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

Adelle Ferris*
adelleferris309@gmail.com
Los Alamos National Laboratory
Los Alamos, New Mexico, USA

Evelyn Needham
evelynmneedham@gmail.com
Los Alamos National Laboratory
Los Alamos, New Mexico, USA

Nikole Grandez
ncgrandez@gmail.com
Los Alamos National Laboratory
Los Alamos, New Mexico, USA

Jesse Martinez
jmartinez@lanl.gov
Los Alamos National Laboratory
Los Alamos, New Mexico, USA

Doug Egan
wegan@lanl.gov
Los Alamos National Laboratory
Los Alamos, New Mexico, USA

Abstract

The readiness of new 400Gbps Ethernet hardware was evaluated for potential production use in High-Performance Computing (HPC) environments over a Local Area Network (LAN) and Wide Area Network (WAN). The approach explored a range of data movement strategies, including parallelized transfer tools, in which Warp-speed Data Transfer (WDT) yielded optimal results. Furthermore, communication protocols were tested such as Transmission Control Protocol (TCP) and Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE). Performance tests were conducted on bandwidth and latency to understand potential bottlenecks. Stress tests were run in Message Passage Interface (MPI) and other HPC-relevant environments. The research examined whether a 400Gbps pipeline can be saturated using current tools and methods, both locally and across geographically distributed environments. The findings provided recommendations for enhancing high-throughput data workflows in HPC settings.

Keywords

High Performance Computing (HPC), 400Gbps, Ethernet, Hardware, Data Transfer, Local Area Network (LAN), Wide Area Network (WAN)

ACM Reference Format:
Adelle Ferris, Evelyn Needham, Nikole Grandez, Jesse Martinez, and Doug Egan. 2025. Between the NIC and a Hard Place: Evaluating 400 Gb/s Ethernet for HPC Data Transfers. In *The International Conference for High Performance Computing, Networking, Storage, and Analysis.* ACM, New York, NY, USA, 3 pages.

1 Introduction

Innovations cannot occur without the framework to support them, High Performance Computing (HPC) seeks to rectify that. Through rapid processing of large amounts of data and complex calculations, breakthroughs are made daily in numerous fields. However, line rate has become a bottleneck, especially with the development of AI chips [1]. Thus, advances cannot occur without the support of proper bandwidth.

1.1 Overview

The analysis of 400Gbps hardware capabilities aims to assist HPC and trailblazers. Prioritizing performance, iperf2 tested bandwidth using multiple streams and threads. Rather than general system improvements, data transmission throughput was prioritized. Progression is typically not localized to one area, thus an examination was launched on a variety of data transfer tools over a Local Area Networks (WAN) and a Wide Area Network (WAN).

1.2 Testbed Specifications

1.2.1 *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

1.2.2 *Software.*

- OS: Rocky Linux 9.5 (kernel 5.14.0)

1.2.3 *Drivers and Libraries.*

- DOCA OFED: v25.04-0.6.1
- ConnectX-6 Firmware: v0.43.1014

1.2.4 *Hardware.*

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

1.3 Data Transfer Tools

Data transfer tools were scrutinized for compatibility with the hardware's optimal speed. Out of High Performance Networking Secure Shell (HPN-SSH), Pro FTP Daemon (ProFTPD) and Linux File Transfer Program (lftp), BaBar Copy Program (bbcp), and Warp-speed Data Transfer (WDT), findings revealed a clear leader. WDT performed 15 times faster, due to the way it conducts parallel transfers

*All authors contributed equally to this research.

LA-UR-25-28696, SC '25, St. Louis, MO

[3]. While other tools use multiple threads over a single socket pair, WDT opens multiple sockets alongside numerous threads. Running in a 'skip_writes' mode allowed the receiving node to discard incoming data, isolating network performance. Tcpdump, a command-line packet analyzer for network traffic validated this behavior.

1.4 Link Aggregation

To simulate a LAN and WAN, link aggregation, a hardware bonding technique, was deployed, as seen in Figure 1. Each compute node was equipped with two 200Gb Mellanox NICs which were bonded into a single logical interface. Initiated by enabling the bonding mode to IEEE 802.3ad with Link Aggregation Control Protocol (LACP). Occurring over the kernel, LACP conflicts with Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) which bypasses the kernel, forcing the use of Transmission Control Protocol (TCP). LACP was employed using breakout cables from 200Gb NICs to 400Gb optics, simulating 400Gb speeds, and thus, testing and tuning could begin.

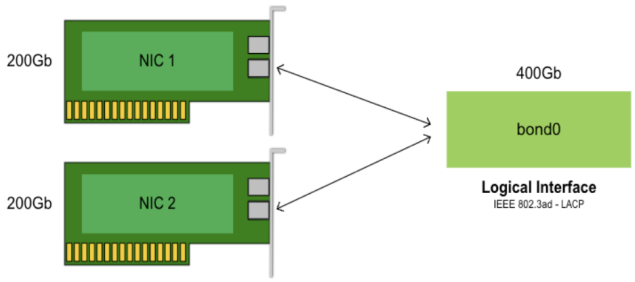


Figure 1: Link aggregation configuration

2 Local Area Network (LAN)

2.1 LAN Environment

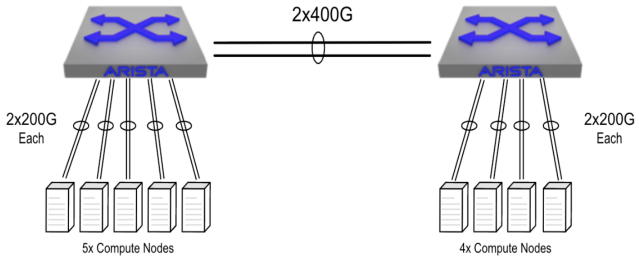


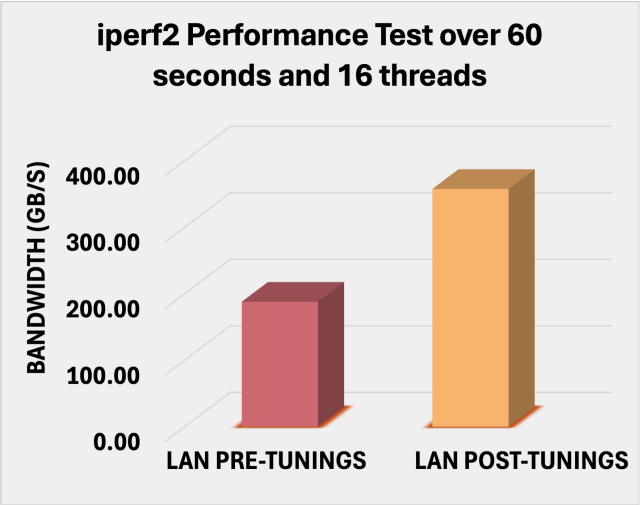
Figure 2

This network was configured as a Data Link Layer (Layer 2) environment, where all nodes were on the same subnet and could communicate directly. The topology accounts for future WAN testing between two locations.

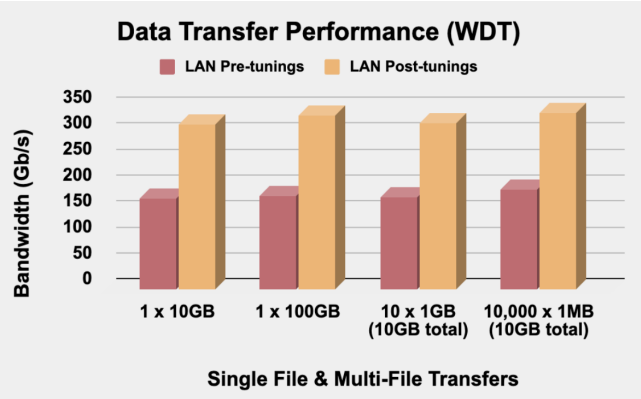
2.2 Network and System Optimization

On the LACP bond and Arista switches in Figure 2, load balancing was set to layer 3+4 for Internet Protocol (IP) and port hashing, to optimize WDT transfers on multiple ports. The Central Processing Unit (CPU) was modified to performance, and Interrupt Request (IRQs) were set to Non-Uniform Memory Access (NUMA). On the kernel, the congestion control algorithm was edited to Bottleneck, Bandwidth, and Round Trip Time (bbr). Additionally, socket buffer sizes were increased along with TCP send and receive buffer sizes. Mellanox Network Interface Cards (NICs) were altered with ethtool and mlnx_tune, which expanded the receive and transmit kernel ring buffers. Jumbo frames were enabled, and the NICs were flashed with optimized Mellanox drivers. Due to the tunings, Message Passage Interface (MPI) decreased twofold in execution times, despite higher data transmission rates.

2.3 Performance Results



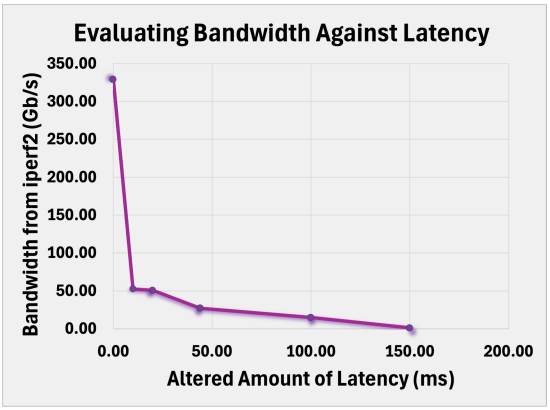
(a) Bandwidth performance (LAN)
After tunings, bandwidth increased 1.9 times relative to the hardware's default.



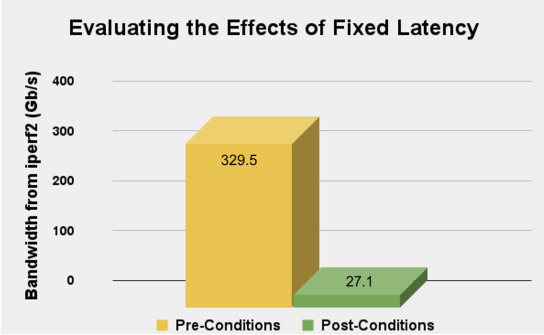
(b) Data transfer performance (LAN)
Tunings increased data transmission 1.8 to 2.1 times. In particular, directories with smaller files achieved better performance.

Figure 3

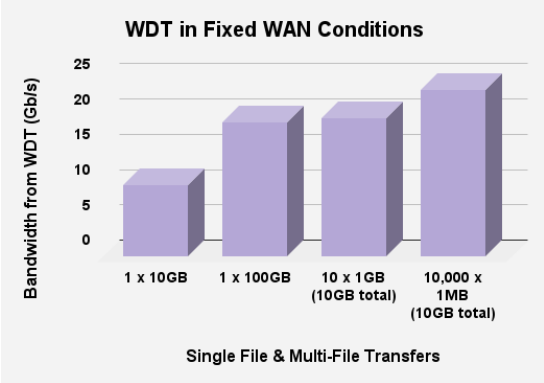
The observed increases in bandwidth and data transmission approached the hardware's expected capabilities, indicating readiness for WAN testing. WDT's performance revealed that collections of



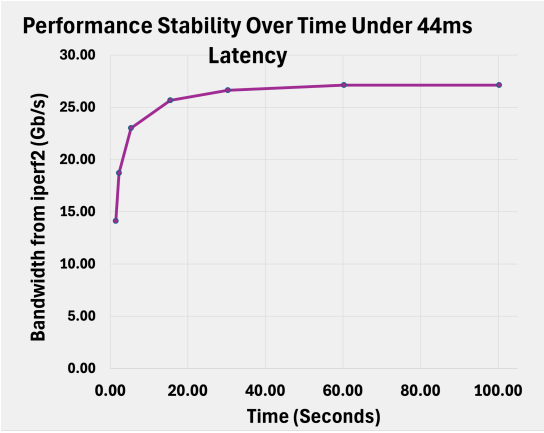
(a) Bandwidth against varying latency (WAN)



(b) Bandwidth against fixed latency (WAN)
Pre-Conditions: WAN configuration with LAN tunings
Post-Conditions: Latency (44ms) & Packet Reordering (0.02%) added



(c) Data transfer performance (WAN)



(d) Length of time bandwidth takes to stabilize (WAN)

Figure 5

smaller files transfer more efficiently, a finding highly beneficial for simultaneously managing numerous workloads.

3 Wide Area Network (WAN)

3.1 WAN Environment

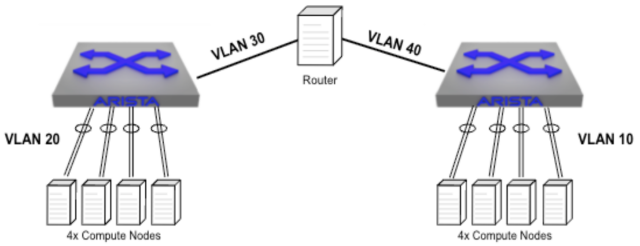


Figure 4
The fifth node became the router with Virtual LAN (VLAN) and static routing. On the router and VLAN bonds, Linux Traffic Controller (tc) and Network Emulator (netem) induced latency and packet reordering to bypass physical distance[2].

3.2 Performance Results

3.2.1 Chart Conditions, unless otherwise stated, 44 milliseconds (ms) latency Round Trip Time (RTT) and Packet Reordering 0.02%. With varying amounts of latency, Figure (5a) displays the bandwidth. A sharp drop is observed when 10ms is introduced, which can be explained by a 26 times increase in the LAN's RTT. Throughput, defined as the TCP window size divided by RTT, shows how the inherently small figures in a LAN make the TCP window a WAN bottleneck without scaling. As stated above, the 12-fold decrease in Figure (5b) when latency is incurred is due to the persisting LAN tunings in a WAN, as stated above. Comparable to LAN Figure (3b), the smaller files directories outperformed other transfers by a factor of two in Figure (5c). In Figure (5d), bandwidth is analyzed over time under 44ms of latency. Emphasizing the need for sustained transfers in a WAN to reach stabilized throughput, this experiment illustrates the conditions needed for optimal throughput.

4 Conclusion

Evaluating 400Gbps hardware capabilities and the ability to transfer data over a LAN and WAN dons insight into potential HPC improvements. Thus, allowing innovators to make their ideas a reality and drive meaningful change.

References

[1] Stuart Daudlin, Anthony Rizzo, Sunwoo Lee, Devesh Khilwani, Christine Ou, Songli Wang, Asher Novick, Vignesh Gopal, Michael Cullen, Robert Parsons, Alyosha Molnar, and Keren Bergman. 2023. 3D photonics for ultra-low energy, high bandwidth-density chip data links. (Oct. 2023). doi:10.48550/arXiv.2310.01615 arXiv:2310.01615 [physics].

[2] Stephen Hemminger. 2005. Network emulation with NetEm. *Linux Conference (2005)* (2005), 9. https://www.researchgate.net/publication/228619146_Network_emulation_with_NetEm

[3] Raimondas Sirvinskas, Preeti Bhat, Harvey Newman, Caătălin Iordache, and Justas Balcas. 2024. Scientific Community Transfer Protocols, Tools, and Their Performance Based on Network Capabilities. *EPJ Web of Conferences* 295 (2024), 04036. doi:10.1051/epjconf/202429504036