Skyway My Way! InfiniBand to Ethernet Optimization and Benchmarking

Joe Downs, Brian Hsu, Jeremy Mao (HPC-DO) | Mentors: Doug Egan, Jesse Martinez (HPC-INF)

Introduction

High-performance computing environments increasingly require bridges between InfiniBand compute fabrics and Ethernet storage infrastructure. The NVIDIA Skyway is a gateway designed to facilitate Internet Protocol over InfiniBand (IPoIB) communication between these two networks.

Scope

This work focused exclusively on network performance and gateway functionality. We established baseline InfiniBand performance, characterized Skyway gateway routing efficiency, validated multi-tenant partition security boundaries, and identified optimization strategies for HPC deployments.

Testbed Specifications

Hardware
IB Switch: NVIDIA MQM8700-HS2F
Eth. Switch: Arista 7060DX5-64S 400G
NICs: ConnectX-6 HDR (MT4123)
NVIDIA Skyway: MGA100-HS2
Cables: ENET 400G/200G cables/optics

Tools
perftest suite: v6.29
iperf: v2.1.6
mpi v5.0.10
fio: v3.35
GPCNeT: v1.3

Software
OS: Rocky Linux 9.7 (kernel 5.14.0)
NVIDIA UFM: 6.25.1-7.el9
NVIDIA Skyway OS: v8.2.2500

Drivers
DOCA OFED: v26.04-0.8.6
ConnectX-6 FW: 20.39.3560

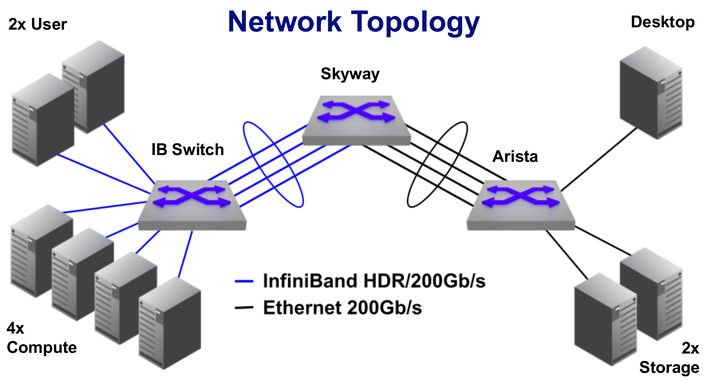


Fig 1. *9-node compute cluster configuration consisting of an ultra-low latency (<2 µs), high bandwidth (200G) InfiniBand fabric and a 200G Ethernet network.*

Skyway & IPoIB

The NVIDIA Skyway serves as a layer 3 gateway between high-performance InfiniBand and Ethernet storage networks, facilitating large data transfers via IPoIB. Skyway eliminates the need for dual NIC I/O server farms, reducing infrastructure costs.

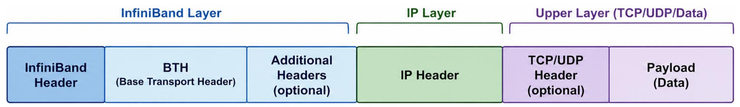


Fig 2. *High-level diagram of an IPoIB packet.*

Baselines & Optimizations

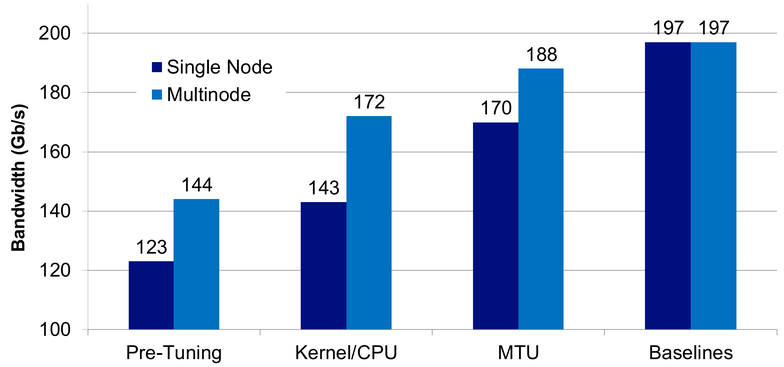


Figure 3. *iPerf bandwidth test results displaying incremental bandwidth optimization of the IPoIB data path, increasing single-source by 38% and multi-source by 30%, ultimately achieving bandwidth close to nominal HDR InfiniBand baseline rate.*

Network Optimizations

- Configured kernel and OS settings
  - ❖ Set CPU governor to performance mode
  - ❖ Increased socket/TCP buffer sizes and kernel ring buffers (512 → 4096 bytes)
  - ❖ Activated jumbo frames on bonded network interfaces
- MTU Optimization: Aligned IB and Ethernet interfaces (2044 → 4090 bytes) to minimize packet overhead.

VF Hash Bottleneck: Single-source limited to 170 Gb/s by source-IP-only VF hashing, lacking layer3-layer4 hashing for single-source distribution

Congestion Performance

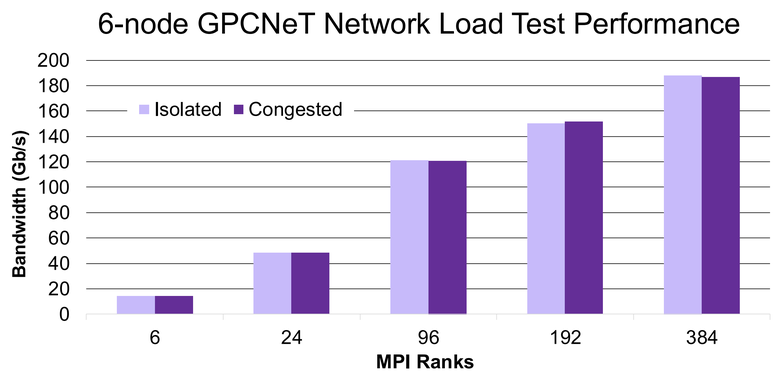


Figure 4. *Modified GPCNeT congestion test bandwidths for 1, 4, 16, 32, and 64 processes per node.*

A modified GPCNeT network load test, to isolate cross-Skyway traffic, was used to test Skyway performance under congestion. With 6 nodes active, 2 as tests and 4 as congestors, simulated traffic had a negligible effect on total bandwidth, even as the tests scaled in ranks and processes/node.

End-to-End Network Partitioning

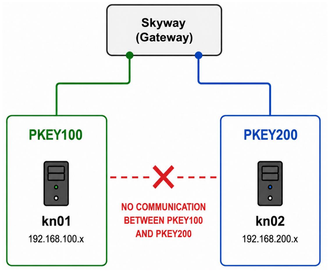


Fig 5. *InfiniBand **partitions** create isolated user environments within the shared fabric. Testing confirmed that this separation remained effective as traffic crossed the Skyway gateway into the Ethernet network.*

The approach provides consistent end-to-end isolation across both IB and Ethernet while allowing shared infrastructure to be used securely.

NFS Performance

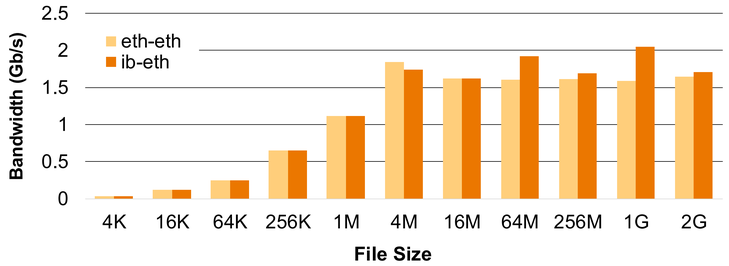


Figure 6. *Cumulative bandwidth measured with fio using random reads, one job, iodepth of four and blocksize of 4K. Absolute bandwidth utilization was poor, but transfers through the Skyway saw no performance hit compared to pure Ethernet transfers.*

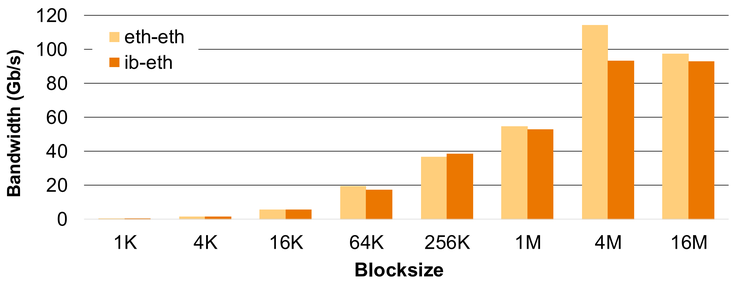


Figure 7. *Bandwidth measured with fio using random reads, one job, iodepth of four and file size of 2G. Increasing blocksizes produced better speeds; max speed at blocksize of 4M: 114 Gb/s and 93 Gb/s for IB and Ethernet, respectively.*

Discussion

Future LANL AI/ML, MPI, and parallel file system workloads depend on high-bandwidth, low-latency InfiniBand fabrics. Our work explored the capabilities and optimizations of the Skyway as a gateway between InfiniBand and Ethernet networks. Results verified expected Skyway functionality, with room for improvement in single-flow performance.

Future work

- High Availability: Evaluate Skyway HA configurations for increased bandwidth and multi-gateway operation.
- Layer 3/4 Hashing: improve load balancing and single-flow bandwidth utilization.

Los Alamos NATIONAL LABORATORY
NNSA National Nuclear Security Administration
Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
LA-UR# 26-26585
UNIVERSITY OF LOUISVILLE
UNIVERSITY OF MICHIGAN
PURDUE UNIVERSITY