



Los Alamos
NATIONAL LABORATORY

EST. 1943

Evaluating Hardware Compression Offload in a Lustre File System

Mariana Hernandez

Mentors: Dominic Manno, Jarrett Crews, and Brian Atkinson

August 13, 2020



Introduction

- Solid-state storage has inspired more efficient bandwidth tiers
- SSDs are now affordable
 - an increase in speed
 - reduction in capacity per device
- Compression is vital
 - This project focuses on compression out of several storage operation offloads



Introduction

- Current compression algorithms either
 - Reduce throughput, efficient compression
 - Good throughput, inefficient compression
- Offloaded compression to specialty hardware
 - High compression ratios and high throughput
 - Eideticom's NoLoad FPGAs
 - Programmable hardware for complex projects
 - Minimal impact on performance
 - High compression efficiency



Objective

- Determine feasibility of implementing NoLoad FPGAs in a production environment
 - ZFS
 - Lustre
- Show that compression done through the NoLoad FPGA could achieve GZIP levels of compression at or above LZ4 throughput



Test Setup

- ZFS implemented these compression algorithms
 - No compression
 - LZ4
 - GZIP
 - GZIP-NoLoad
 - Provided by custom Eideticom ZFS
- ZFS backed Lustre file system
 - 9 NVMe SSDs
 - 8+1 RaidZ1
 - Single combined Lustre OSS/MDS node



Test Setup

- CentOS 7.8
- ZFS 8.3
- Lustre 2.13.54
- 5.0.3 Kernel

When testing NoLoad kernel

- Custom kernel required patches for its FPGAs
 - 4 NoLoad CSP U.2s
 - Patches for the 5.3.1. kernel were provided by Eideticom
 - Unable to use 5.3.1. kernel due to incompatibility with ZFS and Lustre
 - Eideticom provided custom ZFS version with compression offloading handling



Method

Storage Throughput Benchmarks

- XDD ~ written data 30% compressible
 - Raw NVMe
 - ZFS
 - Lustre Client

Network Benchmarks

- ib_send_bw
 - Raw infiniband
- Inetselftest
 - Lustre network



Results – Baseline Benchmarks

ib_send_bw – 21363 MiB/s

Inetselftest - 14314 MiB/s

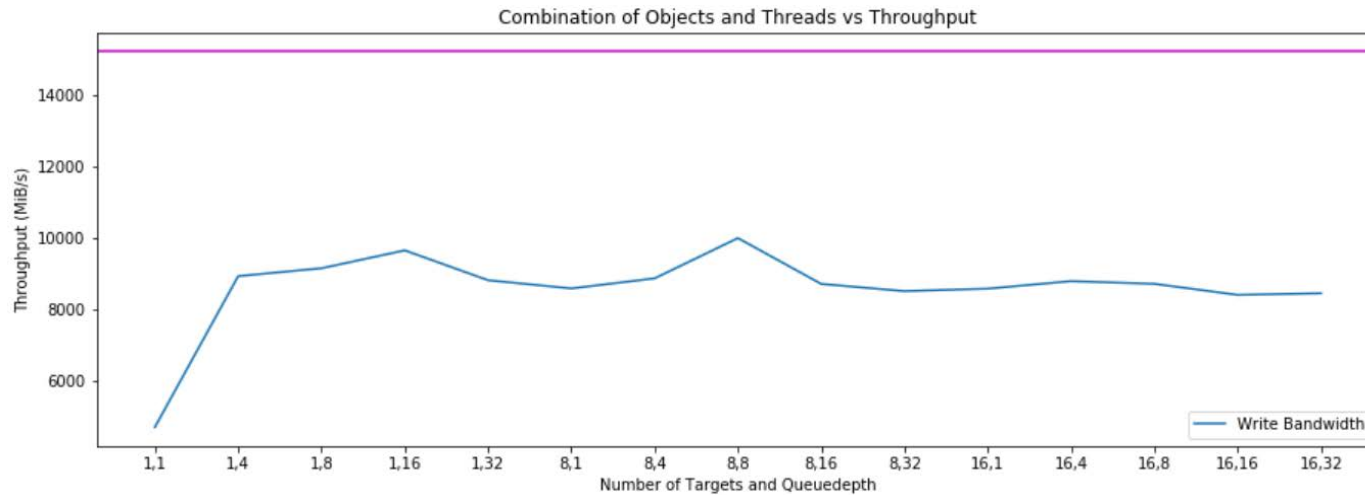
Single NVMe - 1952.021 MiB/s

8+1 Raidz1 – 15200 MiB/s

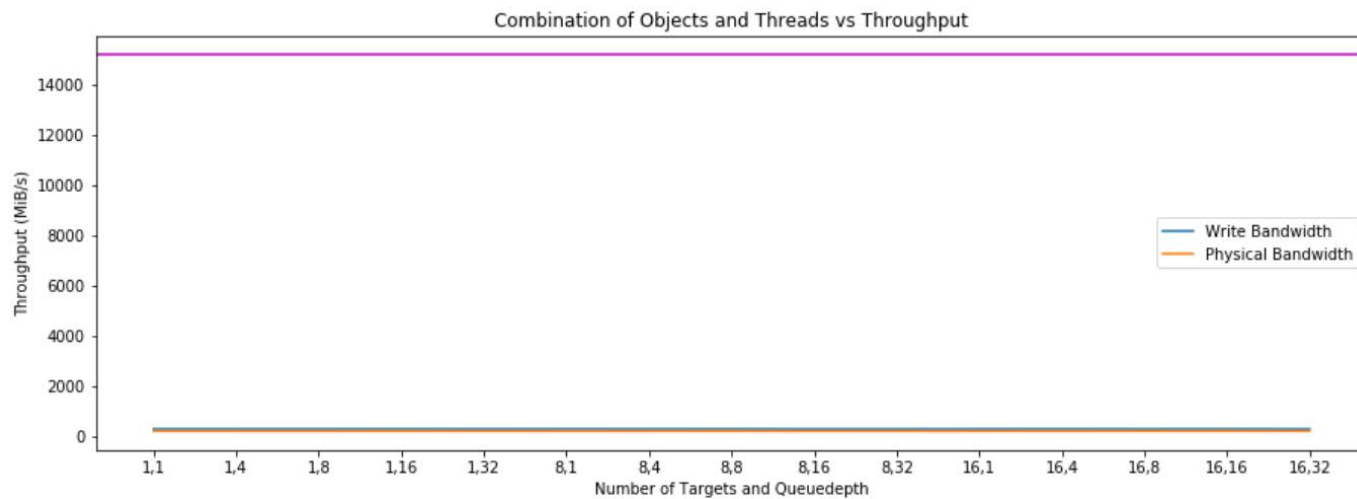


Results – ZFS Performance

ZFS



No
compression

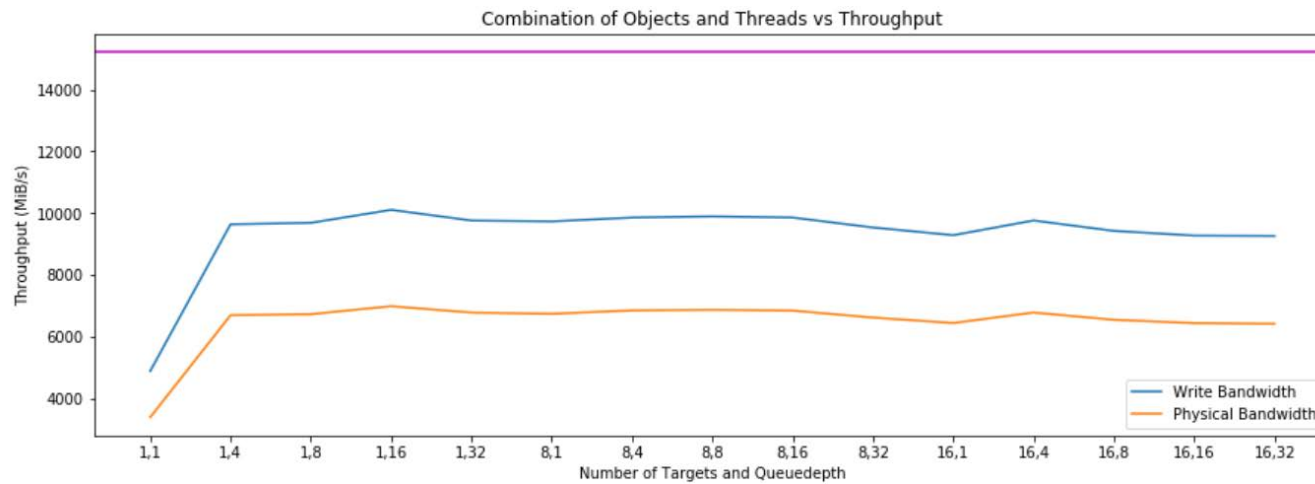


GZIP
compression

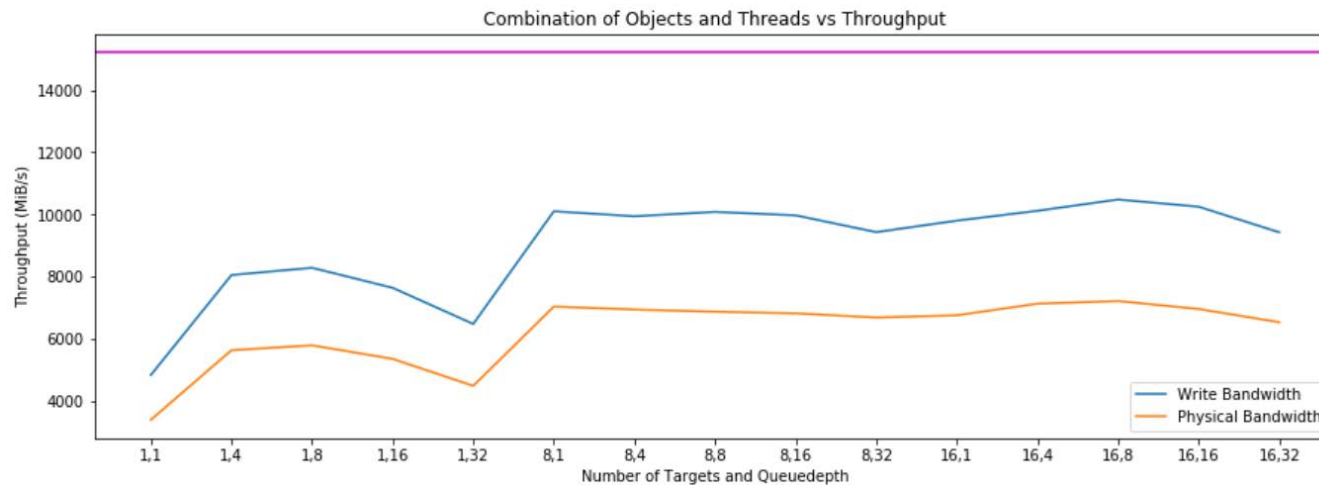


Results – ZFS Performance

ZFS



GZIP-NoLoad
compression

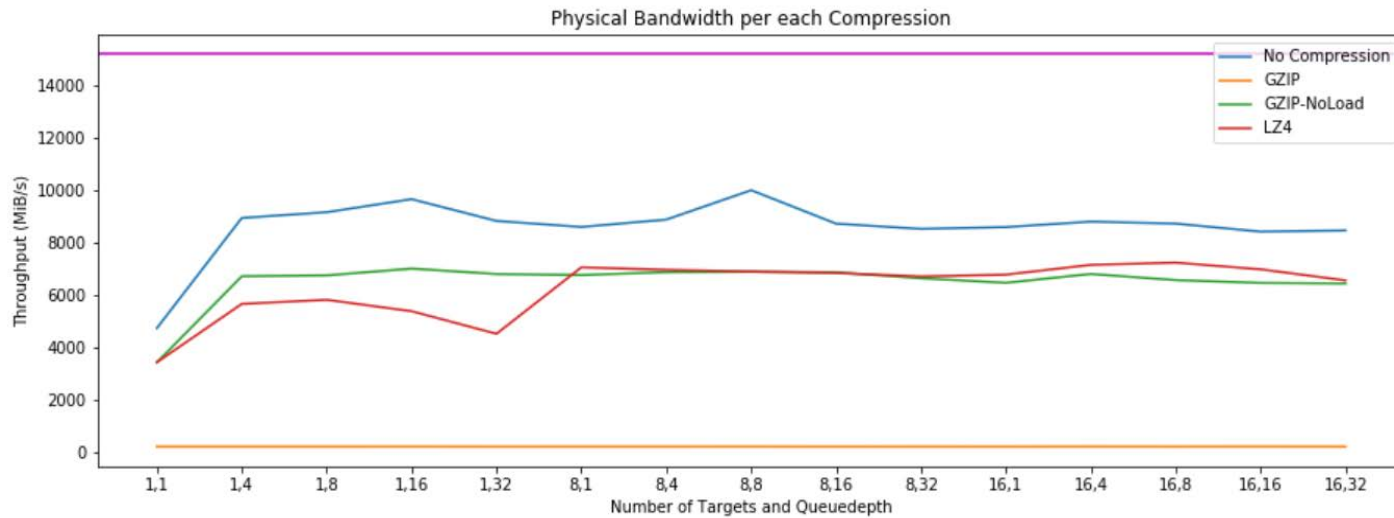
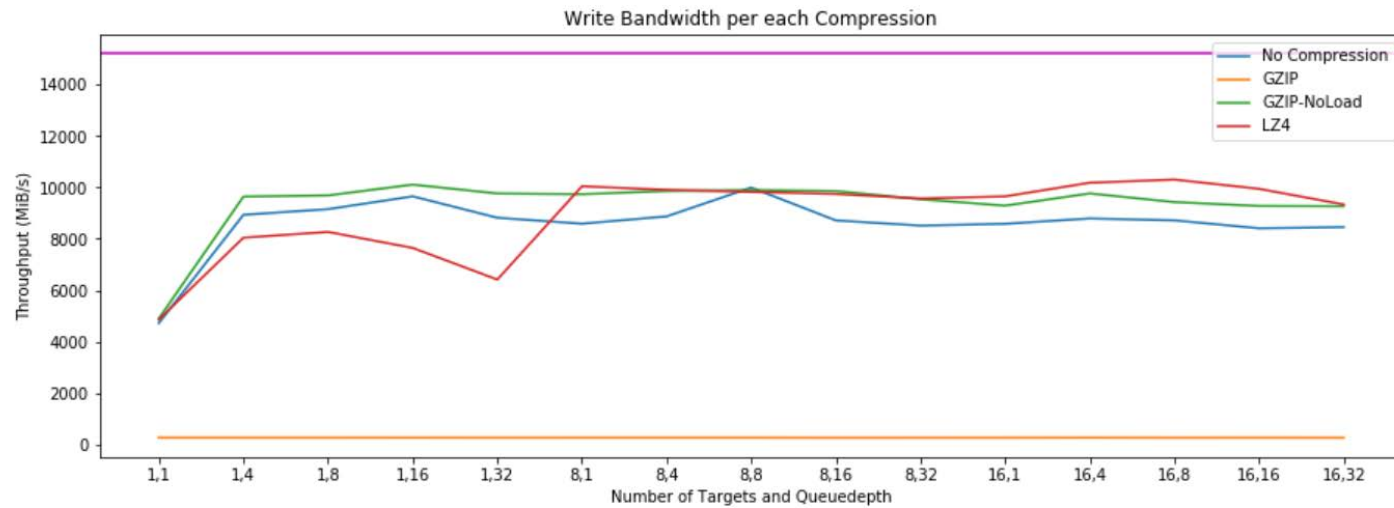


LZ4
compression



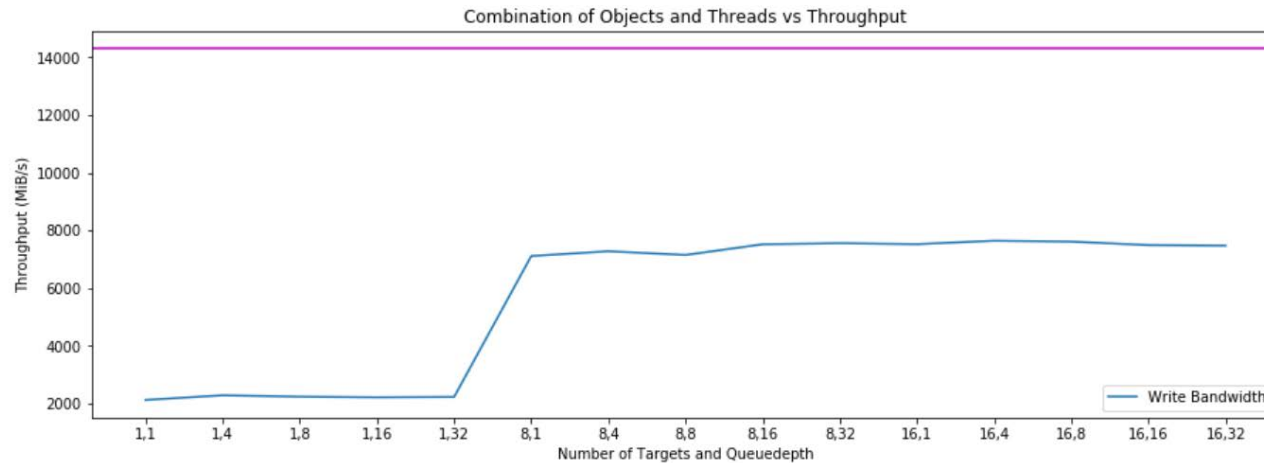
Comparing Compression on the Server

ZFS

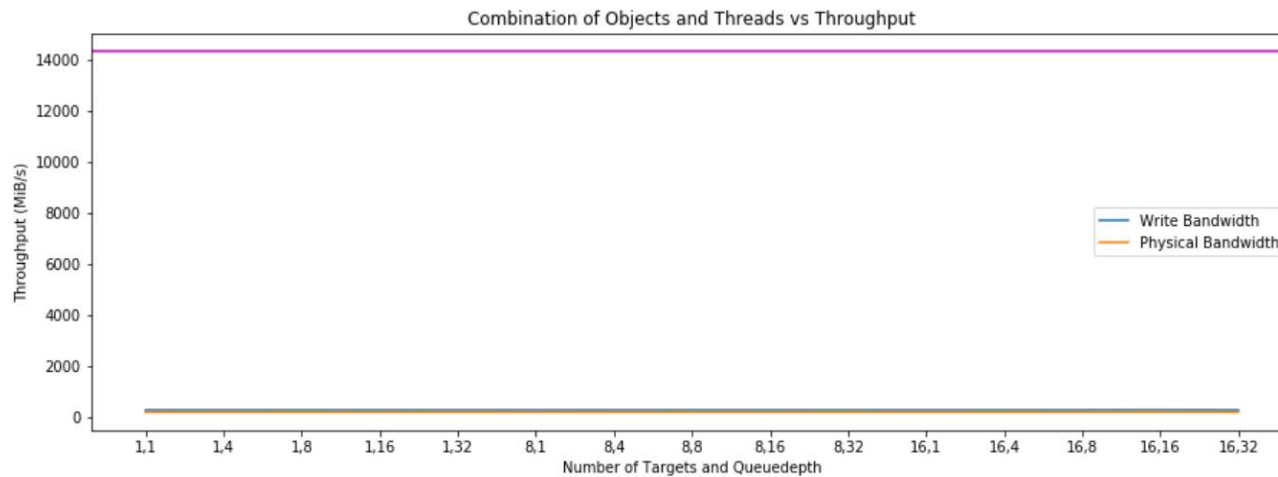


Results – Lustre Client Performance

Lustre Client



No
compression

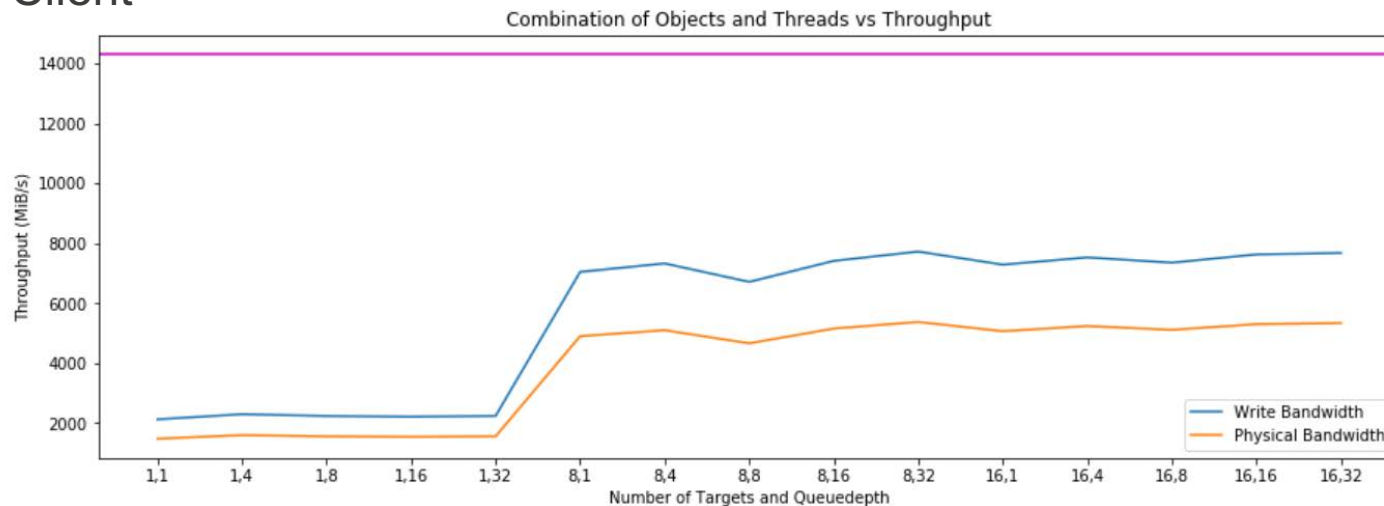


GZIP
compression

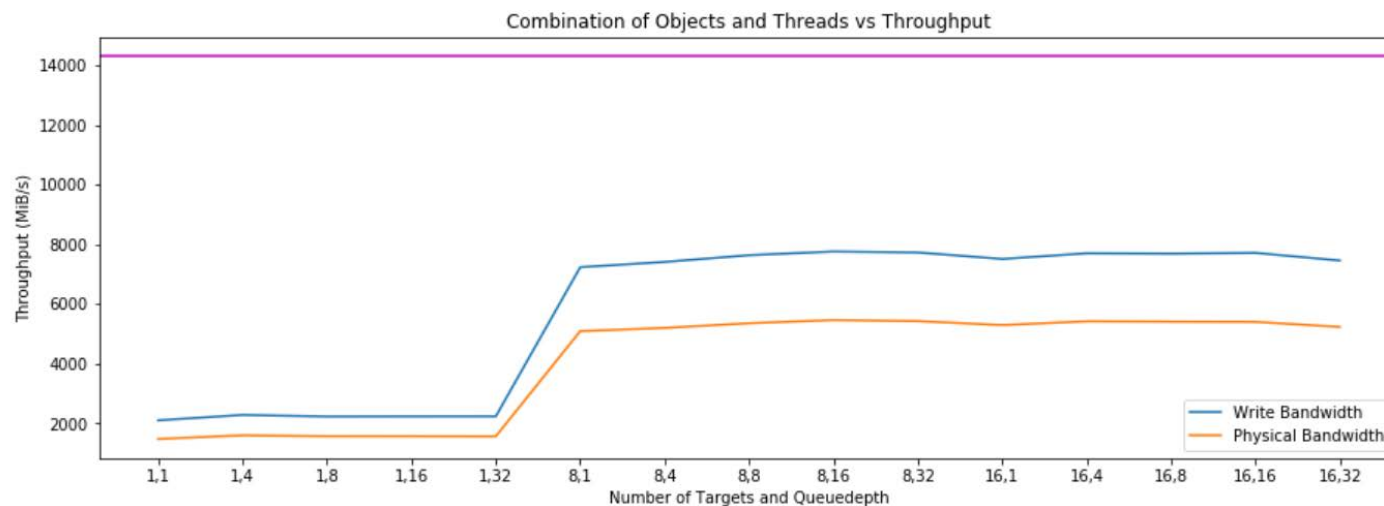


Results – Lustre Client Performance

Lustre Client



GZIP-NoLoad
compression

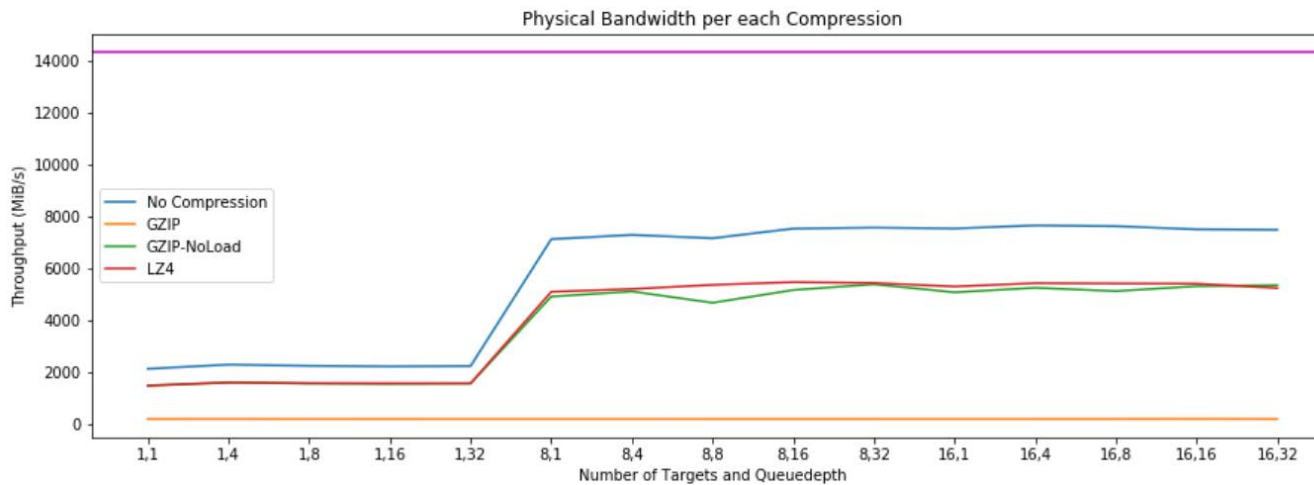
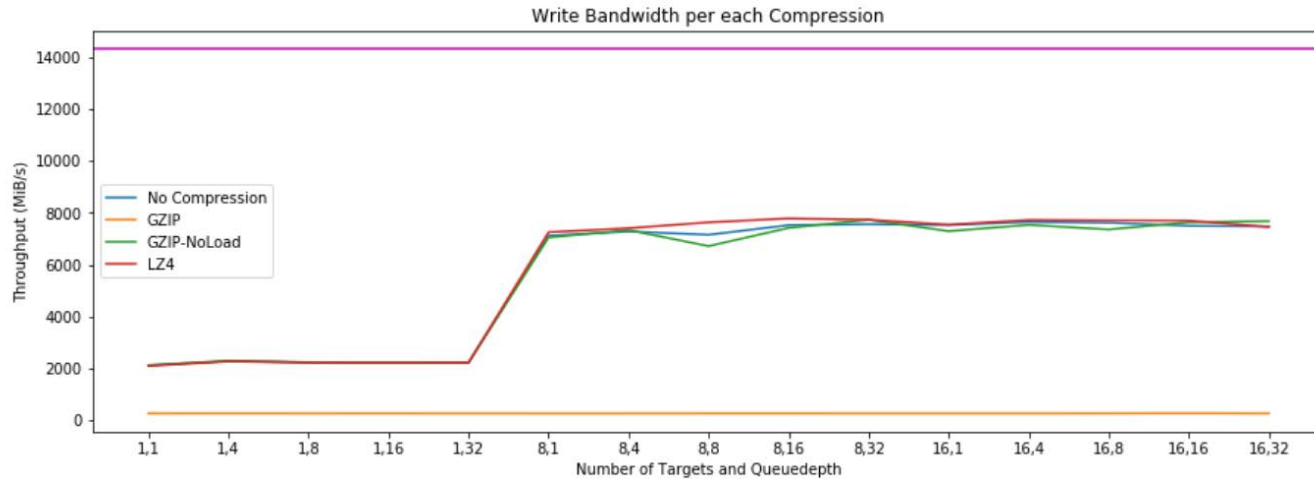


LZ4
compression



Comparing Compression on the Client

Lustre Client



Conclusion

- We have confirmed the viability of using Eideticom NoLoad devices to offload compression in ZFS and a ZFS-based Lustre filesystem.
 - A result from this is a high maintenance cost that should be further evaluated
- Lustre seems agnostic to the fact that we added NoLoads to offload compression
 - We plan to evaluate multiple Lustre clients to confirm we meet no compression performance provided by the server
- Multiple efforts to continue removing ZFS bottlenecks with NVMe devices.



Questions?



Over 70 years at the forefront of supercomputing



Over 70 years at the forefront of supercomputing