Los Alamos NATIONAL LABORATORY

Leveraging **Computational Storage** for **Simulation Science** **Storage** System Design

Qing Zheng, Scientist, Los Alamos National Laboratory

3/4/2024

LA-UR-24-21992

Today's Agenda: 2 Computational Storage Projects @ LANL

ABOF (Eideticom, Aeon, Nvidia, SK hynix)

H/W accelerated ZFS write pipeline

KV-CSD (SK hynix)

H/W accelerated KV storage

OCS (Versity, SK hynix, Airmettle, Neuroblade)

H/W accelerated columnar data lake

Tomorrow's Talk

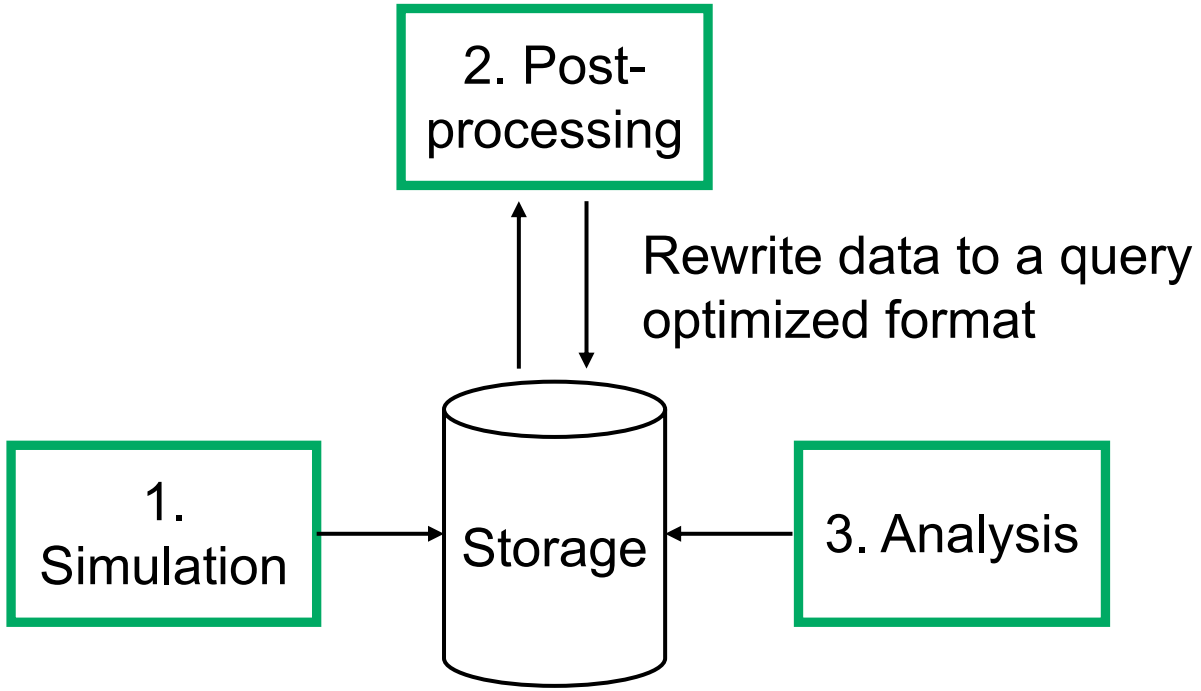
Background image generated by Bing AI Image Creator

Background: HPC Simulation Workflow

A 3-step process: simulation, **post-processing** (may be skipped), and **analysis**

Performance maximized when:

- Storage bandwidth fully utilized during data insertion
- Data transfer minimized during analysis (especially when query selectivity is high)
- Lowest possible data post-processing overhead

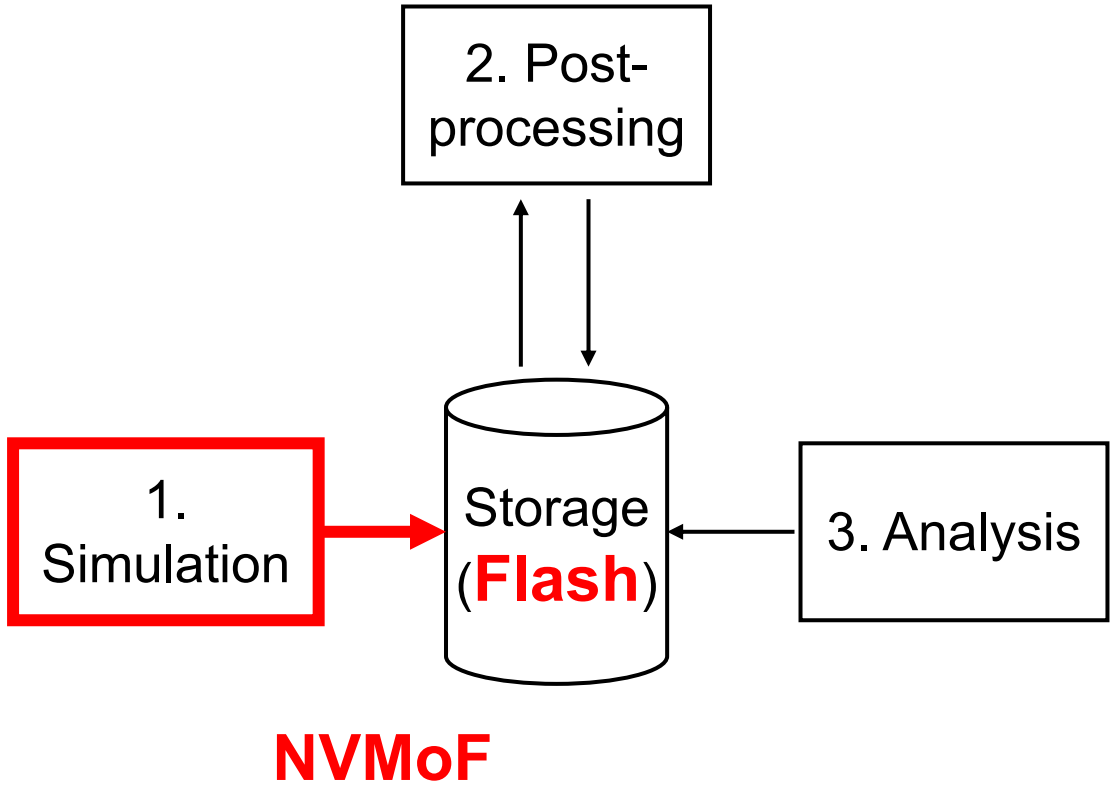


Today's HPC data centers are having problems achieving any of these

Part I – ABOF: Accelerated Box of Flashes

Problem: Today's host CPU fails to compress data as fast as storage can absorb it

- Compression necessary for frugal use of SSD storage space
- High-entropy scientific data requires heavy compression methods (such as gzip)
- CPU-only processing prevents apps from fully utilizing available SSD bandwidth



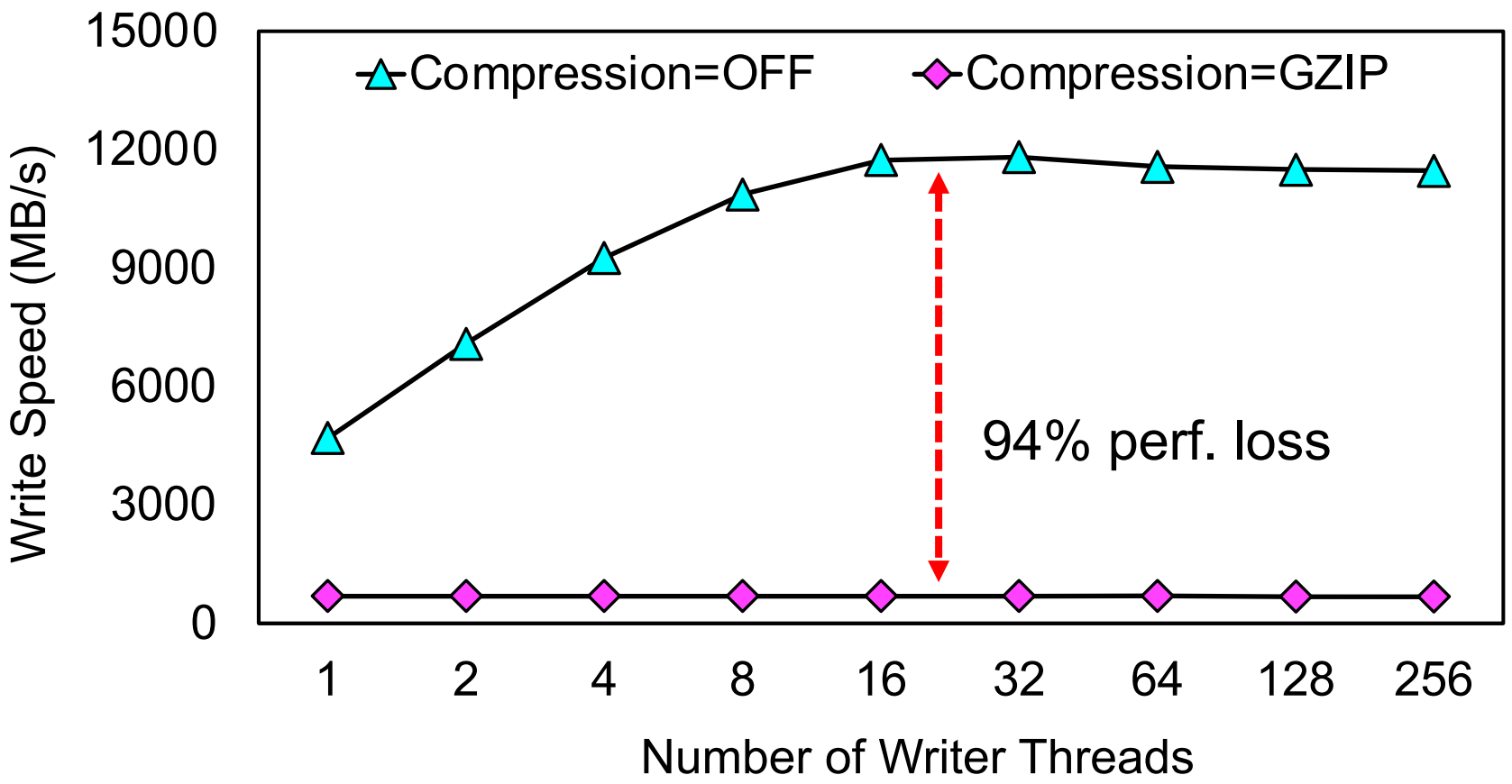
Impact of Shifting to All Flash

	Trinity		Crossroads
	2016		2022
Memory	2PB		0.75PB
Platform Storage	78PB	→	10PB
Platform Storage B/W	1.45TB/s		1.29TB/s

Crossroads does not excel at FLOPs or memory capacity. It excels at memory bandwidth not shown in the table

Under equal bandwidth, flash yields much lower capacity than HDDs, making data compression necessary

Up to 94% Perf. Loss When Compression is ON

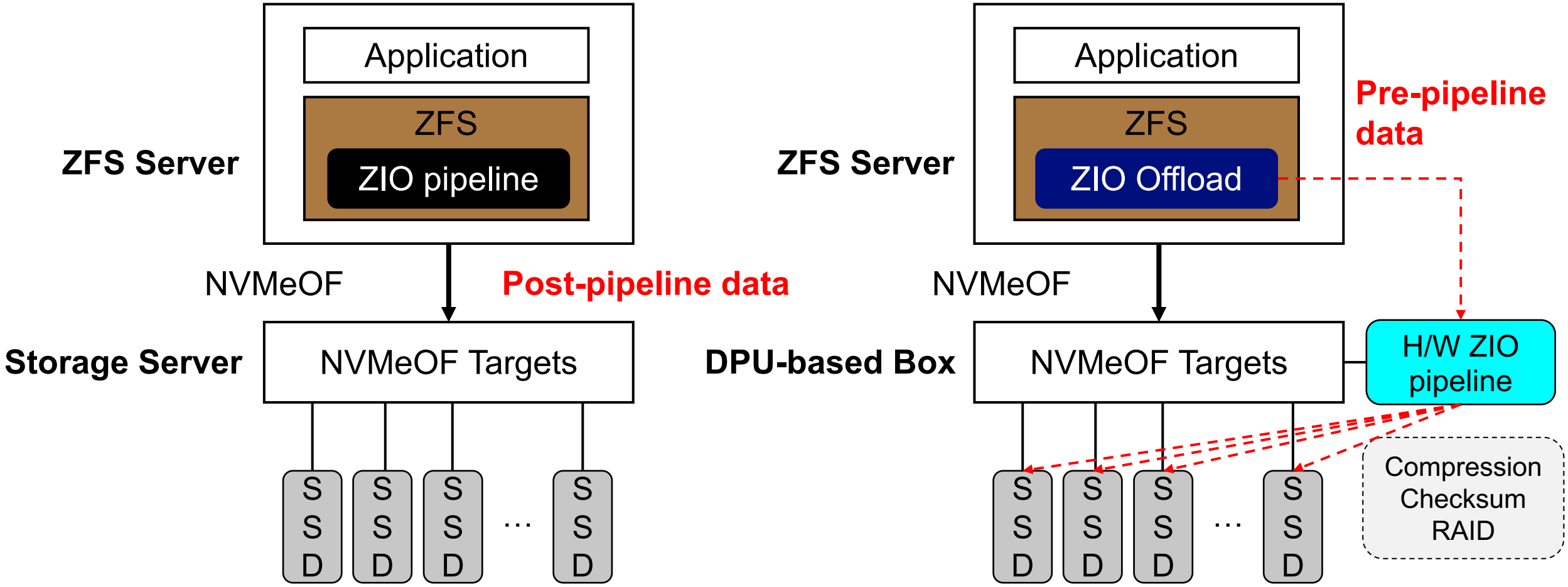


Streaming data into a 10+2 all-flash ZFS pool

- Concurrent 1MB writes to a single file
- 1 ZFS host
- 12 NVMeOF flash SSDs

ABOF: offload compression to dedicated FPGA/ASIC on the storage I/O path to overcome host bottlenecks

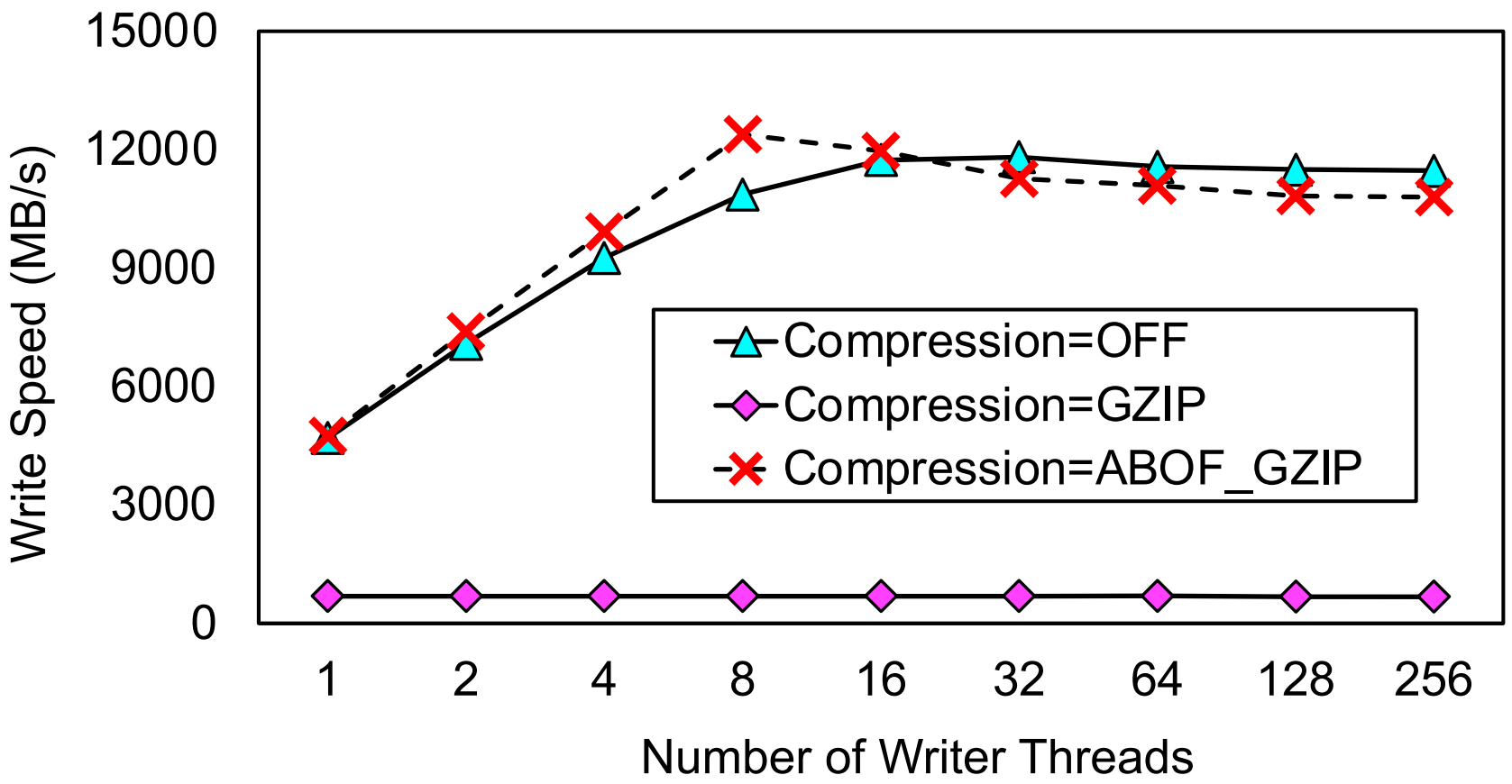
ABOF: Accelerated ZFS Writes



ZFS Software Pipeline Execution

Hardware Accelerated Pipeline

Result: GZIP at Line Rate

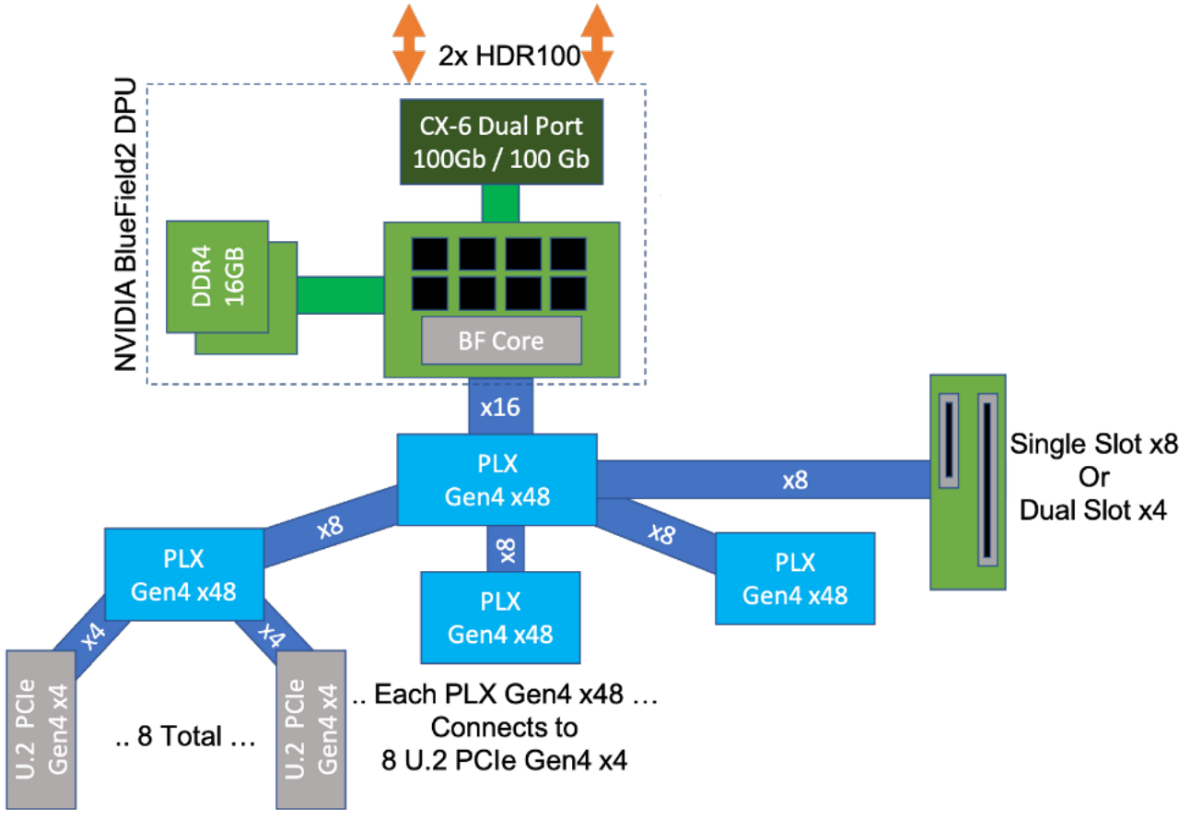


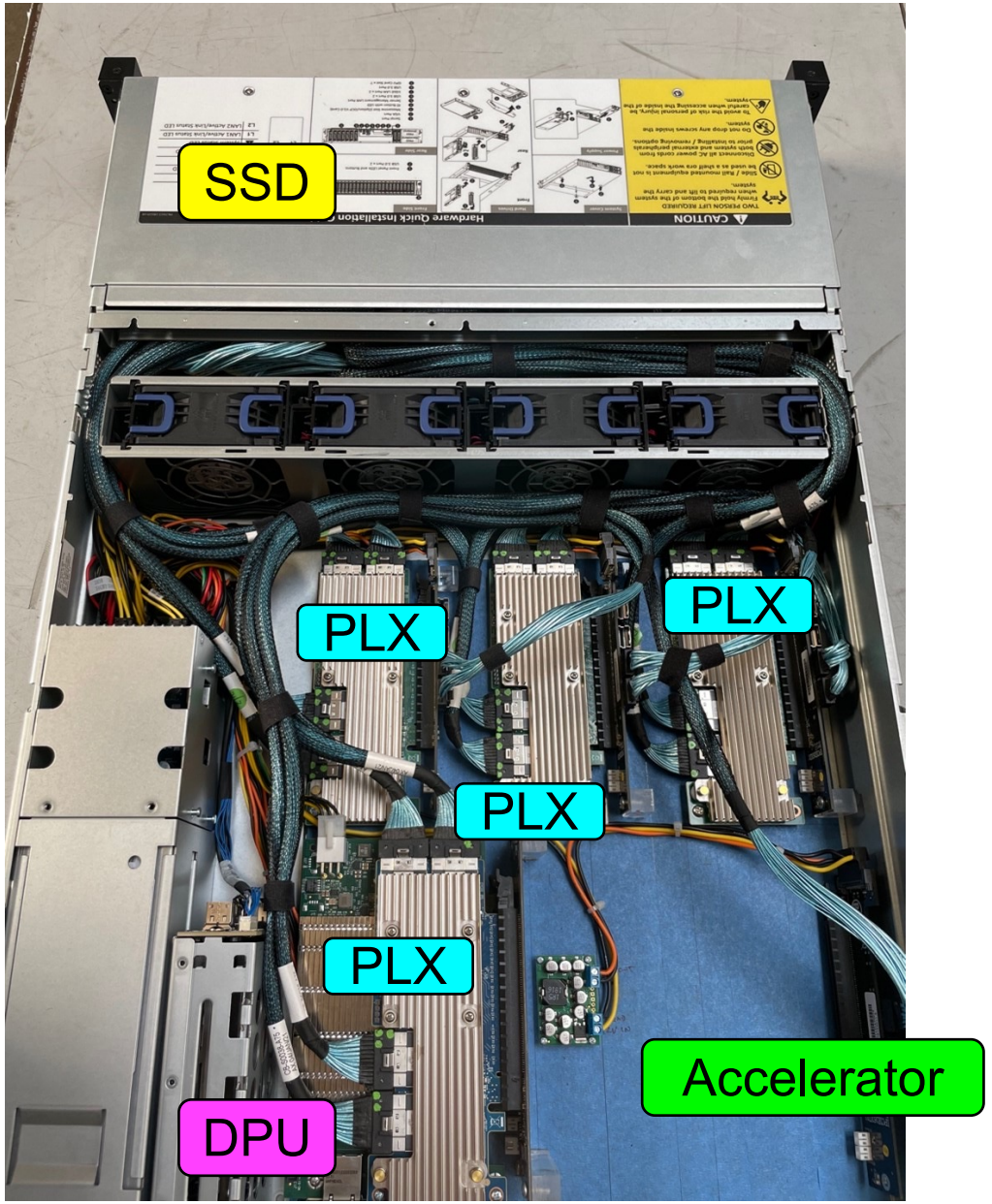
Streaming data into a 10+2 all-flash ZFS pool

- Concurrent 1MB writes to a single file
- 1 ZFS host
- 12 NVMeOF flash SSDs

16x faster than CPU based gzip, comparable with writing raw data

ABOF in Real World





This is the DPU box we mentioned
2 slides before

Final Note: Direct I/O

ABOF alone is **insufficient** to enable full media bandwidth utilization

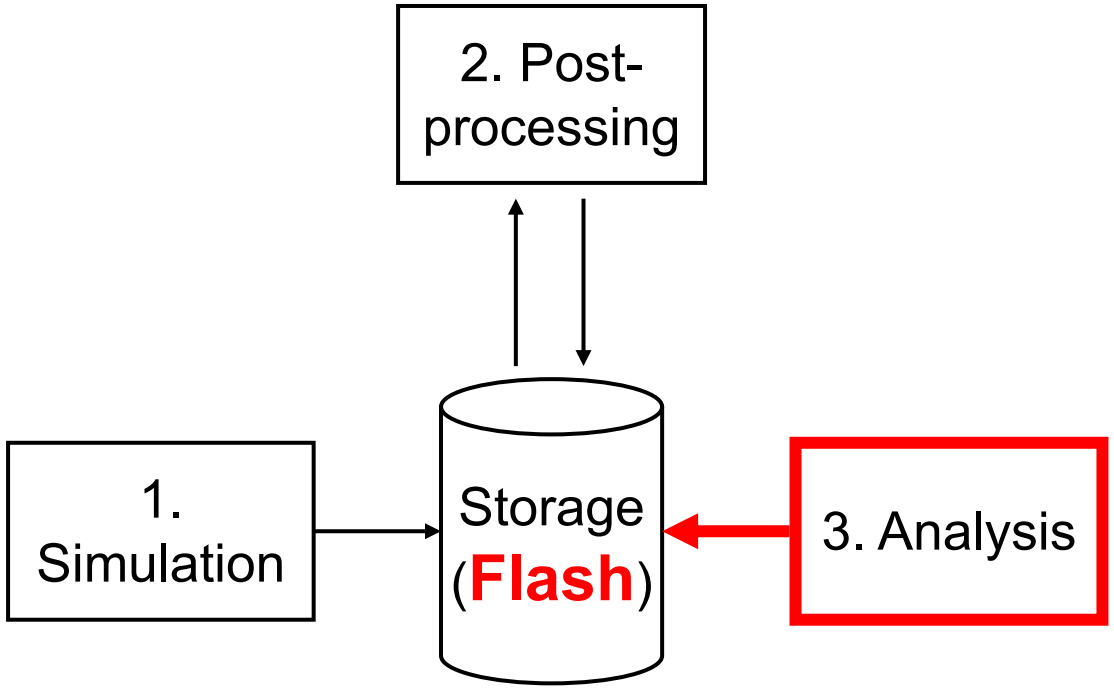
Needs to be combined with direct I/O to reach its full potential

See Brian's OpenZFS developer summit talk for details

Part II – KV-CSD: KV Computational Storage Device

Problem: Scientific analysis is often slowed down by unordered, unindexed data access

- Scientific apps write data without necessarily considering the performance of the queries that follow
- Data may not be persisted in the same order as queries, leading to full data scans
- Pre-sorting data prior to queries is time consuming

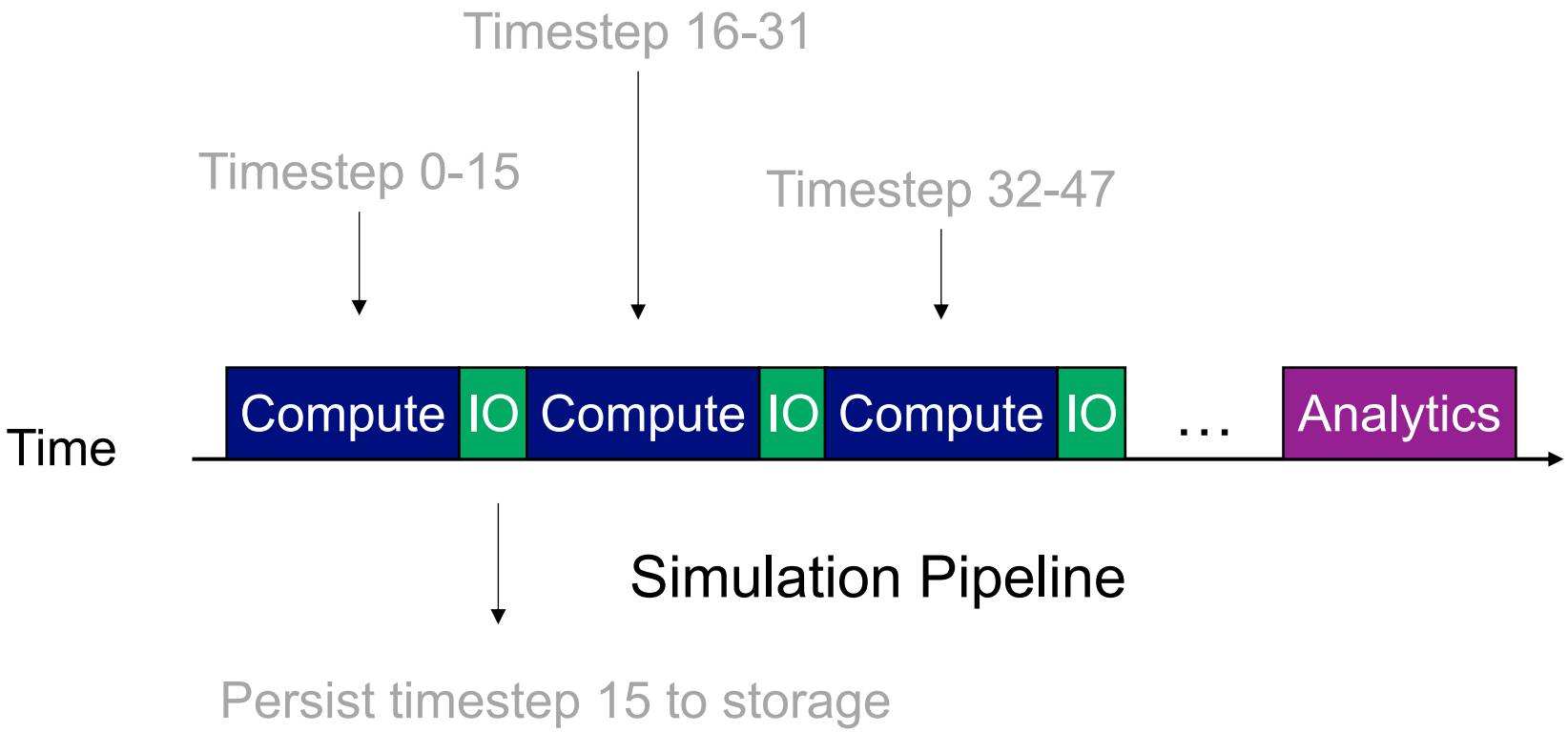


How Scientific Simulations Run

Time based bulk-synchronous parallel programs

Iterate between compute & I/O phases

Analytics occur after simulation

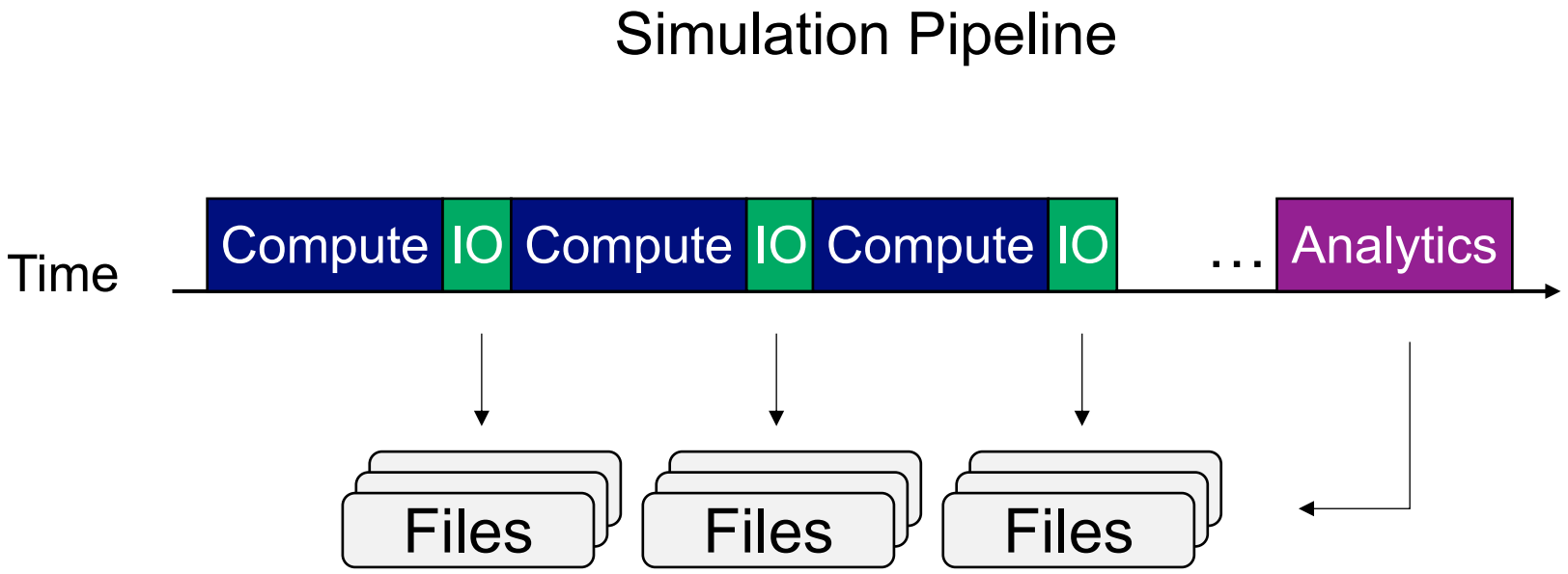


How Data is Stored Today

Through filesystems

Data stored as one big or many small files per timestep

Analysis may have to scan all files



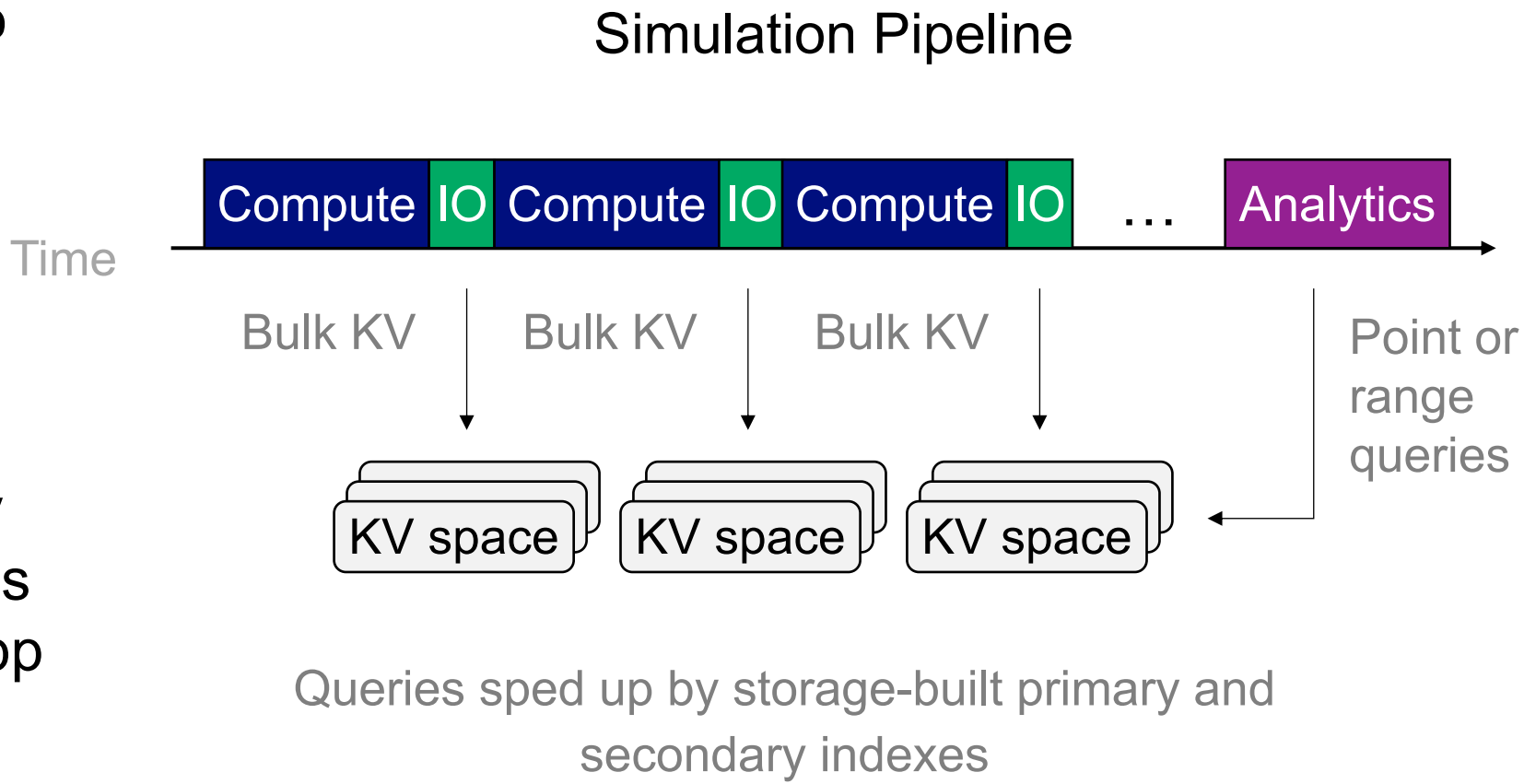
Full data scans

Toward Ordered, Computational KV Storage

App converts data to KV pairs and bulk inserts them into storage

One KV space per app process per timestep

Storage sorts data by key asynchronously and builds secondary indexes per app query needs



Why Hardware Acceleration?

Software KV stores (such as RocksDB) rely on background processing to hide data sorting latency

Insertion is suspended when background jobs cannot keep up

Hardware acceleration allows for more aggressive latency hiding

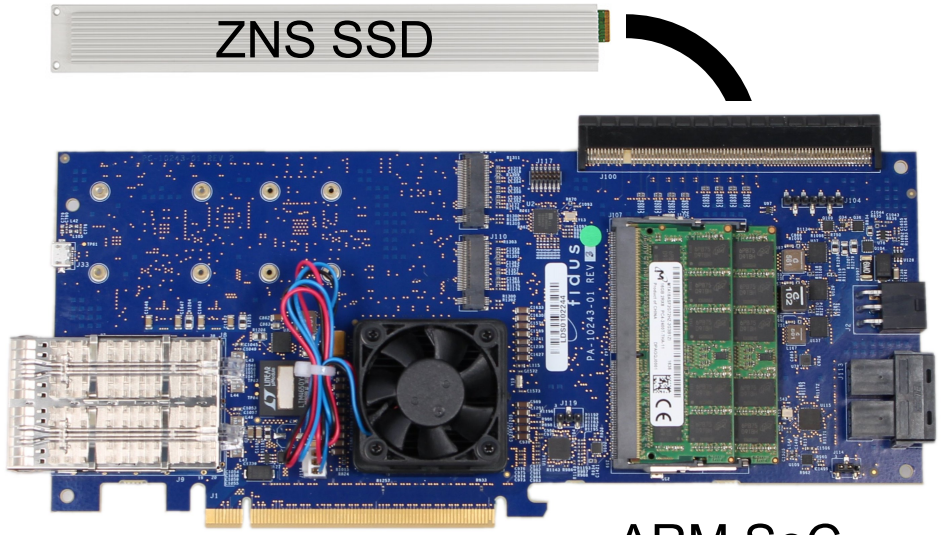
By deferring background work until after insertion concludes and by performing it within a computational storage device

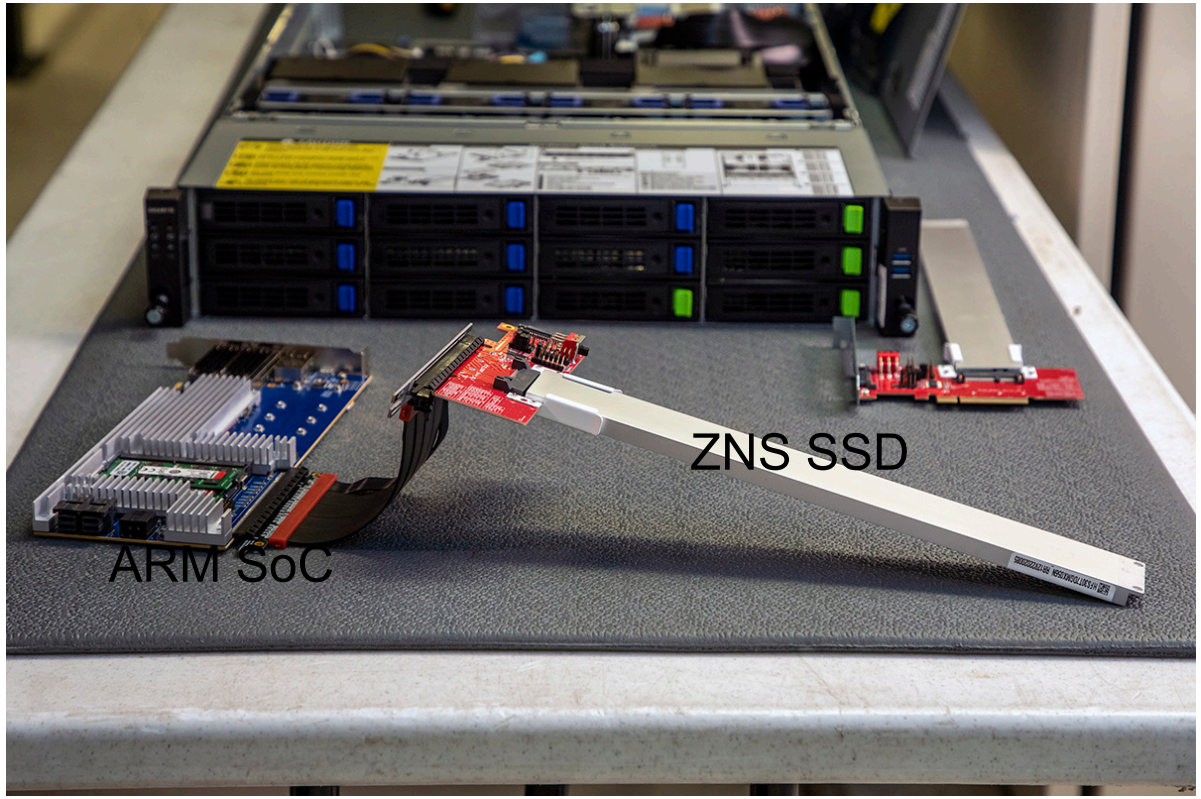
Results

	Filesystem (Baseline)	RocksDB (Software KV)	KV-CSD (Hardware KV)
Simulation I/O Path	Fast	Slow	Fast
Analytics Path	Slow	Fast	Fast

KV-CSD in Real World

Current Prototype





Quick Recap:

ABOF (Eideticom, Aeon, Nvidia, SK hynix)

H/W accelerated ZFS write pipeline

KV-CSD (SK hynix)

H/W accelerated KV storage

OCS (Versity, SK hynix, Airmettle, Neuroblade)

H/W accelerated columnar data lake

Tomorrow's Talk

Background image generated by Bing AI Image Creator

Conclusion

Large-scale data analytics is a core element of scientific discovery

Computational storage provides new ways of accelerating data-intensive analytics workloads

Preliminary results are promising

More work/collaboration/integration is needed for production deployment

Background image generated by Bing AI Image Creator