Enabling And Evaluating Lustre File System Over Trinetra HPC Network

Evancer John, Mahesh Chaudhari, **Yogeshwar Sonawane**, Sanjay Wandhekar

C-DAC, Pune, India

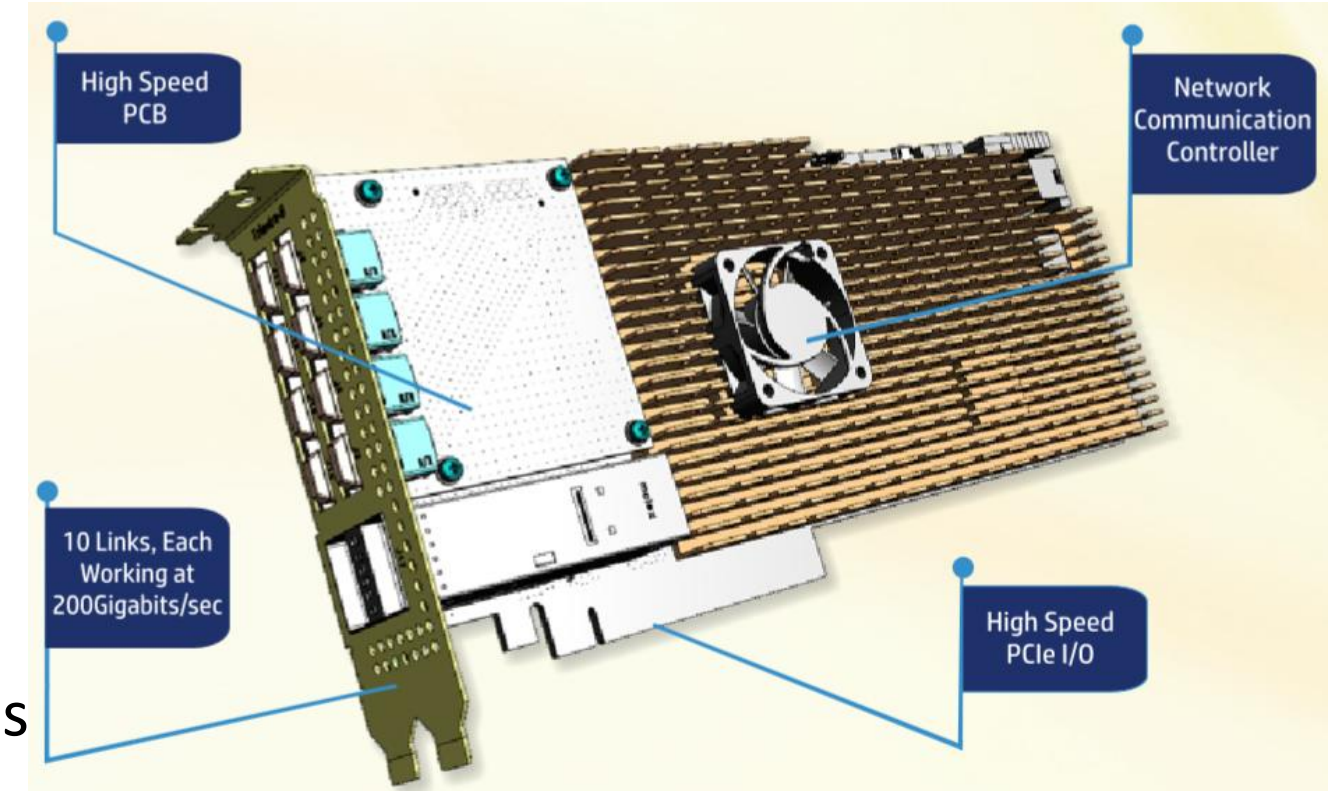
June 26, 2026

Outline

- Introduction - Trinetra
- Motivation
- Lustre Parallel File System
- Problem
- Solution
- Methodology
- Experimental Testbed
- Results
- Conclusion
- Future Work

Introduction - Trinetra HPC Interconnect

- High-speed HPC interconnect developed by C-DAC
- FPGA-based
- Integrated switching capabilities
- Torus and its derivative topology
- Trinetra-A: 6 ports, each 100 Gbps
- Trinetra-B: 10 ports, each 200 Gbps
- Software interface - OFI supported for HPC/AI applications



Trinetra-B NIC

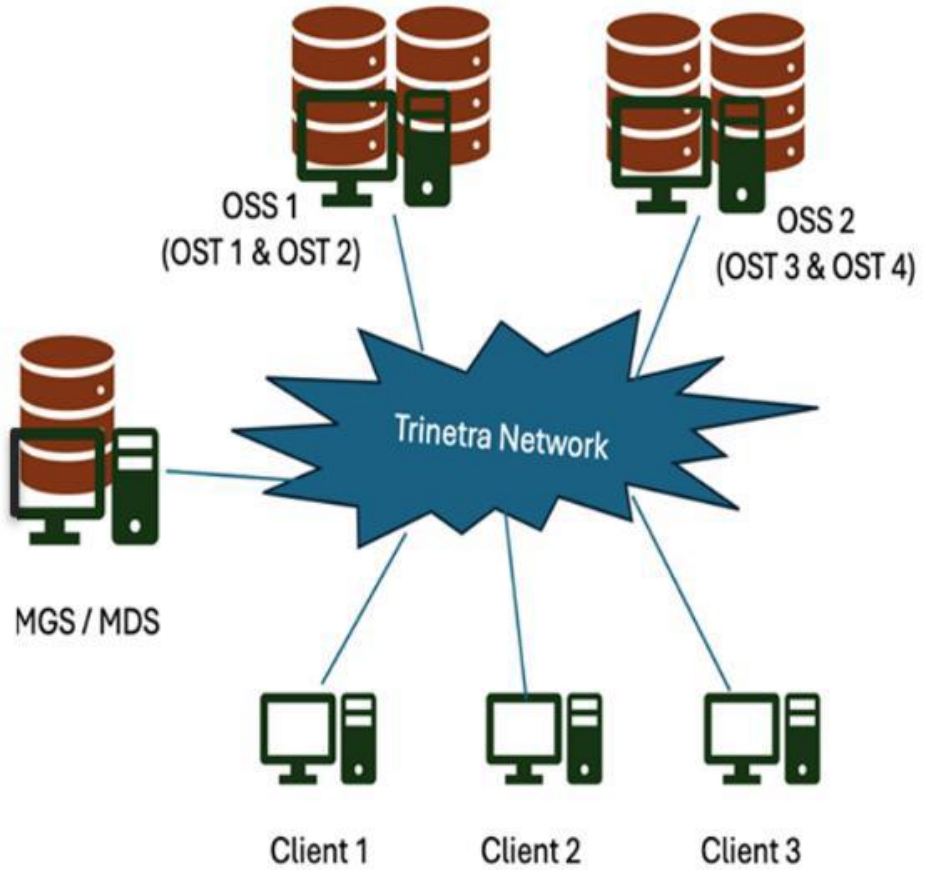
Motivation

Enable Parallel file systems over the Trinetra HPC Network:

- Traditional file systems
  - Designed for serial access patterns
  - Struggle with large-scale concurrent I/O
- Parallel file systems
  - Enable simultaneous access to shared data across many clients
  - Example: Lustre, widely deployed in supercomputers
- **File system performance is also influenced by the network**
  - Deploying Lustre over non-IB/IP networks is not straightforward
  - Requires modifications to Lustre or underlying drivers to support new networks

Lustre Parallel File System

- Delivers high aggregate bandwidth for HPC workloads
- Distributed storage with separate metadata and data management
- **Lustre Networking (LNet)** - flexible layer supporting interconnects like InfiniBand and Ethernet
  - LNet typically communicates using:
    - Verbs Interface - RDMA-based communication
    - Sockets Interface - TCP/IP-based communication



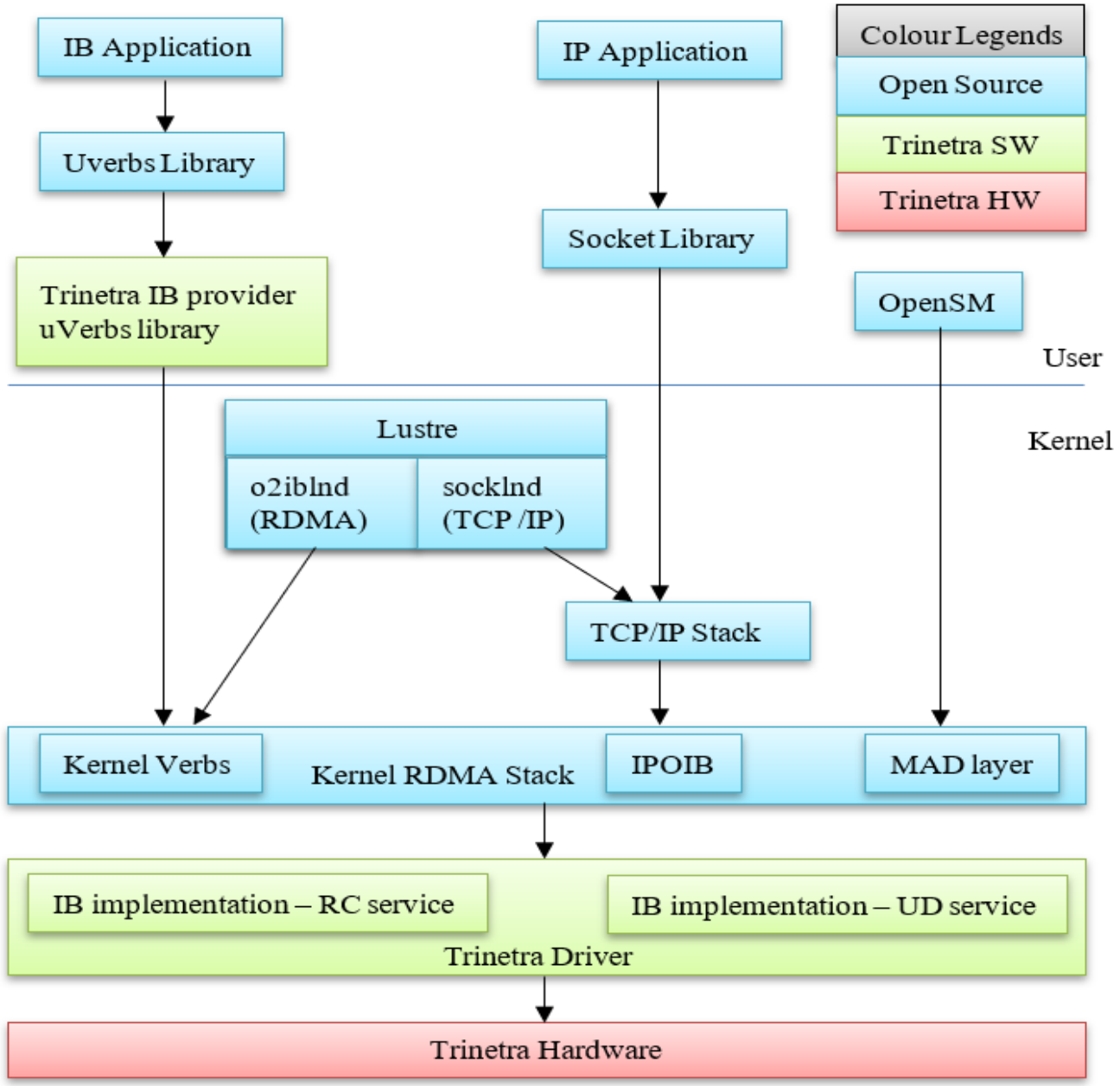
Problem

- Running Lustre over Trinetra requires a compatibility mechanism
  - Modify Lustre – Add LNet for Trinetra
  - Extend Trinetra Driver to support Verbs or Socket based communication

Solution

- Enabling IB (Verbs) and IP based communication over Trinetra
  - Added Verbs support in Trinetra driver to enable Lustre IB/RDMA communication
  - Added IPoIB support in Trinetra driver to enable Lustre TCP communication
- Avoids modifications to Lustre

Software stack for Trinetra

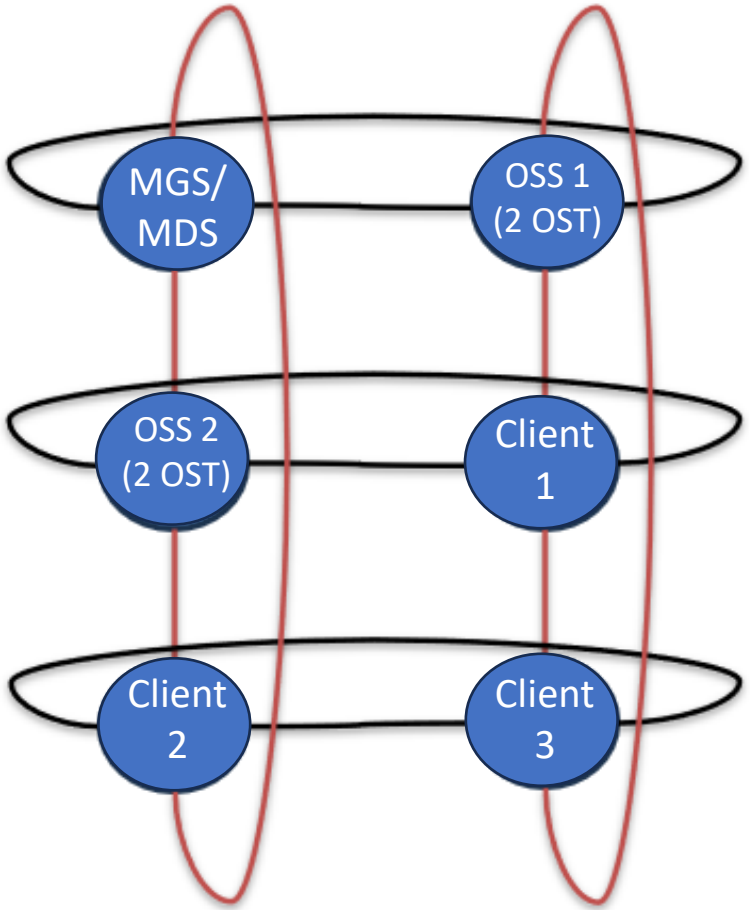


Methodology : Driver Enhancements

- Verbs Support
  - Implemented IB communication semantics to expose Verbs interface
  - MAD support for OpenSM interoperability
- IPoIB Support
  - Added UD, RC services, and IB multicast support
- RDMA Compatibility
  - Emulated RDMA WRITE Verbs operation using send/recv semantics
- IB over Trinetra
  - Encapsulated IB packets within Trinetra link fragments for Trinetra routing

Experimental Testbed

- Cluster of **6 RUDRA** servers
- Interconnected with **Trinetra-B** network
- Connected in a **2-D torus** fashion with configuration of 3x2
- OS is AlmaLinux 8.9, Lustre-patched kernel
- **Lustre version 2.15.4**
- IB/RDMA stack of Linux Kernel 4.18
- 800 GB **Micron 7300 MAX** U.2 NVMe SSD for local storage



Results: Network Performance on Trinetra-B

- Establishes baseline network performance metrics
- Observed bandwidth falls short of theoretical 200 Gbps link rate
  - Constrained by PCIe Gen3 x16 (128 Gbps) and FPGA frequency limits

Tests	Verbs	IPoIB	Native
IMB Pingpong bandwidth (MB/s)	7869	3201	10324
IMB Pingpong latency (µs)	8.35	12.85	2.4

Results: Single Disk Performance

- Assessed NVMe storage performance using FIO benchmarks
- Tested a single 800 GB Micron 7300 MAX U.2 NVMe SSD
- Established baseline reference figures for storage capability

FIO Operation	Performance Result
Sequential Read (MB/s)	1433
Sequential Write (MB/s)	901
Random Read (KIOPS)	185
Random Write (KIOPS)	188

Results: Lustre Performance over Trinetra-B

IOR

- Measures parallel I/O throughput across Lustre clients and OSTs

Lustre Network	IOR Write bandwidth (MB/s)	IOR Read bandwidth (MB/s)
Verbs	3732	5118
Sockets	3642	4634

Results: IO500

- Runs multiple phases to stress different aspects of a filesystem

Operations	Verbs	Sockets
ior-easy-write	2.989073 GiB/s	2.513573 GiB/s
ior-easy-read	3.068740 GiB/s	2.059966 GiB/s
ior-hard-write	0.410399 GiB/s	0.413255 GiB/s
ior-hard-read	2.671684 GiB/s	1.733779 GiB/s
mdtest-easy-write	36.089491 KIOPS	29.314652 KIOPS
mdtest-hard-write	11.843909 KIOPS	10.271125 KIOPS
MDTest Score	41.379278 KIOPS	36.314486 KIOPS
Total Score	8.584253	7.099216

Conclusion

- Provides a method for deploying Lustre over custom HPC interconnects
- Driver extensions enable compatibility with standard Lustre transports (Verbs and Sockets)
- No modifications required to Lustre
- Benchmark results demonstrate feasibility and efficacy

Future Work

- Extend experiments to a larger cluster with increased storage capacity
- Performance optimization by tuning Lustre configuration parameters
- Evaluate other parallel file systems like BeeGFS, DAOS
- Evaluate Lustre on next-generation Trinetra with hardware RDMA support

Thank You

yogeshwars@cdac.in

evancervj@cdac.in