Flash Memory Summit

Making Real File Systems Faster with Applied Computational Storage

Dominic Manno
Los Alamos National Laboratory

LA-UR-22-27739

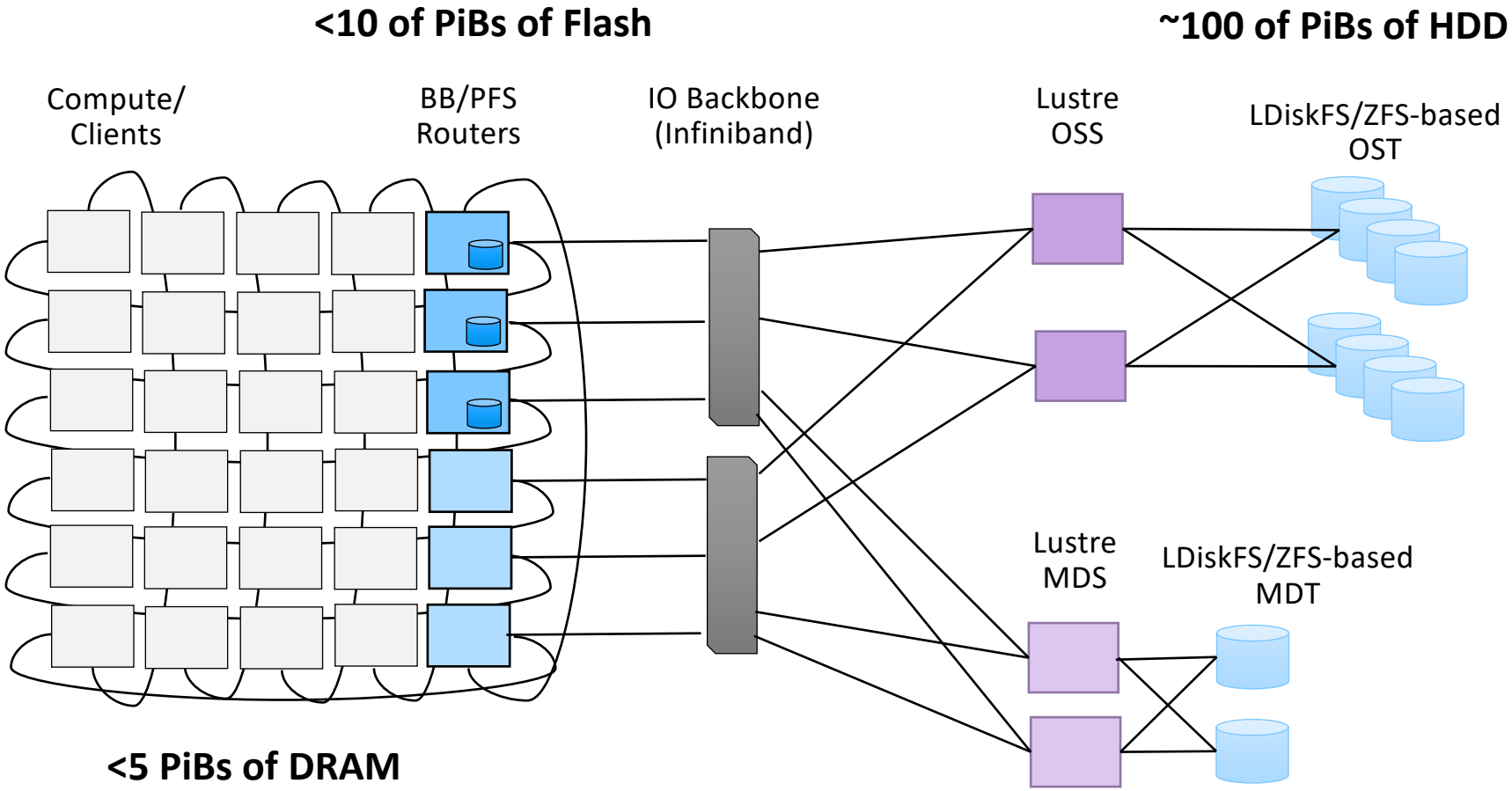
Acknowledgements

- 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
  - Stephen Bates, Roger Bertschmann, Sean Gibb - Eideticom
  - Jeff Johnson, Doug Johnson - Aeon Computing
  - Dominic Manno, Gary Grider, Jason Lee, Brian Atkinson - LANL

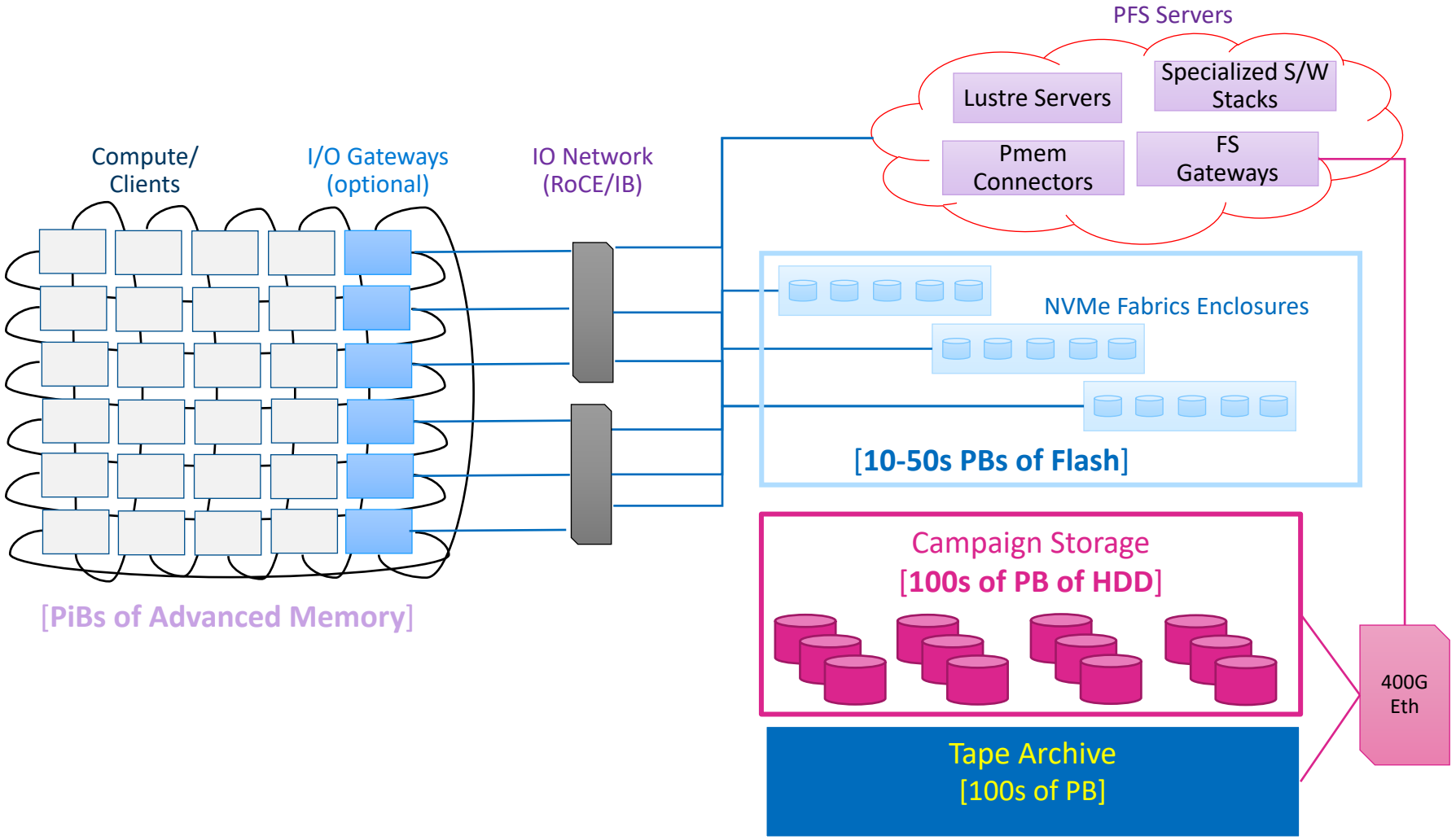
Overview

- File System Performance Challenges
  - All-flash File Systems
  - HPC Datasets
- Enabling A Flexible Design – Accelerated Box of Flash (ABOF)
  - HW overview
  - SW overview
- Performance Review
- Outlook

Traditional HPC Storage



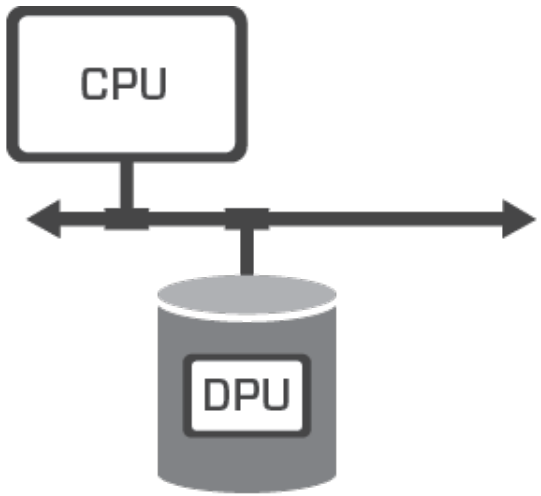
Redesign Opportunity Thanks to NVMe



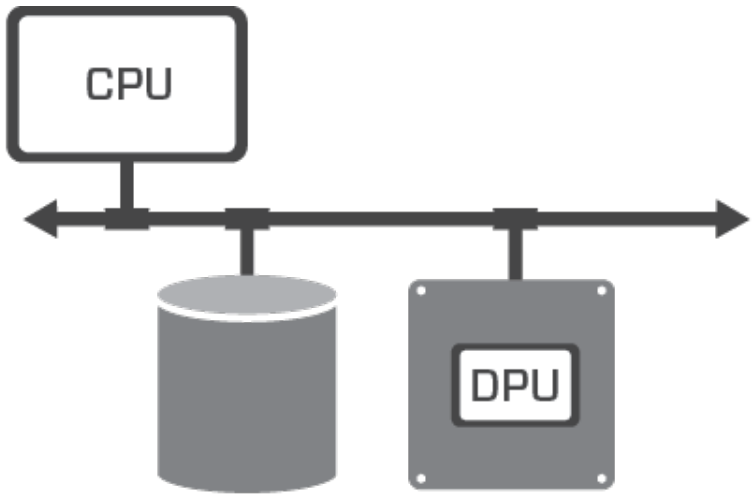
All Flash File Systems

- Require high performing storage server endpoints
  - Otherwise – disaggregated isn't as important cost wise
- Current generation server memory bandwidth limitations observed relatively quickly
- With a budget, buying BW often doesn't result in high capacity
  - Compression is important
  - Compressing simulation data is hard!

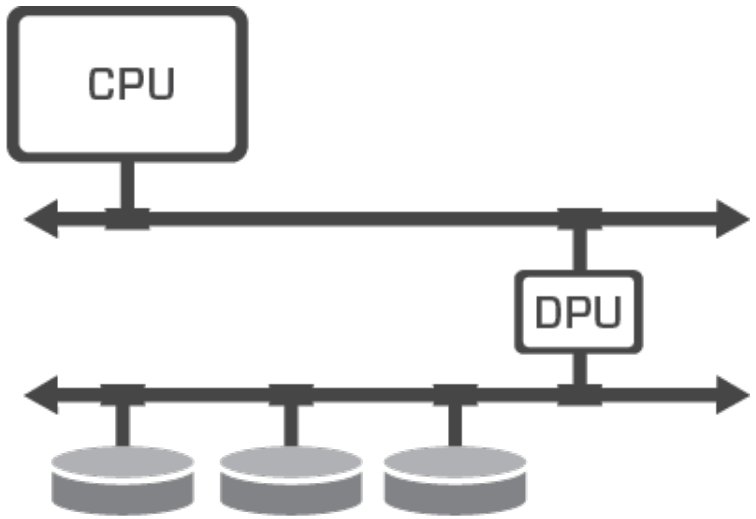
Computational Storage Can Help!



Computational Storage Device (CSD)

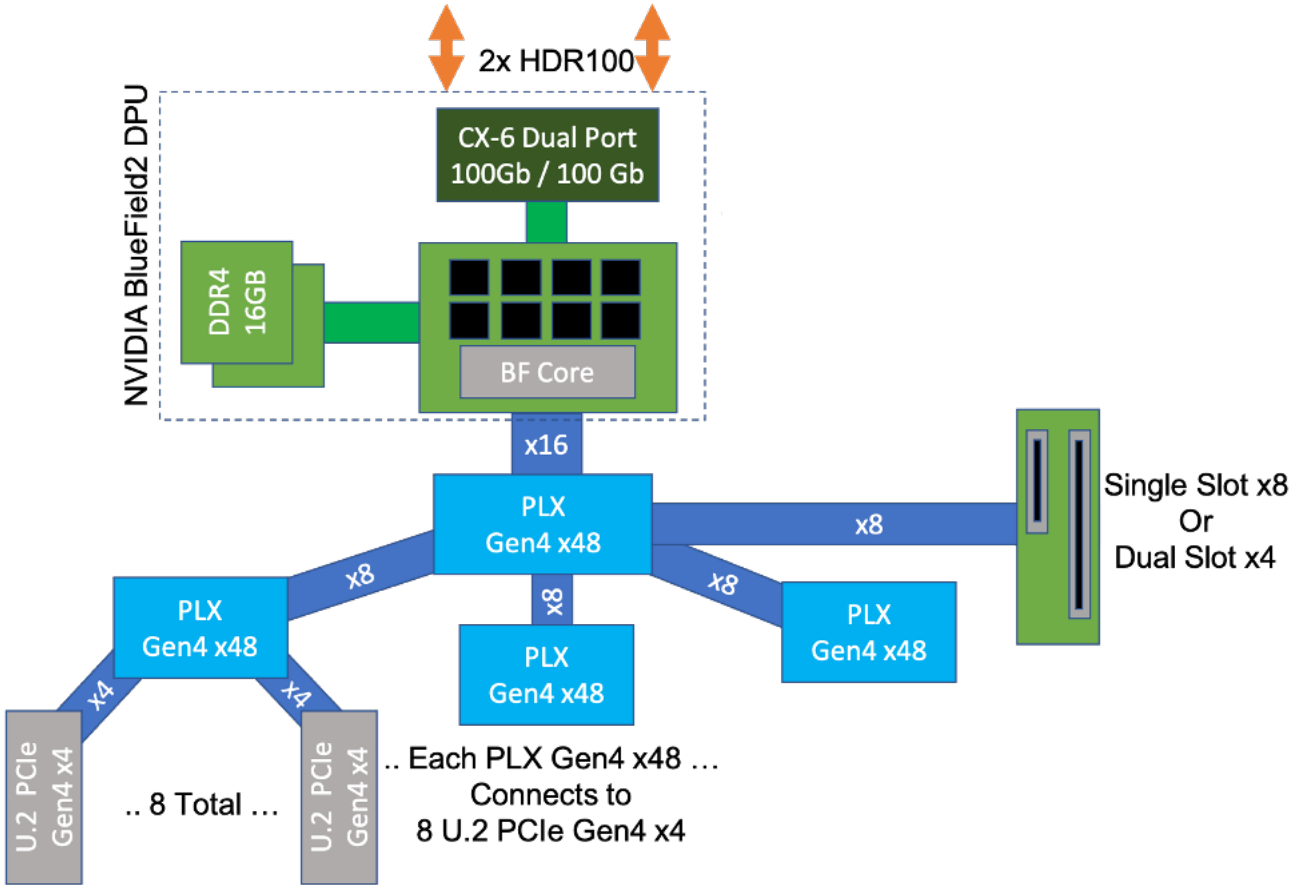


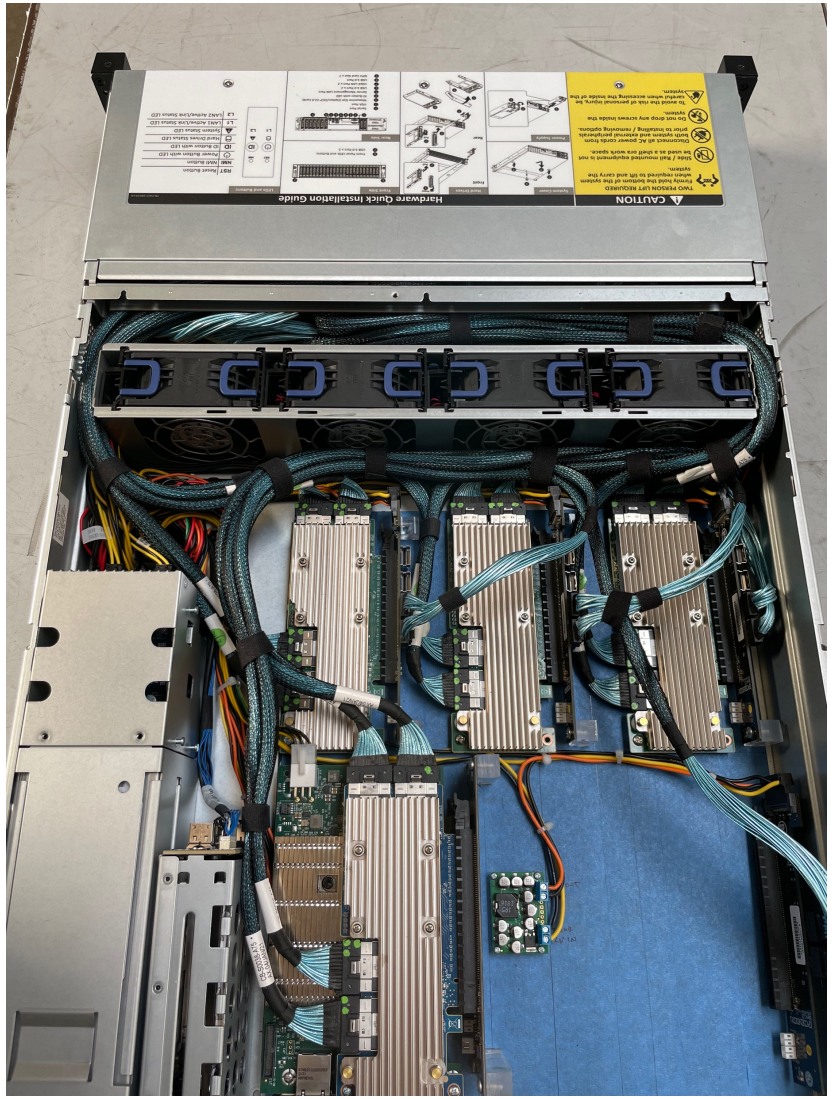
Computational Storage Processor (CSP)



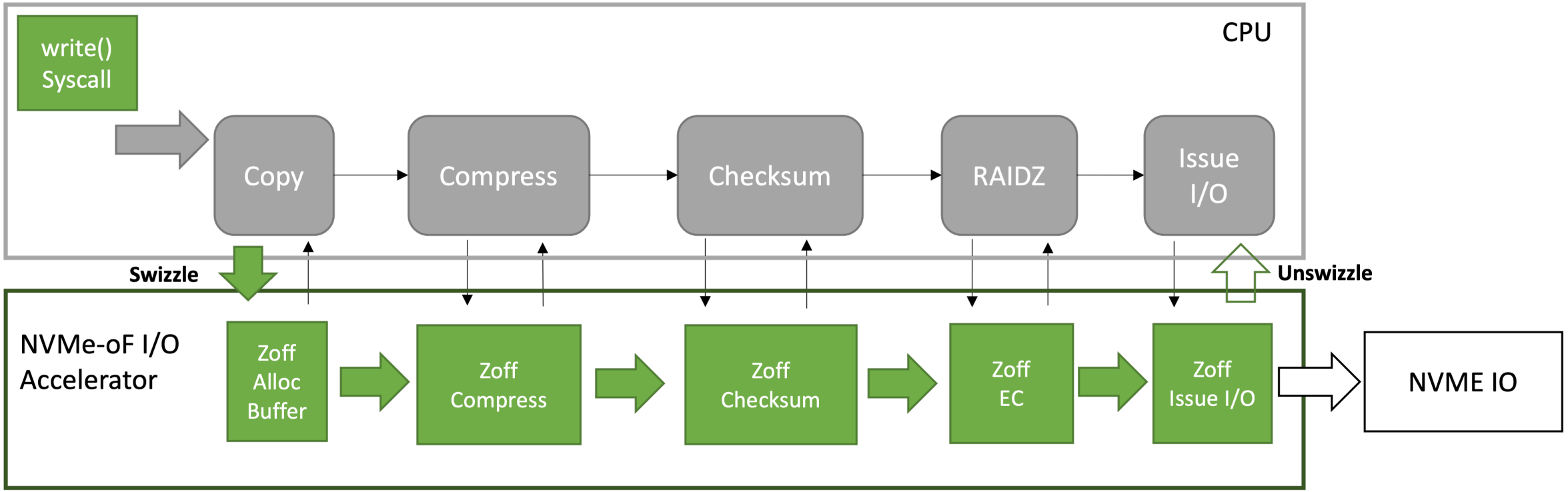
Computational Storage Array (CSA)

ABOF - Hardware Overview

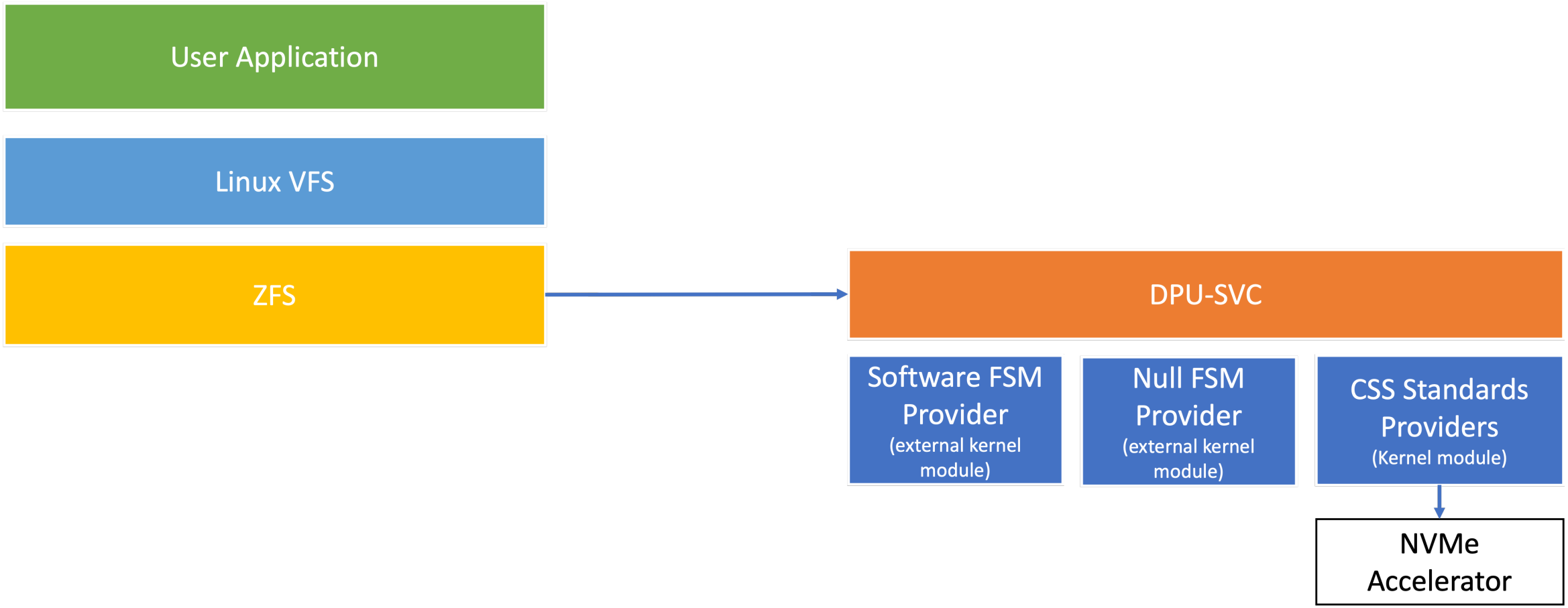




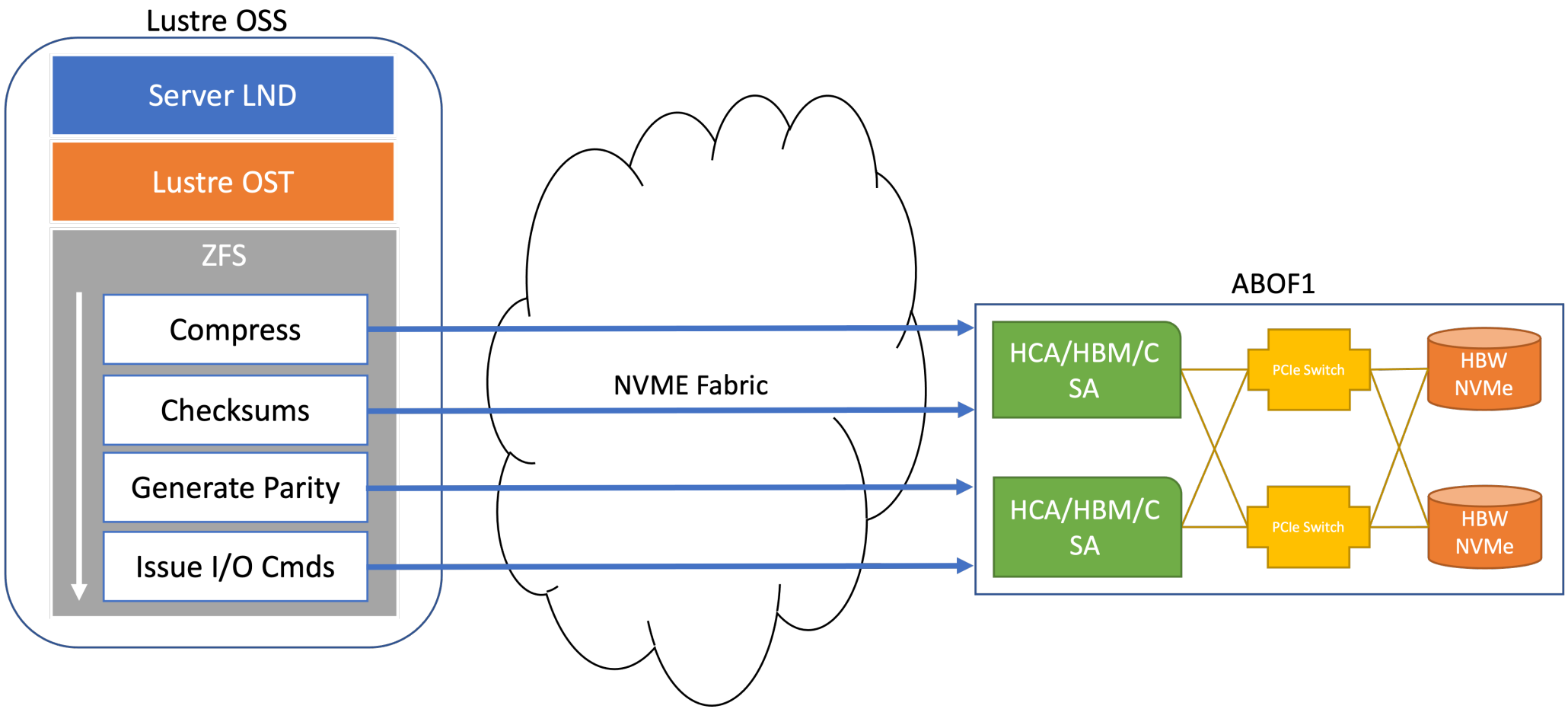
ZFS Interface for Accelerators (Z.I.A)



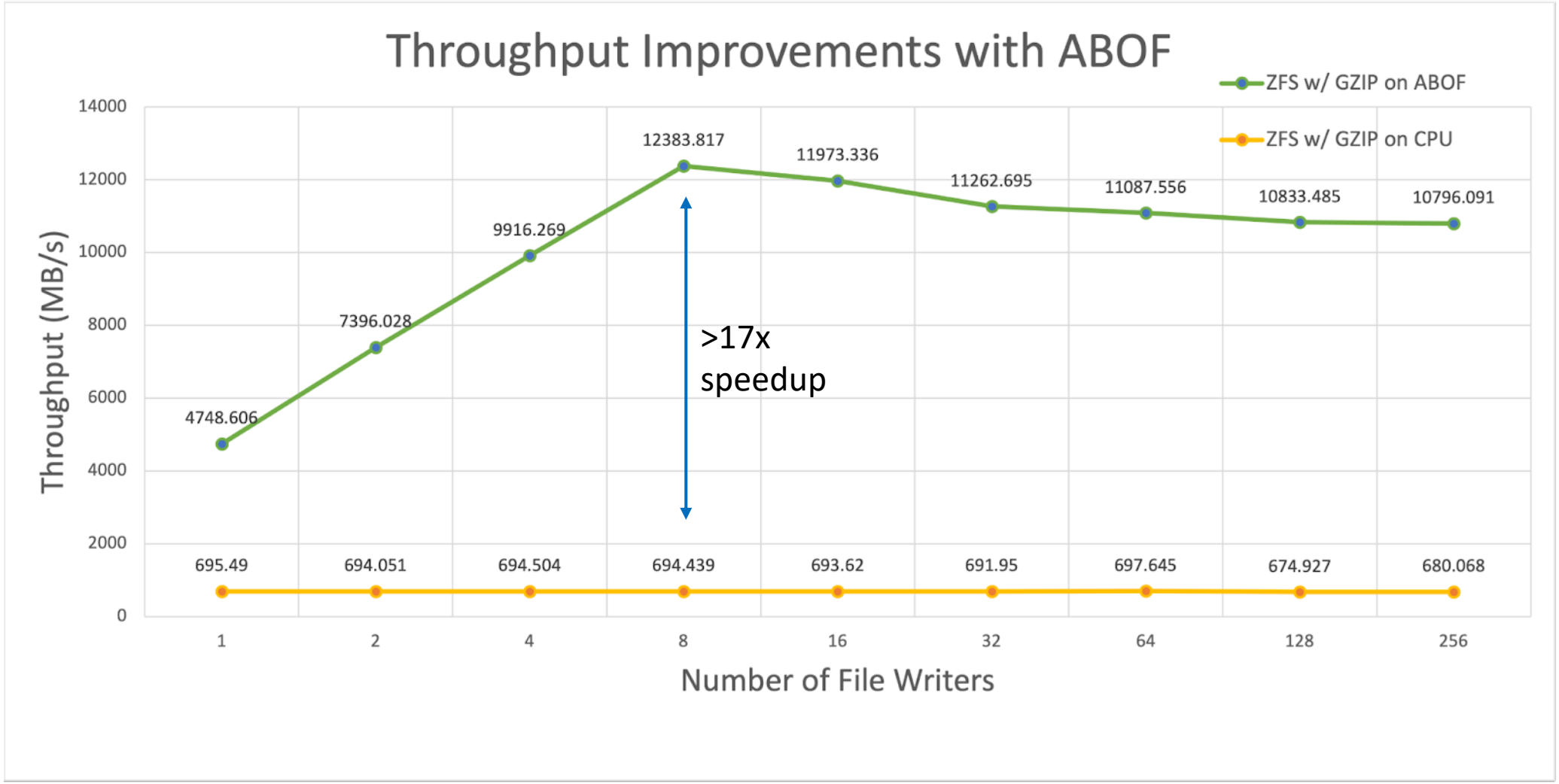
Data Processing Unit Services Module (DPU-SVC)



Theory of Operations



Performance Analysis



Follow-on Work

- Exploring “data-aware” offloads to enhance analytic capabilities without requirement large amounts of data movement
- Continuing performance analysis and improvements
  - Hardware upgrade
- Determining optimal location of offloads

