Between the NIC & a Hard Place: Evaluating 400 Gb/s Ethernet for HPC Data Transfers

Evelyn Needham, Adelle Ferris, Nikole Grandez | HPC-DO
Mentors: Jesse Martinez, Doug Egan | HPC-INF

Overview

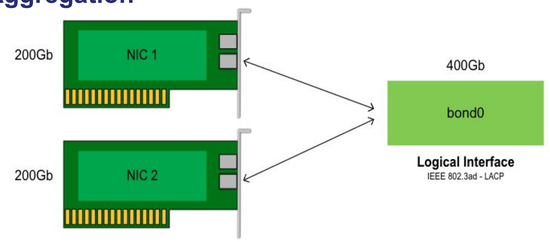
Introduction

As high-performance computing (HPC) workloads continue to scale in complexity and volume, efficient data movement is essential. This project assess the production readiness of 400Gb/s Ethernet by benchmarking performance locally and in a simulated long-distance environment to explore whether parallelized transfers can effectively fill a 400Gb pipe.

Scope

All tuning efforts targeted the goal of optimizing data transfer performance, rather than general system improvements. This was evidenced by **MPI** job performance actually decreasing after our tunings, despite data transfer improvement increasing by nearly **2x**.

Link Aggregation



Each compute node was equipped with two 200Gb Mellanox **Connect x6** NICs which were bonded into a single logical interface by setting the bonding mode to IEEE 802.3ad with **LACP** enabled.

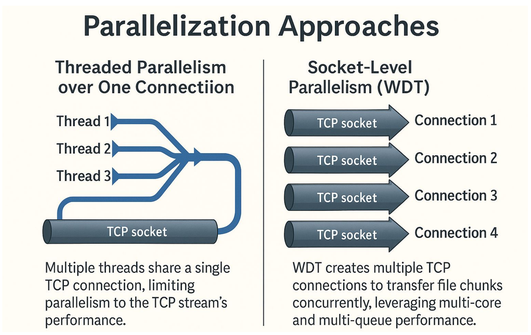
Data Transfer Tools

We evaluated several data transfer tools to find the best performance.

Key Requirements:
- Open-source
- Parallel
- Fully utilize bonded NICS

Tools Tested:
- HPN-SSH
- ProFTPD + lftp
- bbcp
- WDT

Results:
After network tuning, WDT significantly outperformed the other tools, achieving transfer speeds up to **15x faster**. While the other tools use multiple threads over a single socket pair, WDT opens multiple socket connections to enable true parallelism.To avoid I/O bottleneck (~100 Gb/s), we ran WDT in a 'skip_writes' mode where the receiving node discarded incoming data, allowing us to isolate network performance. This behavior was validated with tcpdump.



Testbed Specifications:

Hardware
Switch: Arista DCS-7060DX5-64S-F
CPU: Intel(R) Xeon(R) Gold 6438Y+
NIC: Nvidia MCX653106A-HDAT
Cables: 16 Fiber Y breakout (400gbps)

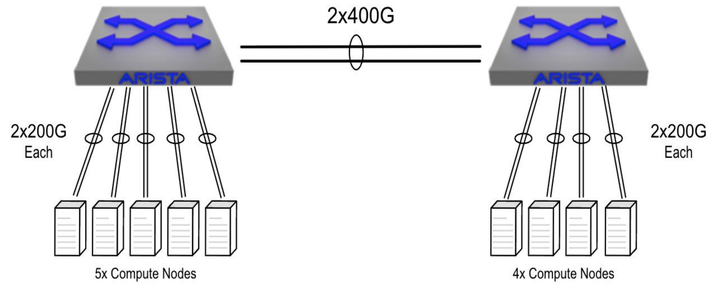
Software
OS: Rocky Linux 9.5 (kernel 5.14.0)

Tools
iperf2: v2.1.6
HPN-SSH: v18.6.2
ProFTPD: v1.3.8d
lftp: v4.9.2
bbcp: v15.02.03.00.1
WDT: v1.27

Drivers/Libraries:
DOCA OFED: v25.04-0.6.1 | **ConnectX-6 Firmware**: v0.43.1014

Local Area Network (LAN)

LAN Environment



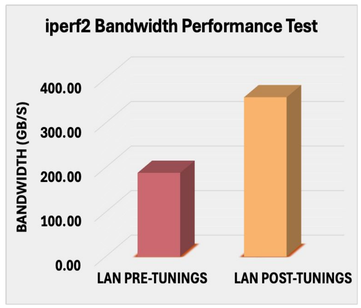
This network was configured as a Layer 2 environment; all nodes were on the same subnet and could communicate directly.

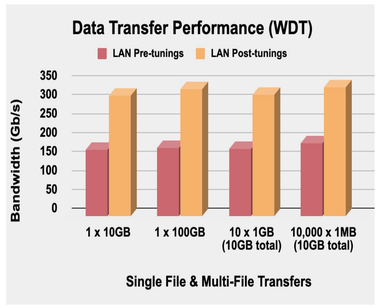
Network and System Optimization

Targeted performance optimizations to maximize throughput:

- ★ Set load balancing (bond & switches) to layer3+4 for IP + port hashing to support WDT's multi-port transfers
- ★ Set CPU governor to *performance* and pinned IRQs to set NUMA
- ★ Tuned kernel via **sysctl**
  - Increased socket buffer size & TCP send and receive buffer sizes
  - Switched to the 'bbr' congestion control algorithm
- ★ Mellanox NIC tuning via **ethtool** & **mlnx_tune** tool
  - Expanded receive and transmit kernel ring buffers
- ★ Enabled jumbo frames on bonded interfaces
- ★ Flashed NICs with optimized Mellanox driver

Performance Results

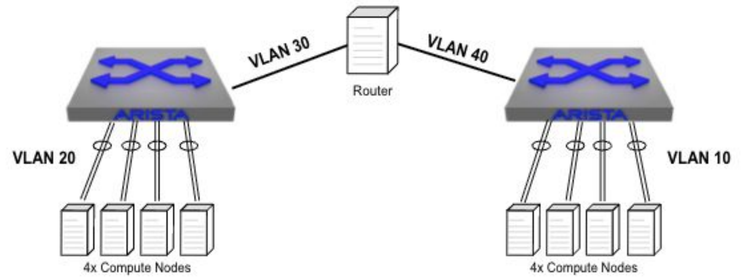




Wide Area Network (WAN)

Using Virtual Local Area Networks (VLANs), we logically separated network paths and configured a node as a router using static routes to evaluate how our tuned network performed under wide-area conditions. We used Linux's **tc** (traffic control) with **netem** to introduce latency and packet reordering on links without needing physical distance.

WAN Environment



Performance Results

Conditions (Unless Stated Otherwise): 44ms Latency RTT, 0.02% Packet Reordering

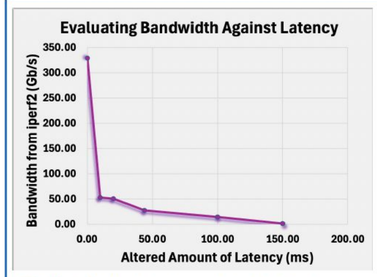


Figure 1: *Performance sharply drops when latency is introduced at 10ms (**25x** the prior RTT).*

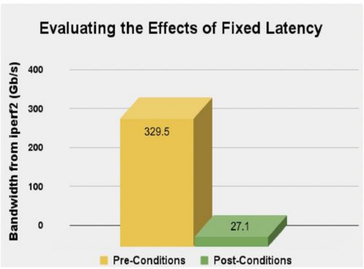


Figure 2: *Performance dropped **12x** with latency introduced. This impact is likely due to a LAN-scaled TCP window size on a WAN.*

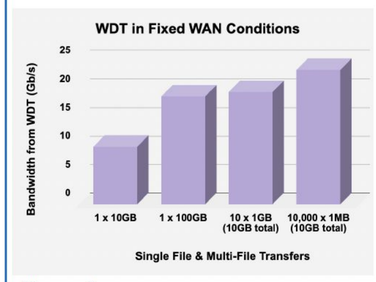


Figure 3: *Small-file directories outperform a single file by over **2x** in a WAN environment.*

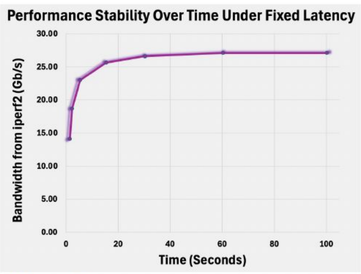


Figure 4: *Emphasizes the need for sustained transfers to reach full, stabilized throughput in a WAN.*

Future Work

- Compare TCP with LACP bonding to RDMA using multi-rail support.
- Integrate high-performance file systems to evaluate performance with NVMe over Fabrics (NVMe-oF).
- Experiment with WAN-specific tunings and transport optimizations.

Los Alamos NATIONAL LABORATORY
National Nuclear Security Administration
Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
LA-UR-25-27873
COLORADO STATE UNIVERSITY
UNIVERSITY OF MARYLAND
COLORADO COLLEGE