EIDETICOM

Accelerate Everything

Computational Storage Solutions Over Fabrics for ZFS

Kelly Ursenbach

Massive Storage Systems and Technology Conference – 23 May 2023

Acknowledgements

This work is all a part of a successful partnership between:

- Aeon Computing
- Eideticom
- Nvidia
- Los Alamos National Laboratory (LANL)
- SK hynix

Much of the content provided in this talk can be attributed to:

- Brad Settlemyer – Nvidia
- Roger Bertschmann, Sean Gibb, Andrew Maier, Martin Oliveira – Eideticom
- Jeff Johnson, Doug Johnson - Aeon Computing
- Dominic Manno, Gary Grider, Jason Lee, Brian Atkinson - LANL

Overview

- Motivation
- A flexible solution
  - Accelerated Box of Flash (ABOF)
- Performance Analysis
- Outlook

Motivation

Get maximum milage from flash storage

- Capacity
- Bandwidth

Memory bandwidth limitations observed

Get the benefits of compression without losing performance

Why ZFS?

Feature Rich

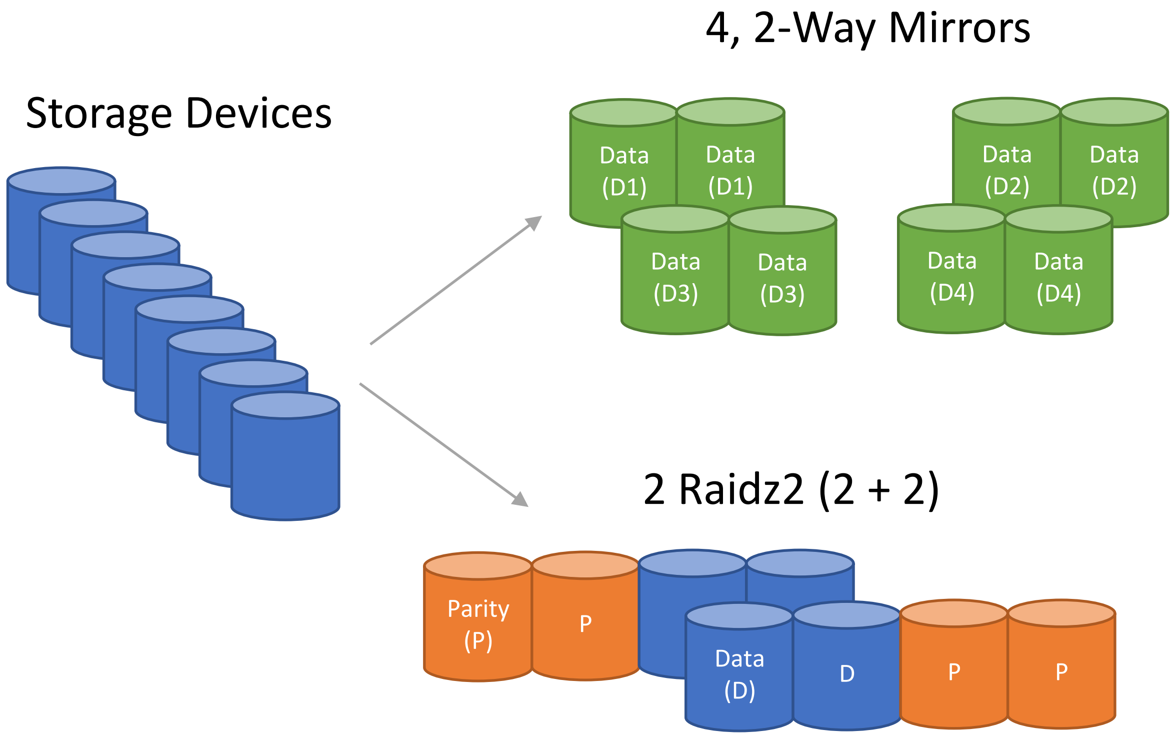
- Compression
- Deduplication
- Encryption

High Integrity

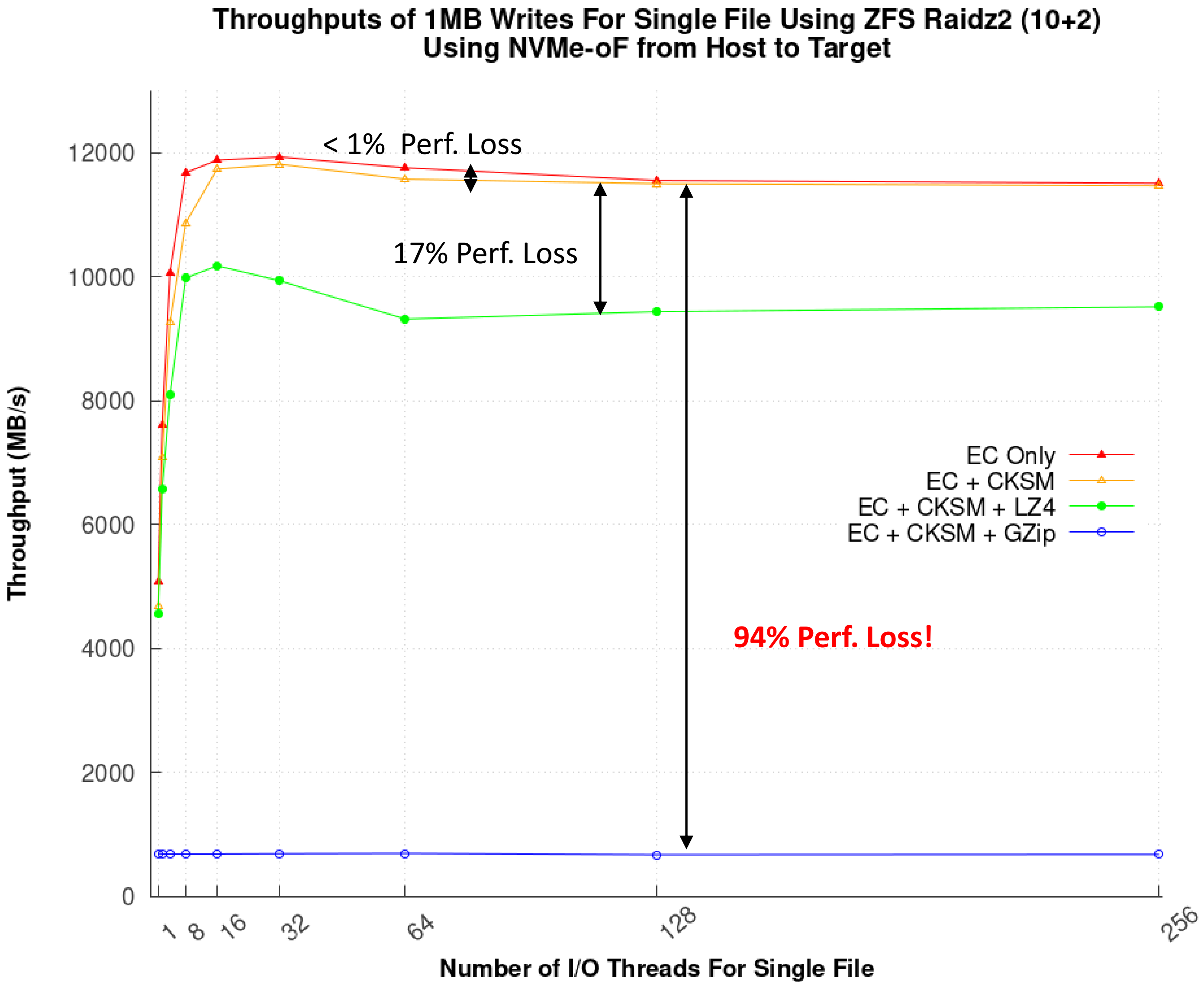
- Erasure Coding (EC)
- Mirrors
- Snapshots
- Checksums

Lustre over ZFS

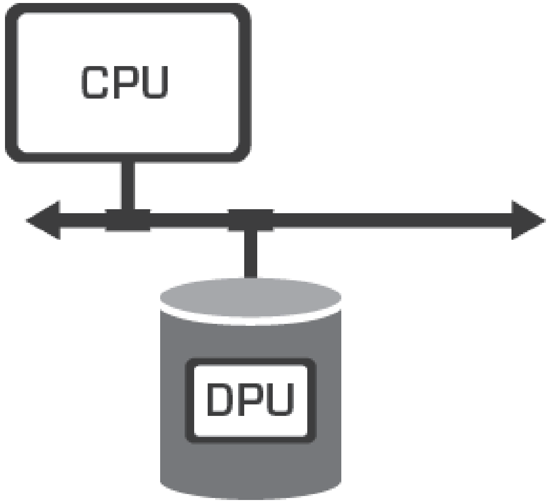
OpenZFS



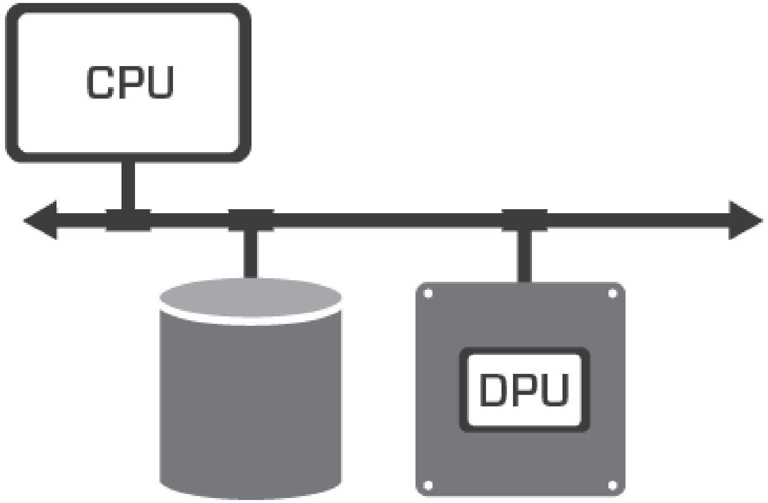
GZIP is Expensive on CPU



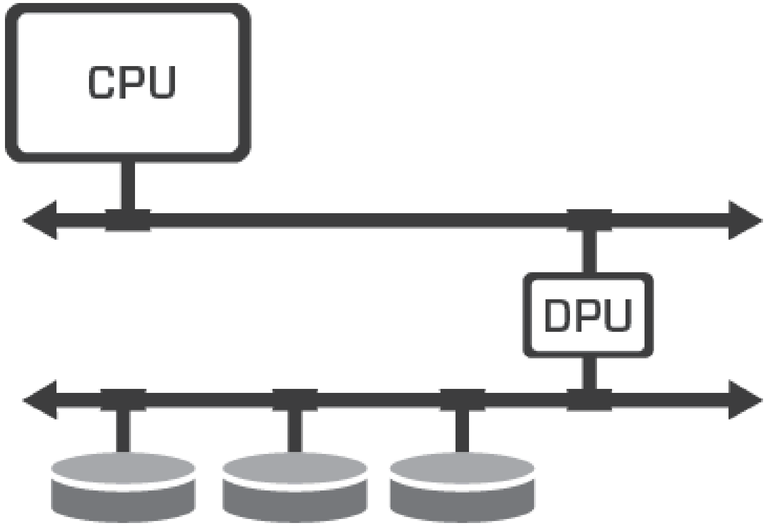
Computational Storage can help



Computational Storage Device (CSD)



Computational Storage Processor (CSP)



Computational Storage Array (CSA)

Benefits of Compression Offload

Improved storage bandwidth

- Dedicated hardware performs near PCIe line rate

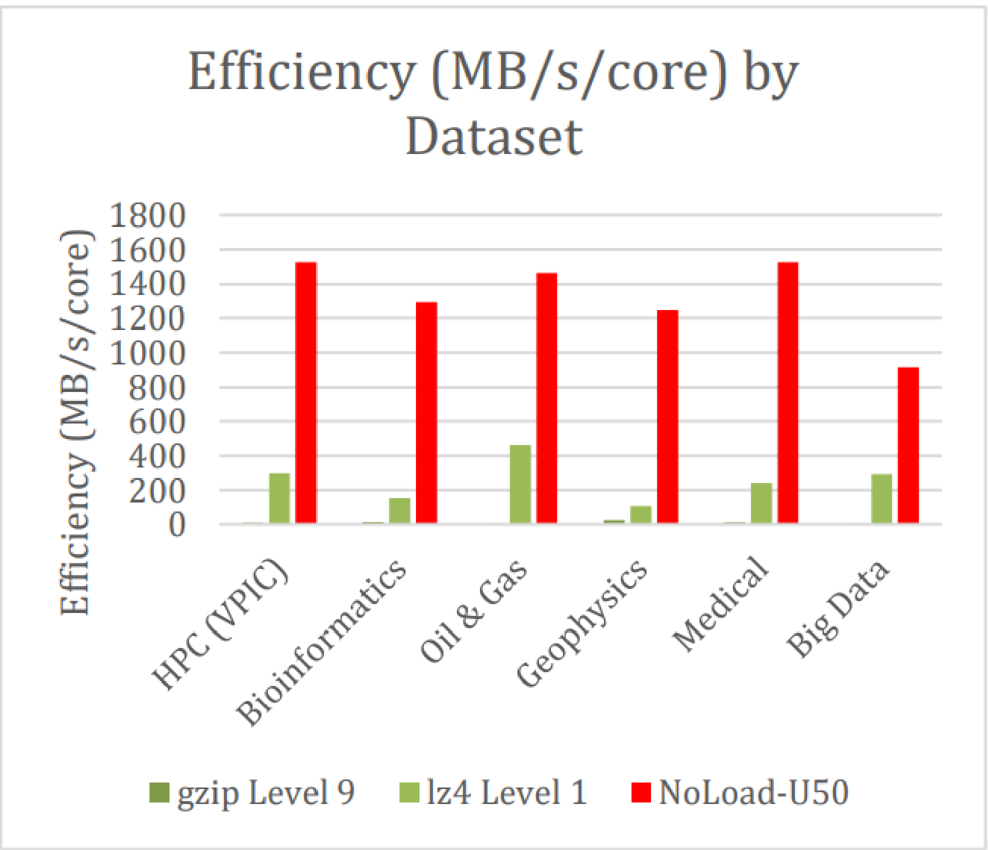
Reduced Storage Cost

- Lower power
- Increase effective storage via compression

Scalability

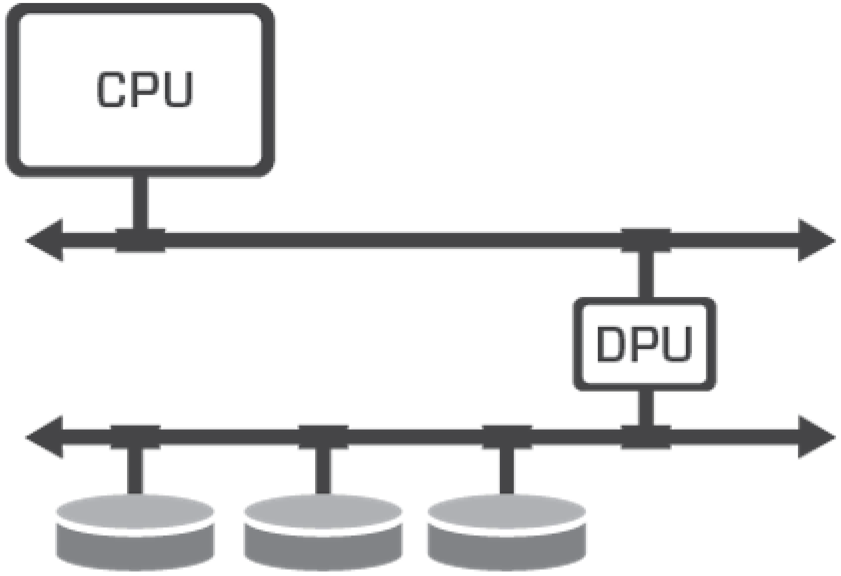
- Disaggregating compute and storage into independently scalable resources

Save CPU cycles on compute nodes



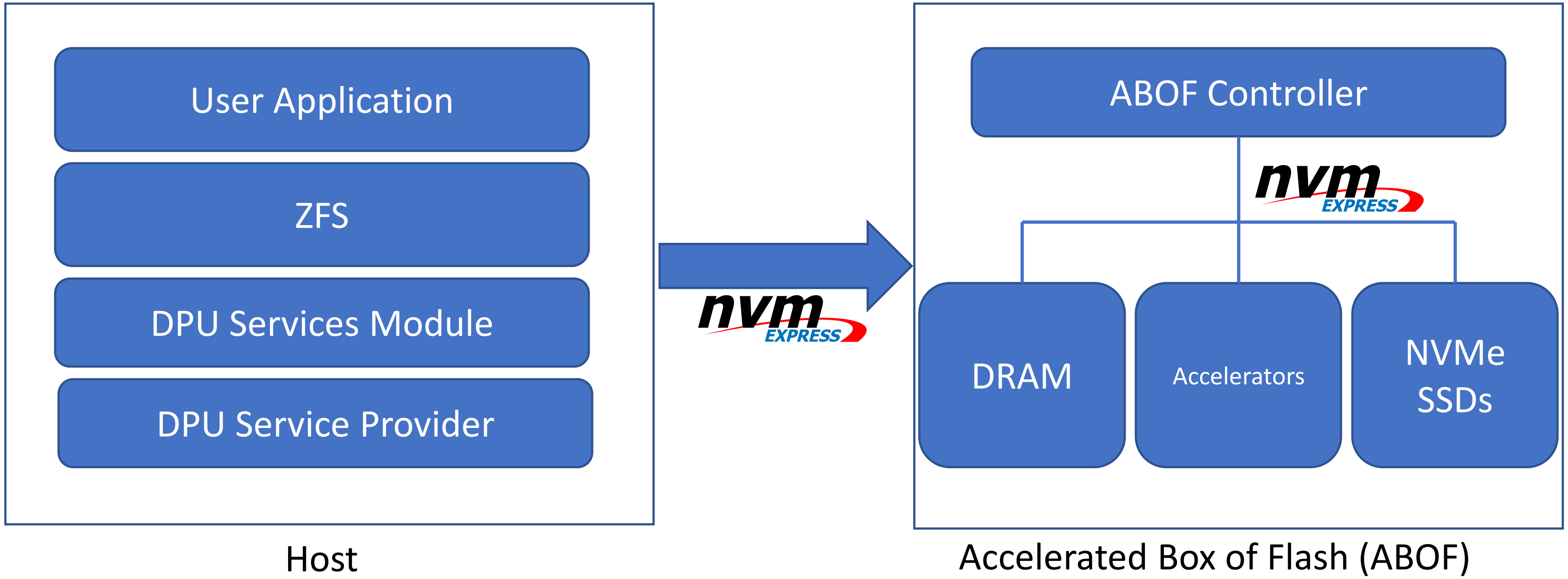
Encapsulate Storage and Storage Specific Compute (CSA)

- Contain storage software stack
- Eliminate stranded resources
- Free memory bandwidth on compute nodes

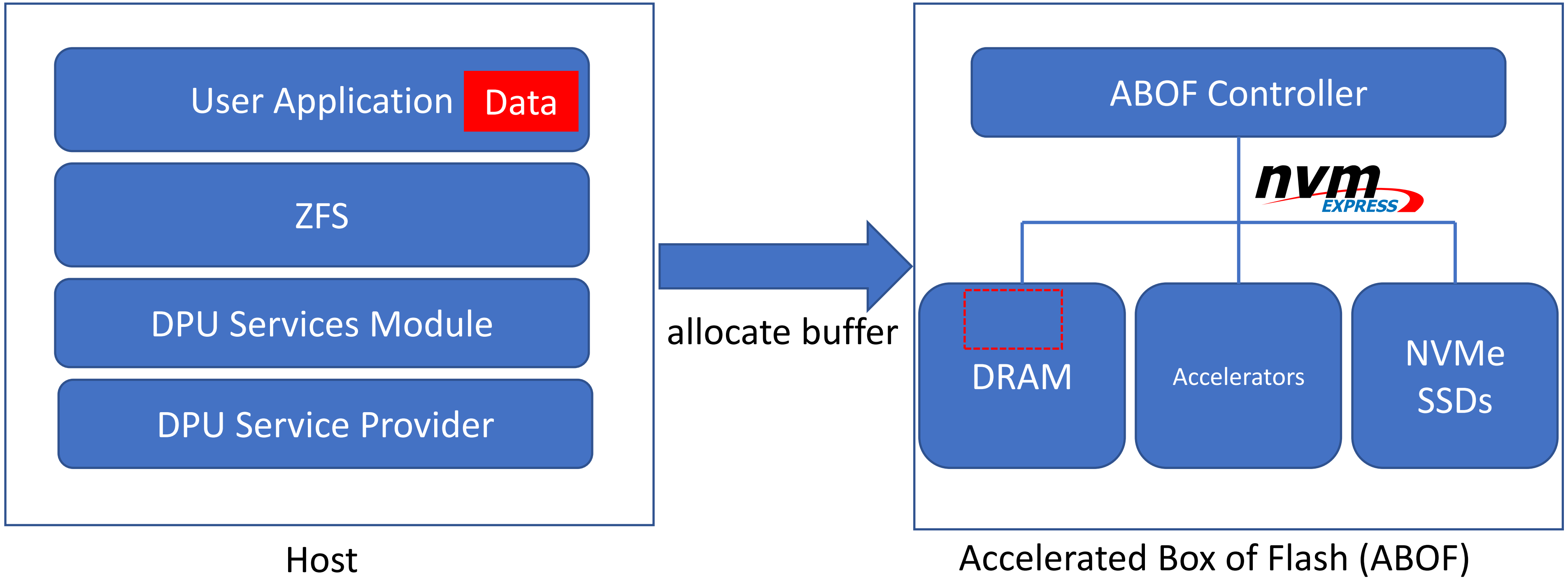


Computational Storage Array (CSA)

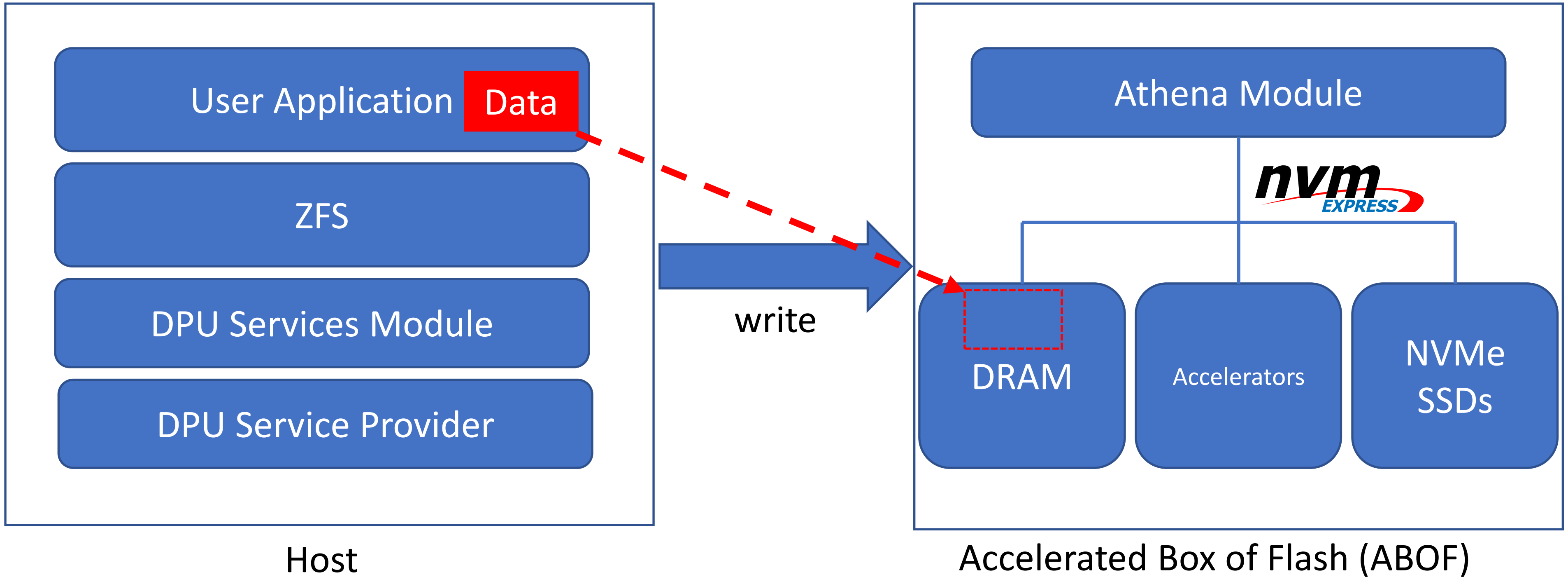
Accelerated box of flash (ABOF) Operation Overview



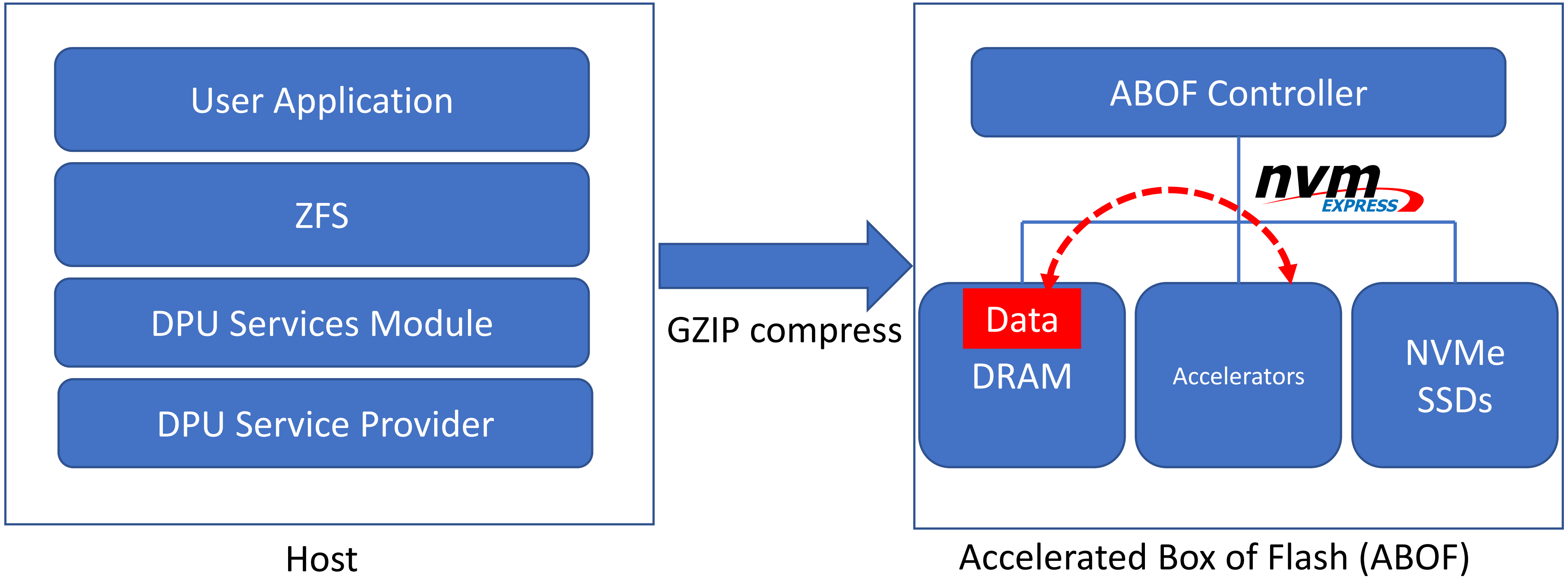
ABOF Operation Overview (alloc)



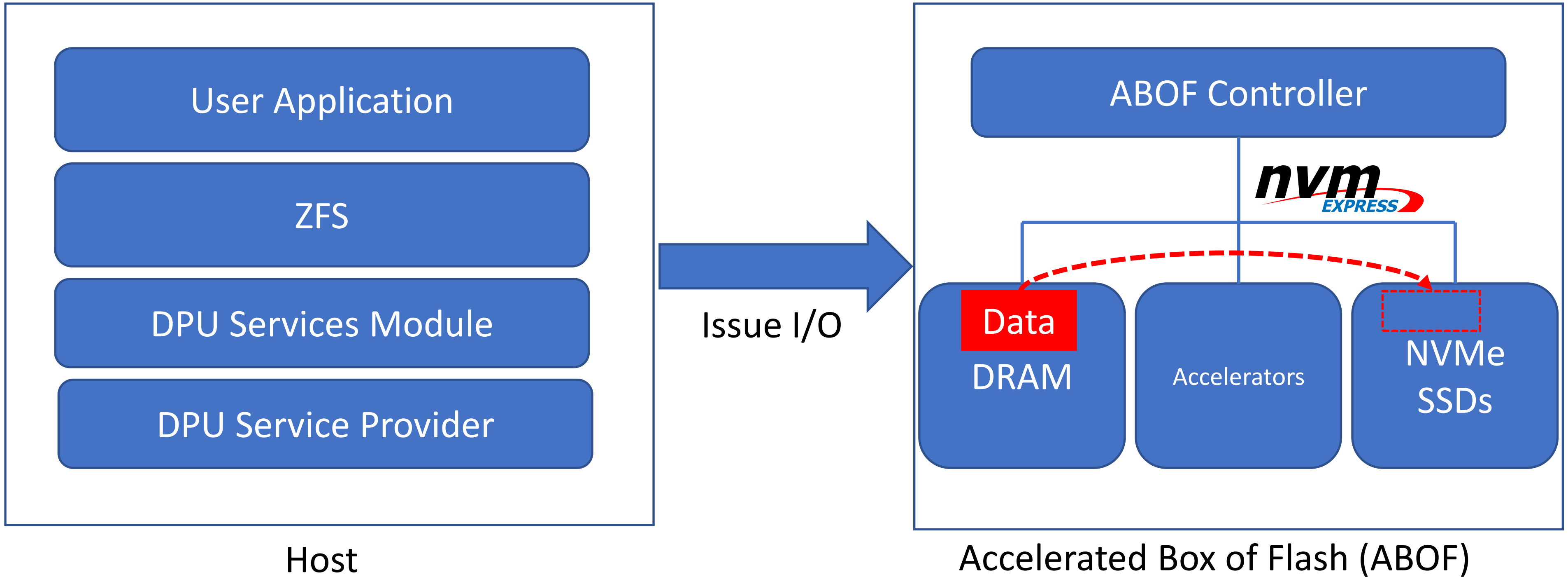
ABOF Operation Overview (write)



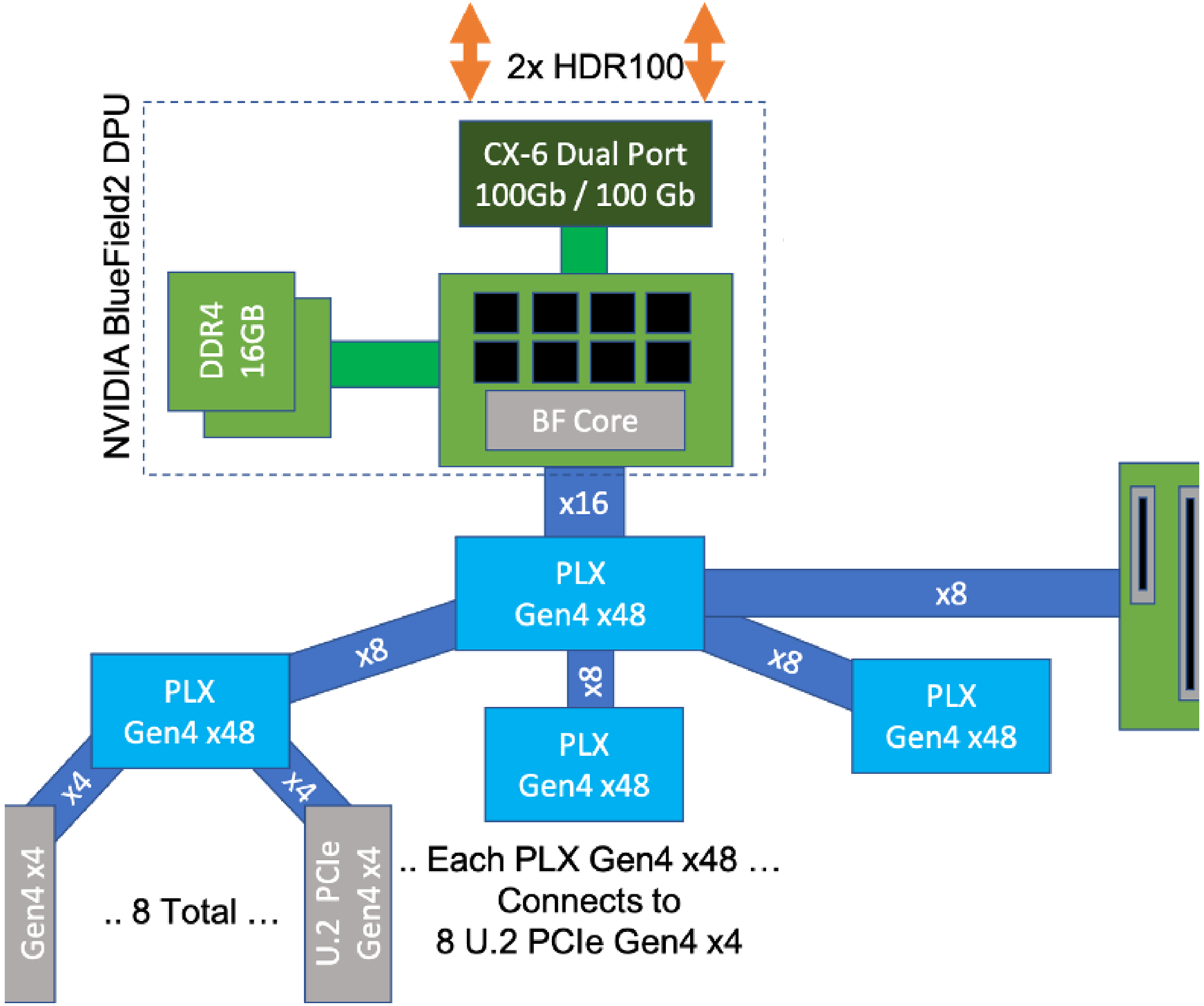
ABOF Operation Overview (compress)

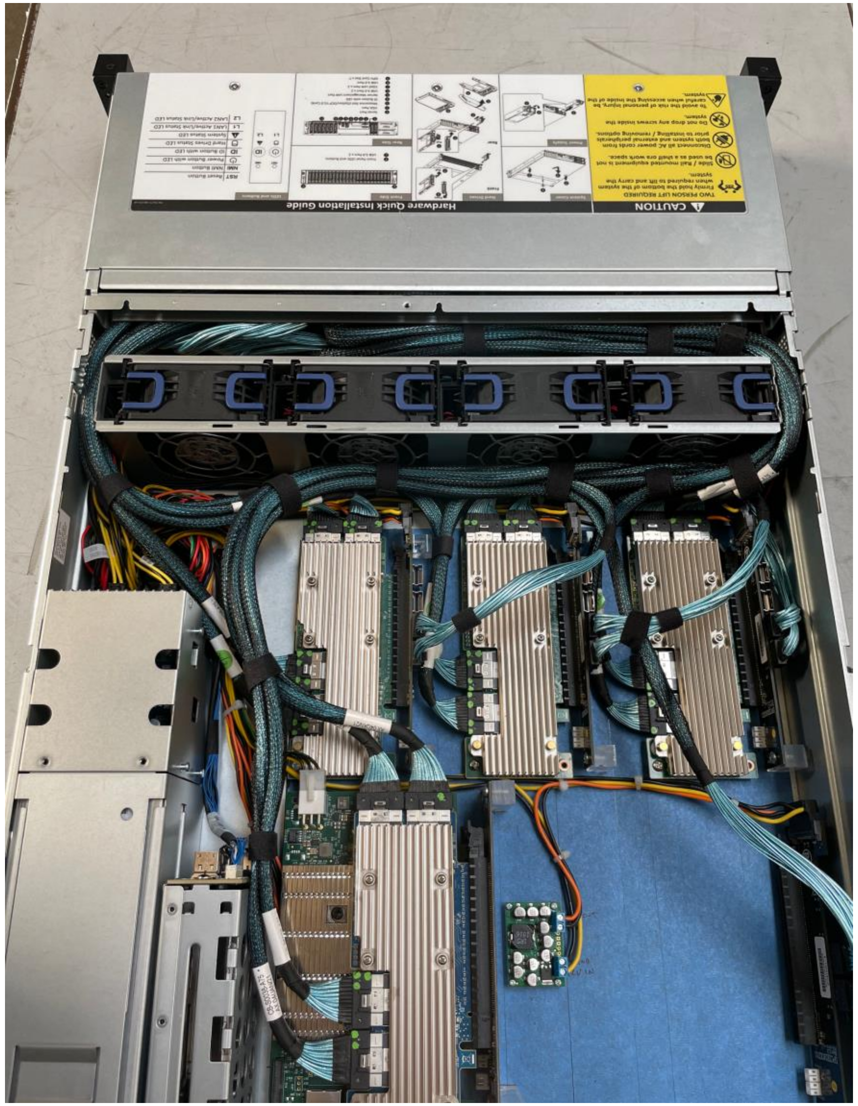


ABOF Operation Overview (issue I/O)

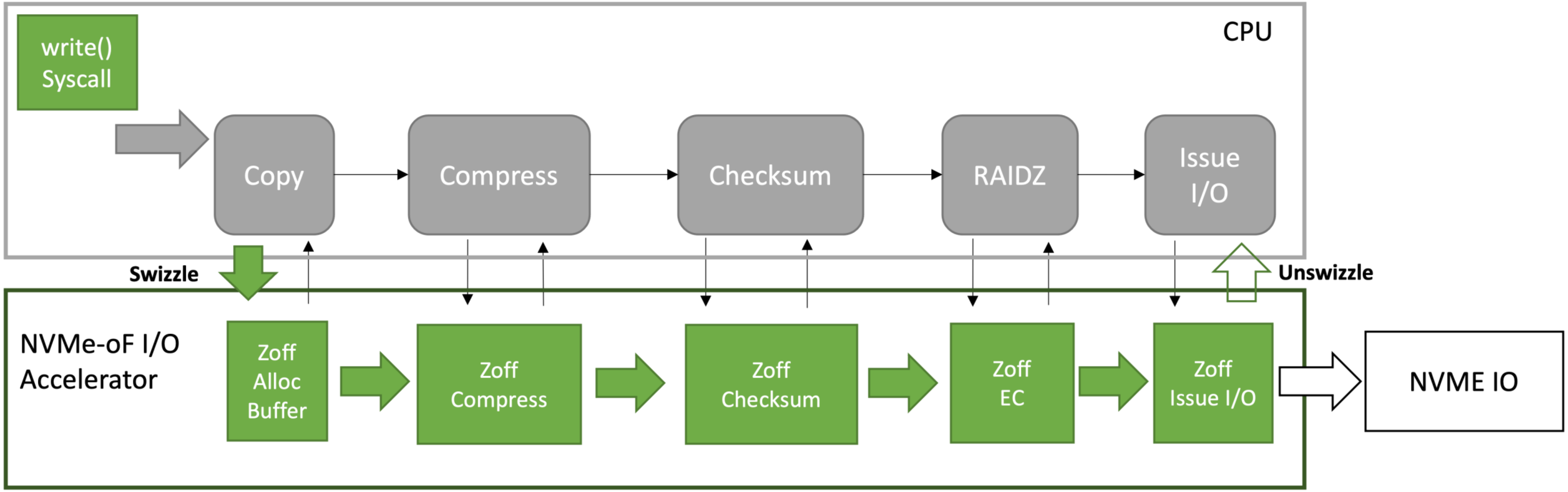


ABOF Hardware



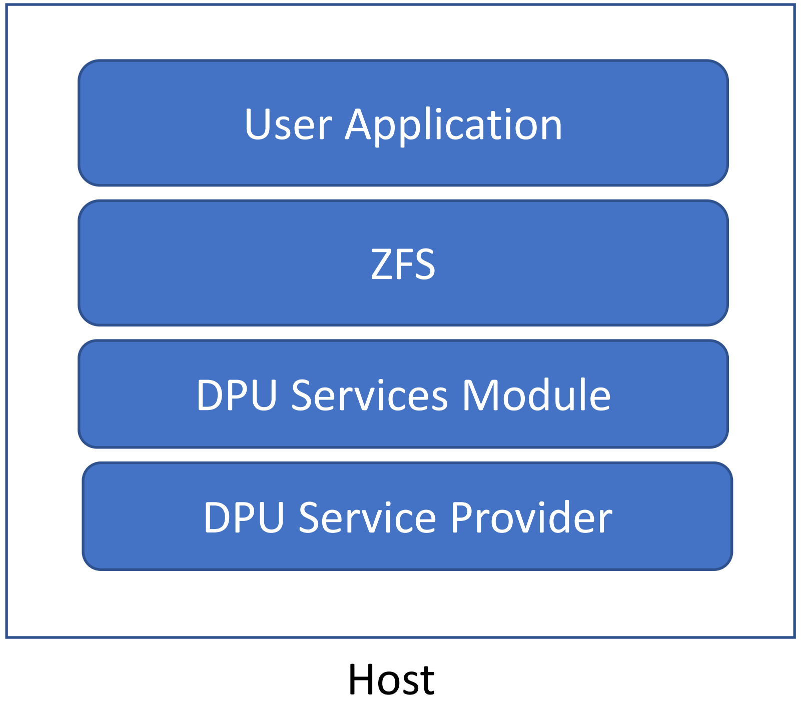


Host Software: ZFS Interface for Accelerators (Z.I.A.)



Host Software: DPUSM

- DPUSM is an API bridge to access computational storage features via DPUSM providers
- DPUSM providers links requested operations to available architecture
- ABOF provider offloads Accelerator requests via NVMe-oF

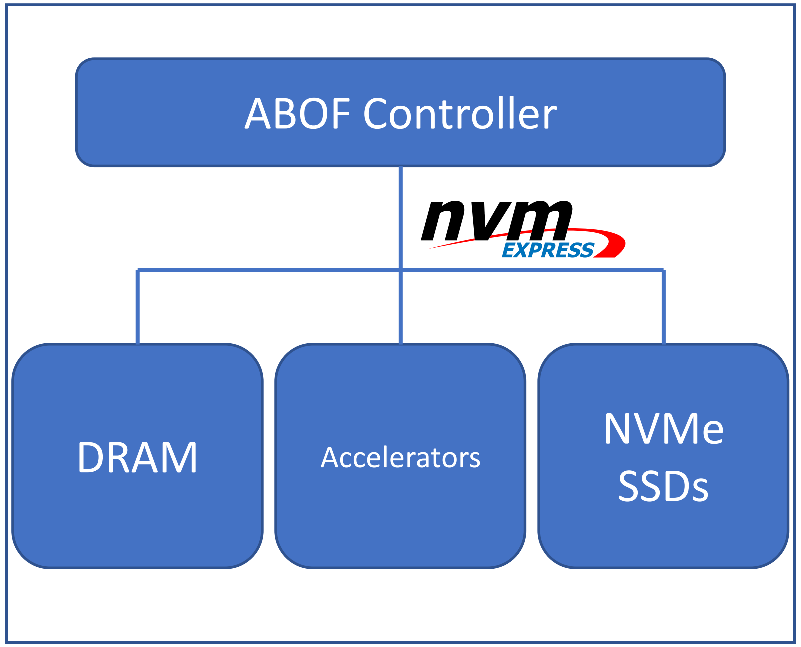


ABOF Software

Use a set of vendor specific op codes via NVMe-oF to:

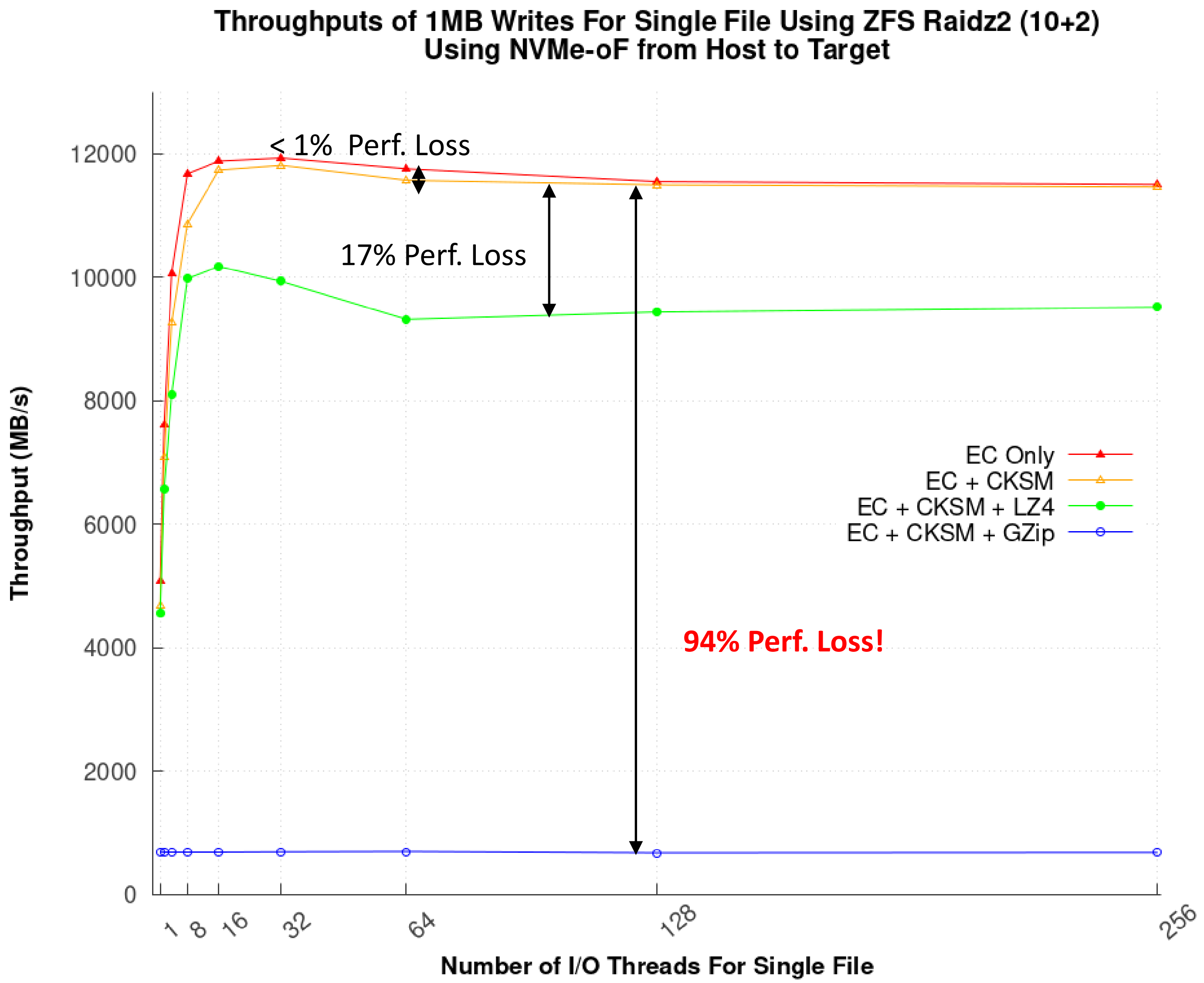
- create/free buffer
- load/Store buffer from disk
- read/write buffer
- Perform operation on buffer
  - Compress/Decompress
  - Checksum
  - EC

Kernel Module or SPDK (userspace) implementation

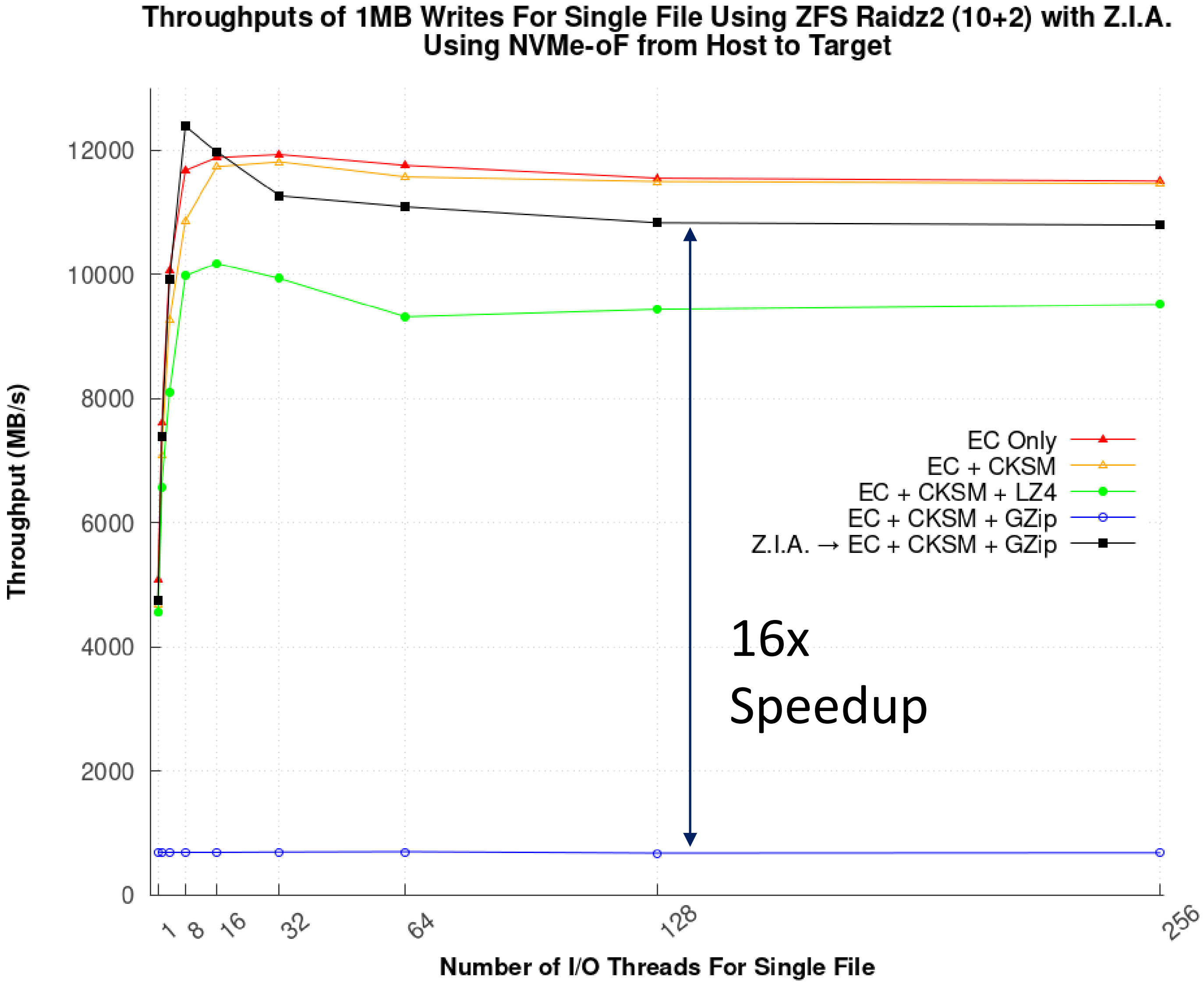


Accelerated Box of Flash (ABOF)

Performance Analysis



Performance Analysis



16x

Speedup

Conclusions and Future

- Standardization
  - Reduce the cost of integration
- Upcoming ZFS direct IO feature
- Beyond ZFS
  - Additional computational offloads (Analytics, AI, etc.)
- Faster, Higher, Stronger
  - PCIe Gen5
  - 400G+ Networking