MVAPICH3 over C-DAC's Trinetra Network

Yogeshwar Sonawane
C-DAC, India

August 20, 2024

12th Annual MVAPICH User Group (MUG)
2024 Conference

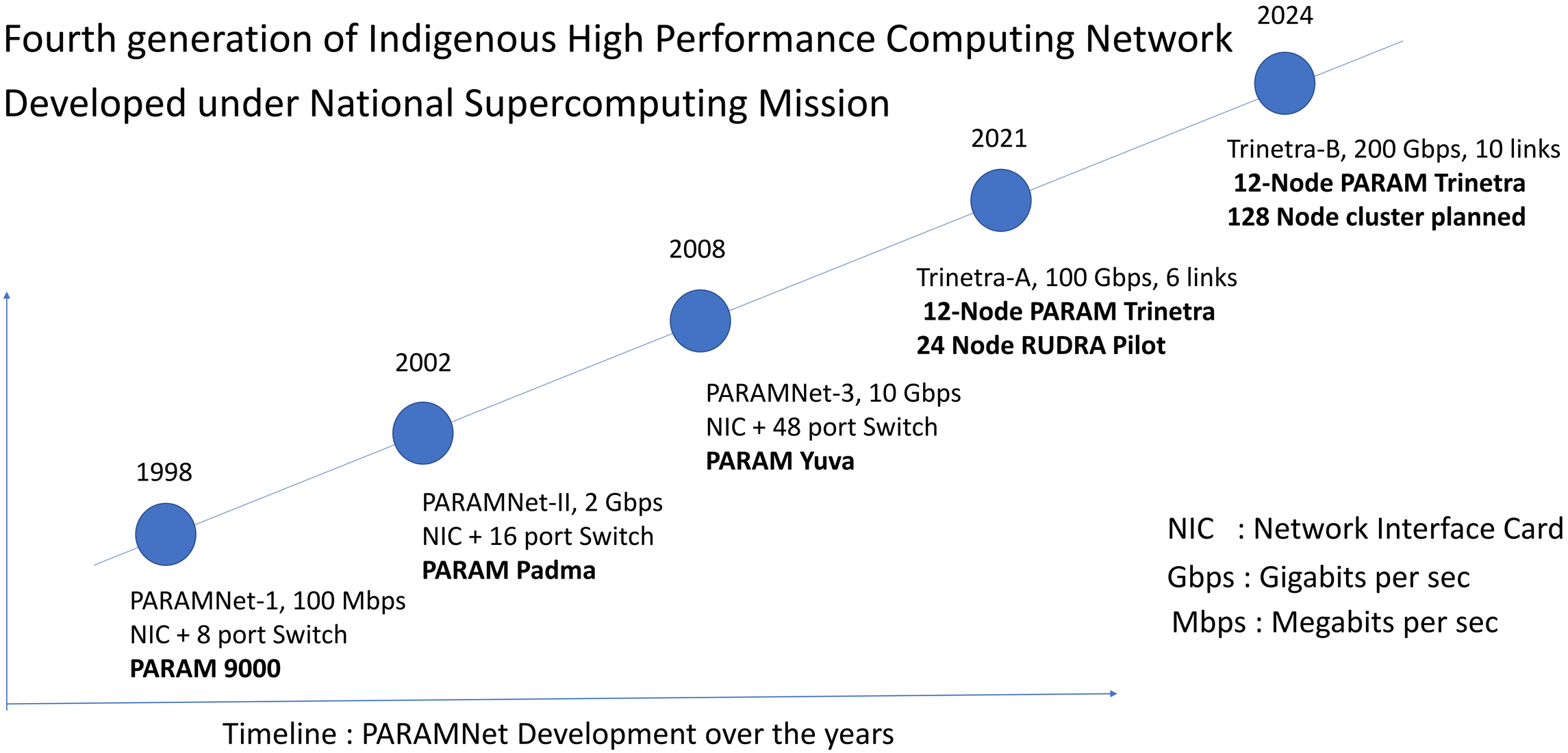
National Supercomputing Mission

Talk Outline

- Introduction to Trinetra HPC Interconnect
- Trinetra Software Stack
  - Features
  - MVAPICH3 support
  - Support for Legacy / Non-MPI applications
  - Performance Evaluation
- Future Work
- Concluding Remarks

C-DAC's HPC Networks

- Fourth generation of Indigenous High Performance Computing Network
- Developed under National Supercomputing Mission



Timeline : PARAMNet Development over the years

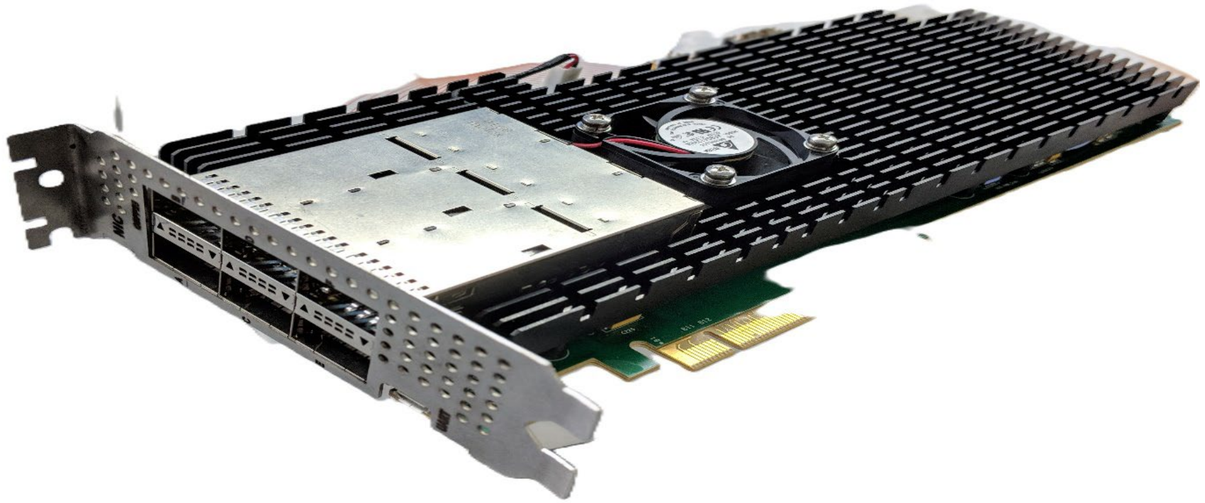
NIC : Network Interface Card

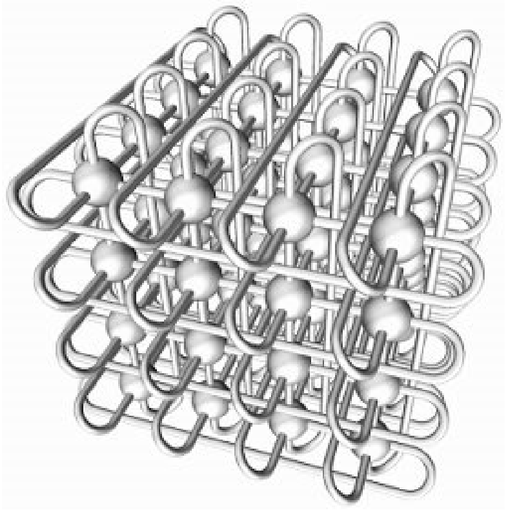
Gbps : Gigabits per sec

Mbps : Megabits per sec

Trinetra-A Network

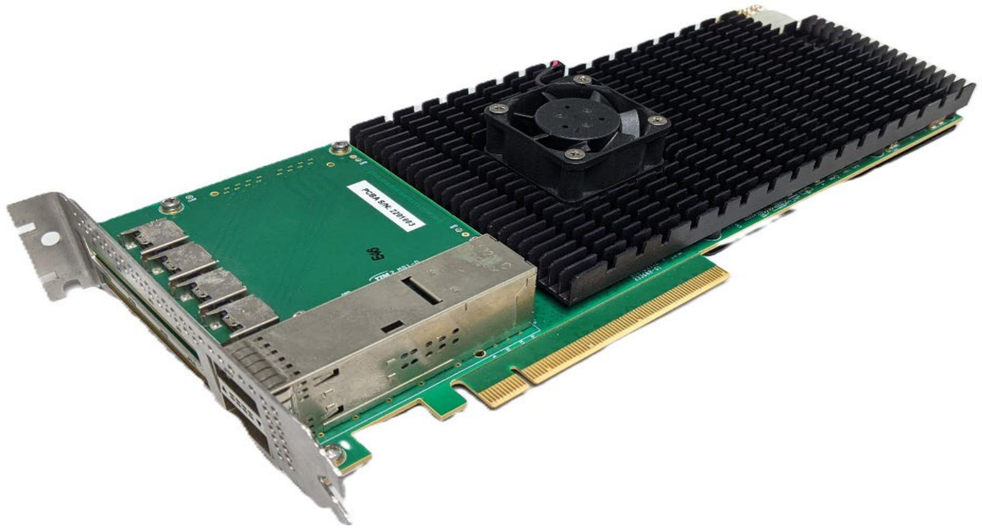
- Physical link Speed : 100 Gigabits per sec
- No. of links : 6 links
- Host Interface : PCIe Gen3 x8
- FPGA : Xilinx Virtex Ultrascale VU095
- Topology : Switchless 3D Torus

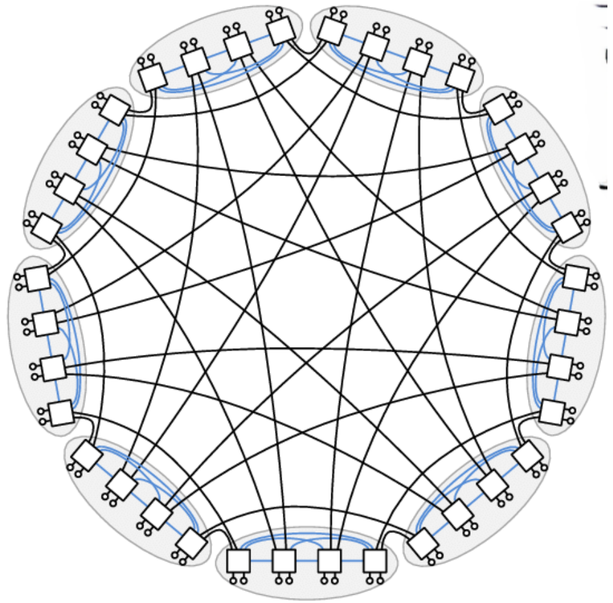




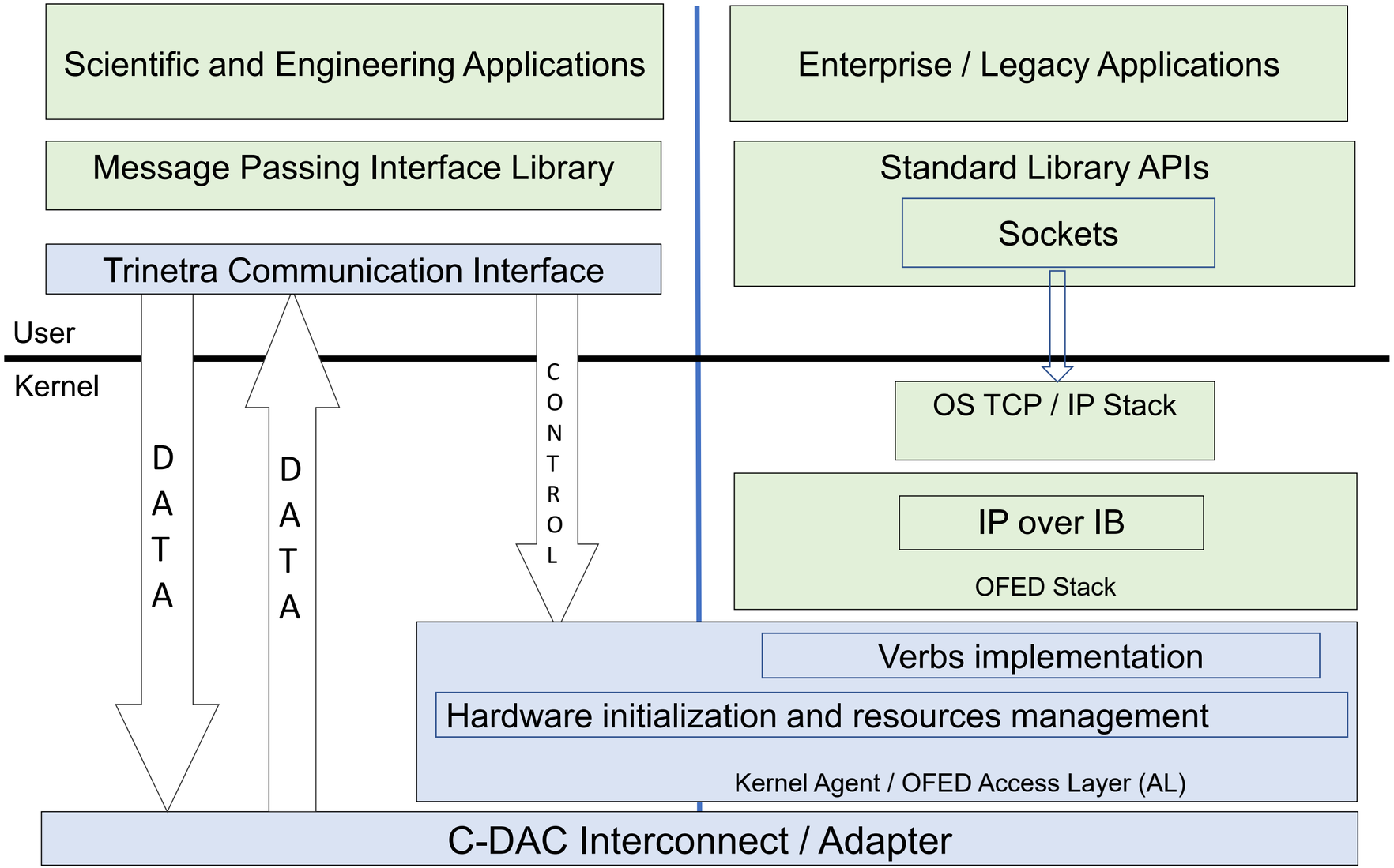
Trinetra-B Network

- 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)





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
  - Adaptive flow control
  - Packetization and assembly of packets
  - Out-of-order packet handling
  - In-order message delivery
- Two mechanisms for data transfer
  - SFQ (for small messages) and EAGER (for medium/large messages)
- Latency optimized path for tiny and small length messages.
- Tag matching.
- Kernel bypass mechanism

MPI Support - MVAPICH Preferred

- Multiple variants for additional features
  - MVAPICH-PLUS, MVAPICH2-X etc.
  - Support one variant, other variants can be supported faster

- Long association with MVAPICH team
  - Using MVAPICH since 2005
  - Excellent interaction, help from the team

- Exploring to support INAM over Trinetra
  - Helpful to corelate network data with MPI data.

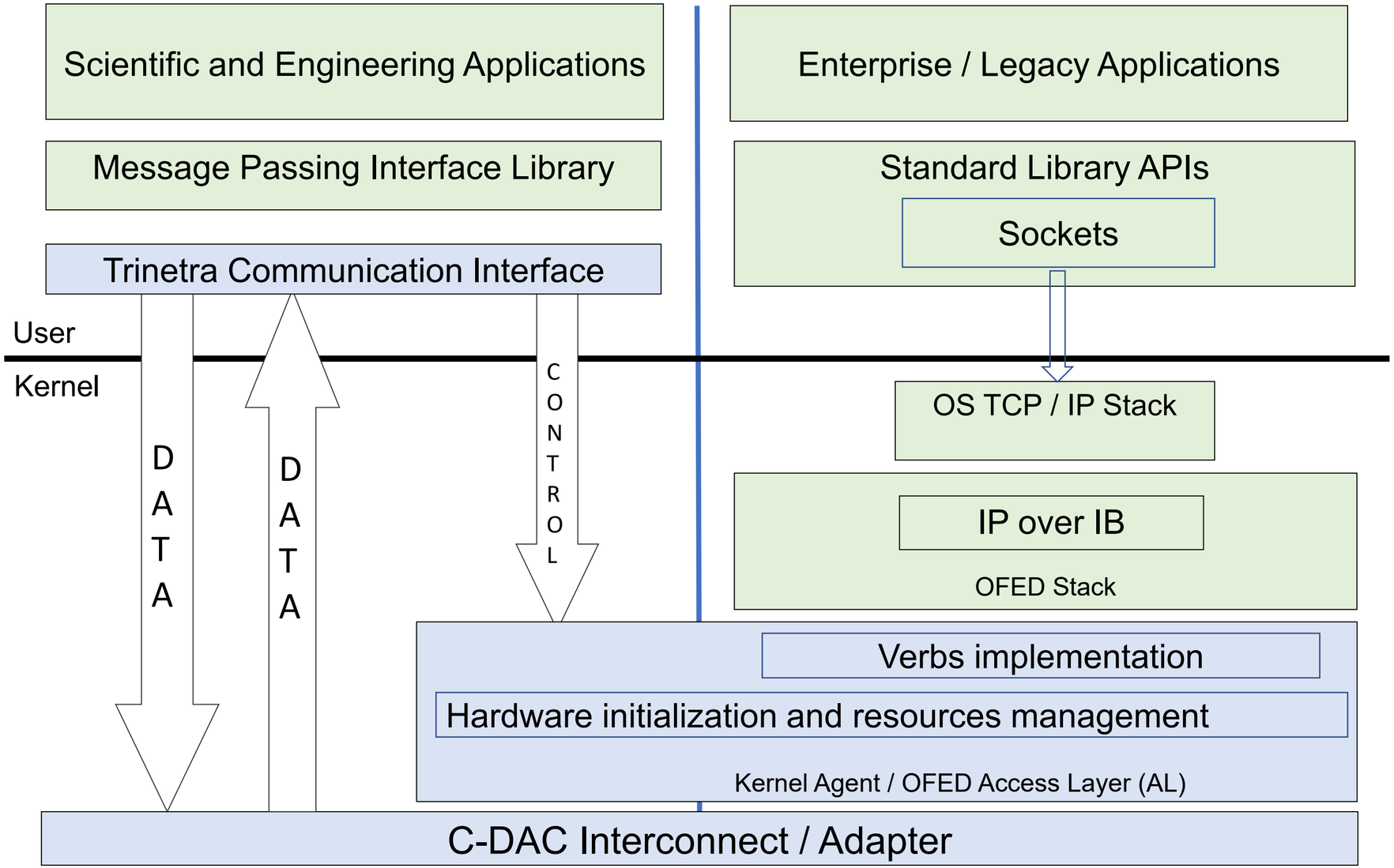
MVAPICH3 support

- MVAPICH3 is supported for Trinetra.
  - Open Fabric Interface (OFI) support
  - Supports CH4 device interface

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

- HPC applications executed using Trinetra
  - GROMACS, OpenFOAM, WRF, LAMMPS

Software Stack for Trinetra-A & Trinetra-B



IPoIBoT: Enabling Sockets over Trinetra

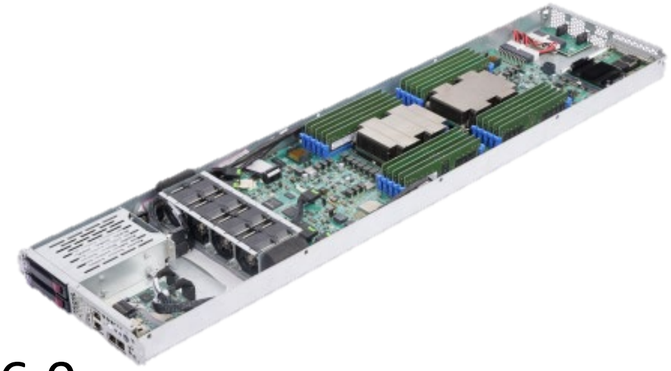
- Enables legacy (socket/IP based) apps
- For non-MPI loads i.e. Enables 2 Application classes (Apart from MPI)
  - Socket interface (IP) applications
    - UDP
    - TCP
- Uses IPoIB stack from OFED/Linux
- Enable other domains (Apart from HPC)
  - AI/ML/DL , Big Data/Data centre, storage, database, software load balancer (web)

- Validation using
  - Iperf, Netperf, FIO, NFS, lustre

Performance Evaluation

Trinetra-A Testbed / Environment at C-DAC, Pune

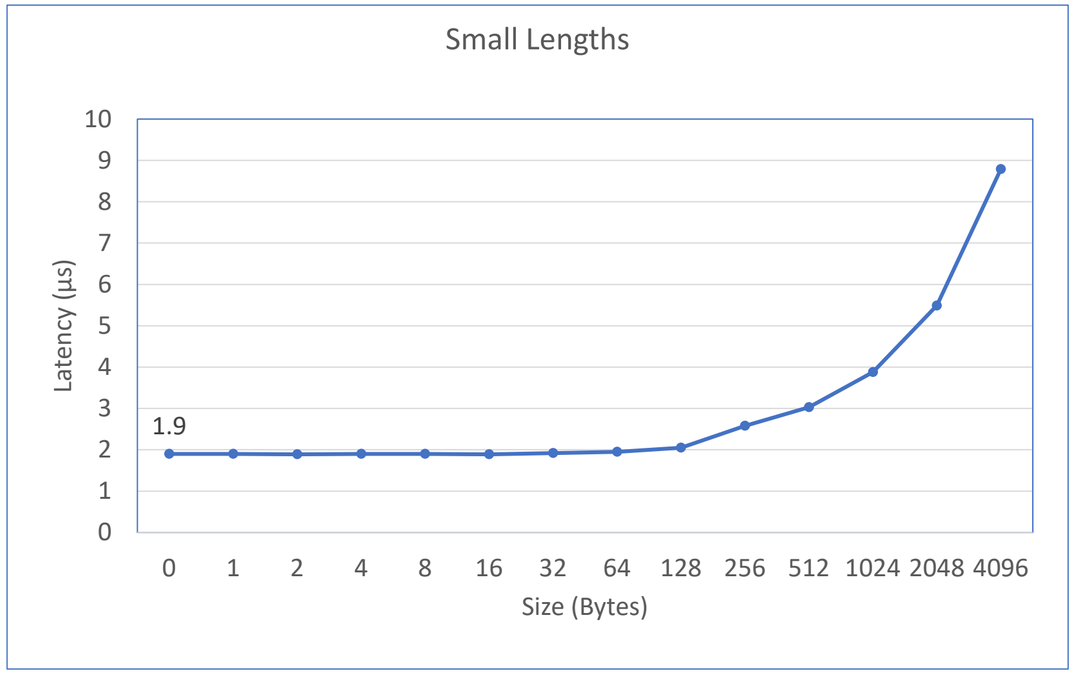
- Cluster – 24 nodes based on C-DAC's indigenously designed **Rudra servers**, Intel Xeon Gold 6240R @ 2.4GHz, 48 cores
- Operation System – CentOS 7.6, AlmaLinux 8.9
- RAM – 196 GB per node
- Primary network – **Trinetra-A 100Gbps (2 x 2 x 6)**
- InfiniBand – Nvidia HDR 100 with Mellanox OFED-24.04-0.6.6.0
- MPI library – mvapich3.0, mvapich2-2.3.7, Libfabric – 1.21
- MPI Benchmark – IMB-2021.8, IMB-4.0.2, OMB (as part of MVAPICH tar)
- HPL – 2.3, OpenBLAS – 0.3.3
- Compiler – gcc-4.8.5, gcc-8.5, icc-2021.4.0, icx-2024.2

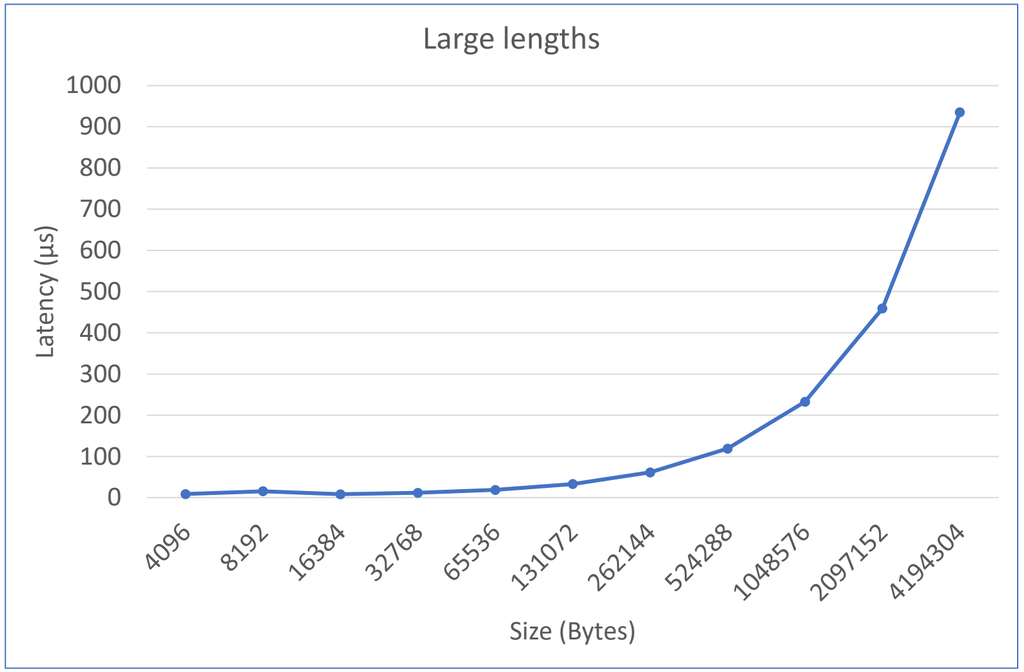


Trinetra-B Testbed / Environment at C-DAC, Pune

- Cluster – 12 nodes, Intel Xeon Gold 5118 @ 2.3 GHz, 24 cores
- Operation System – CentOS 7.6, AlmaLinux 8.9
- RAM – 92 GB per node
- Primary network – **Trinetra-B 200Gbps (4 x 3)**
- InfiniBand stack – OFED-4.17-rc1
- MPI library – mvapich3.0, mvapich2-2.3.7, Libfabric – 1.21
- MPI Benchmark – IMB-2021.8, IMB-4.0.2, OMB (as part of MVAPICH tar)
- HPL – 2.3, OpenBLAS – 0.3.3
- Compiler – gcc-4.8.5, gcc-8.5, icc-2021.4.0, icx-2024.2

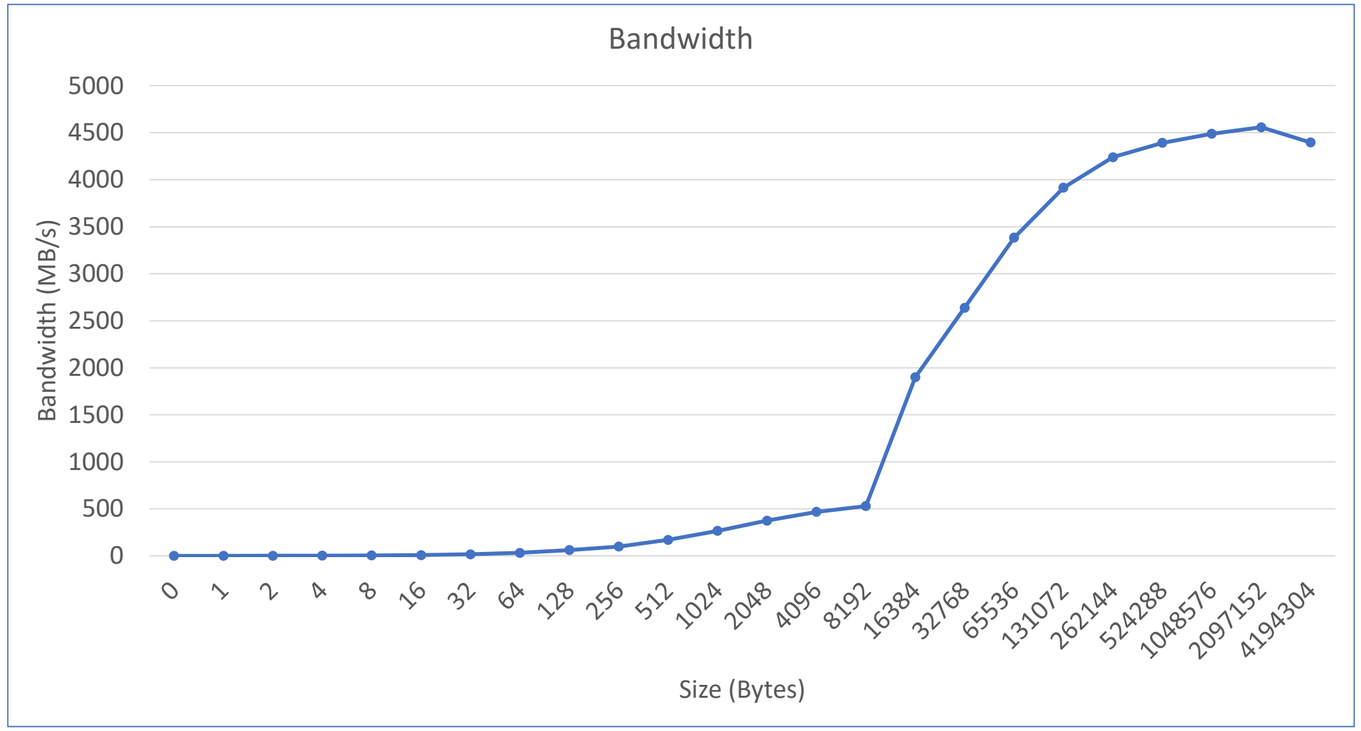
Micro-benchmarks: Trinetra-A Latency





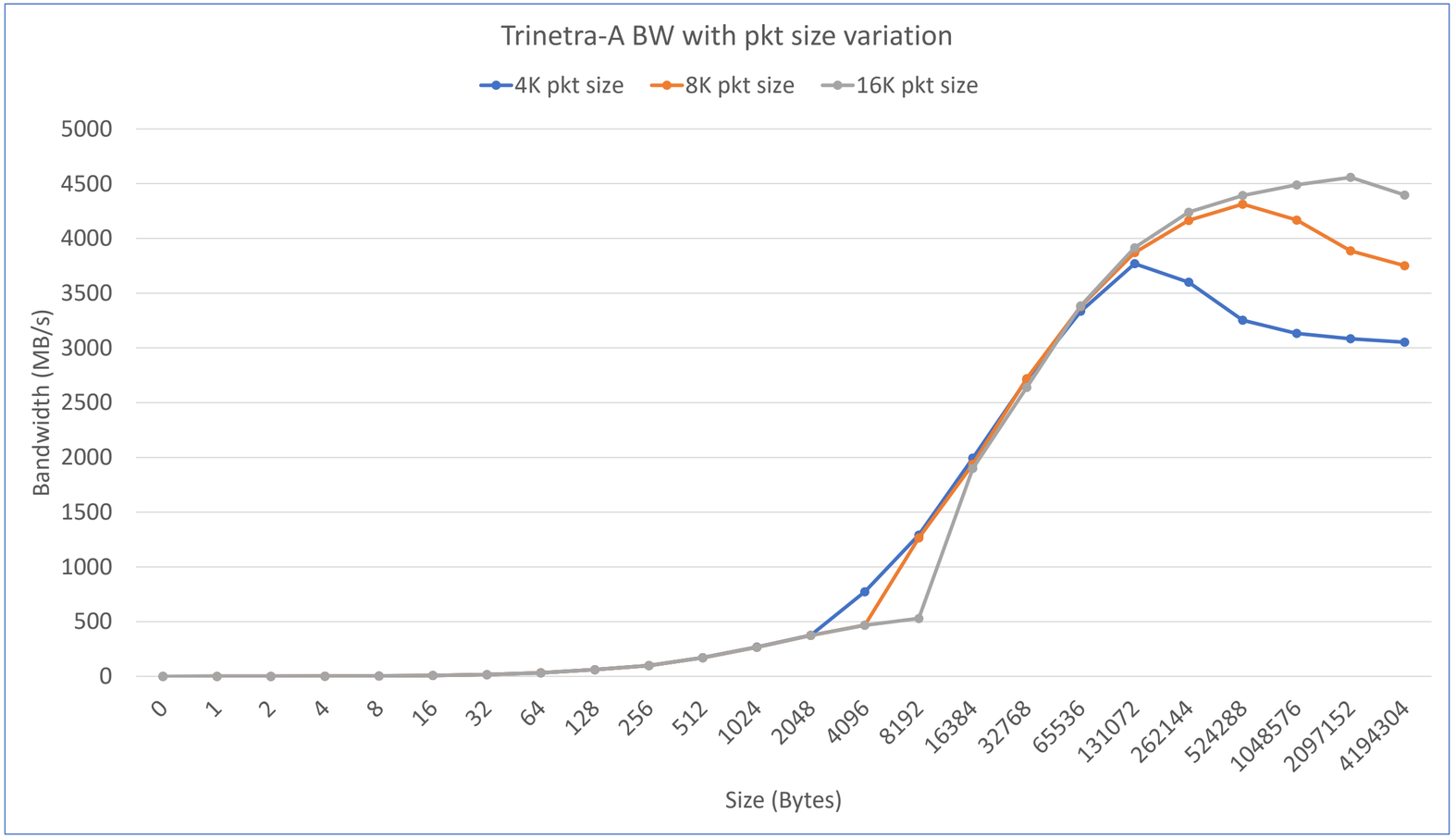
Latency = 1.9 µs

Micro-benchmarks: Trinetra-A Bandwidth

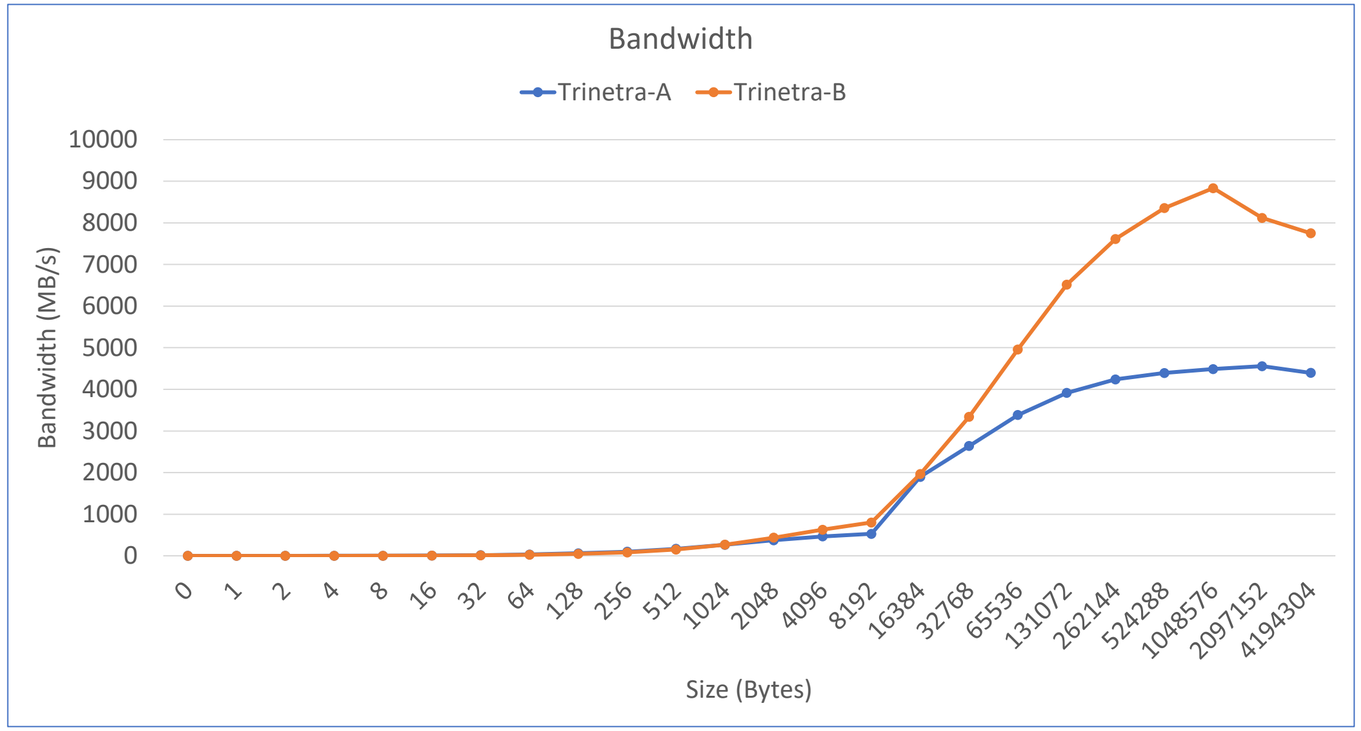


Bandwidth = 4557 MB/s

BW with protocol switch threshold variation



BW comparison between Trinetra-A and Trinetra-B



TrA BW = 4557 MB/s, TrB BW = 8911 MB/s

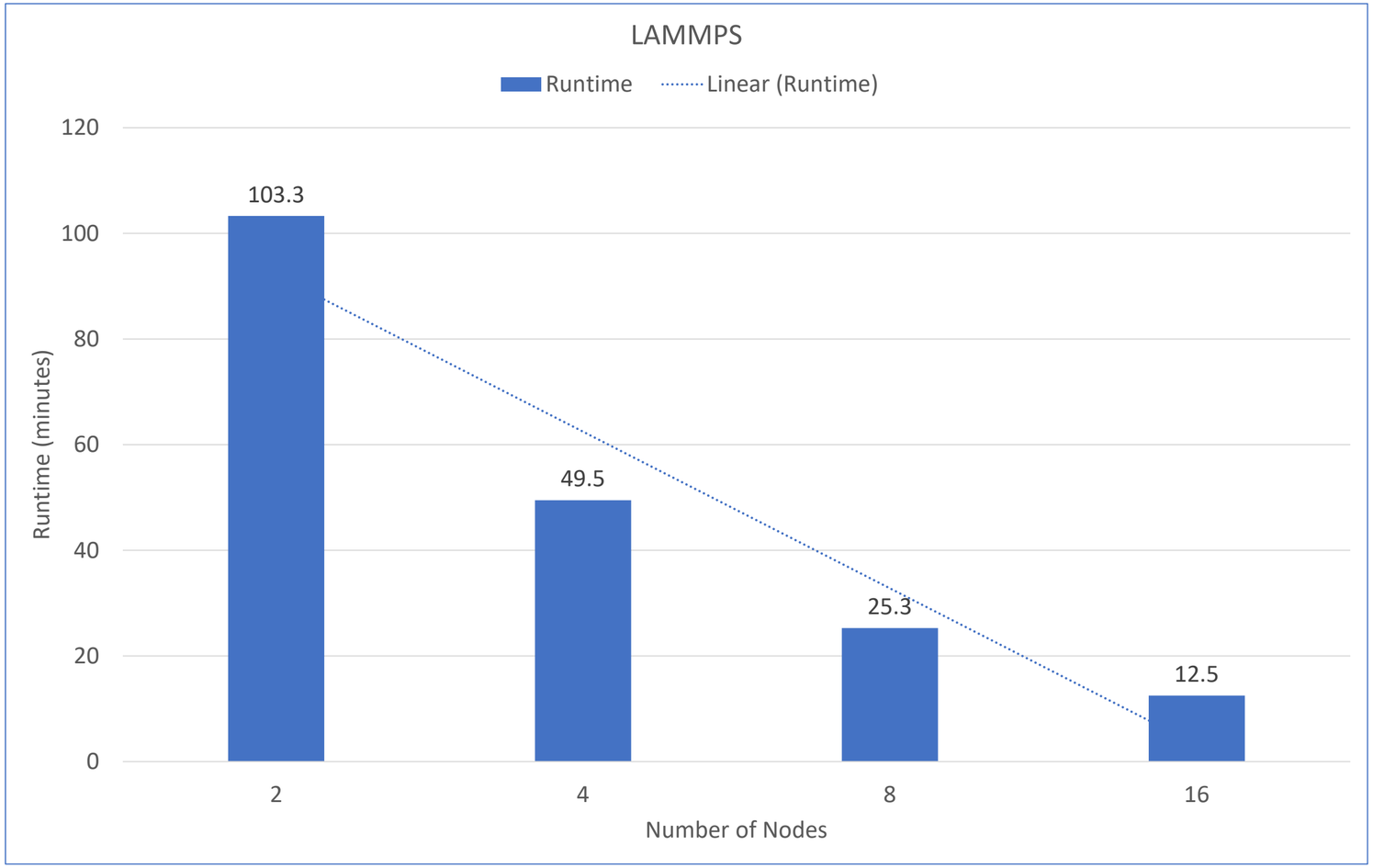
IPoIBoT Performance

Iperf	Netperf
3012 MBPs	2708 MBPs

Network - Trinetra-A

Application Performance

LAMMPS

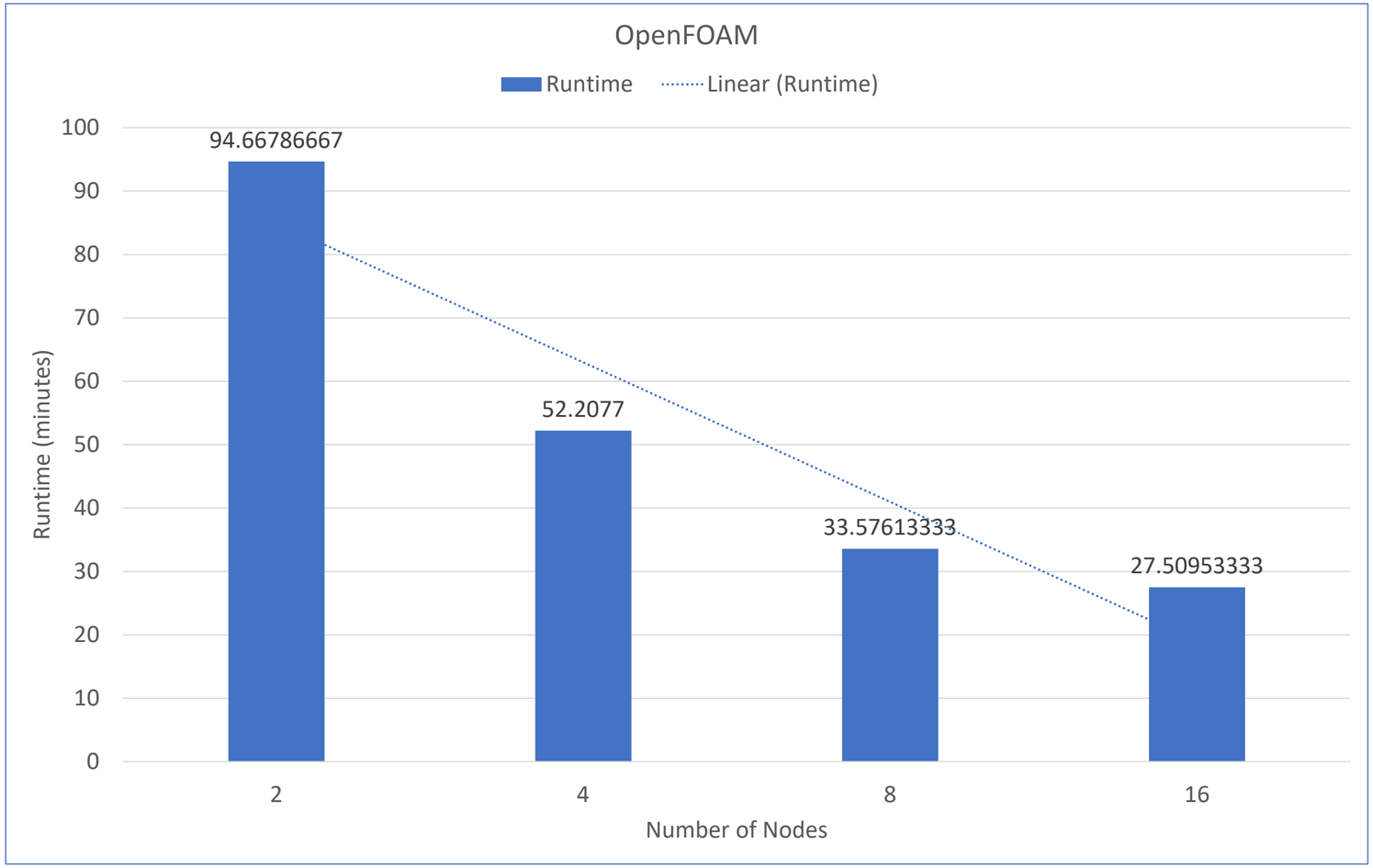


MPI - MVAPICH

Network - Trinetra-A

Version - lammps2@Aug2023

OpenFOAM

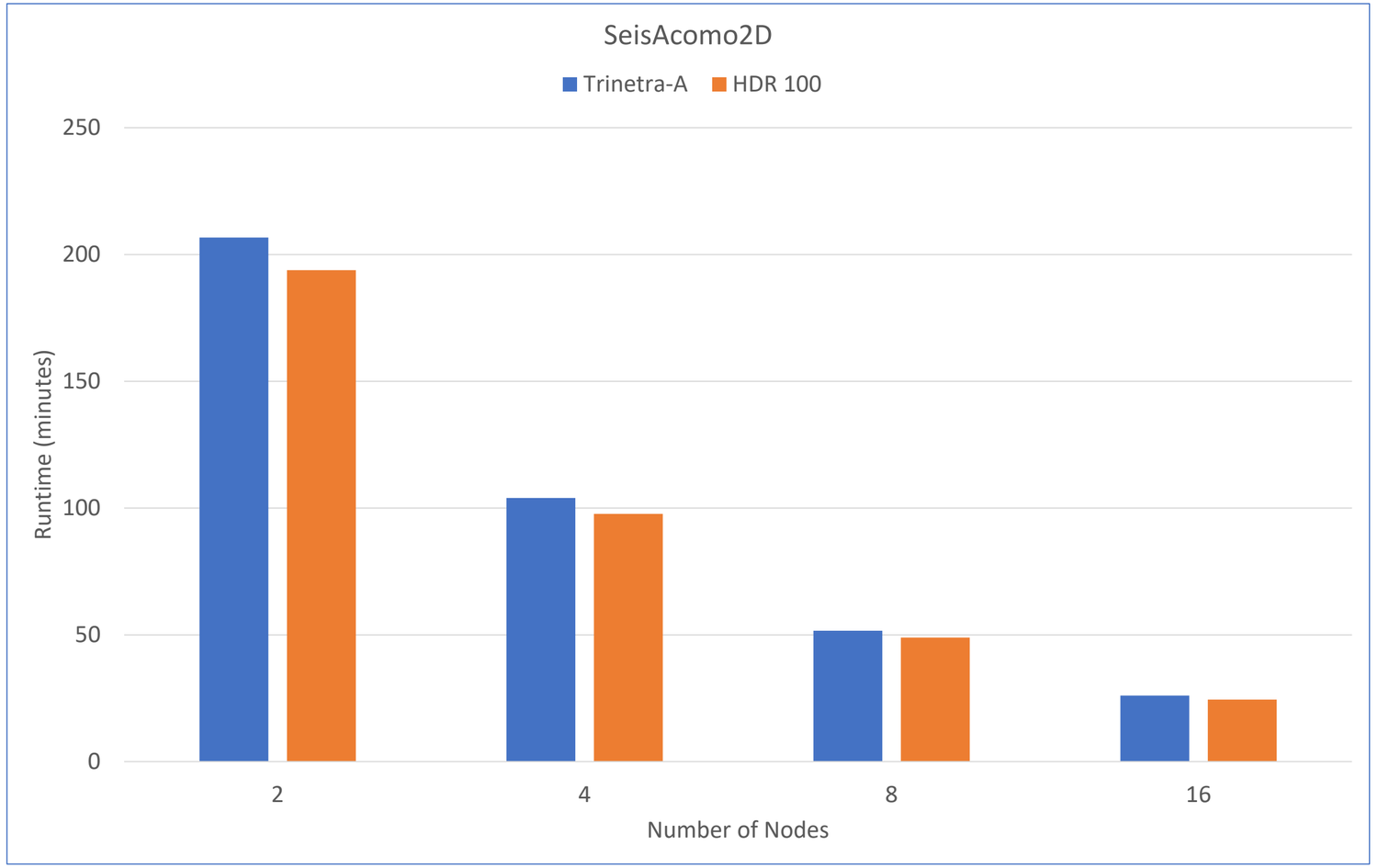


MPI - MVAPICH

Network - Trinetra-A

Version - openfoam@2312

SeisAcomo2D



MPI - MVAPICH

Network - Trinetra-A and Nvidia HDR-100

Future Work

- 128 node cluster based on Trinetra-B network @ C-DAC Bengaluru
- Enabling AI domain applications over Trinetra
- Handholding with application developers for optimal performance
- Supporting INAM over Trinetra

Summary

- Trinetra is an indigenously developed interconnect
- Feature rich software stack
- MVAPICH3 is primary MPI
- Trinetra-B (200 Gbps) with 10 links will provide flexibility of supporting various topologies

Thank you!

yogeshwars@cdac.in