Making Real File Systems Faster with Applied Computational Storage

Presented by Sean Gibb*, Andrew Maier*, Dominic Manno^

*Eideticom, ^Los Alamos National Laboratory

LA-UR-22-29289
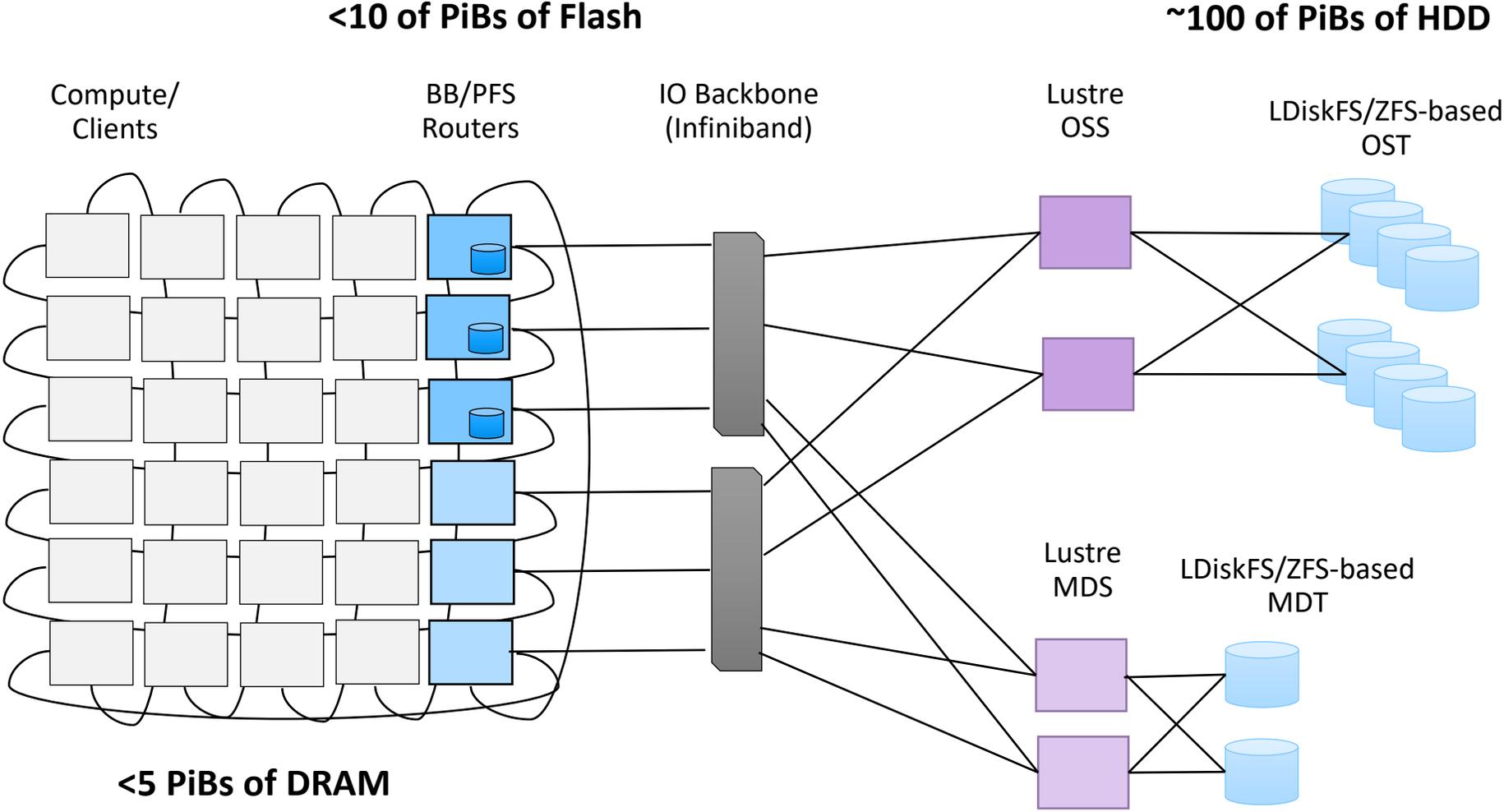
Acknowledgements

- 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, Andrew Maier – Eideticom
    - Jeff Johnson, Doug Johnson - Aeon Computing
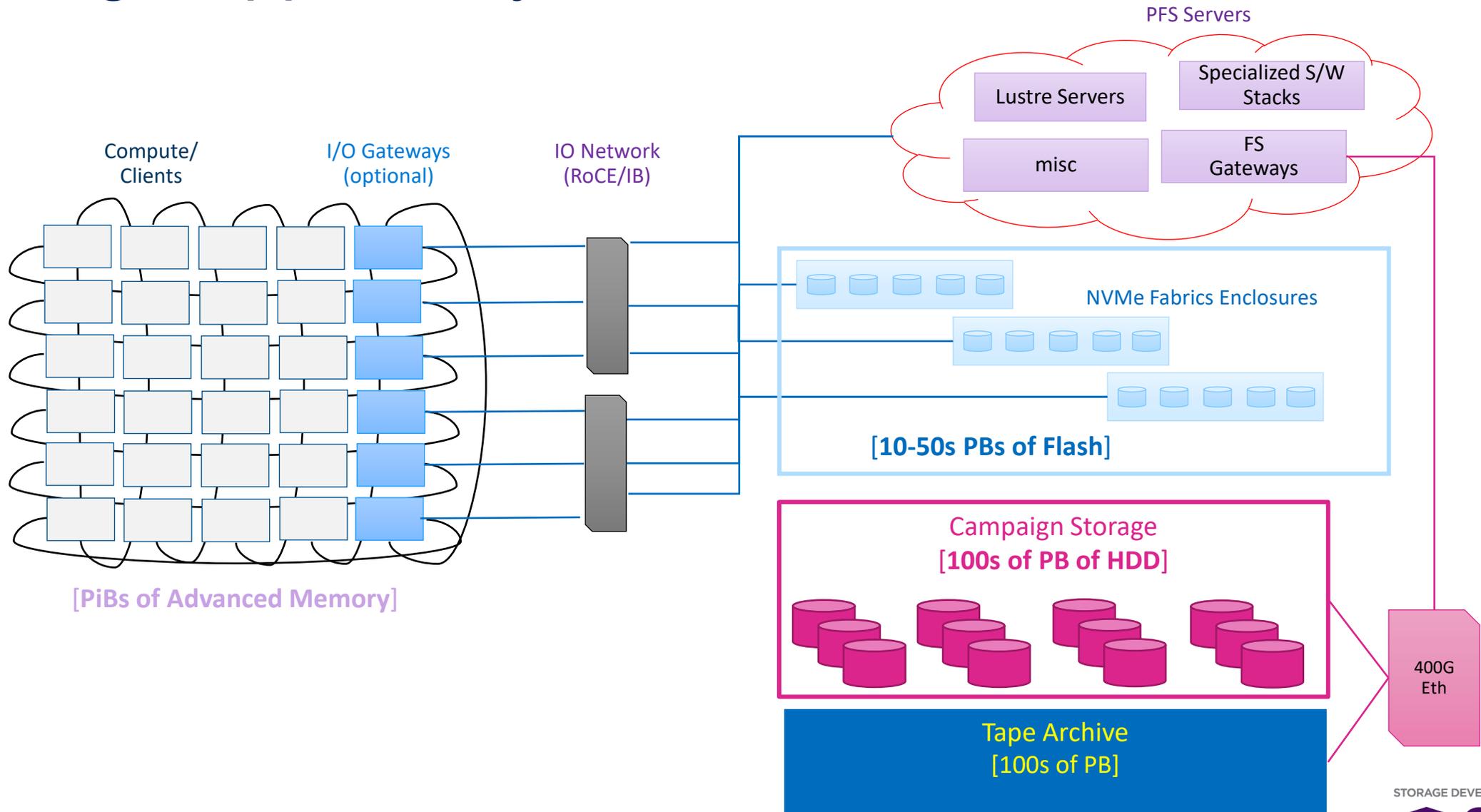   - Dominic Manno, Gary Grider, Jason Lee, Brian Atkinson - LANL

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**

STORAGE DEVELOPER CONFERENCE

SDC 22

Traditional HPC Storage

<10 of PiBs of Flash

~100 of PiBs of HDD

Compute/
Clients

BB/PFS
Routers

IO Backbone
(Infiniband)

Lustre
OSS

LDiskFS/ZFS-based
OST

Lustre
MDS

LDiskFS/ZFS-based
MDT

<5 PiBs of DRAM

STORAGE DEVELOPER CONFERENCE

SDC 22

Redesign Opportunity Thanks to NVMe

STORAGE DEVELOPER CONFERENCE

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
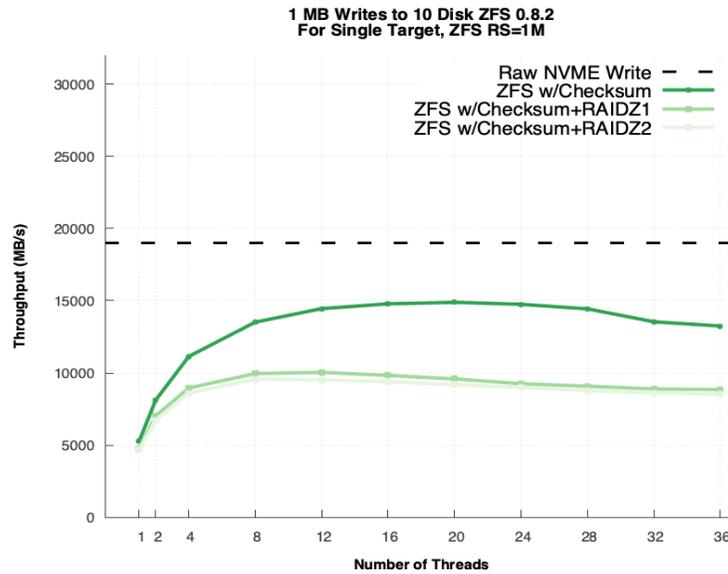 - Compressing simulation data is hard!
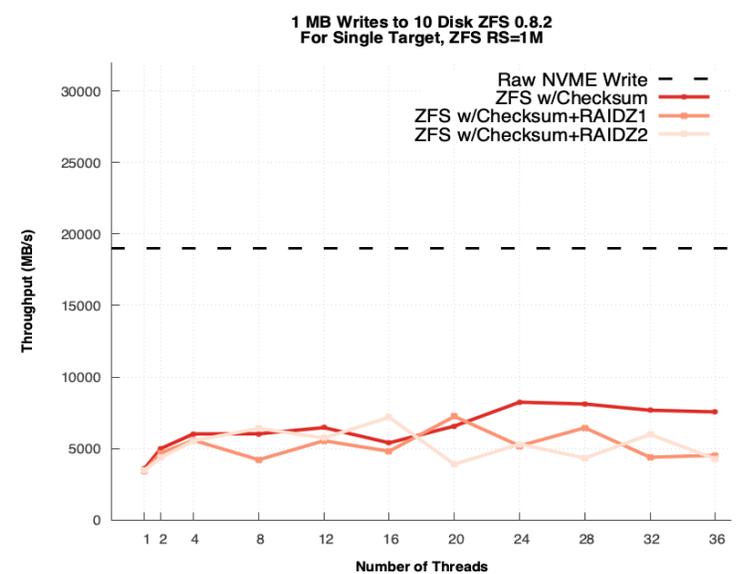
HPC Storage Pipeline
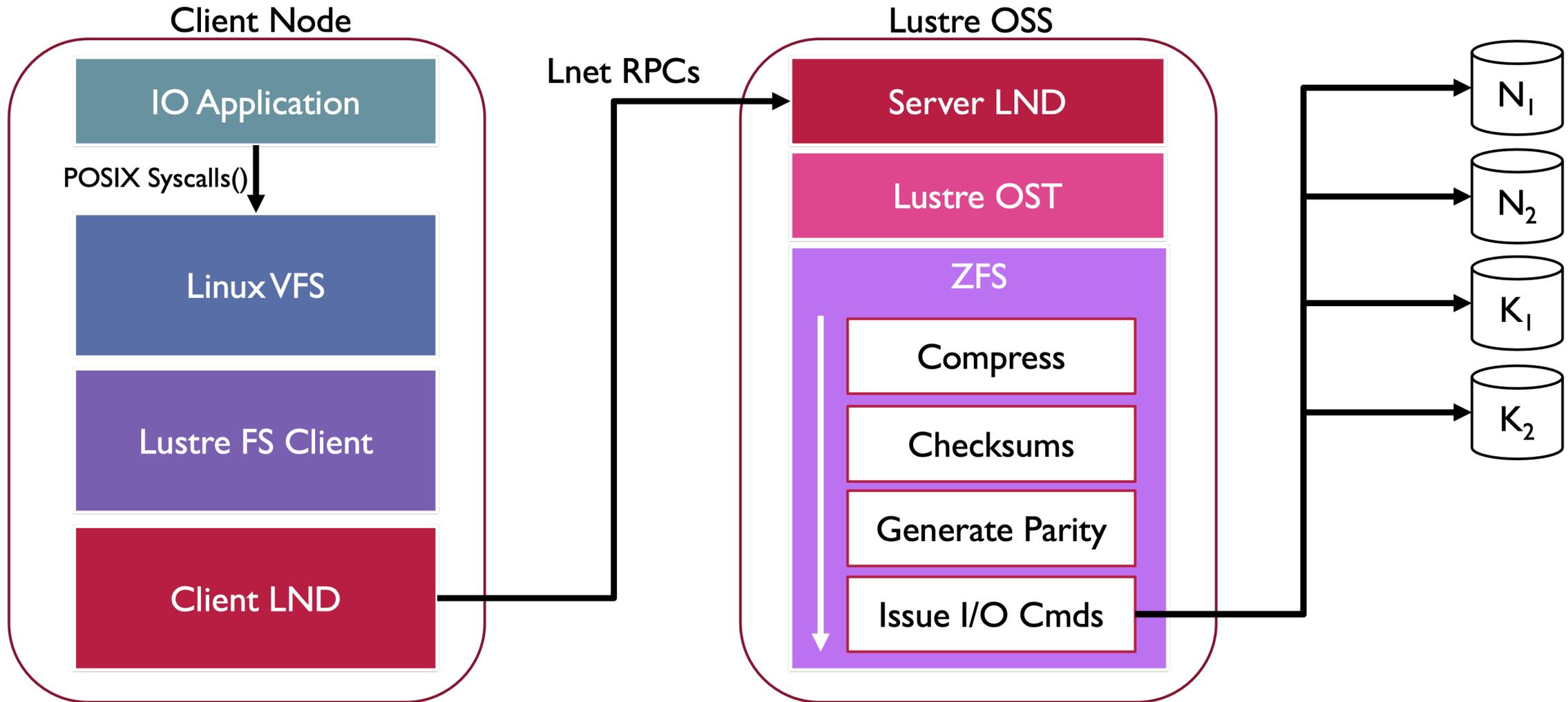
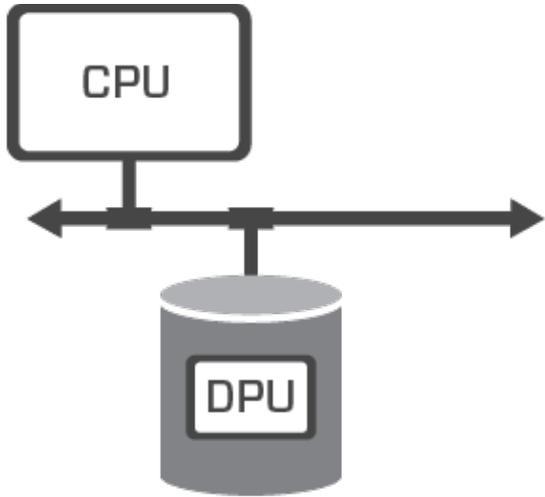# File System Services Limitation
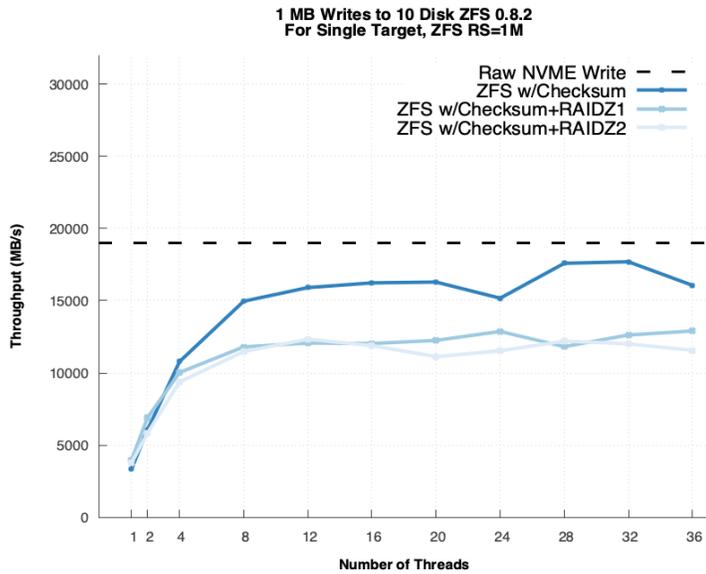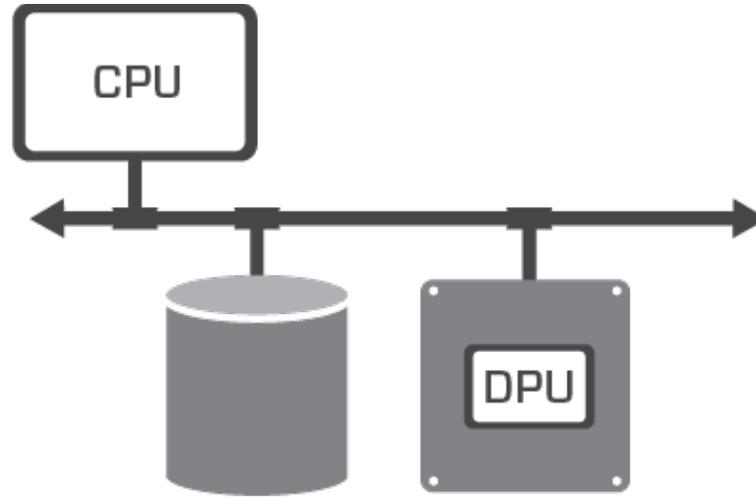

Intel Platinum (Dual Socket)

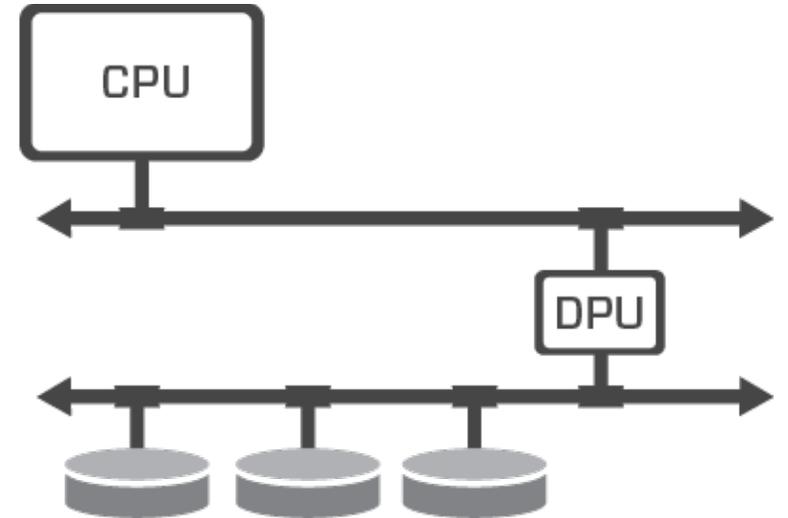AMD EPYC (2nd Gen)

AMD EPYC (1st Gen)

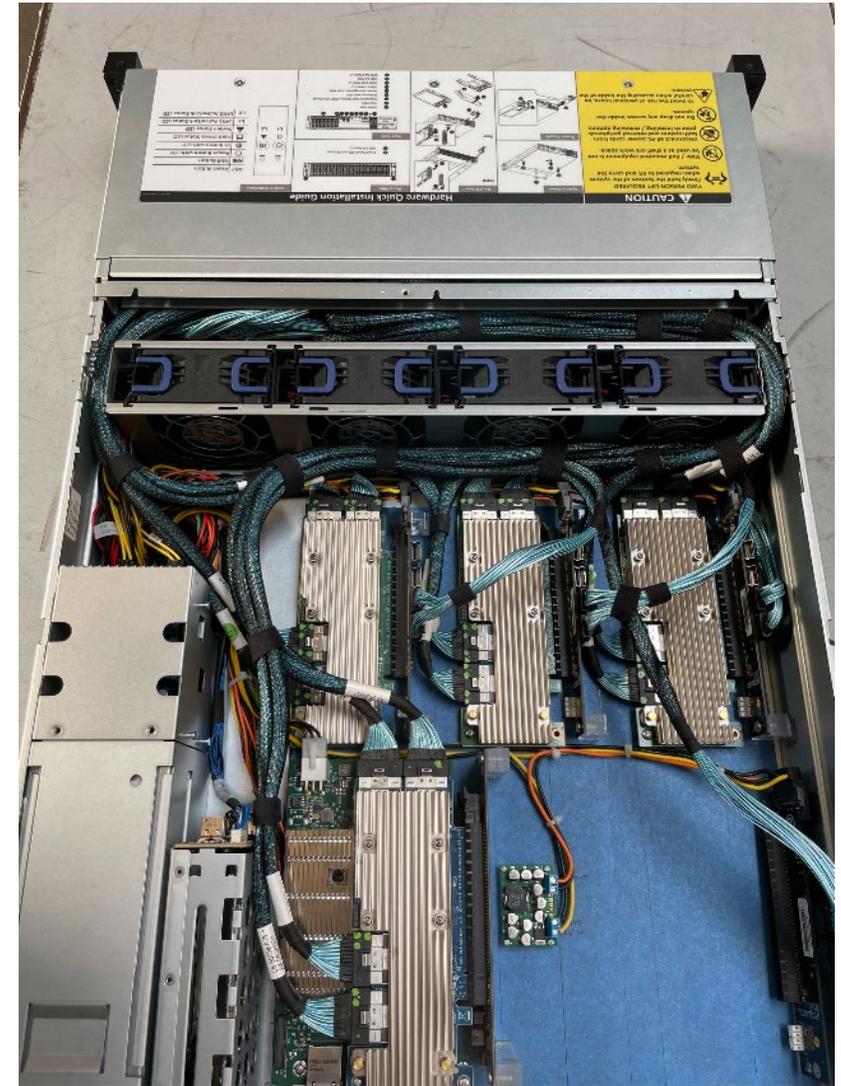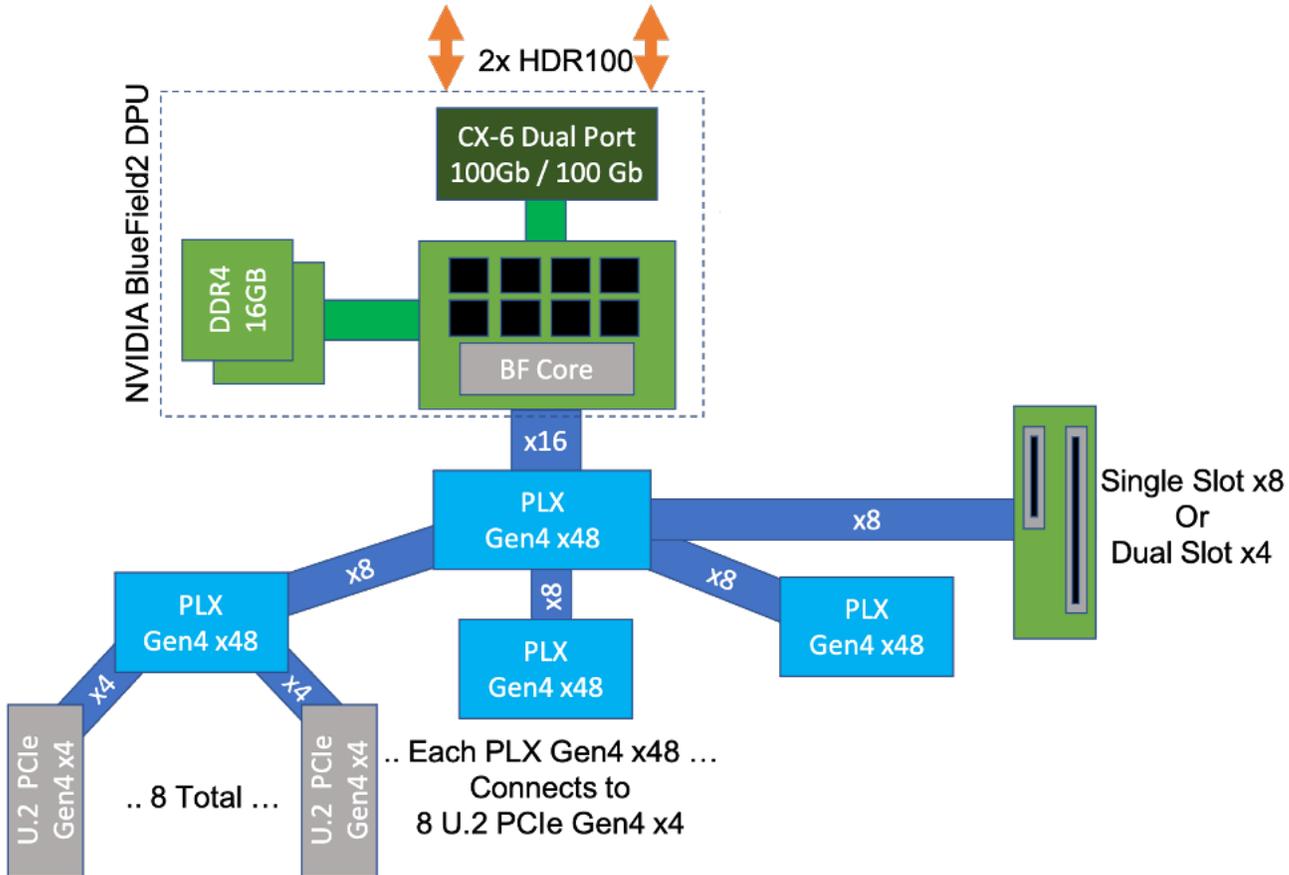# Computational Storage Can Help!

Computational Storage
Device (CSD)

Computational Storage
Processor (CSP)

Computational Storage
Array (CSA)

STORAGE DEVELOPER CONFERENCE
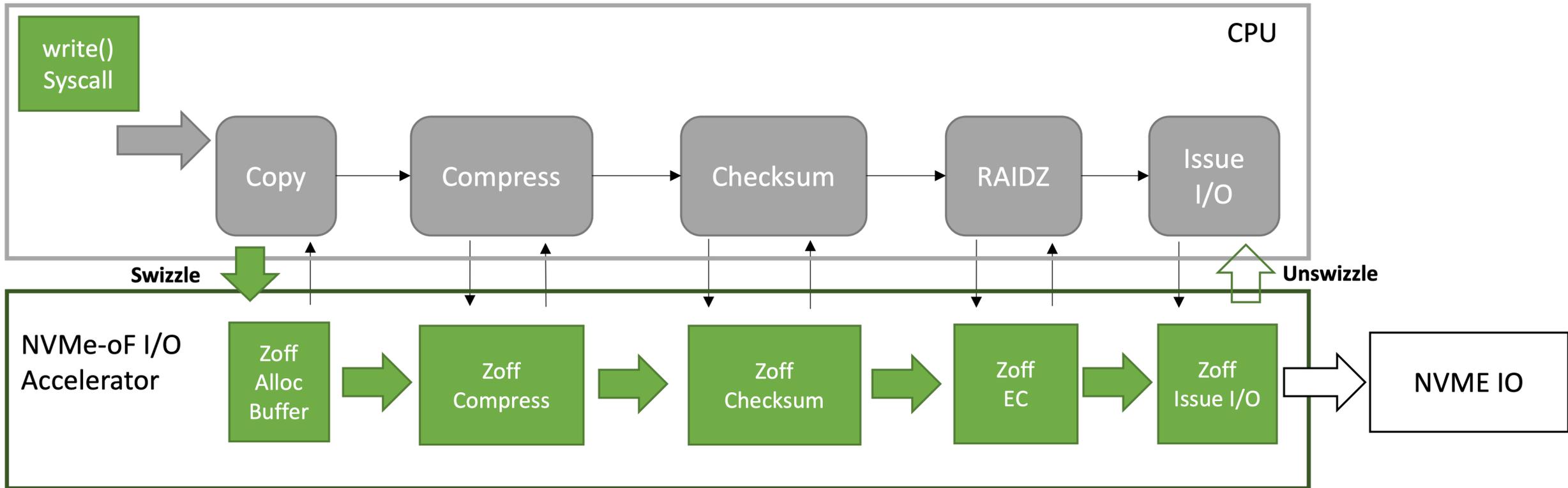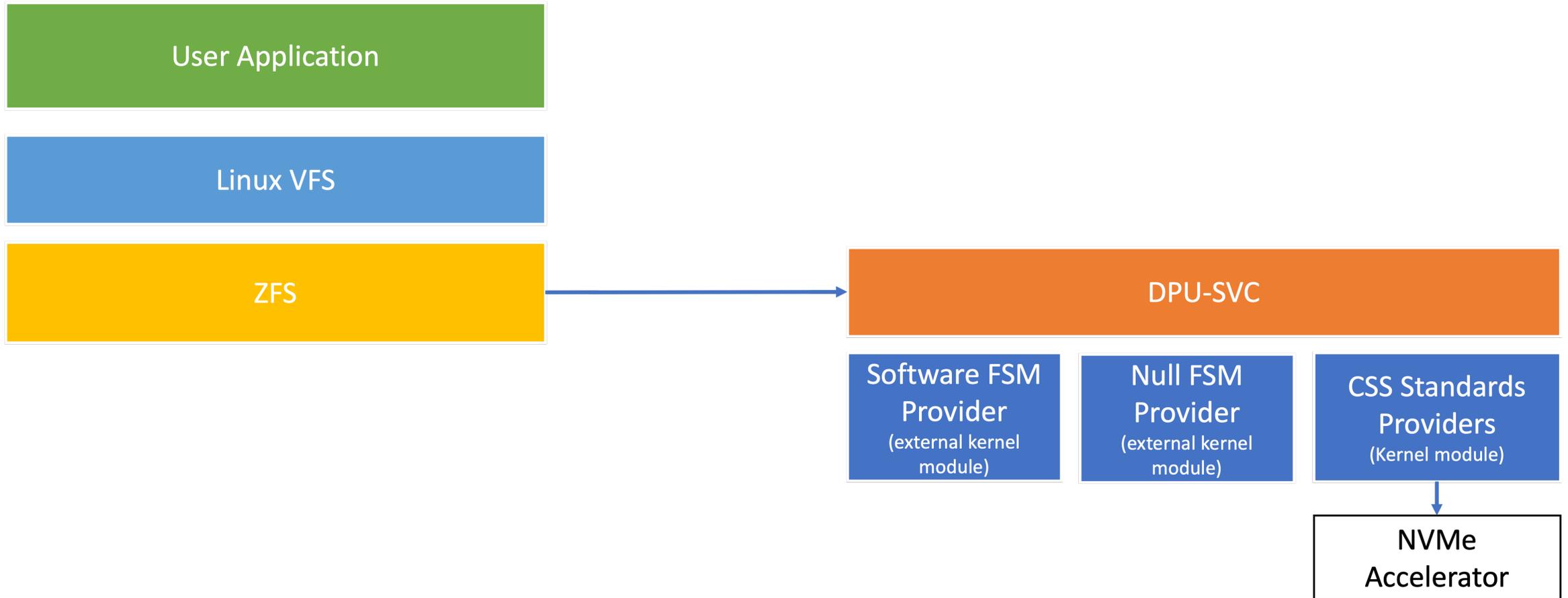
ABOF - Hardware Overview

ZFS Interface for Accelerators (Z.I.A)

Data Processing Unit Services Module (DPU-SVC)

User Application

Linux VFS

ZFS

DPU-SVC

Software FSM Provider (external kernel module)

Null FSM Provider (external kernel module)

CSS Standards Providers (Kernel module)

NVMe Accelerator

Data Processing Unit Services Module (DPU-SVC)

User Application

Linux VFS

ZFS → DPU-SVC

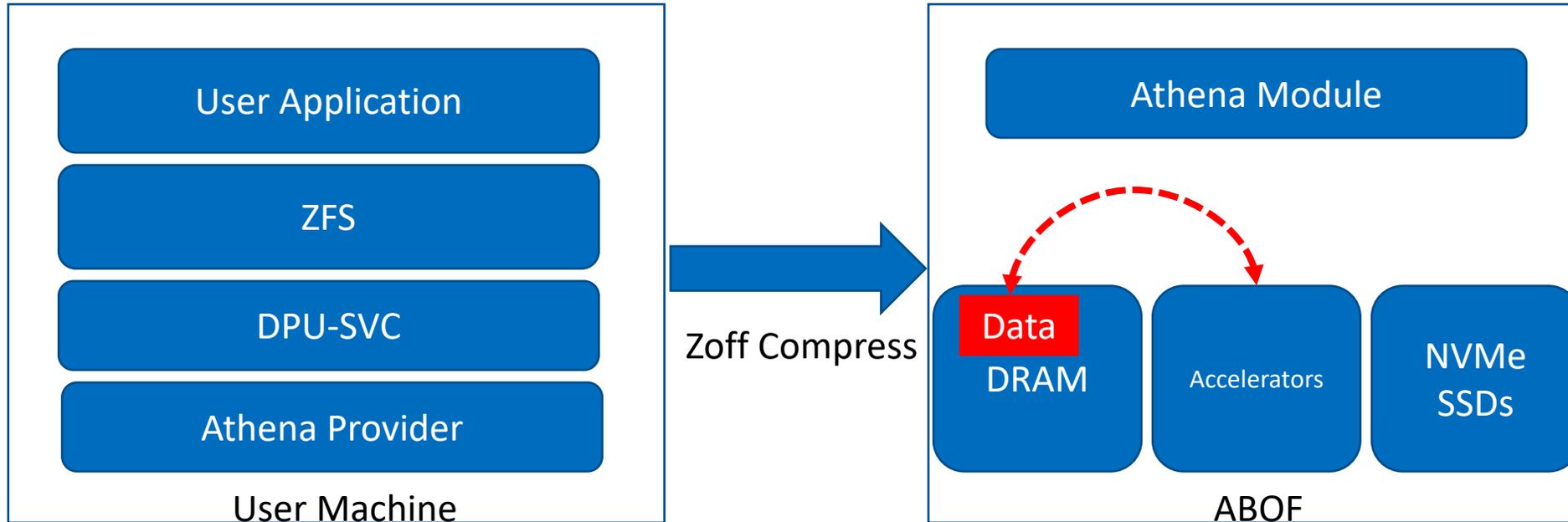| Software FSM Provider (external kernel module) | Null FSM Provider (external kernel module) | CSS Standards Providers (Kernel module) |

NVMe Accelerator
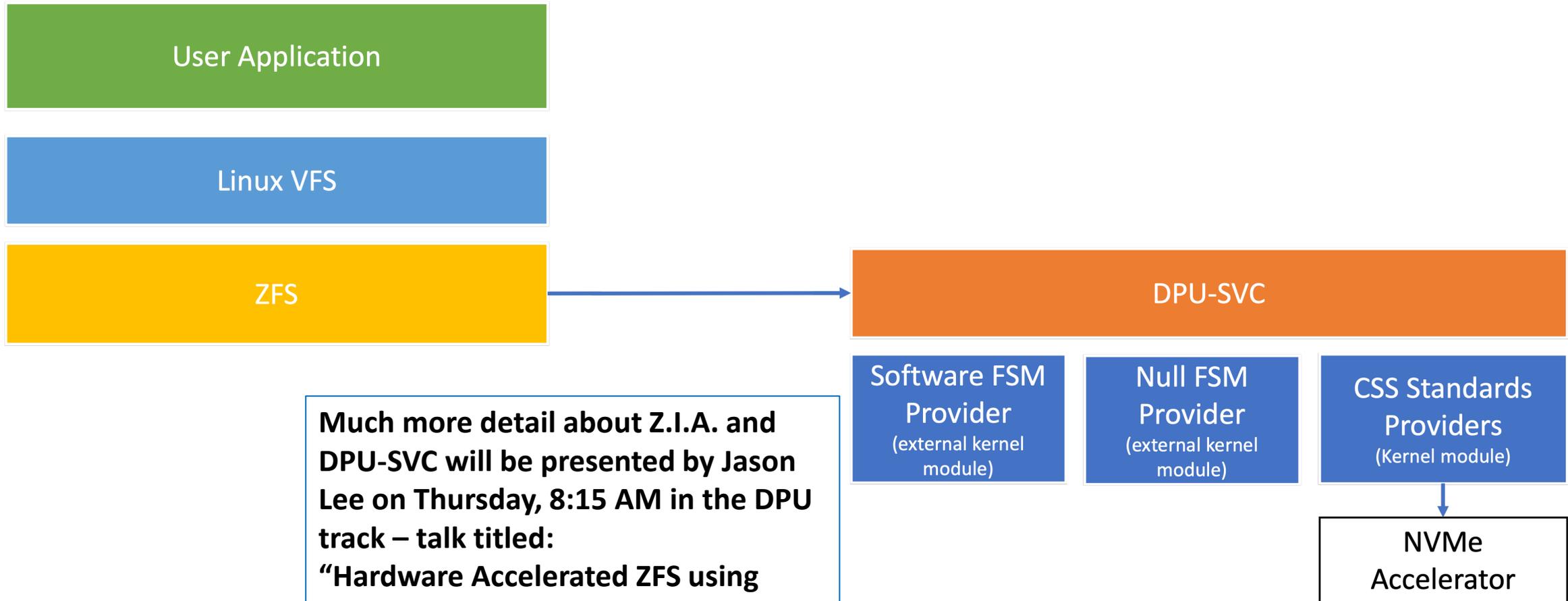
Much more detail about Z.I.A. and DPU-SVC will be presented by Jason Lee on Thursday, 8:15 AM in the DPU track – talk titled:
"Hardware Accelerated ZFS using Computational Storage"
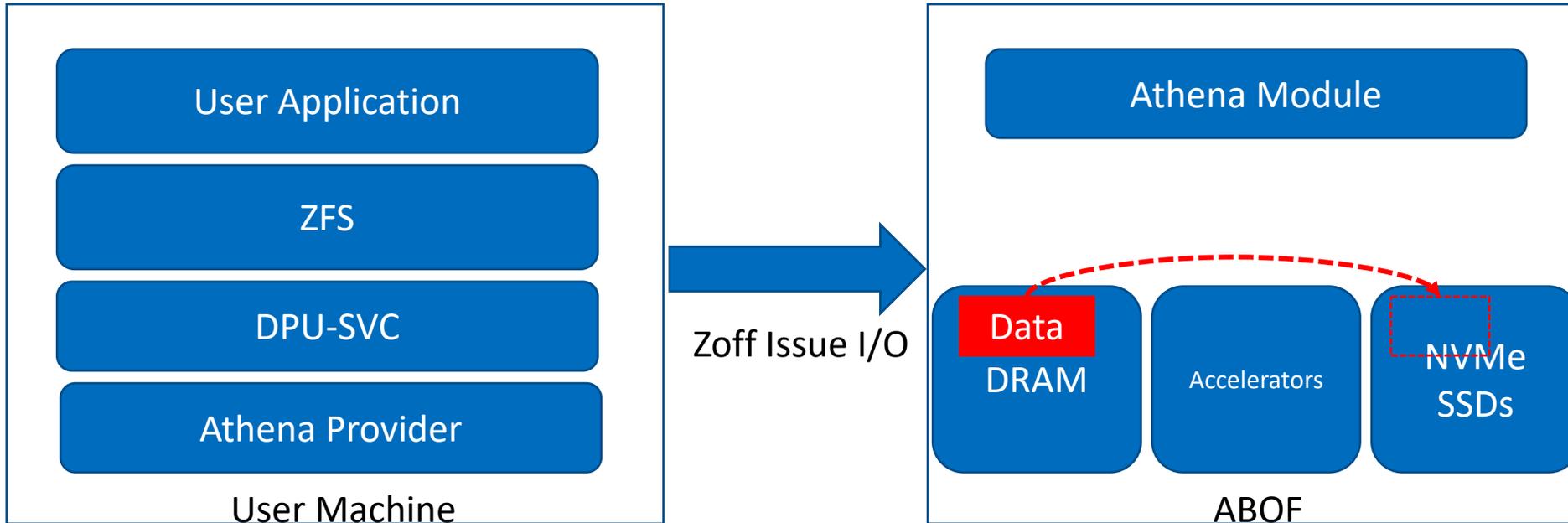
STORAGE DEVELOPER CONFERENCE
SDC 22
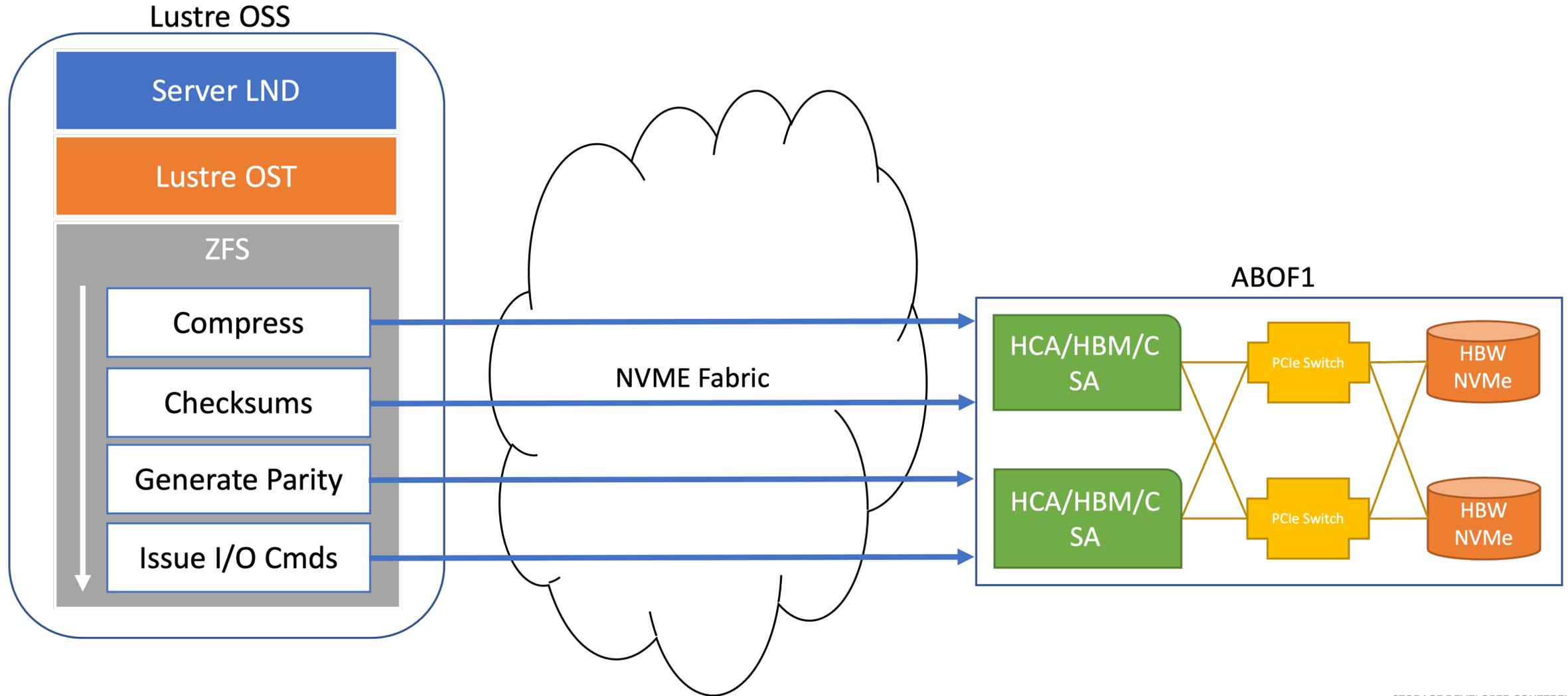
Data Movement – Zoff Alloc Buffer

User Machine
- User Application | Data
- ZFS
- DPU-SVC
- Athena Provider

Zoff Alloc →

ABOF
- Athena Module
- DRAM
- Accelerators
- NVMe SSDs
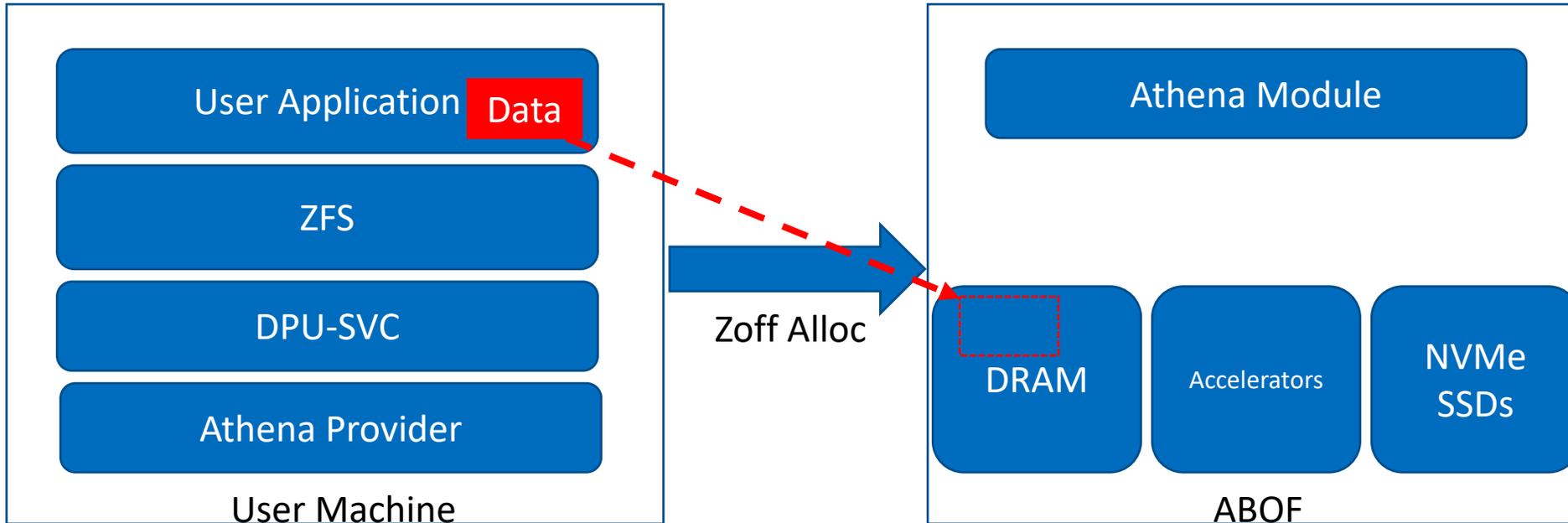
Data Movement – Zoff Compress/Zoff EC

User Machine

- User Application
- ZFS
- DPU-SVC
- Athena Provider

Zoff Compress →

ABOF

- Athena Module
- DRAM (Data)
- Accelerators
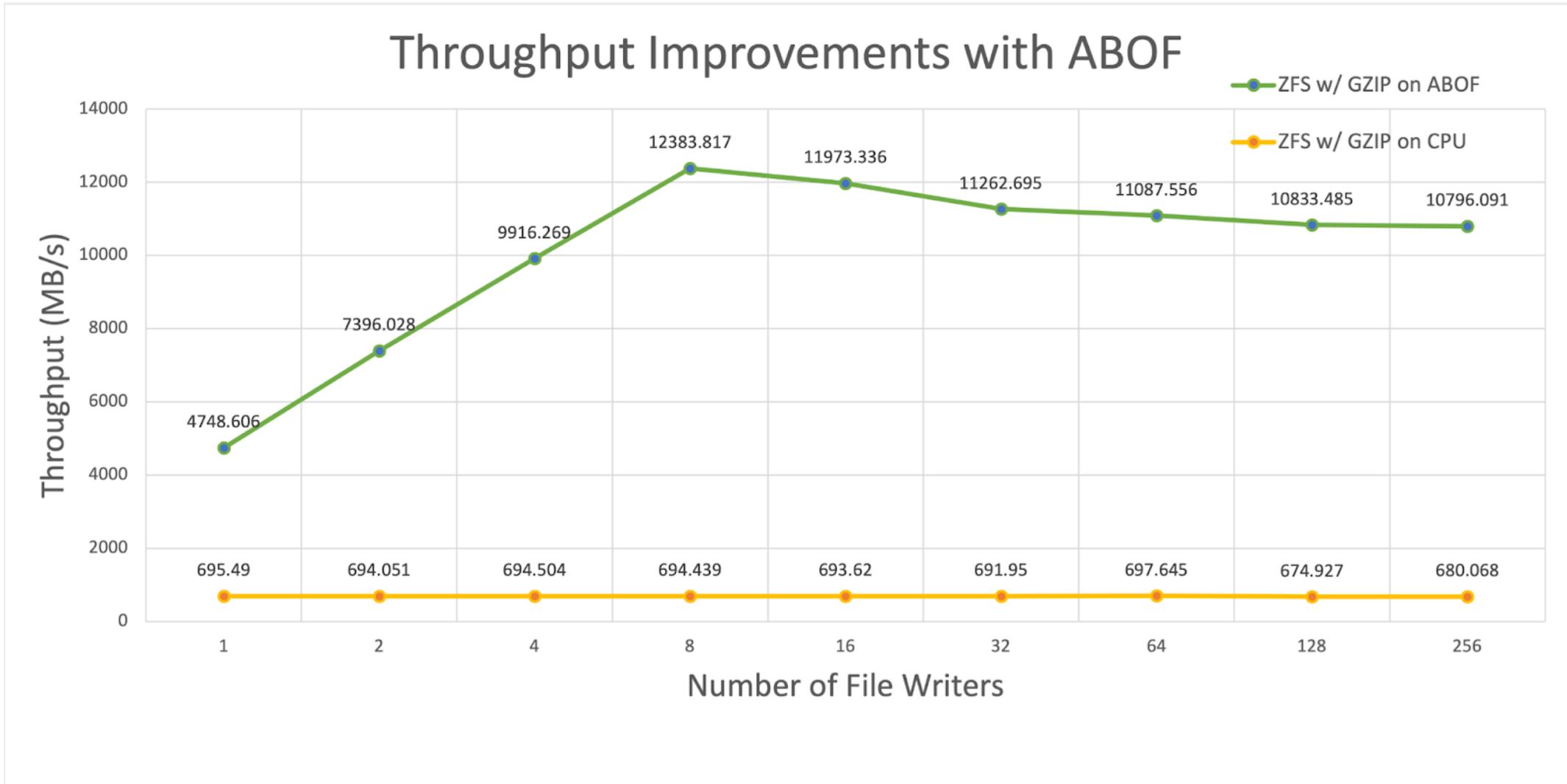- NVMe SSDs

STORAGE DEVELOPER CONFERENCE

SDC 22

Data Movement – Zoff Issue I/O

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

Please take a moment to rate this session.

Your feedback is important to us.

STORAGE DEVELOPER CONFERENCE

SDC 22