Los Alamos
NATIONAL LABORATORY

Supporting Scientific Simulation and Analysis: Computational Storage and Efficient Storage System Design

Dominic Manno

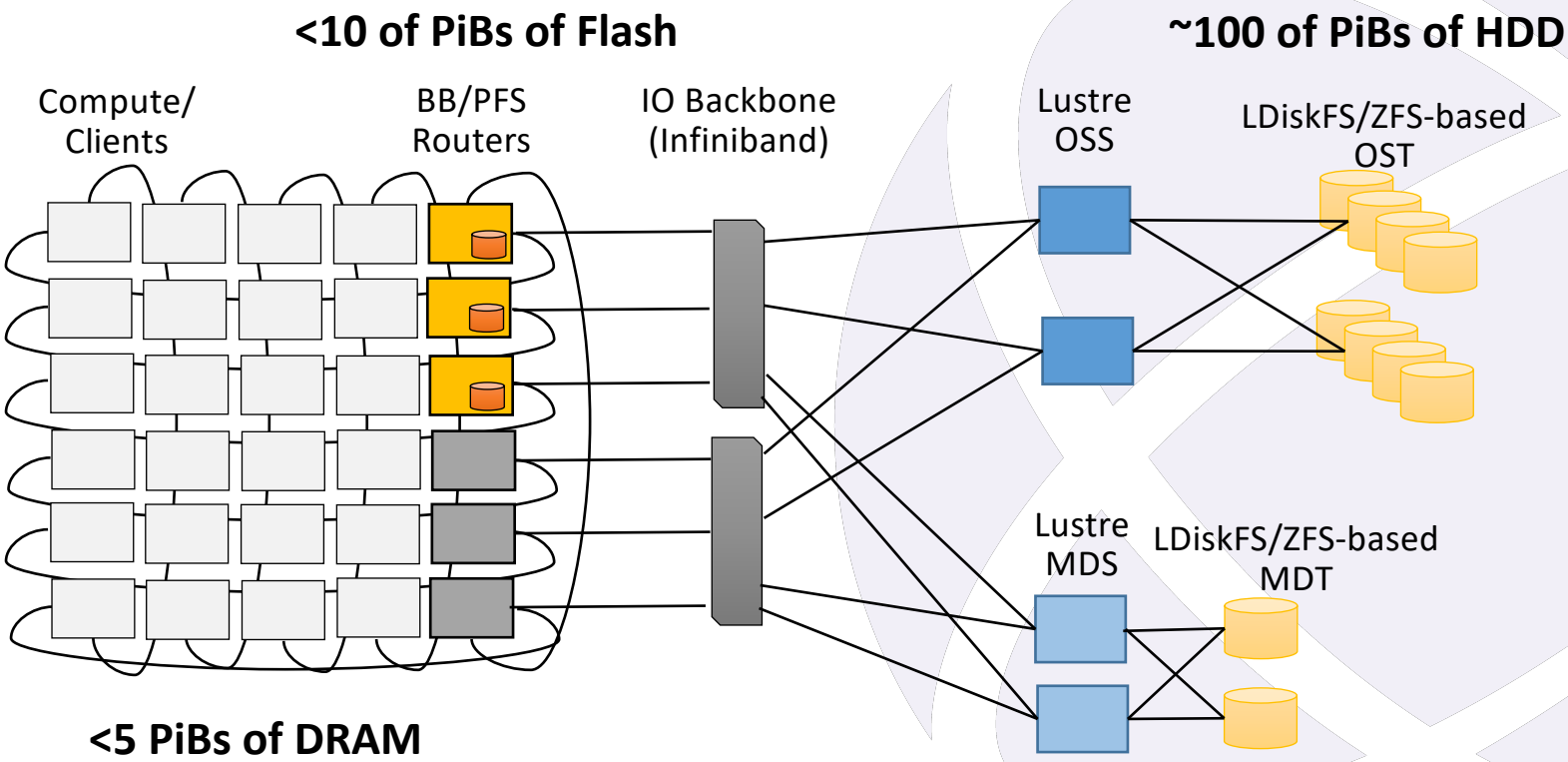
April 26, 2022

LA-UR-22-23340

Outline

- HPC Storage Background
  - Examples
- Current Performance Challenges
  - Review new technologies at play
  - Existing bottlenecks
  - Potential for new designs
- Challenges for emerging workloads
  - Without resorting to islands!

Recent Generation HPC Platform



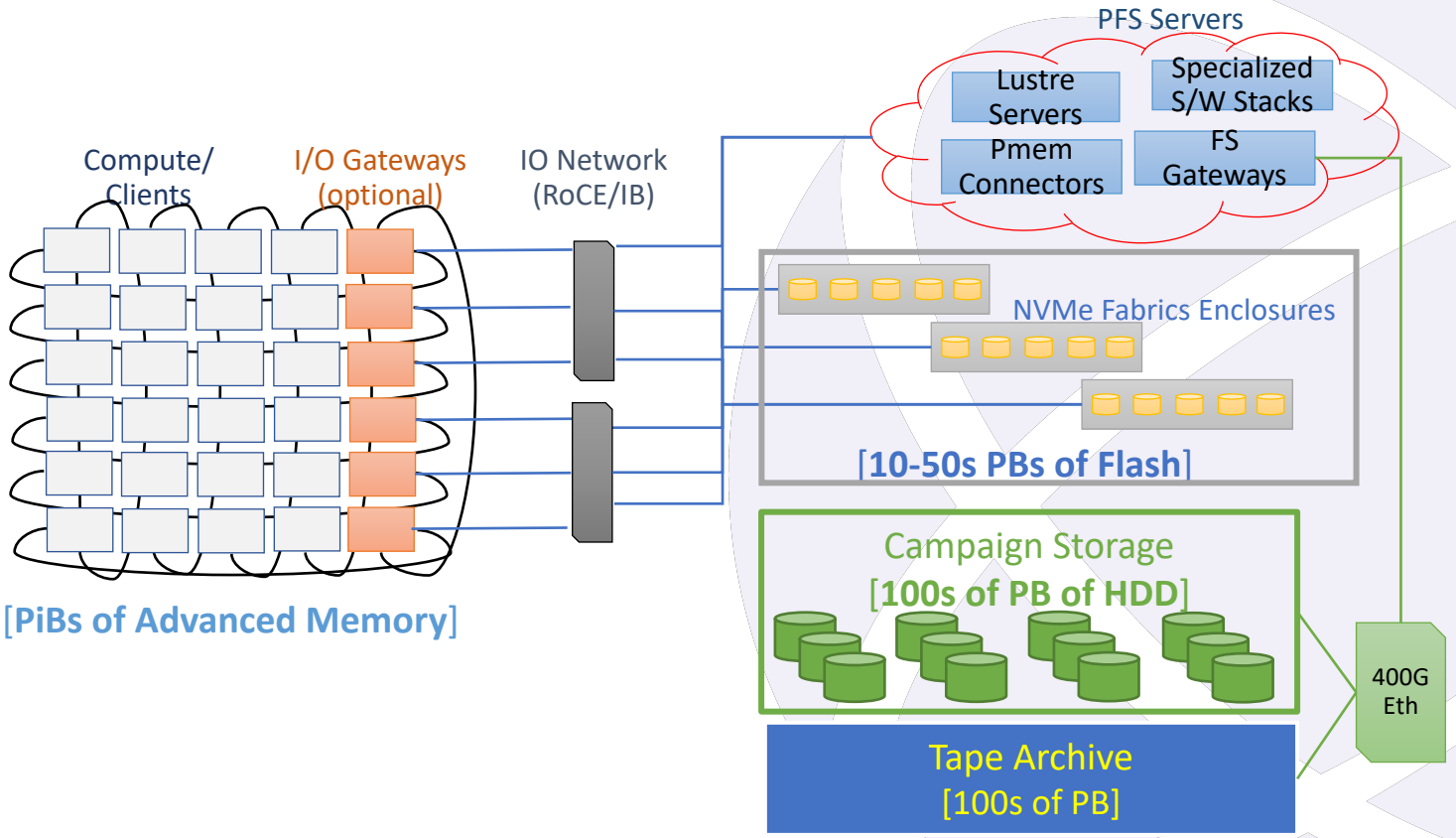
Disaggregated Storage

- Disaggregated Storage might be preferred if
  - Duty cycle is an uncertain mix of reads and writes
  - Storage workload is throughput-oriented
  - Changing capacity over time is necessary
  - Fast interconnection network is available
- Converged Storage (sometimes hyperconverged) might be preferred if
  - Duty cycle is organized into linear load-analyze phases
  - Workload requires large numbers of small IOPs
  - Expense of dedicated storage servers is too great
  - Storage devices faster than interconnection network
- **This is a trade space**

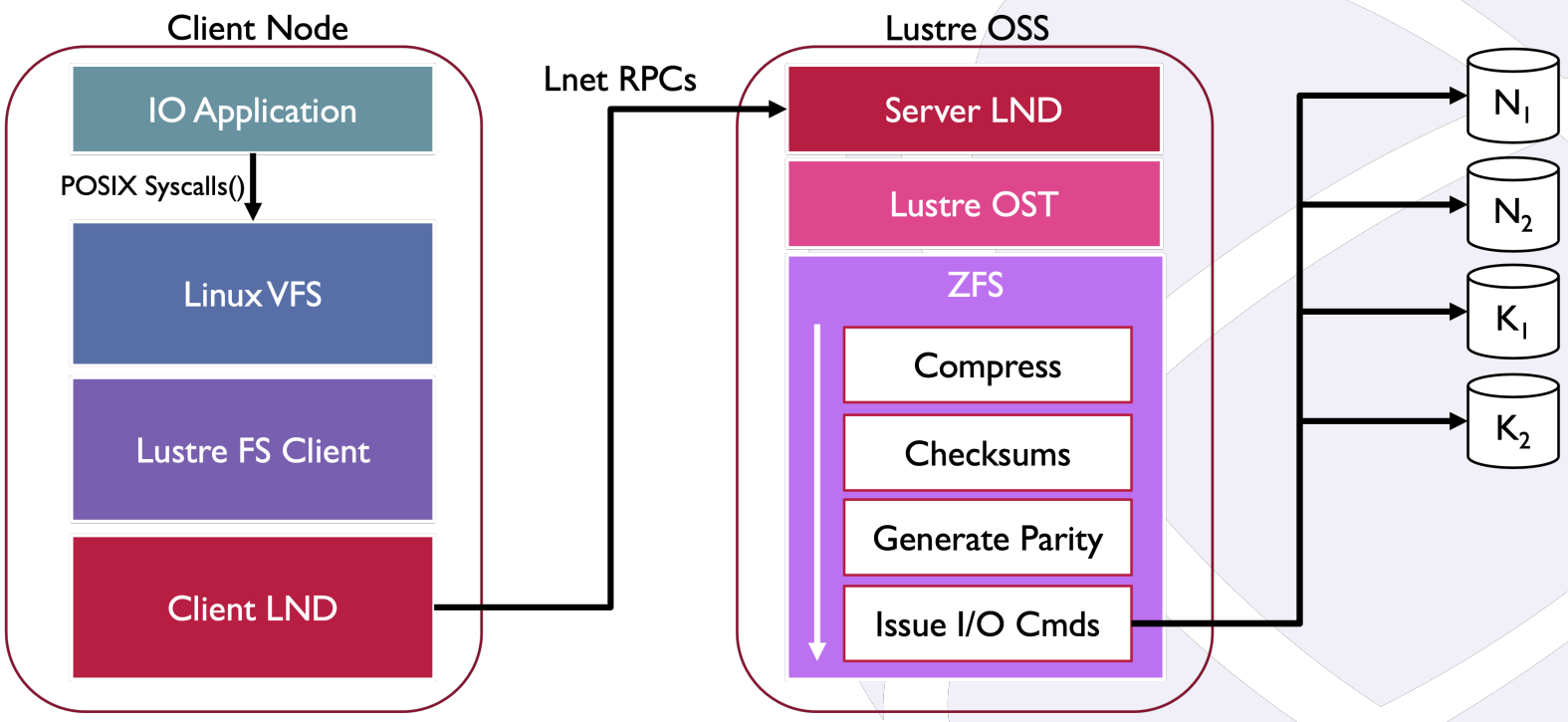
Ongoing Challenges:

Building a fast disaggregate storage endpoint

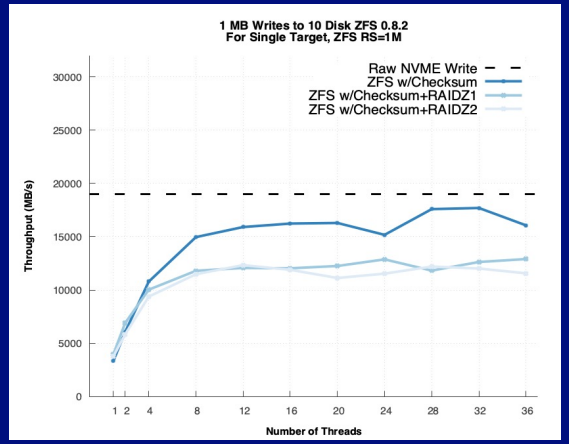
A Next Generation HPC Platform



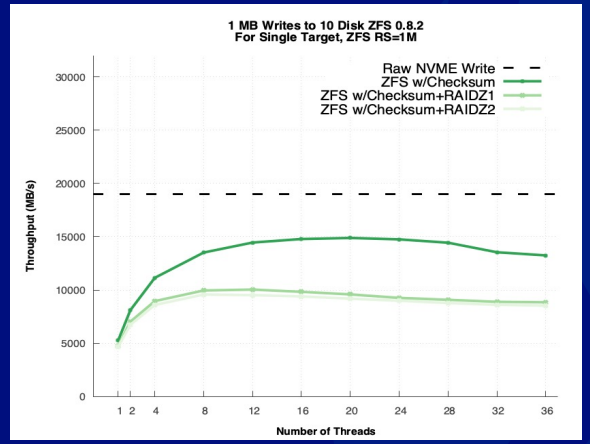
HPC Storage Pipeline



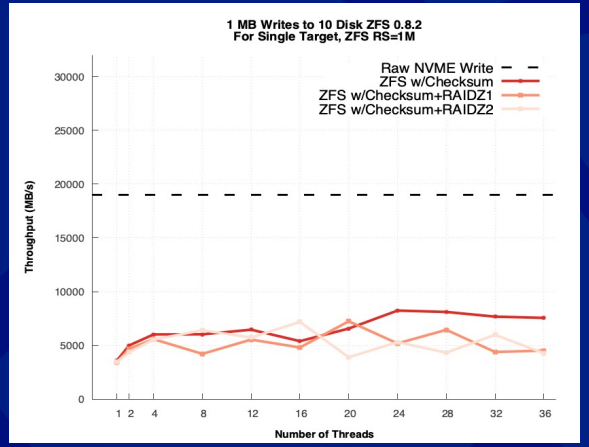
File System Services Need Help First



Intel Platinum (Dual Socket)



AMD EPYC (2nd Gen)

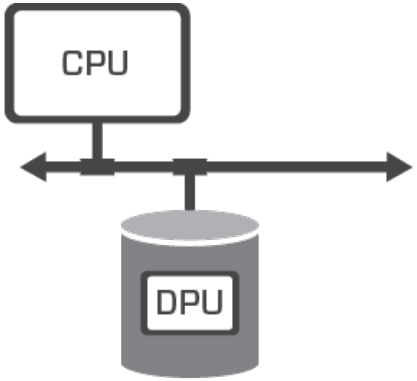


AMD EPYC (1st Gen)

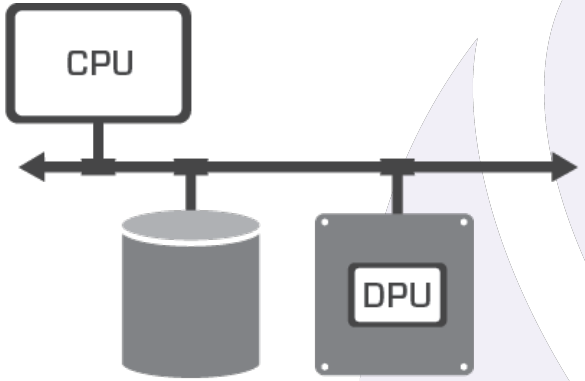
Data Processing Units Can Help!

- Tailored hardware design
  - Storage servers idle much of the capability provided by server CPUs
  - Even more so when waiting on memory subsystem
- Emerging processor architectures
  - Hardware EC, compression, checksum, …
  - PE gets trimmed down
  - This is economical using today's costs
- Data agnostic offloads

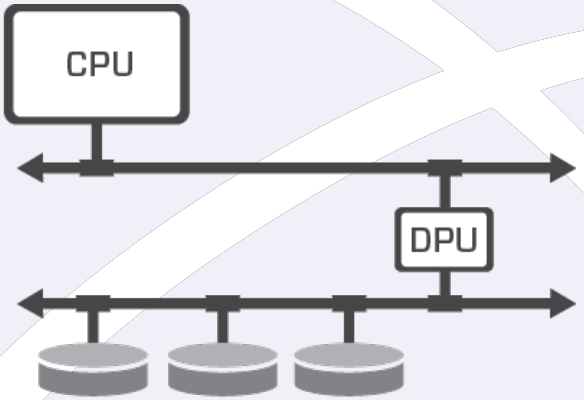
Integrating DPUs Into Storage



Computational Storage Device (CSD)

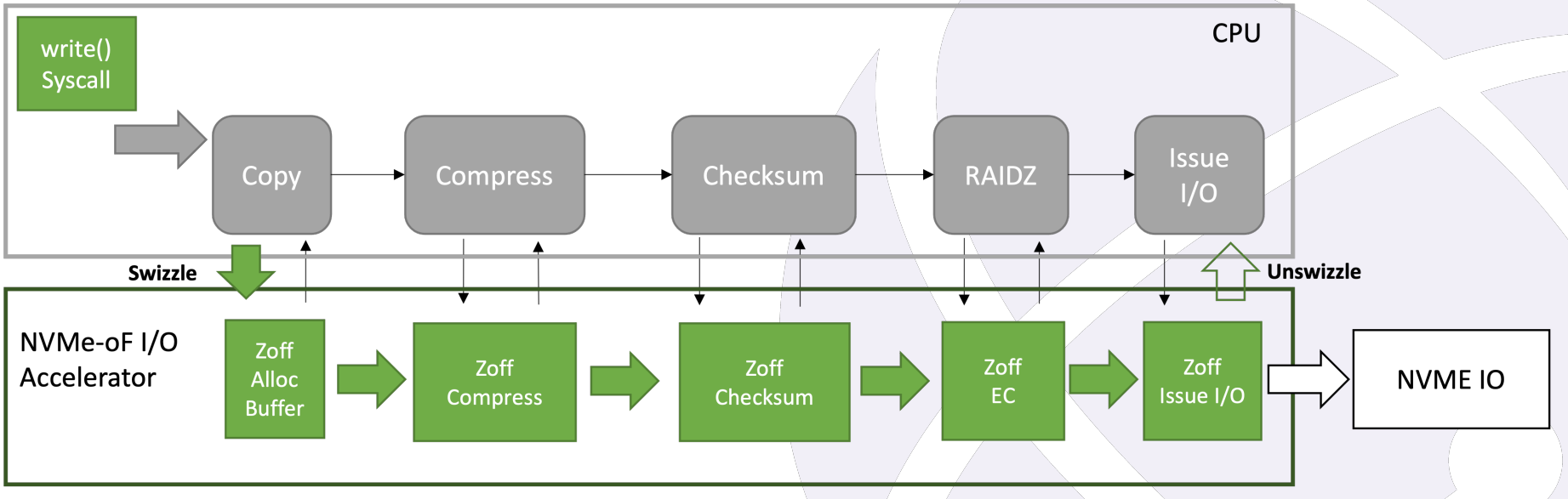


Computational Storage Processor (CSP)

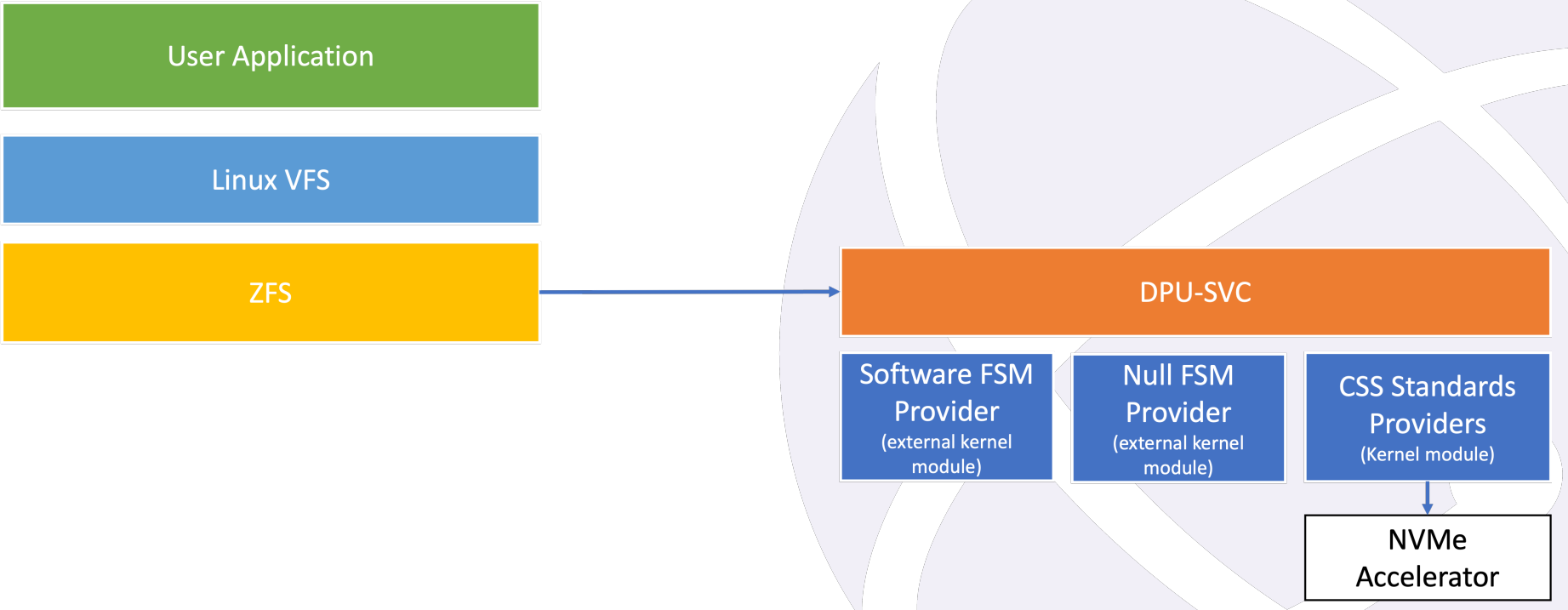


Computational Storage Array (CSA)

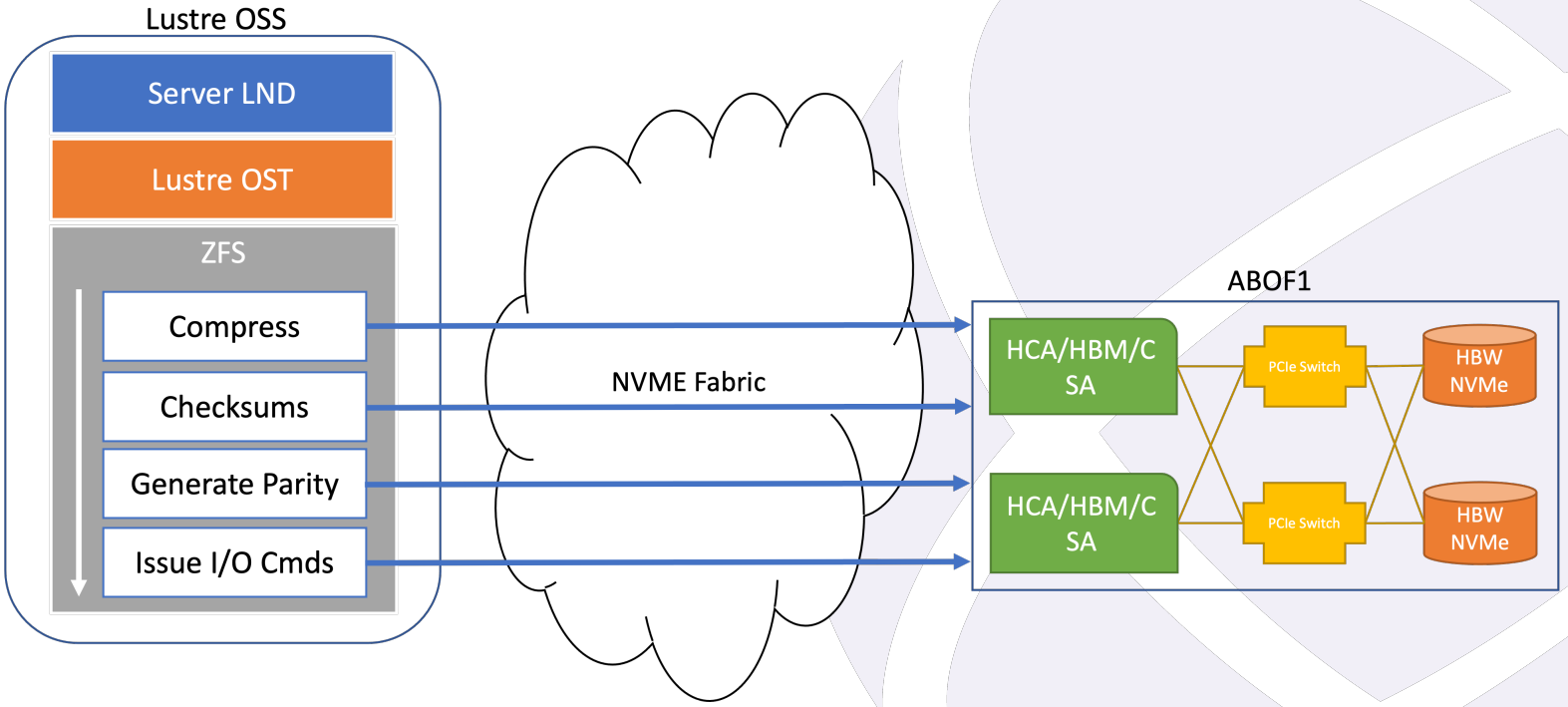
ZFS Interface for Accelerators



Data Processing Unit Services Module



ABOF Theory of Operations



Promising Results

- Performing high-value compression 7x faster than high-end processor
  - Compression increases from 1.06:1 to 1.26:1
- Ability to reconfigure write path acceleration to dedicate more resources
  - Think high priority rebuild
- End-to-end results are proving that placement decision of offloads is essential
- Tuning performance for EC is ongoing

Enabling New Workloads:

Accelerating Small, Random I/O with Disaggregated Storage

Non-traditional HPC Workloads

- Supporting small,random I/O emerging workloads
  - Random sample data for training
  - Random sub-graphs for inference
- The physical media we purchased to solve the throughput problem is great at these workloads too
  - SW not so much
- Leverage computational resources along the data path
  - Data aware offloads
- Leverage design and technological features (NVMe namespaces) to allow for the exploration of emerging technologies (KV, new FS software, etc)
- Enable analysis in multiple tiers (scratch and campaign)

Other Interesting Problems:

Data Management

Data Management Topics

- Metadata indexing
  - Providing high performing query capability to users and storage administrators
- Capacity management
  - Migrating intermediate data becomes more important than ever
  - Simplify this process and tooling

Wrap-up

- Our storage designs evolve with ever changing tradeoffs
  - Generally impacts old assumptions
- New technology drives these changes
  - NVMe is a transformational technology
  - Storage, network performance outpacing memory bandwidth
- Iterative approach
  - Make sure we don’t break what works today, hopefully improve upon it
  - Enable, explore, and integrate