Evaluating TCP Protocol Performance on High-Speed Networks

Noah Jones; Jerrod Parten; Lucas Ritzdorf | Mentors: Jesse Martinez; Thomas Areba; Chase Harrison | Supercomputer Institute team EternalBlue; HPC-ENV

LOUISIANA TECH UNIVERSITY NM STATE MONTANA STATE UNIVERSITY

Introduction

BACKGROUND

- High-performance computing relies on high-bandwidth, low-latency networks to maximize usable computation time.
- These networks (including InfiniBand, Slingshot, Omni-Path, Aries, and RoCE) make a supercomputer more than simply a collection of nodes.
- In addition to their remote direct memory access (RDMA) protocol, these systems can also host traditional IP networks.
- Many systems within the HPC environment require IP communication.
- **IP over InfiniBand (IPoIB)** is a popular solution that removes the need for extra interfaces and infrastructure.
- No need for extra ethernet networks, cables, and interfaces.
- IPoIB tends to be slower and have more overhead than native communication due to the overhead of IP emulation.

GOALS

- Benchmark "out-of-the-box" IPoIB bandwidth.
- Perform system and firmware tuning to approach vendor throughput estimates.
- Ensure that IPv6 does not degrade network performance.
- Test throughput when routing traffic between an IPoIB network and a traditional Ethernet network.

Specifications

HARDWARE

- **CPU:** AMD EPYC 7502 32-core processor
- **RAM:** 128 GB per node
- **NIC:** Intel I350 Gigabit Ethernet
- **InfiniBand HCAs:** Mellanox ConnectX series
    - ConnectX-5 (master node)
    - ConnectX-6 (compute nodes)
- **Switch:** Mellanox SB7700 Series EDR InfiniBand switch
- **Cables:** Mellanox 4X EDR (100 Gbps) InfiniBand cables

SOFTWARE

- **OS:** Rocky Linux 8.6 (kernel 4.18.0)
- **InfiniBand Firmware**
    - HCAs: 20.33.1048
    - Switch: 3.9.2400
- **Benchmarking software**
    - `iperf` (version 2)
    - Intel MPI Benchmarks (IMB), OpenMPI 4.0.5
    - GPCNeT (network congestion test)

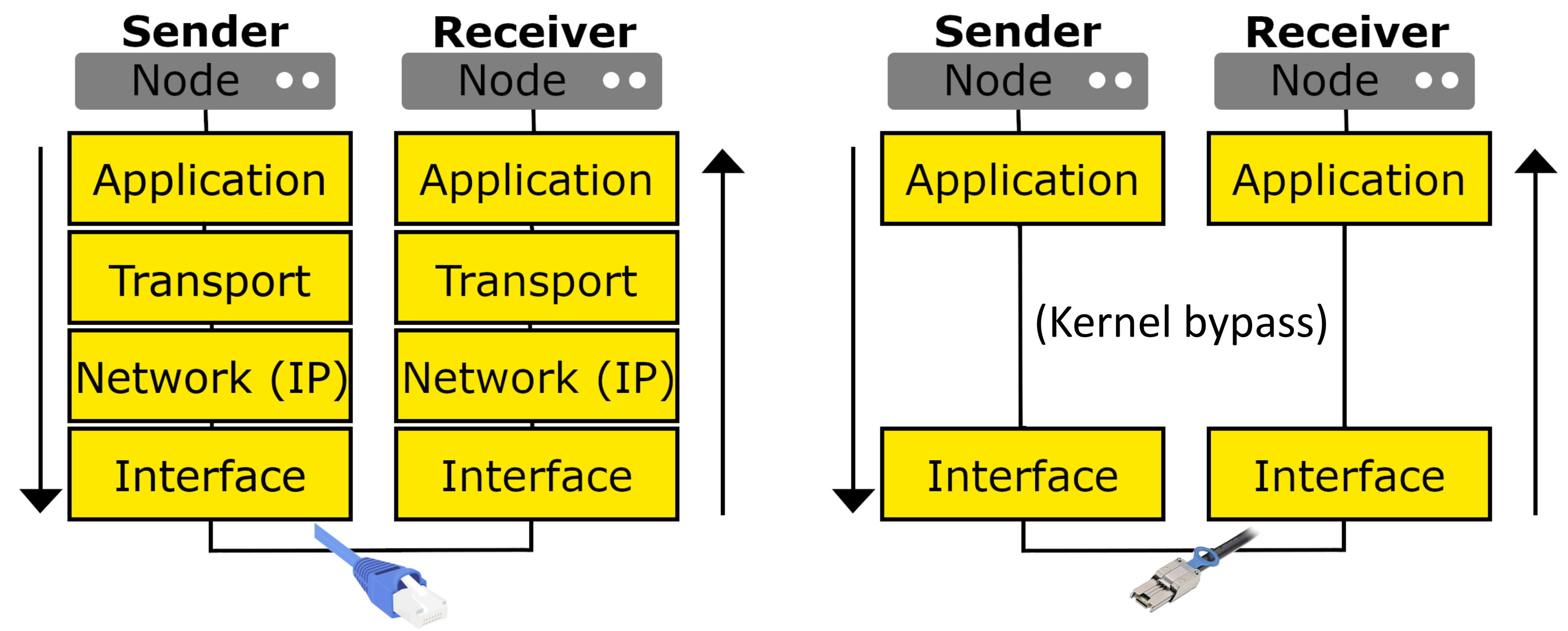


Figure 1: Comparison between a traditional TCP/IP stack (left) and a low-latency RDMA stack (right)

Methods and Results

TUNING AND CONFIGURATION

- MLNX_OFED proprietary driver stack
- Kernel parameter tuning via `sysctl`
    - Increased socket buffer size
    - Increased TCP send and receive buffer sizes
    - Support for additional simultaneous connections
    - Further advanced parameters
- InfiniBand NIC tuning via `ethtool`
    - Expanded receive and transmit kernel ring buffers

TEST RESULTS

Figure 2: Bandwidth vs. Connection Modes (via `iperf`, 32 threads)

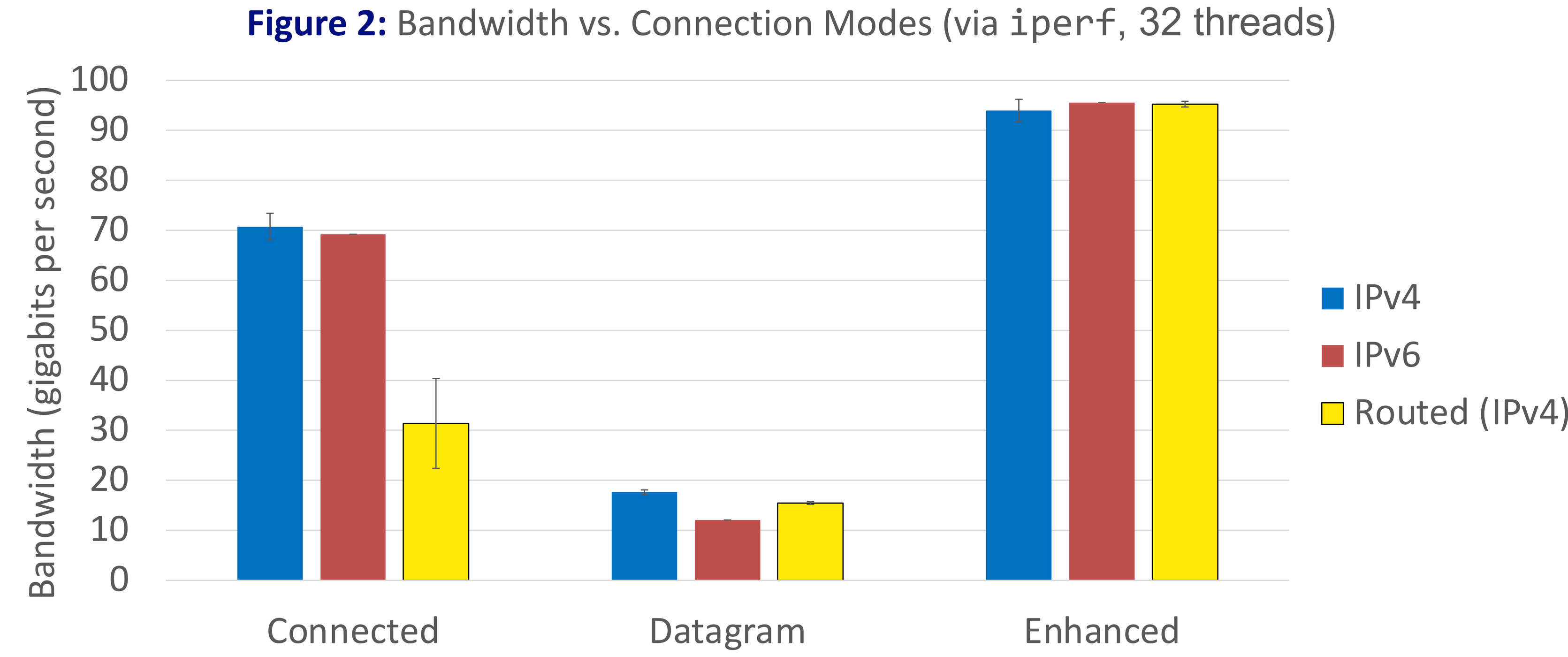
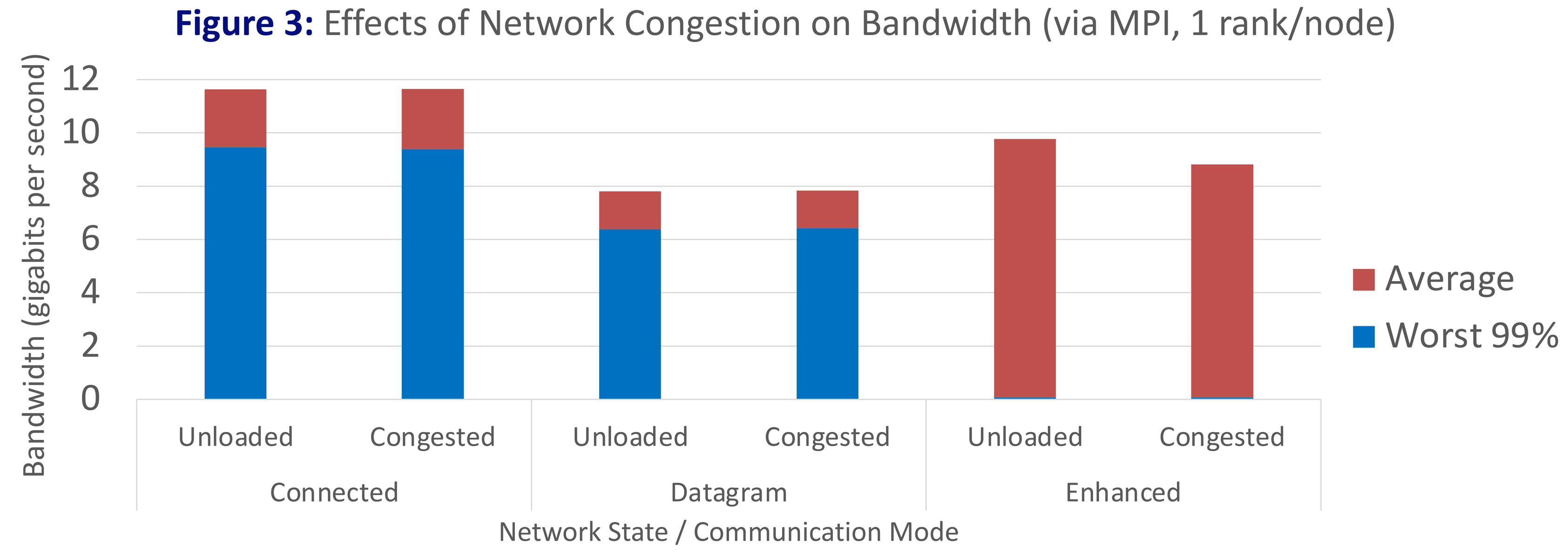


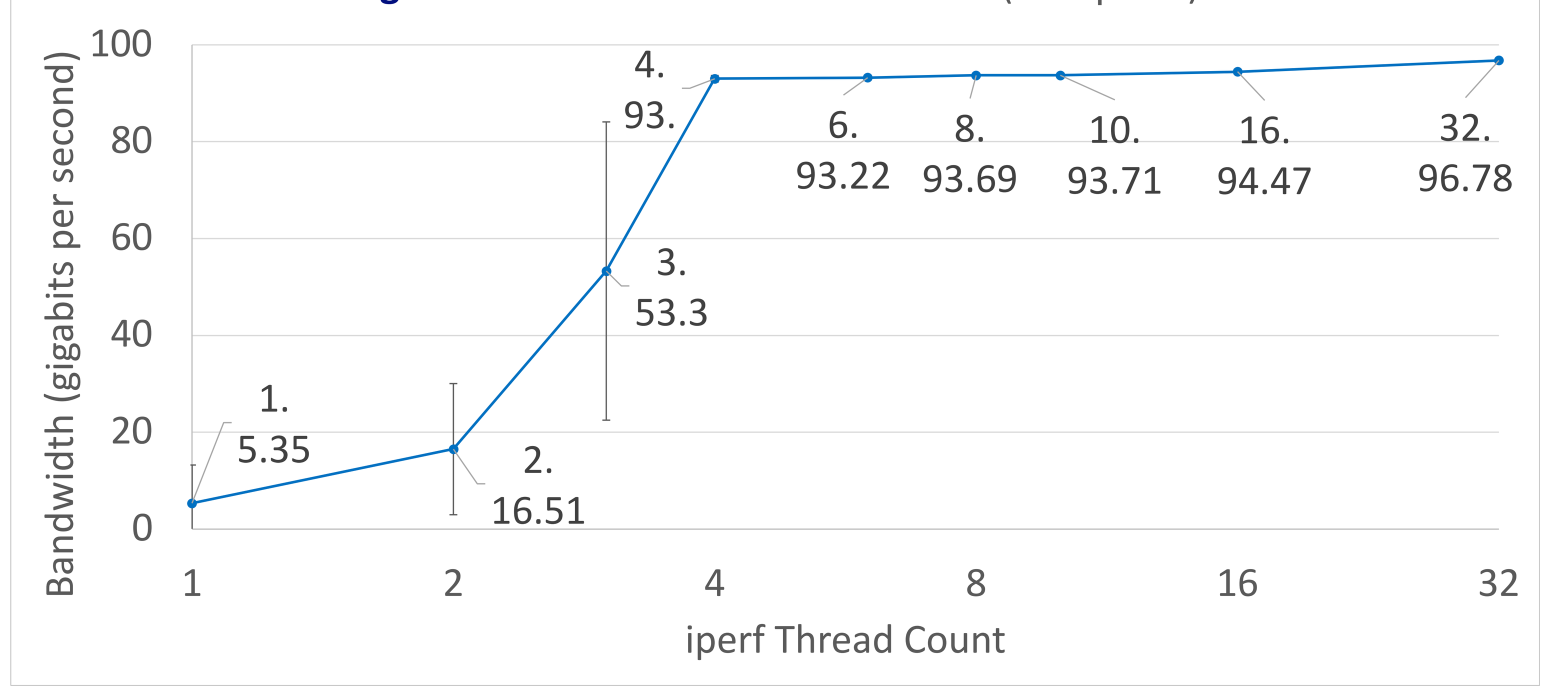
Figure 3: Effects of Network Congestion on Bandwidth (via MPI, 1 rank/node)



OBSTACLES

- MLNX_OFED installation, especially in an alternate root
- Automation with Ansible: `--skip-broken` bug
- MPI transport parameters: TCP transport layer not available by default
- Lack of multithreading support in `iperf3`
- Outdated Mellanox switch firmware: HCA compatibility issues
- TCP Slow Start: built into TCP stack, difficult to disable. May skew results slightly, but reflects true network performance.

Figure 4: Bandwidth vs. Thread Count (via `iperf`)



Conclusions

SUMMARY

- After applying optimizations, manufacturer throughput expectations were consistently met.
- Gathered additional insight regarding InfiniBand communication modes and their performance for various applications.
- Determined that higher thread counts correlate with increased performance, up to link capacity. On the systems tested, four threads were sufficient to saturate an EDR InfiniBand link.
- Observed little significant effect on throughput due to IPv6 addressing.
- Observed identical performance with all combinations of IPoIB-compatible ConnectX cards (CX-4 through CX-6).

FUTURE RESEARCH

- Investigate low 99% performance in MPI-based congestion tests
- Evaluate MPI benchmark performance with IPv6 addressing
- Investigate potential performance improvements on dual-socket or high-thread-count systems
- Provide further insight into what causes lower speeds for different InfiniBand connection modes and situations

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-22-28115 HIGH PERFORMANCE COMPUTING