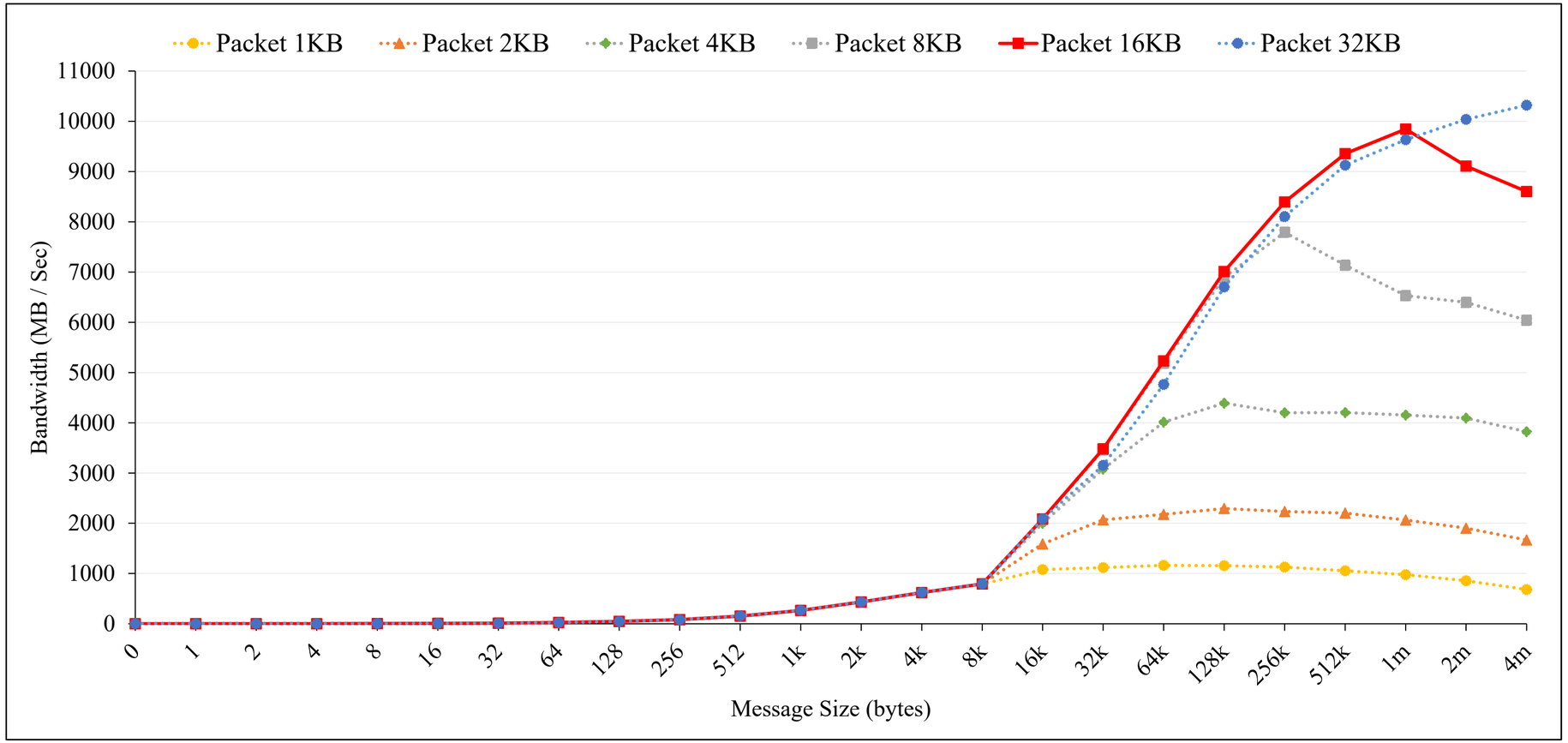
Performance Comparison of Multiple MPIs over Trinetra-B





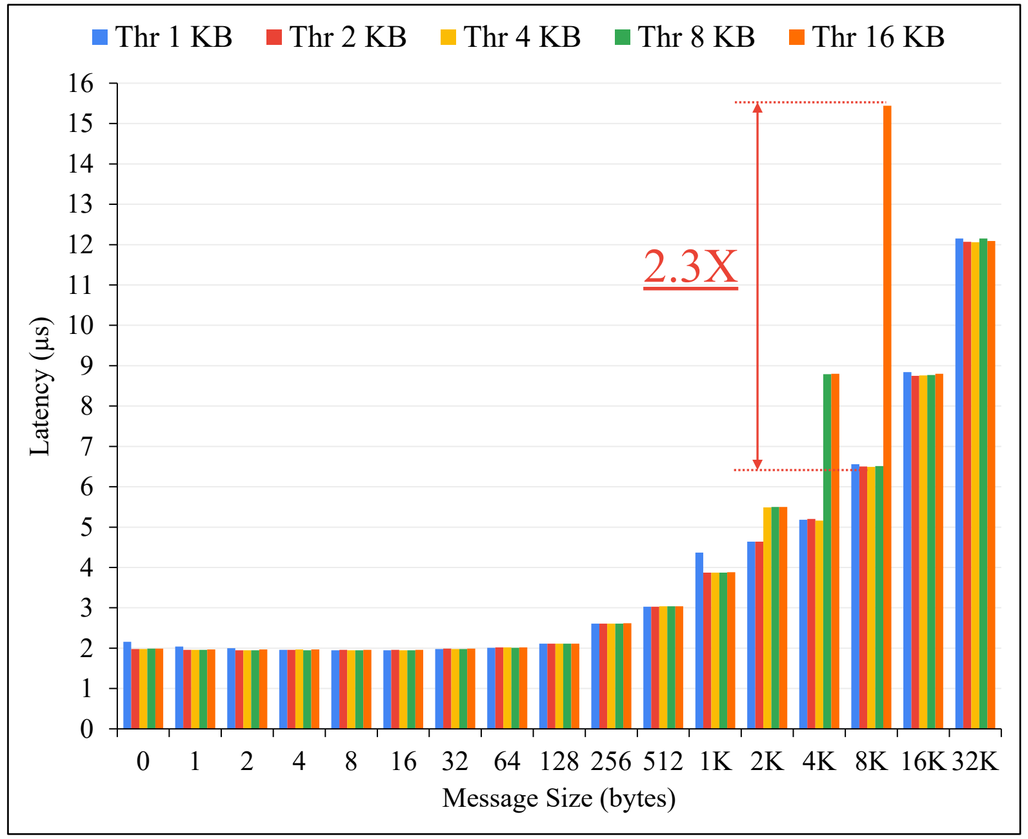
MPI – MVAPICH4, OpenMPI, MPICH, Intel MPI

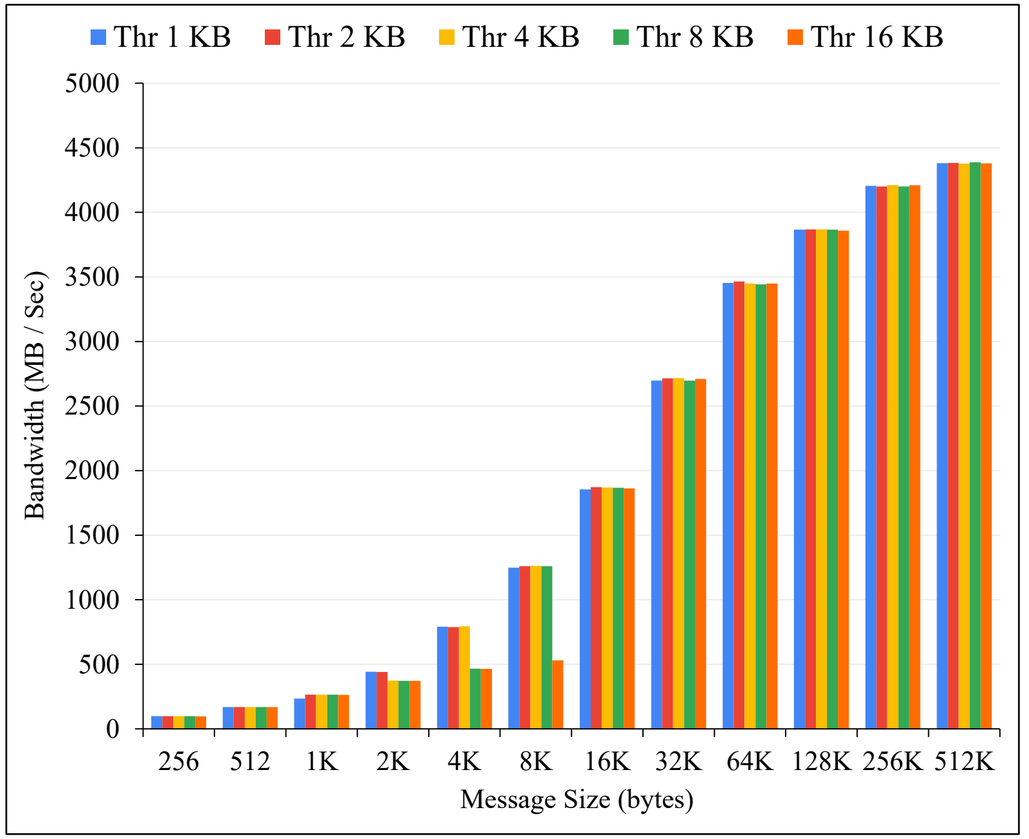
Impact on Bandwidth of Software Packet size Variation on Trinetra-B



MPI - MVAPICH4

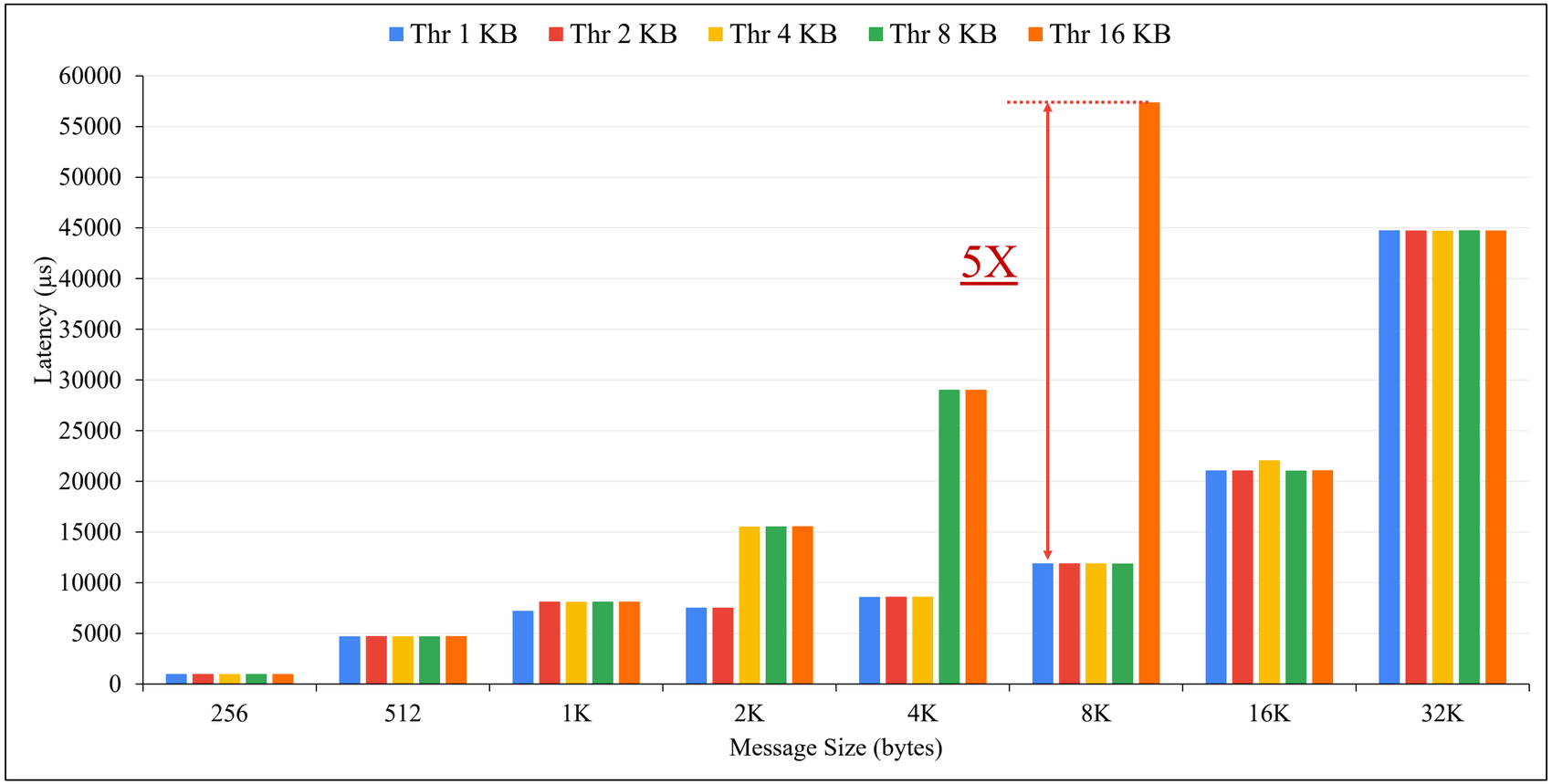
Impact of Protocol Switch Threshold Variation on Point-to-Point Communication





MPI - MVAPICH4, Network - Trinetra-A

Impact of Protocol Switch Threshold Variation on Collective Communication



MPI_Alltoall

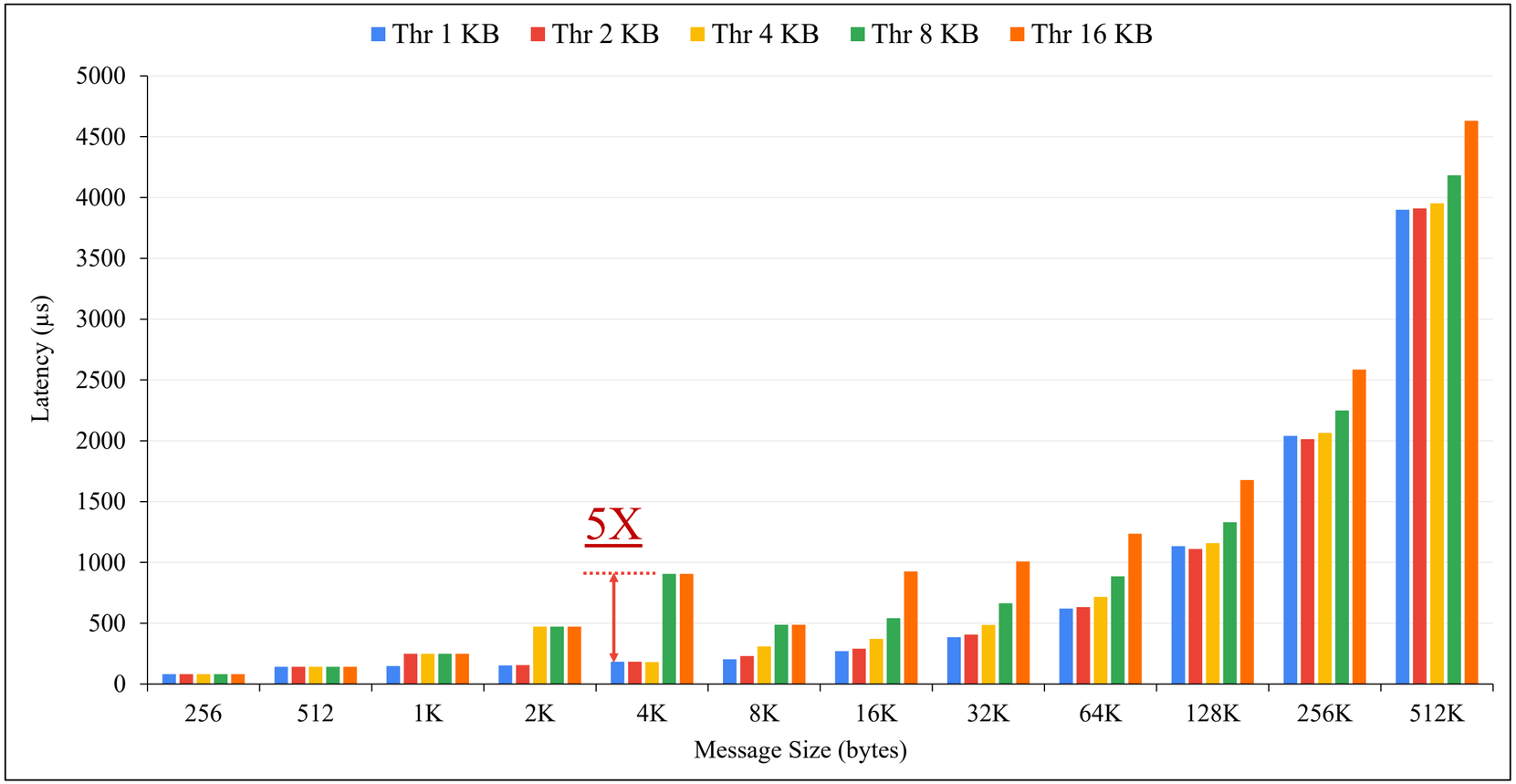
MPI - MVAPICH4

Network - Trinetra-A

Processes – 16 PPN

Nodes - 16

Impact of Protocol Switch Threshold Variation on Collective Communication



MPI_Allreduce

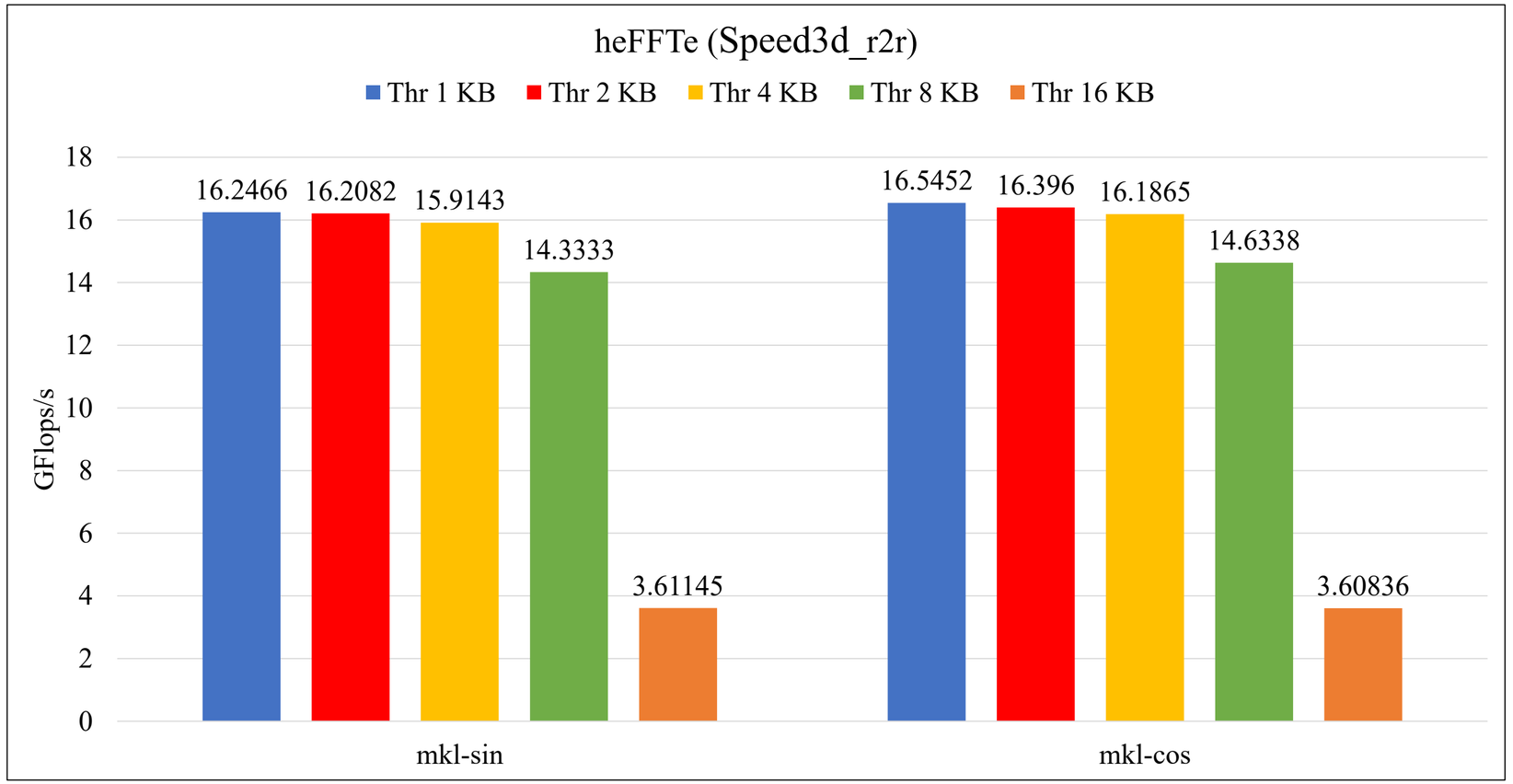
MPI - MVAPICH4

Network - Trinetra-A

Processes – 16 PPN

Nodes - 16

Impact of Protocol Switch Threshold Variation on heFFTe



heFFTe

heFFTe backend – Intel MKL

MPI - MVAPICH4

Problem Size – 128*128*128

Network - Trinetra-A

Processes – 16 PPN

Nodes - 16

IPoIBoT Performance

FIO over NFS	
Randread	Randwrite
1823 MBPs	1096 MBPs

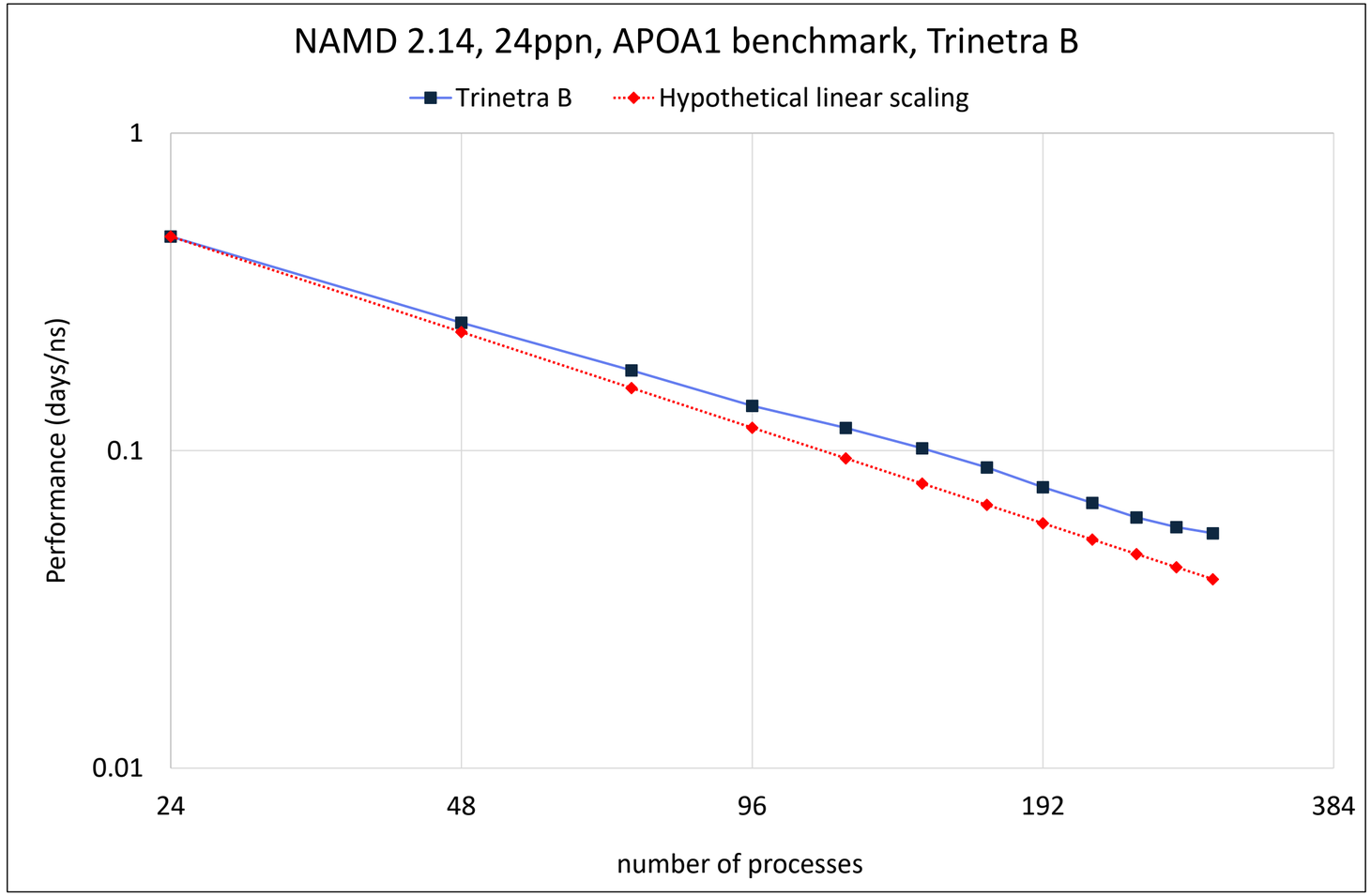
OMB using MVAPICH4	
OSU Latency	OSU Bandwidth
14.70 usec	3785 MBPs

Iperf	Netperf
3179 MBPs	3234 MBPs

Network - Trinetra-B

Applications Performance on Trinetra

NAMD



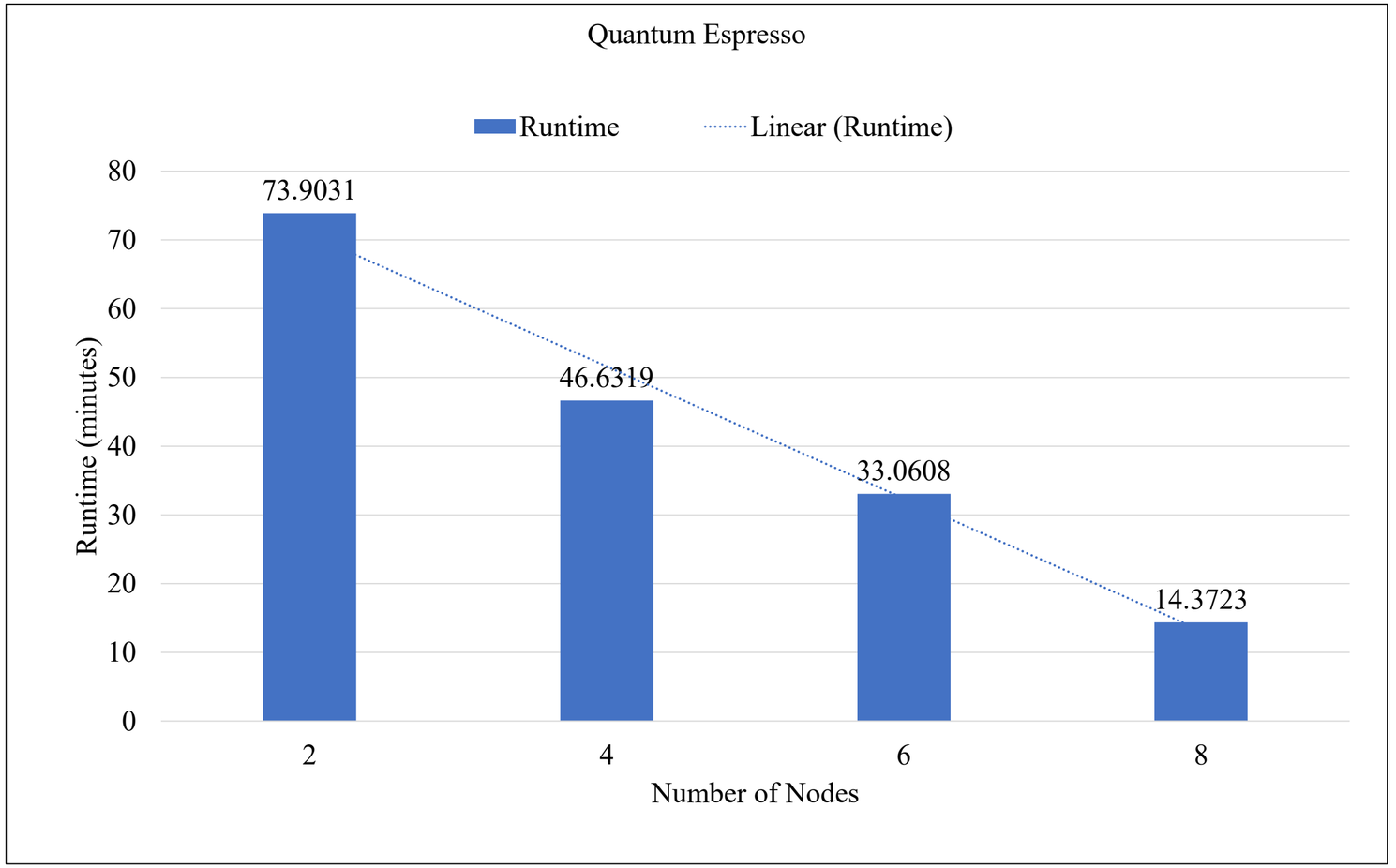
MPI - MVAPICH4

Network - Trinetra-B

Domain – Molecular Dynamics

Version - NAMD@214

Quantum Espresso



MPI - MVAPICH4

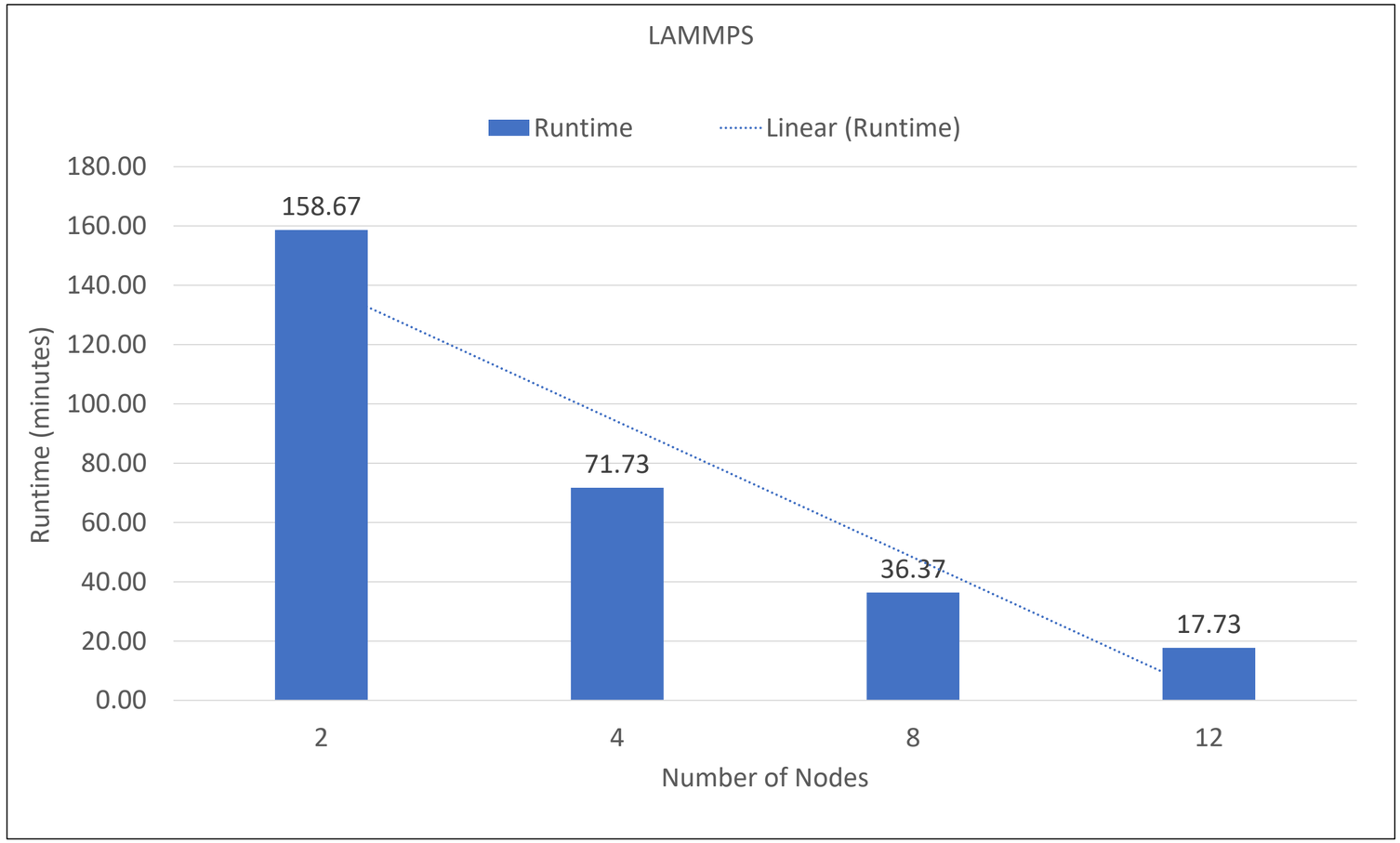
Network - Trinetra-B

Domain – Quantum mechanical

Used for – Nanoscale modelling of materials

Version - QE@714

LAMMPS



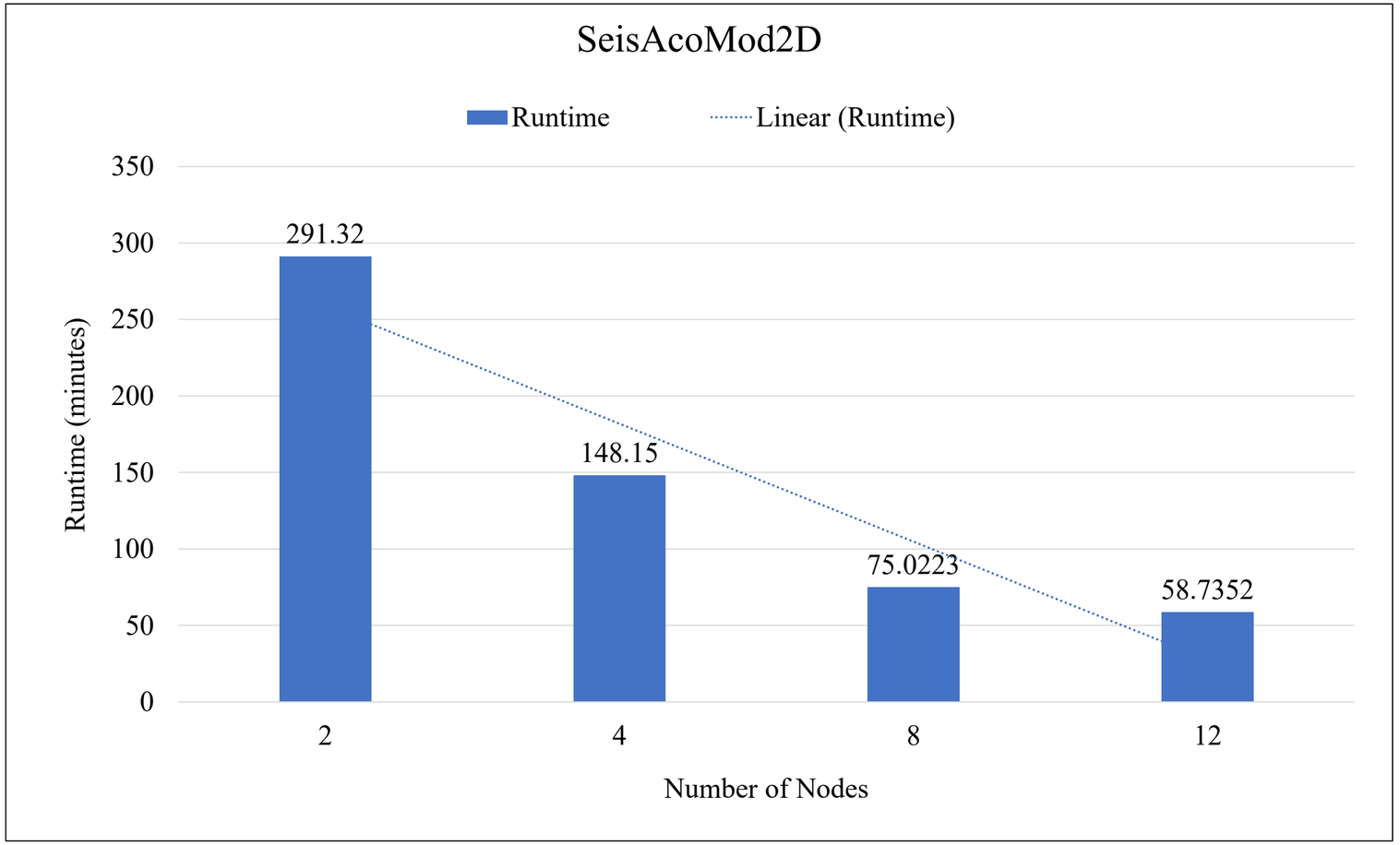
MPI - MVAPICH4

Network - Trinetra-B

Domain – Molecular Dynamics

Version - lammps@Aug2024

SeisAcomo2D



MPI – MVAPICH4

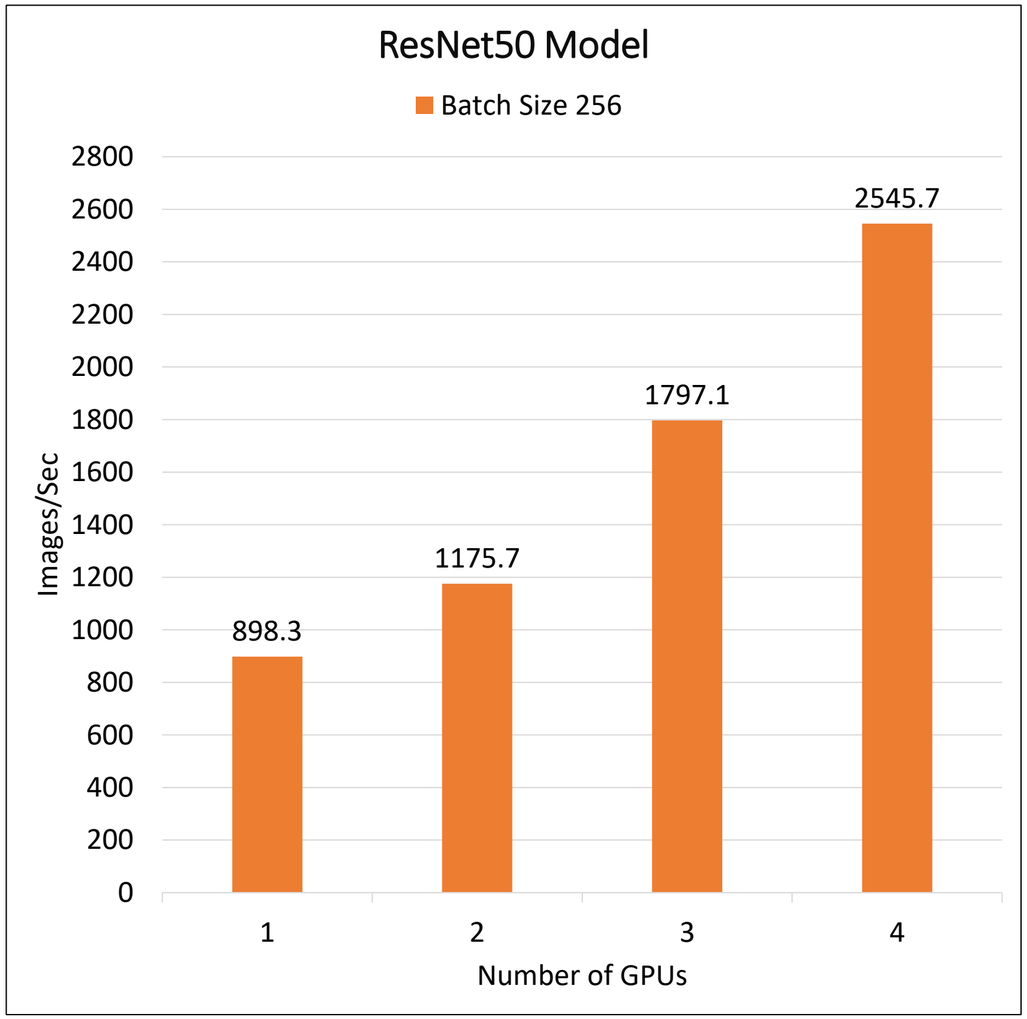
Domain – Seismology

Network - Trinetra-B

Experiments with AI Frameworks on Trinetra

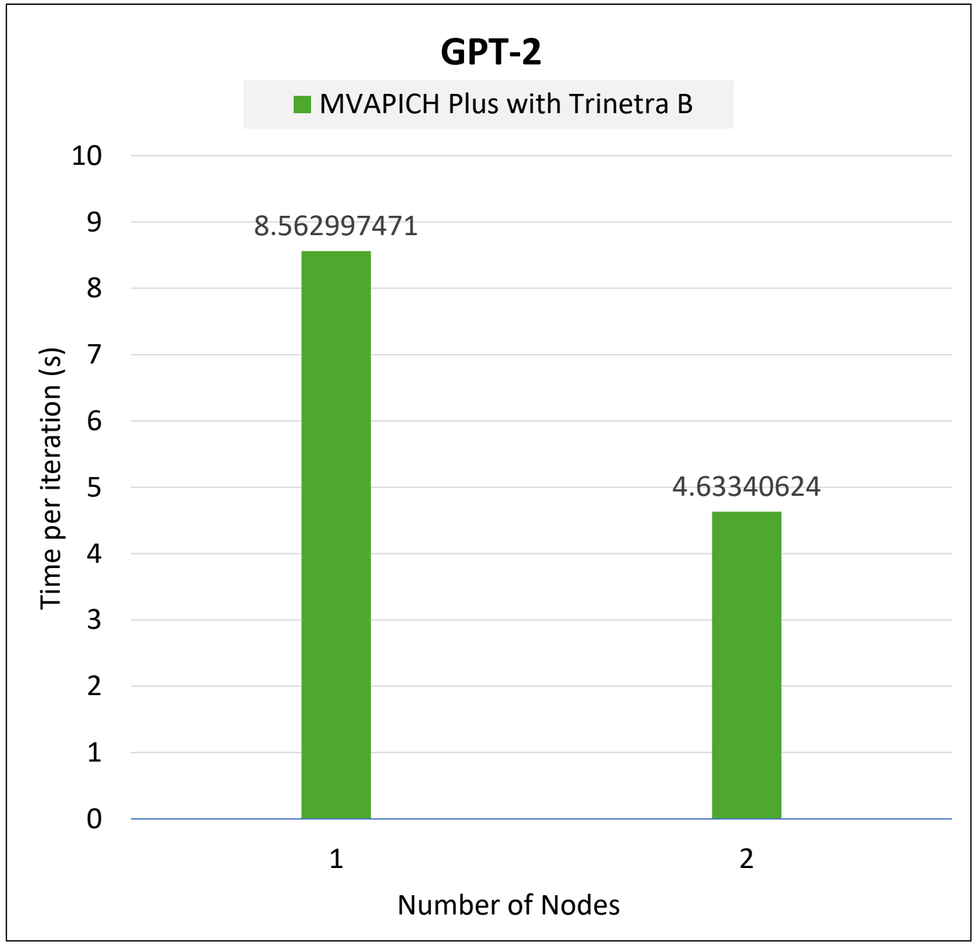
Horovod with MPI over Trinetra

- CPU : Intel(R) Xeon(R) Gold 6240R @ 2.40GHz
- GPU : NVIDIA A100 40GB
- GPU PER NODE : 1
- NIC : Trinetra-A (100 Gb/s)
- DL Framework : PyTorch 1.12.1
- Benchmark : pytorch_synthetic_benchmark
- MODEL : ResNet50



PyTorch 2.0 with MVAPICH-Plus on Trinetra

- CPU : Intel(R) Xeon(R) Gold 6240R @ 2.40GHz
- GPU : NVIDIA A100 80GB
- GPU PER NODE : 1
- NIC : Trinetra-B (200 Gb/s)
- DL Framework : PyTorch 2.6.0
- MODEL : GPT-2
- Benchmark: NanoGPT
- Batch Size: 12
- Block Size: 1024
- Dataset: OpenWebText



MVAPICH4 on RISC-V Architecture

MVAPICH4 on RISC-V Architecture

- RISC-V Architecture
  - Open Standard
  - Customizable and Flexible

- Successfully compiled MVAPICH4 on RISC-V architecture
  - Hardware - StarFive VisionFive2 board
  - OS – Fedora for RISC-V

- Benchmarking with
  - HPL
  - HPCG

Future Work

- 128 node cluster based on Trinetra-B network @ C-DAC Bengaluru
- Optimize Trinetra Eco-system to enhance performance of HPC and AI domain applications in heterogeneous environment.
- Collaborate with application developers to achieve optimal performance.

Summary

- Trinetra is an indigenously developed network.
- Software stack can be tuned based on requirement.
- MVAPICH4 is used as the primary MPI implementation.
- Experiments with HPC and AI workloads on Trinetra enabled clusters.

Thank you!

rkyadav@cdac.in

yogeshwars@cdac.in