Performance Analysis of Non-Volatile Memory Express Over Fabrics (NVMeoF) Using Infiniband and Ethernet

Christa Collins (Liberty University), Joseph Sarrao (UC Berkeley), Zachary Wadhams (Montana State)
Mentors: David Bonnie, Jarrett Crews, Dominic Manno, Alex Parga

Abstract

With the increase in complexity of scientific computer codes, the way in which data is transferred and stored on the high-performance computers (HPC) needs to become more dynamic. Non-volatile memory express over fabrics (NVMeoF) allows for this flexibility. While past storage structures have included static storage over the network for each server node, NVMeoF allows access to all storage media over various network types such as Infiniband, RDMA (Remote Direct Memory Access) over Converged Ethernet (RoCE) and Transmission Control Protocol (TCP). Thus, providing us with the capability of dynamically allocating storage pools that are specially designed for running jobs on a subset of worker nodes. This project's primary objective is to analyze the performance of NVMeoF over various high-speed networks such as Infiniband, RoCE and TCP. We analyzed the data throughput and input/output operations per second (IOPS) of NVMeoF with each network type using the IO benchmarking tool FIO. We compared these results with the same test to local NVMe storage as a baseline.

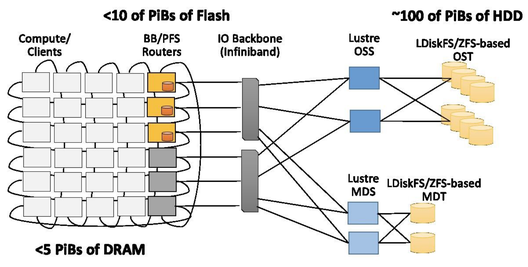

Figure 1. Current static storage configuration. Featuring a set number of servers, with each server pair connected to their own storage media (hard drives, SSDs, etc.) over a protocol. These protocols are grouped into a single pool of storage through a file system (Lustre) which is accessed by the client nodes.

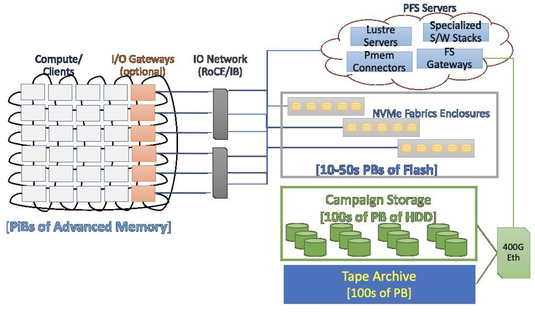

Figure 2. Future dynamic storage configuration. With every server within the cluster having access to all storage media over a high speed network such as Infiniband or RoCE (RDMA over Converged Ethernet). Providing us with the ability to create storage pools designed for specific jobs across the client nodes.

Results

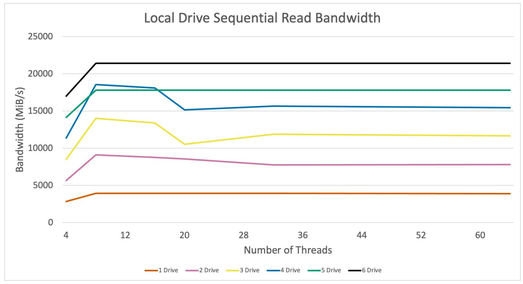

Figure 3. Baseline testing of the sequential reading of data over the base drive local NVMe devices.

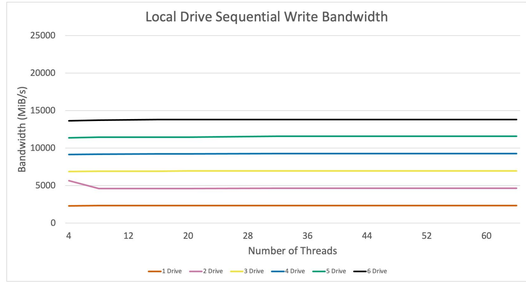

Figure 4. Baseline testing of the sequential writing of data over the base drive local NVMe devices.

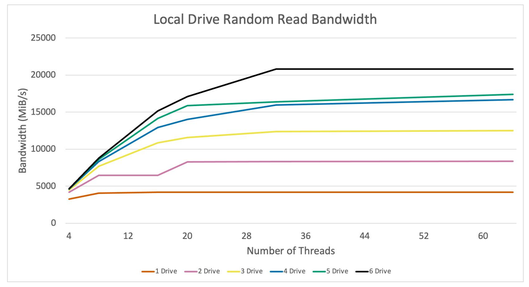

Figure 5. Baseline testing of the random reading of data over the base drive local NVMe devices.

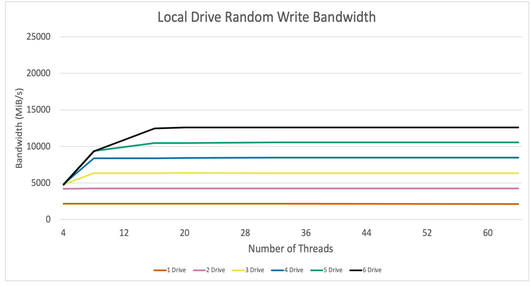

Figure 6. Baseline testing of the random writing of data over the base drive local NVMe devices.

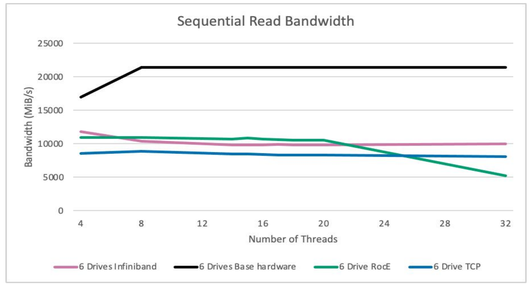

Figure 7. Comparison graph of the sequential read bandwidth both locally and over Infiniband, RoCE, and Ethernet TCP.

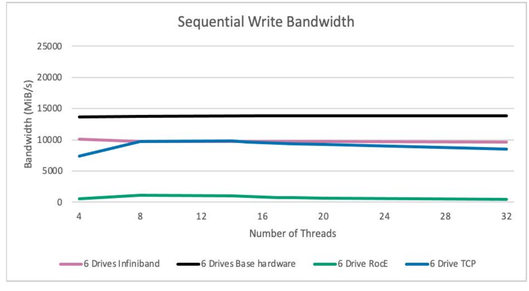

Figure 8. Comparison graph of the sequential write bandwidth both locally and over Infiniband, RoCE, and Ethernet TCP.

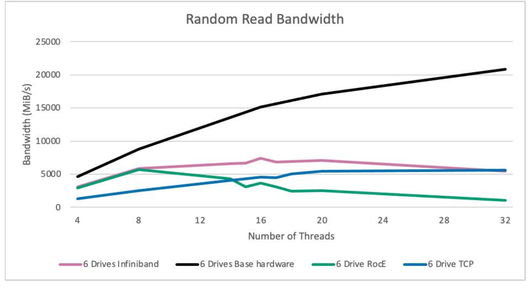

Figure 9. Comparison graph of the random read bandwidth both locally and over Infiniband, RoCE, and Ethernet TCP.

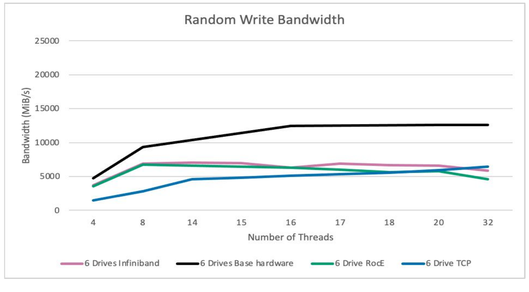

Figure 10. Comparison graph of the random write bandwidth both locally and over Infiniband, RoCE, and Ethernet TCP.

Testing Overview

EDR Infiniband (100 Gb/s) & Ethernet (100 GbE)

1 Target Node with
- 1 AMD EPYC 7502 32-Core Processor
- 131 GB RAM
- 6 Gen 4 NVMe drives

1 Client Node
- 1 AMD EPYC 7502 32-Core Processor
- 131 GB RAM

Step 1: Baseline Testing
- Network Validation (iPerf)
- Local Block Device (fio)
  - blocksize of 4K for rand, 128K seq
  - file size of 16G
  - runtime of 100 seconds per test
  - io depth of 32
  - consistent across ALL tests, filesystem, fabric or not
- Local Filesystem (fio to ext4)

Step 2: NVMe Over Fabrics
- RDMA Infiniband, RoCE, and TCP Ethernet
  - Remote Block Device
  - Remote Filesystem

Conclusion

- Figures 3 through 6 show linear scaling of performance up to 6 NVMe drives
- Fio failed to write data to ext4 filesystem over ROCE due to known kernel bug
- Writing to more than four NVMe drives over ROCE was abnormally slow as shown in figure 8
- Infiniband was generally the fastest and most reliable fabric
- TCP was on average 17.82% slower than Infiniband

Future Work

- Work around or fix ROCE and ext4 filesytem bug
- Locate and fix the issue with greater than 4 drives with NVMEoF and ROCE
- Bandwidth testing from multiple nodes with NVMeoF for the parallelization of the storage
- Test performance with RAID on backend NVMe storage
- Test routing of NVMeoF through a router node between various networking types

LA-UR-22-28158