SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

Improving Storage Systems for Simulation Science with Computational Storage

Dominic Manno
Los Alamos National Laboratory

Background: HPC Scientific Simulation Systems

Trinity – circa 2016

- Haswell and KNL
- 20,000 Nodes
- Few Million Cores
- 2 PByte DRAM
- 4 PByte NAND Burst Buffer ~ 4 TByte/sec
- 100 Pbyte Scratch PMR Disk File system ~1.2 TByte/sec
- 60PByte/year Sitewide Campaign Store ~ 50 GByte/sec
- 60 PByte Sitewide Parallel Tape Archive ~ 3 Gbyte/sec

I know its not Tier1 sized but at LANL its for one job for several years.
10 PB files and 200 PB Campaigns
For a single user/small user team

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Topics: Crawl, Walk, Run - with much help from our partners!

- ABOF 1.0 (Eideticom, Aeon, Nvidia, SK hynix)
  - Format agnostic operations (compression, erasure, encoding)
- DeltaFS->Ordered KV-CSD (CMU and SK hynix)
  - Format aware, record-oriented applications with a single-dimension, easily shard-able indexing
- ABOF 2.0 plans (Eideticom, Aeon, Nvidia, SK hynix, others?)
  - Format aware, column-oriented applications, multi-dimension, difficult to shard indexing
- Data-analysis in disk tier (Seagate)

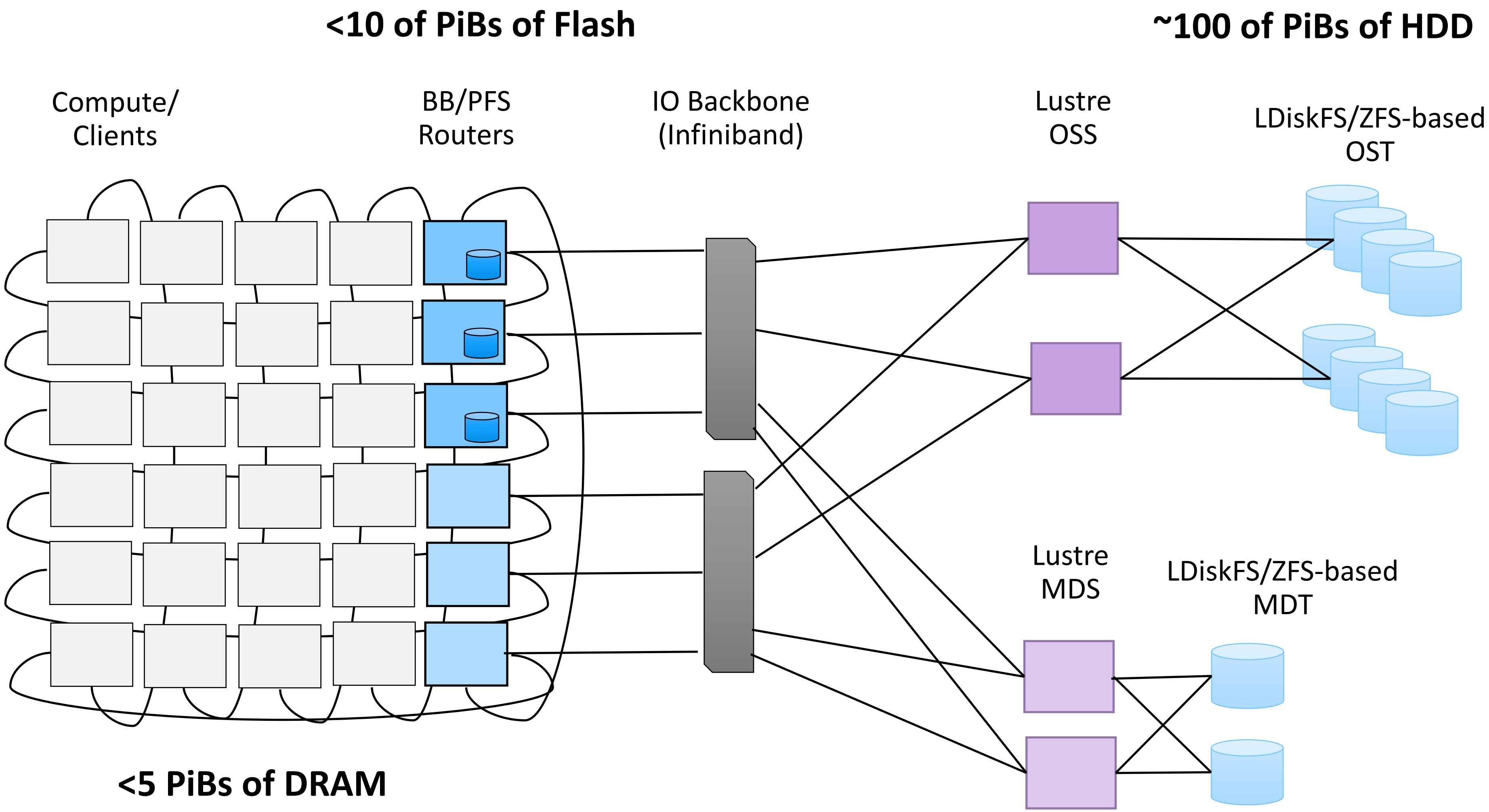
 © SNIA. All Rights Reserved.

ABOF 1.0

- **Format agnostic operations**: compression, erasure, encoding
  - Eideticom, Aeon, Nvidia, SK hynix

 © SNIA. All Rights Reserved.

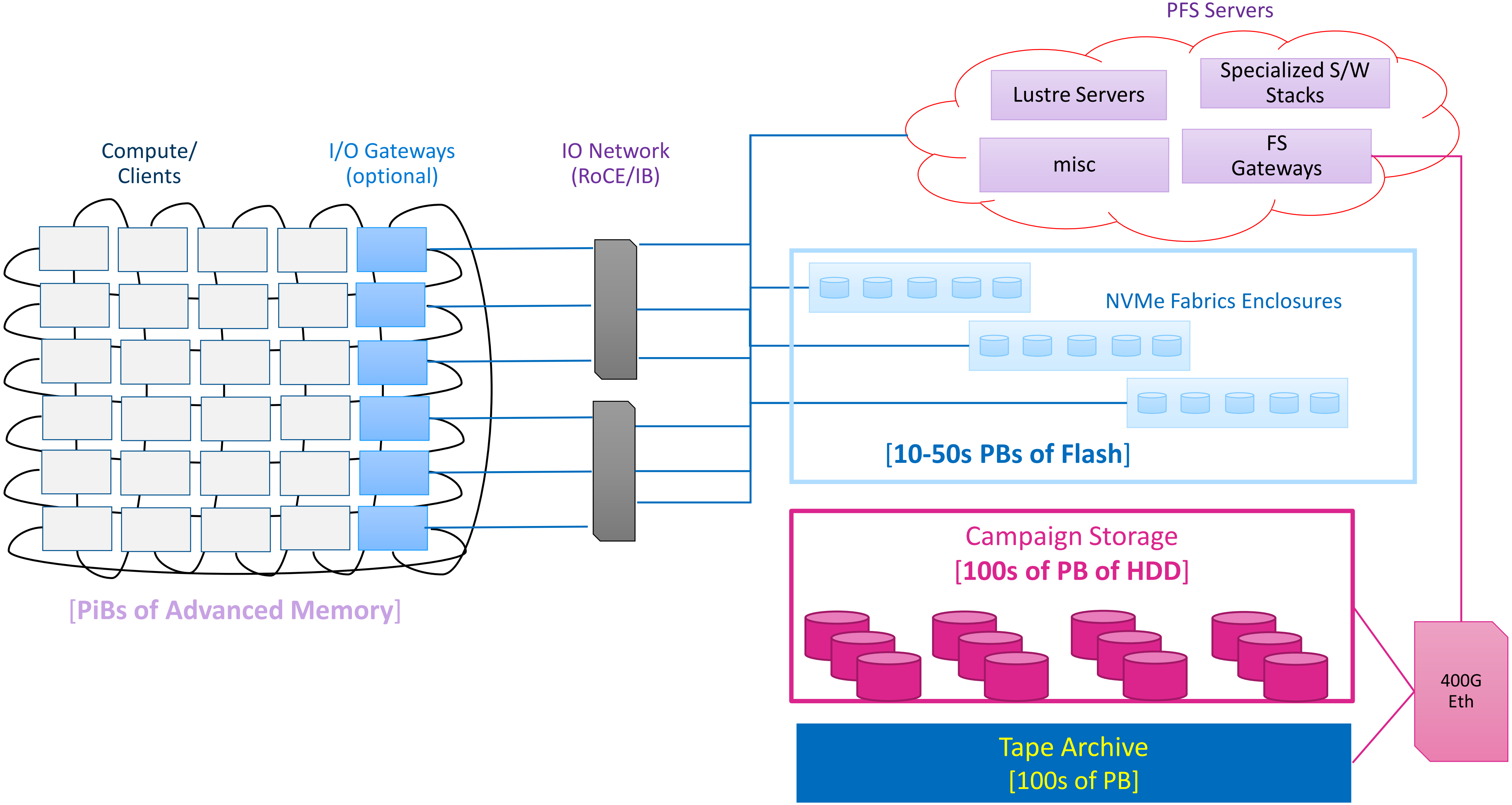
Traditional HPC Storage



 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Redesign Opportunity



 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

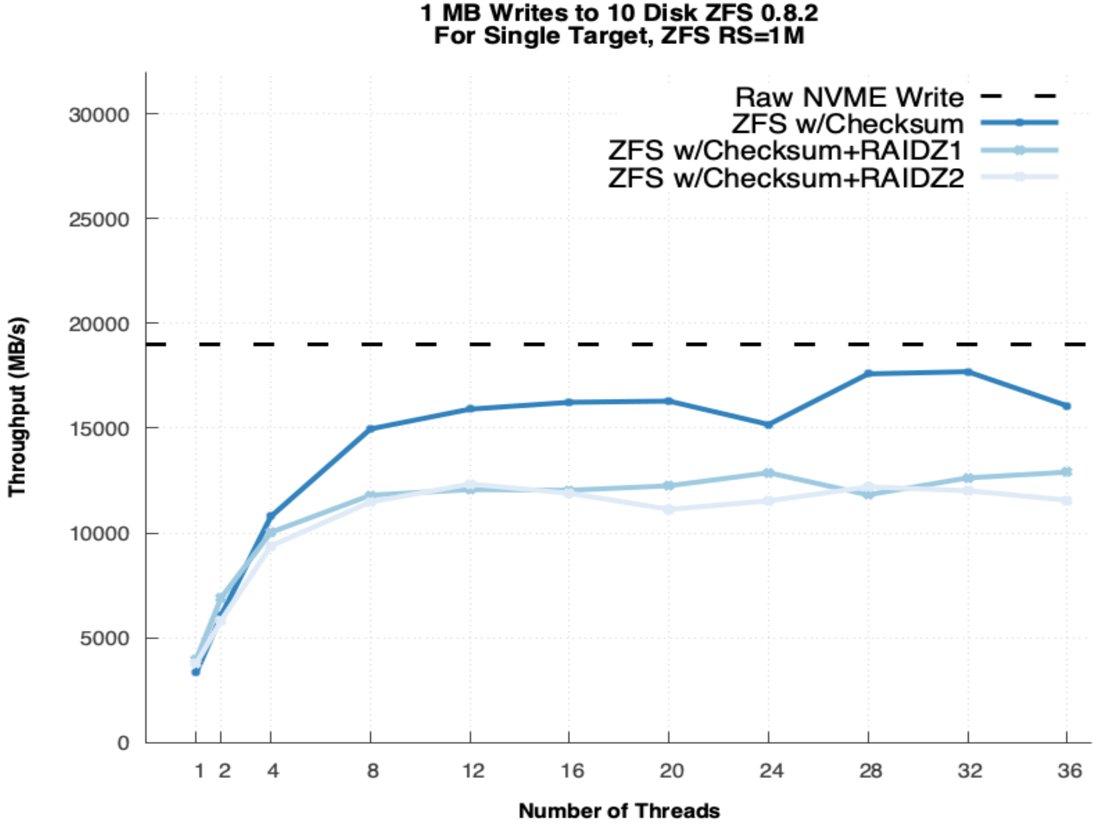
All Flash File Systems

- 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!

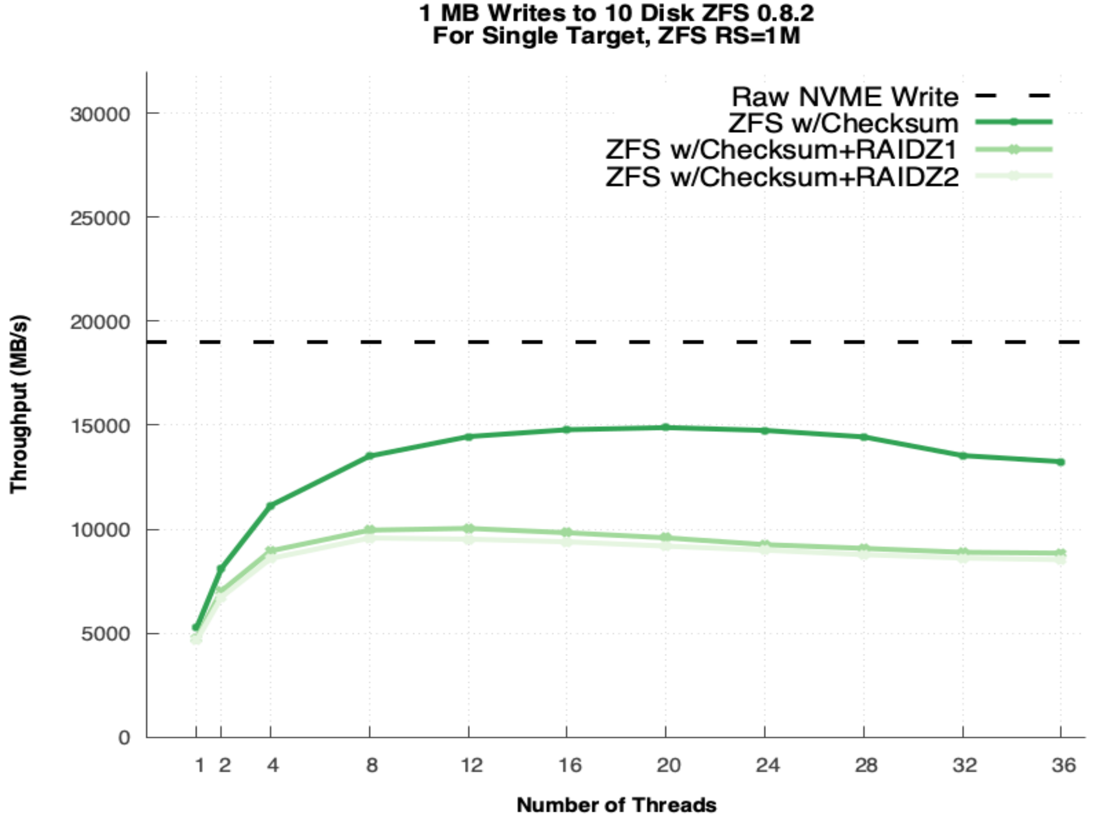
 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Why Offload? ZFS Checksums, Erasure, Compressive
Server memory bandwidth is problematic and expensive



Intel Platinum (Dual Socket)

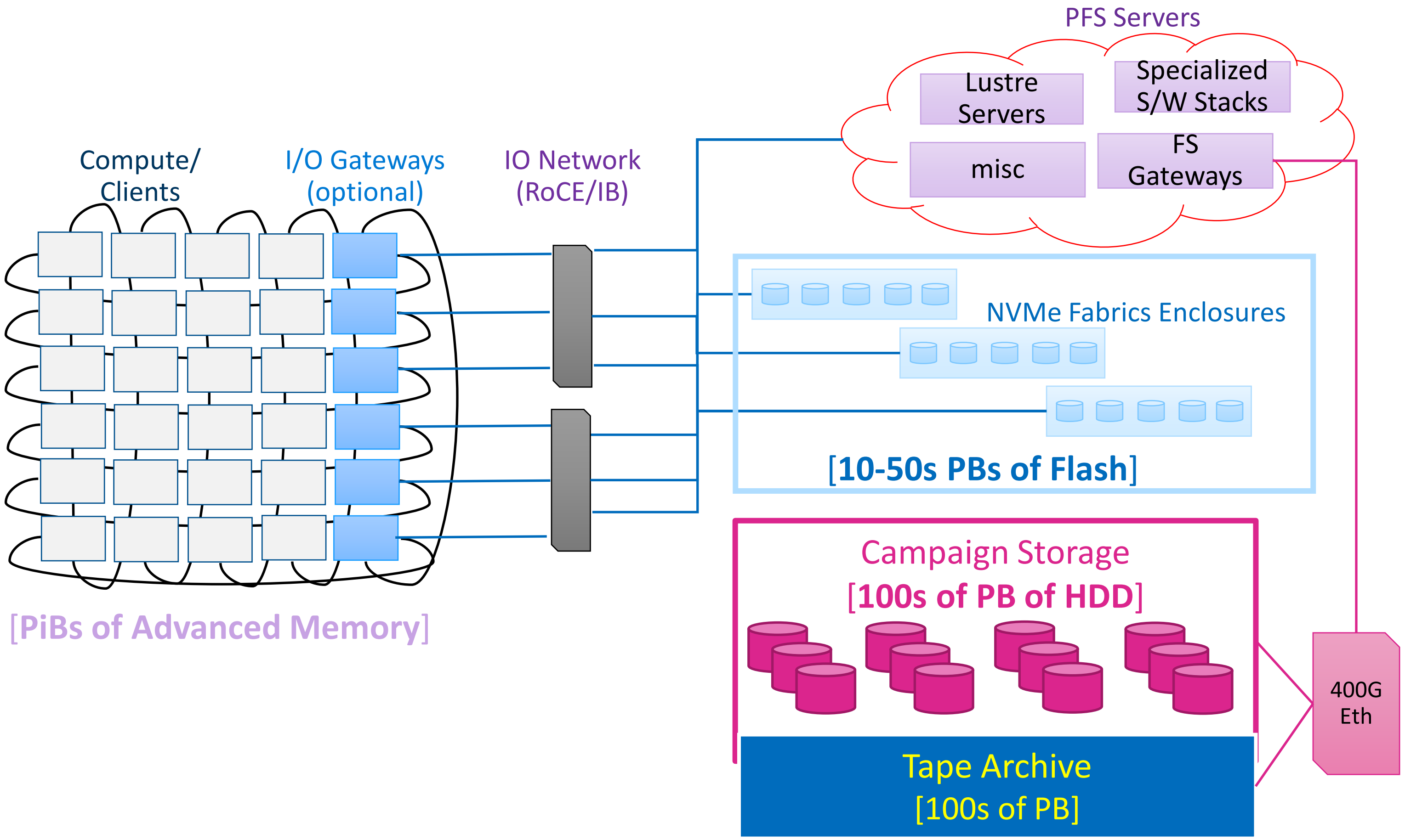


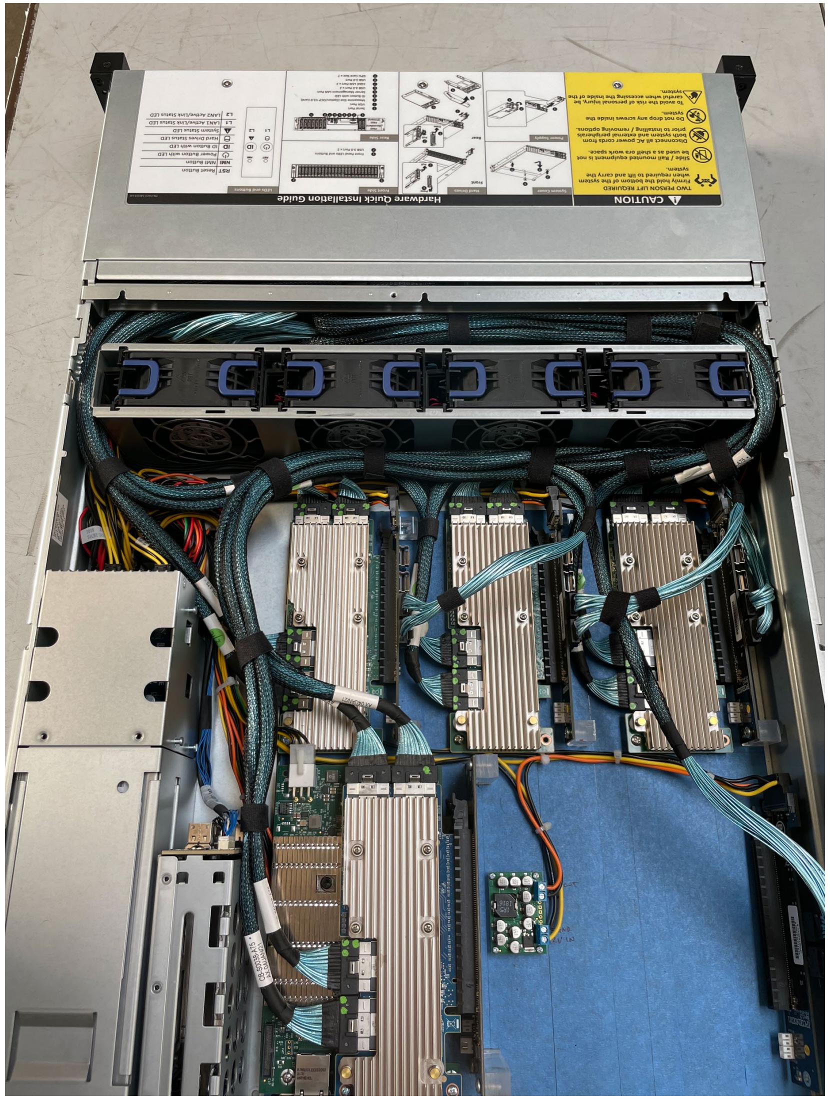
AMD EPYC (2nd Gen)

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

How to Consume: File System Services Offload

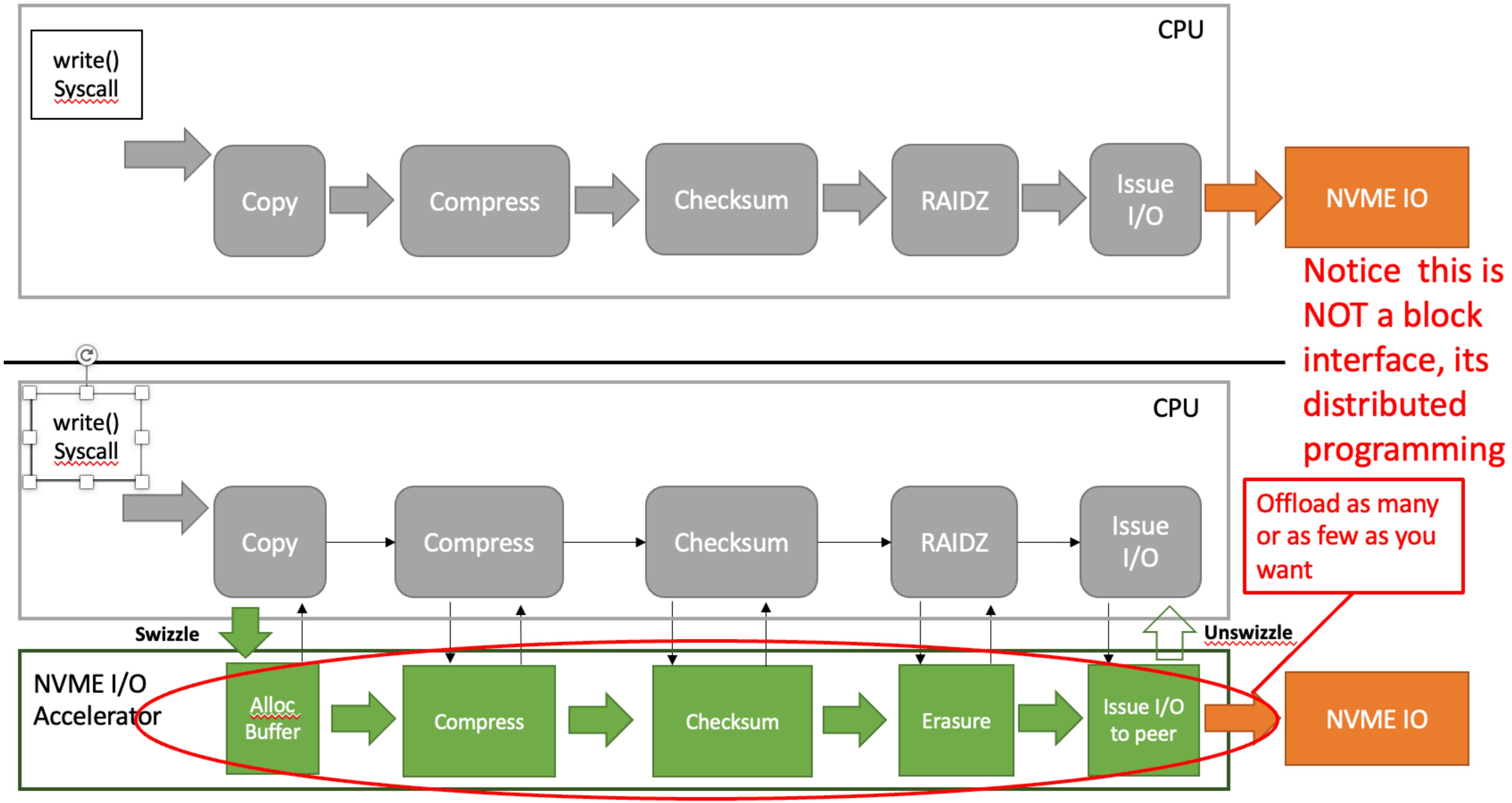




 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Notional fixed function offloads in ZFS



 © SNIA. All Rights Reserved.

COMPUTE + MEMORY + STORAGE SUMMIT

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

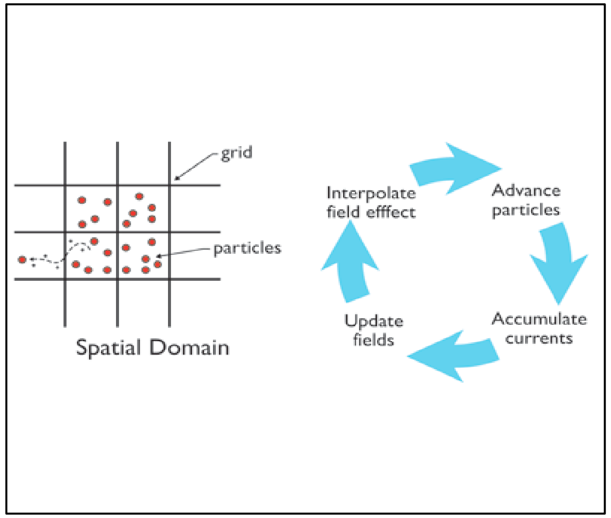
DeltaFS->Ordered KV-CSD

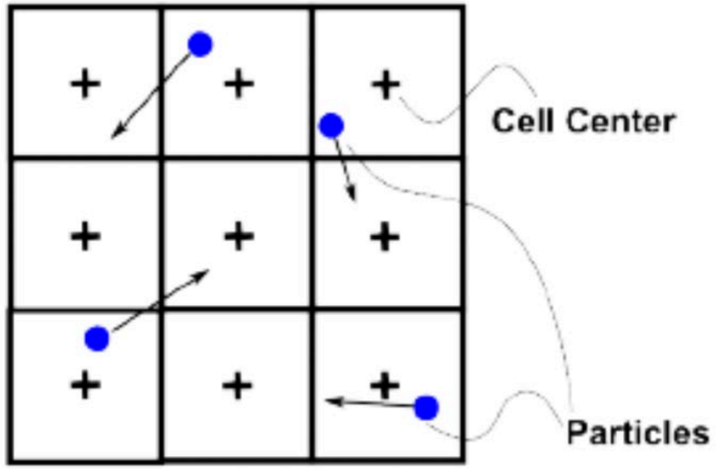
- Format aware, record-oriented applications with a single-dimension, easily shard-able indexing
  - CMU, SK hynix

 © SNIA. All Rights Reserved.

Vector Particle in Cell (VPIC) our Record Based/Single Dimensional Index Application

- Particle-in-cell MPI code (scales to ~100K processes)
  - Fixed mesh range assigned to each process
  - Record: 32 – 64 Byte particles (id, cell id, energy, …)
  - Particles move frequently between processes
  - Million particles per node (Trillions of particles in target simulation)
  - Interesting particles identified at simulation end (say 1000 interesting particles)





Grid for Computational Domain

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Emerging Trends: Analysis Increasingly Selective

- Analysis used to require seeing **all** data records
- Today: queries tend only to hit a small subset of data
- Problem: how to retrieve just interesting rows?

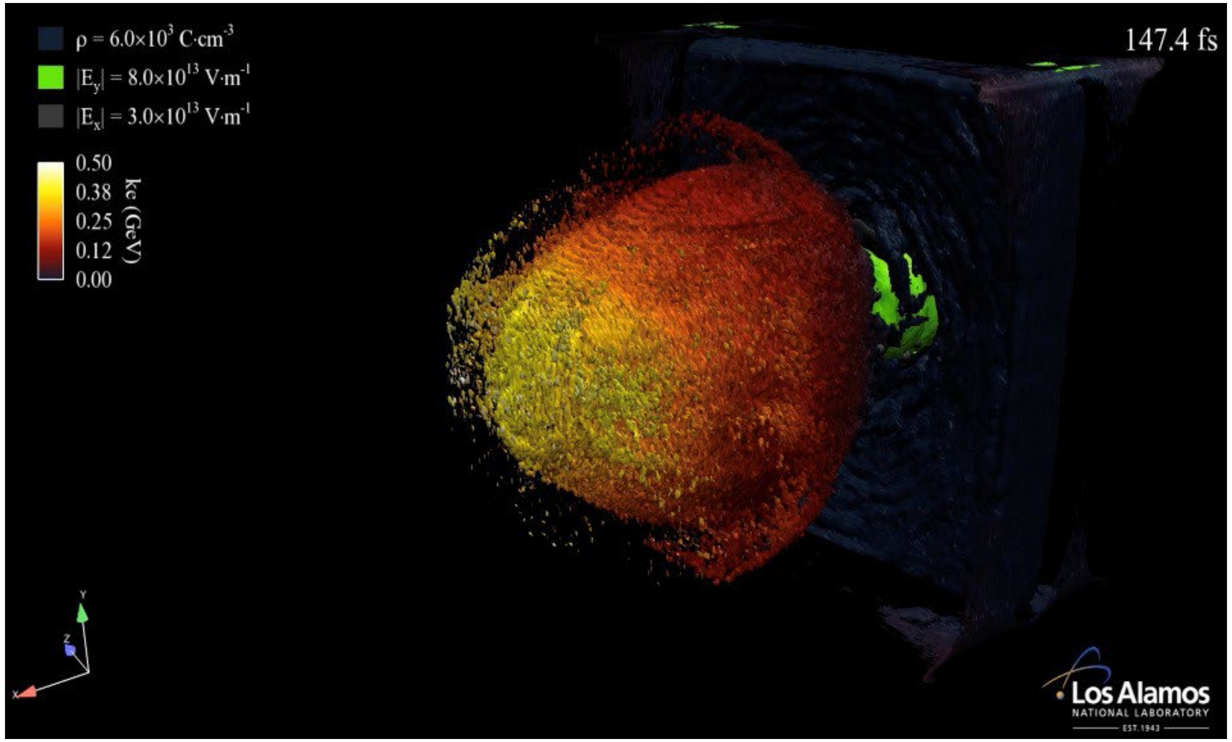


Image from LANL VPIC simulation done by L. Yin, et al at SC10

Example: SELECT X, Y, Z FROM particles **WHERE** E >= 1.5

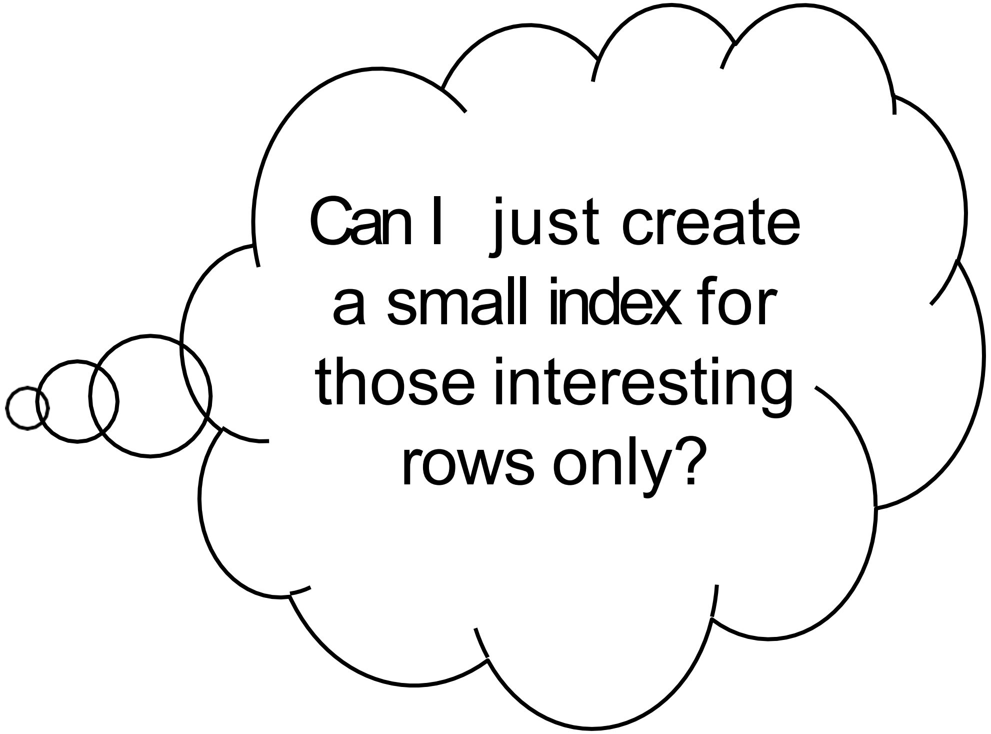
Less than **0.1%** needs to be read from storage

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Reading Back Just Interesting Data is Non-Trivial

- Data known to be interesting only at simulation end
- Indexing only works when all rows are indexed at all timesteps
- Compute node resources are limited
- Sorting only helps one query

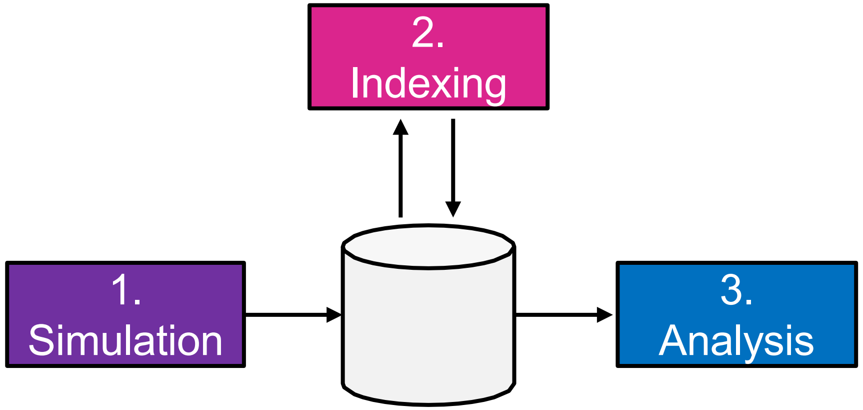


 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

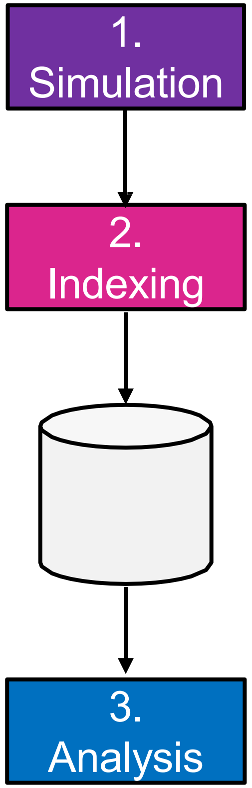
Existing Solutions Fall Short in Different Ways

Post-processing



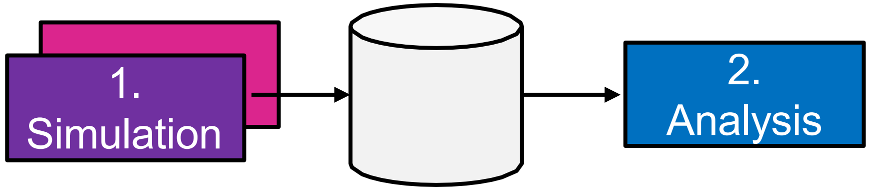
Excessive data movement

In-transit processing



Requires additional compute nodes than the job
Does not work for larger jobs

In-situ processing



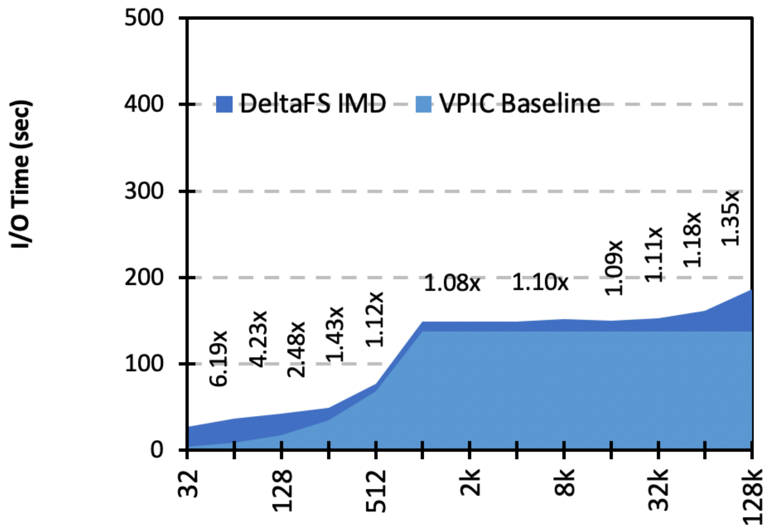
May only produce indexes on 1 or few columns

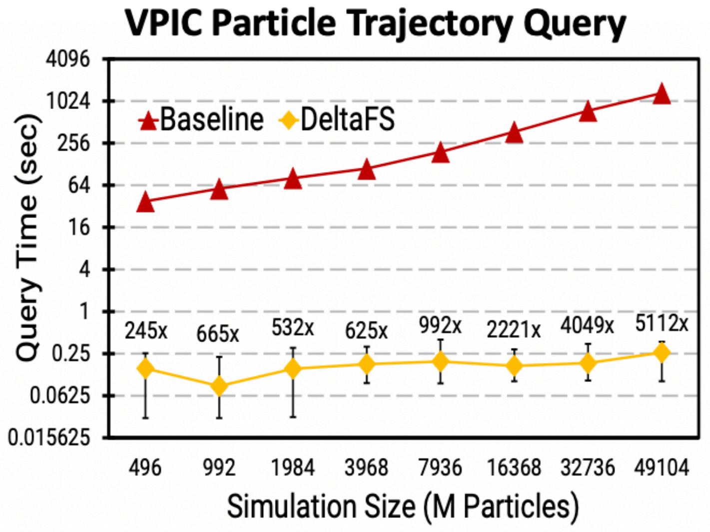
© SNIA. All Rights Reserved.

DeltaFS - Near-device Indexing and Analytics

- Requirements
  - Simulations run under intense memory pressure (app may use 90%)
- Computational Storage Benefits/Opportunities
  - Speedups for post-hoc analysis (1000x speedup demonstrated)
  - Less reliance on massive compute tier as a large merge sort space

Get efficiency and lower time to solution (1000X)





Simulation Processes

(papers at PDSW 15, PDSW 17, SC19 (Best Student Paper)

 © SNIA. All Rights Reserved.

HPC-Driven KV Storage API

- Data insertion:
  - Bulk KV put operations
- Reads:
  - Range queries
  - Secondary indexes
  - Histogram construction
- Management:
  - Compaction control
  - Per key space data export

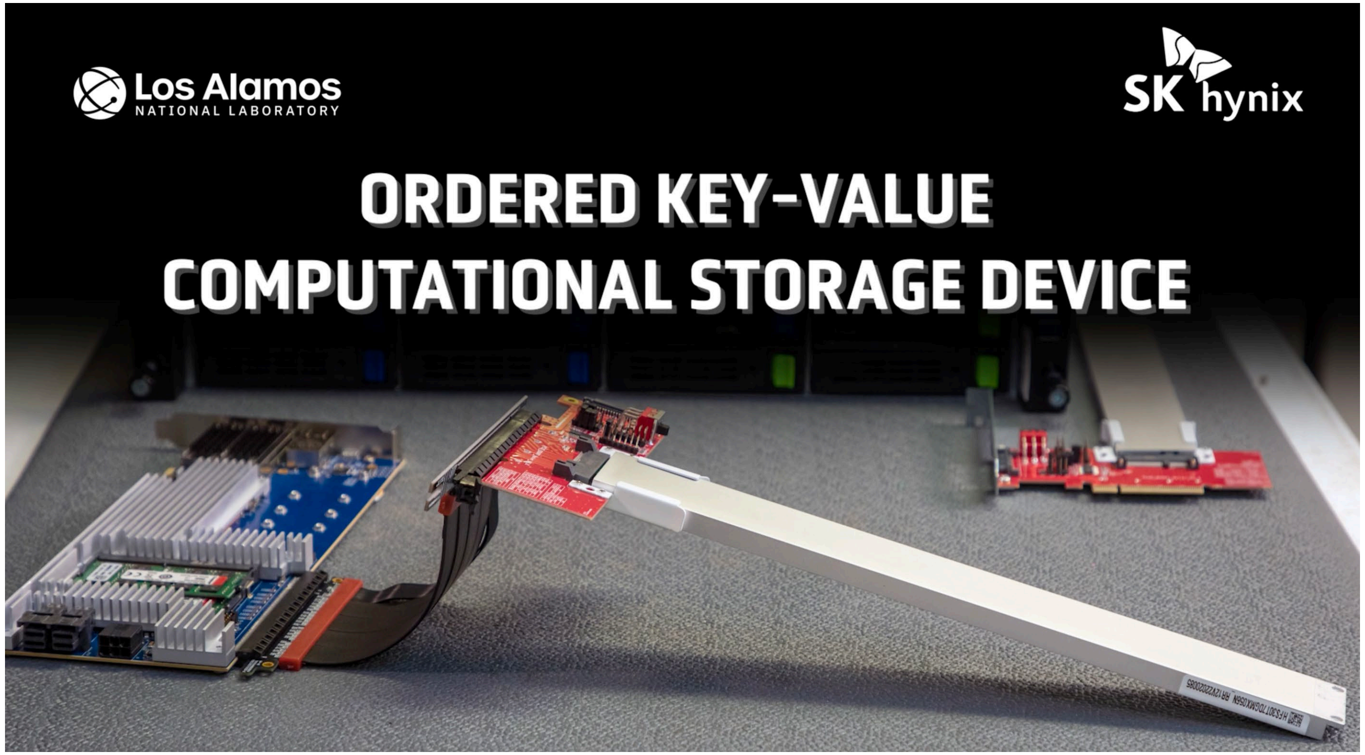
LANL is collaborating with industry for accelerated KV storage that speeds up scientific discovery

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

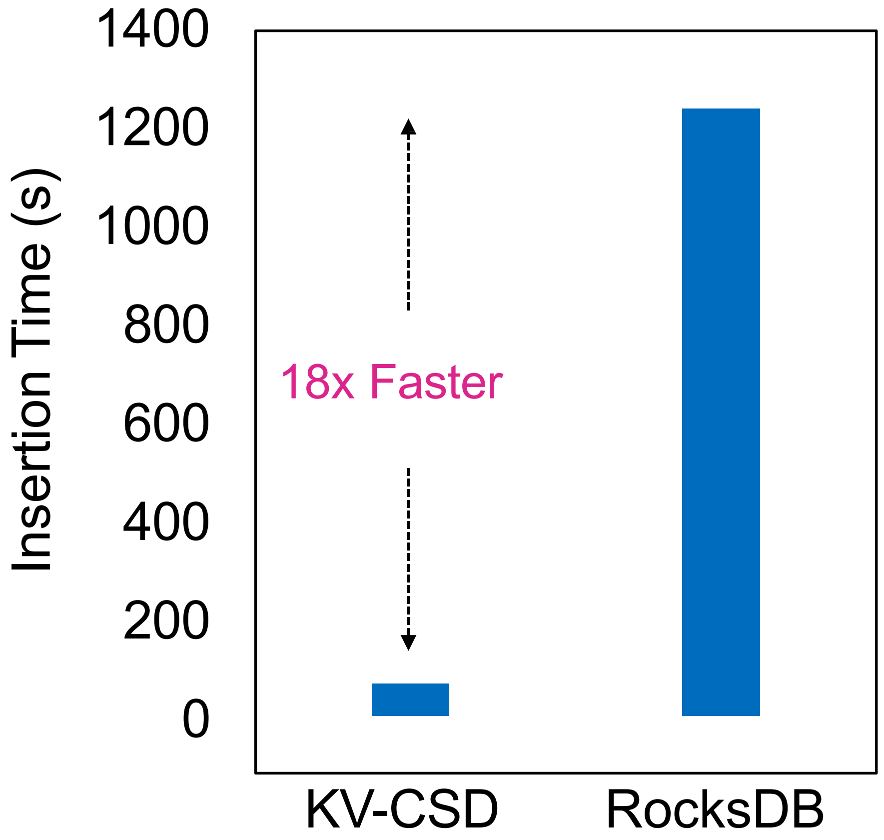
SK hynix Ordered KV-CSD Prototype Revealed at FMS ‘22

- Fully offloaded ordered key value with point and range query capability (put, get, mput, mget, etc.)
- Extensions – control of compaction, and more
- Competitive performance

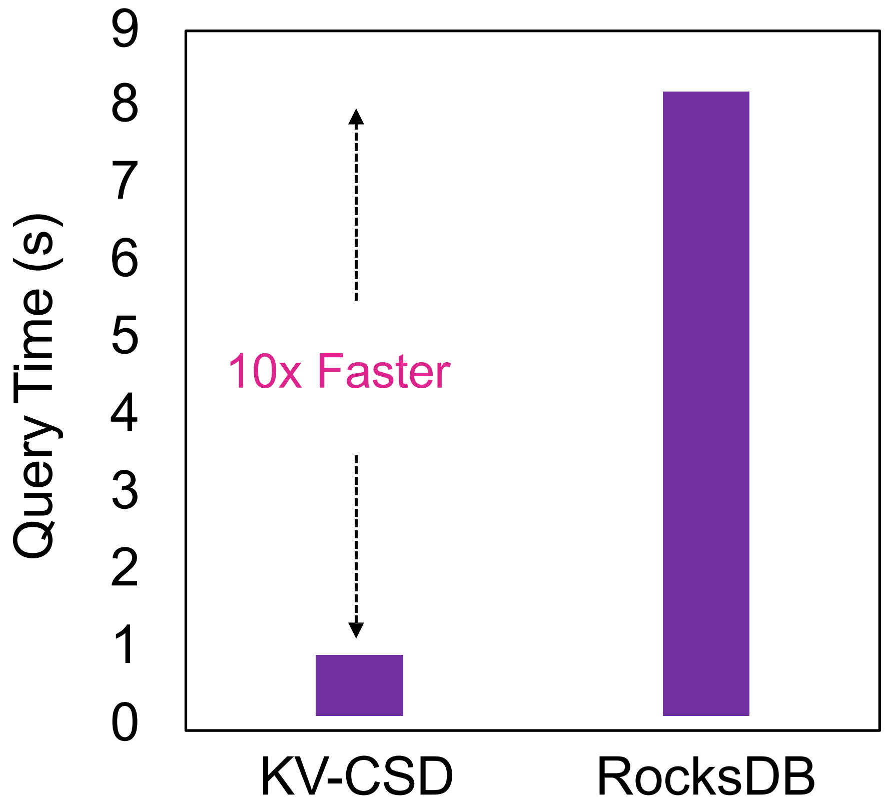


 © SNIA. All Rights Reserved.

Preliminary Results: SK KV-CSD vs RocksDB



Data Insertion: Up to 18x faster



Queries: Up to 10x faster

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

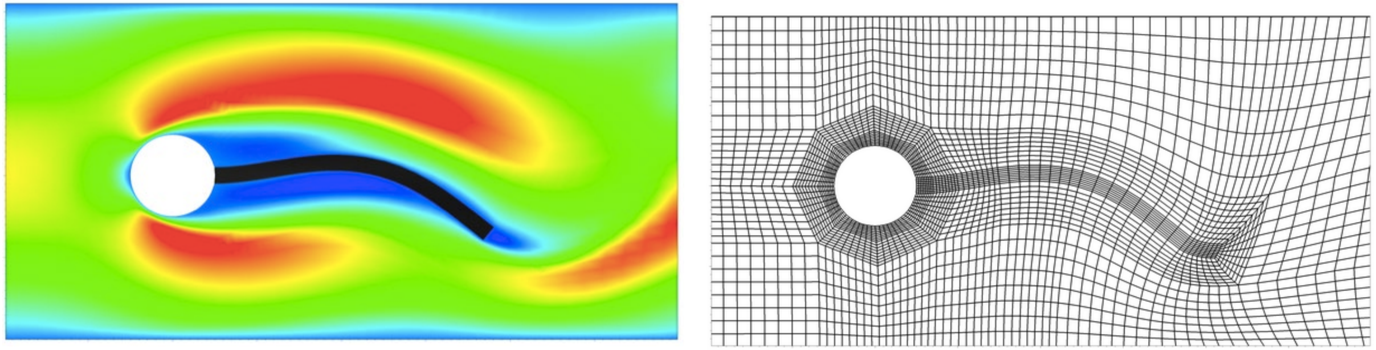
Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

Next Steps

- Format aware, column-oriented applications, multi-dimension, difficult to shard indexing
  - Eideticom, Aeon, Nvidia, SK hynix, Seagate, others?)

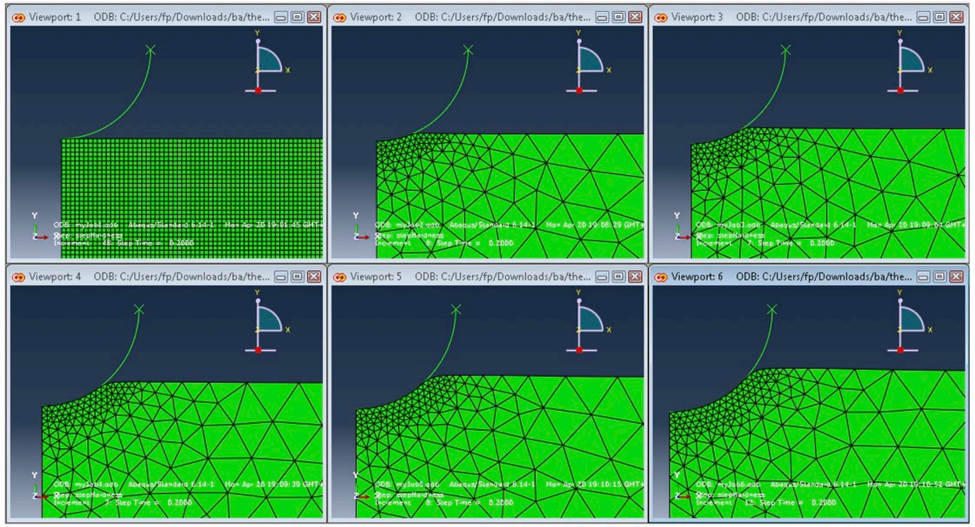
 © SNIA. All Rights Reserved.

What’s a Grid Method and an Adaptive Mesh Refinement (AMR)?



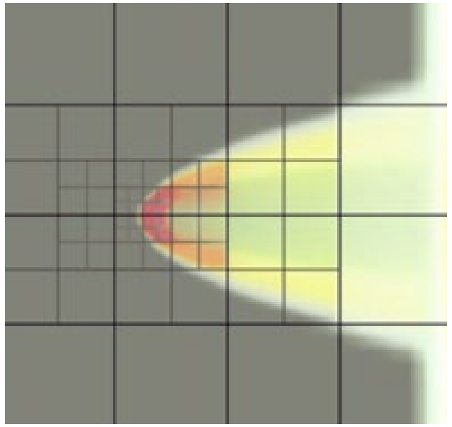
ALE – Advanced Lagrangian Eulerian

http://web.cs.ucdavis.edu/~ma/VolVis/amr_mesh.jpg



AMR

- Lagrangian (mesh deforms)
- Eulerian (mesh doesn’t deform)
- AMR – mesh adapts (refines where the action is)
- Why? – to fit a problem that is way to big for your RAM
- AMR eliminates compression, copy on write, other low hanging fruit

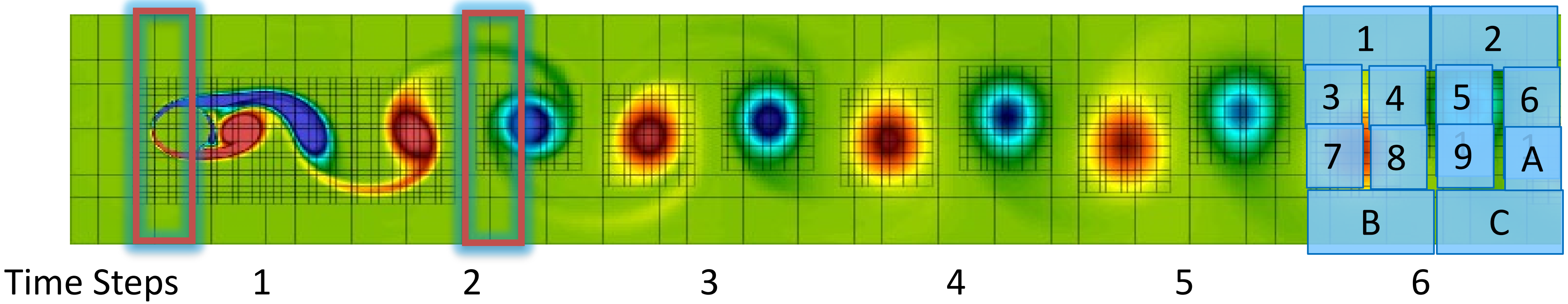


Eulerian AMR

 © SNIA. All Rights Reserved.

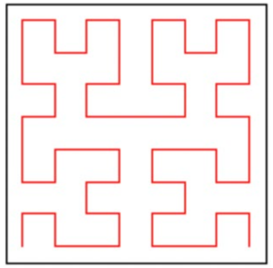
SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Indexing Multi-Dimensional Unstructured Adaptive Meshes

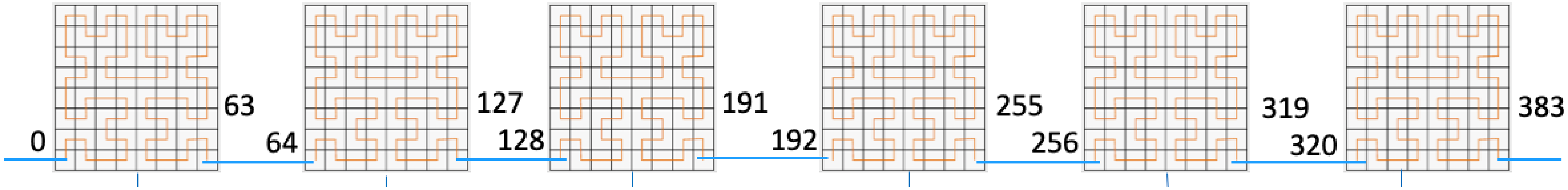


Processes have roughly same number of cells for comp/mem balance but must shuffle cells for AMR

- Time is explicit (a “file” for every time step) and that “file” contains all the state (for restart) (think 1 PB)
- Inside each mesh cell there is 10-100 state variables (64float) (temp, pressure, energy, momentum, …)
- 2D and often 3D – the other dimensions but how do you specify the geometric dimensions?



Single process Hilbert order



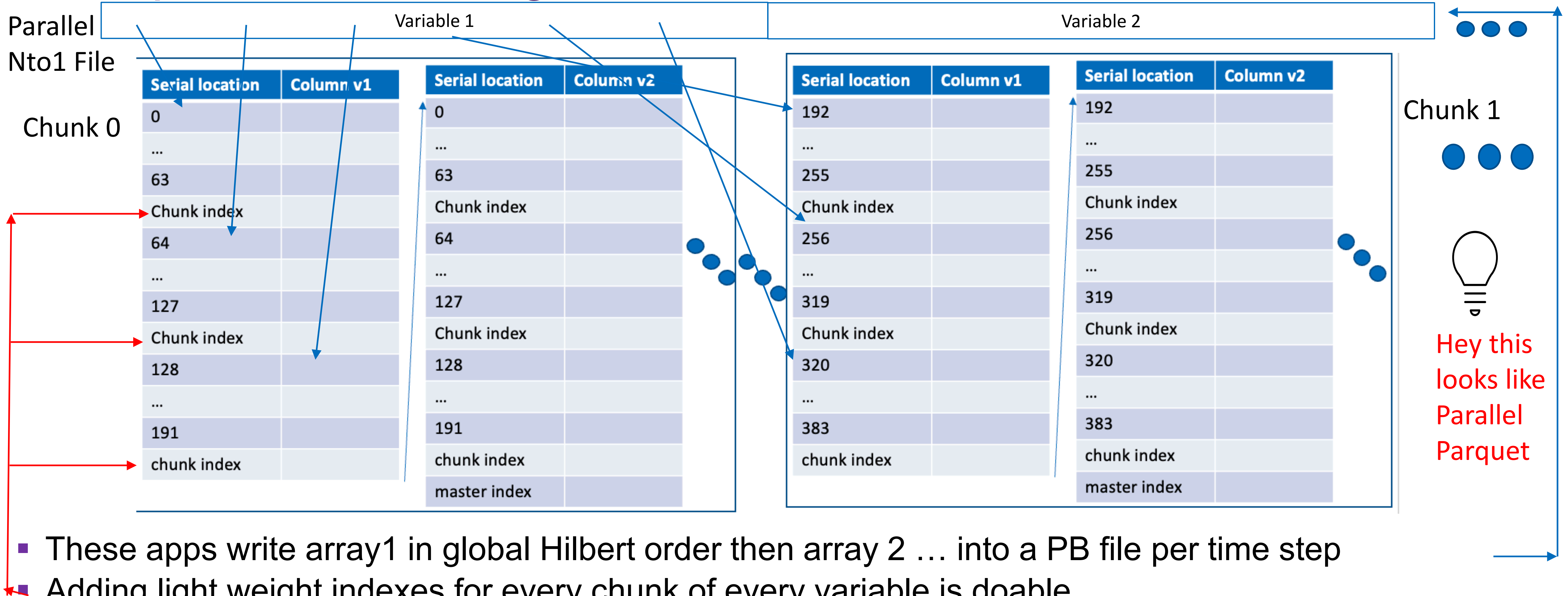
Hilbert space filling curves, we will call this the global Hilbert order, it serializes a value in the cells into distributed array. So we really have 10-100 distributed arrays in Hilbert order ☺

Find the outer edge of the eddy’s (light blue and yellow) <1/100th of the total data, usually less. Can light weight indexing yield 1000X less data and can it be done very near the storage device to save transmission?

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Adding index and offloading columnar analytics into Computational Storage, how would it work?



- These apps write array1 in global Hilbert order then array 2 … into a PB file per time step
- Adding light weight indexes for every chunk of every variable is doable
- Use standard analytics with things like Duckdb or Apache Drill and have the power of SQL and joins on columns and the simple indexes do massive reduction in parallel

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Can this Hilbert Inspired Chunked Parquet Concept Extend to On-Disk Processing, Even with Erasure?

- Parquet ZFS File with Erasure and On-Kinetic Disk Analytics in parallel

N + P

Serial location	Column v1
0	
...	
63	
Chunk index	
64	
...	
127	
Chunk index	
128	
...	
191	
chunk index	

Serial location	Column v2
0	
...	
63	
Chunk index	
64	
...	
127	
Chunk index	
128	
...	
191	
chunk index	
master index	

Serial location	Column v1
192	
...	
255	
Chunk index	
256	
...	
319	
Chunk index	
320	
...	
383	
chunk index	

Serial location	Column v2
192	
...	
255	
Chunk index	
256	
...	
319	
Chunk index	
320	
...	
383	
chunk index	
master index	

Parity chunk

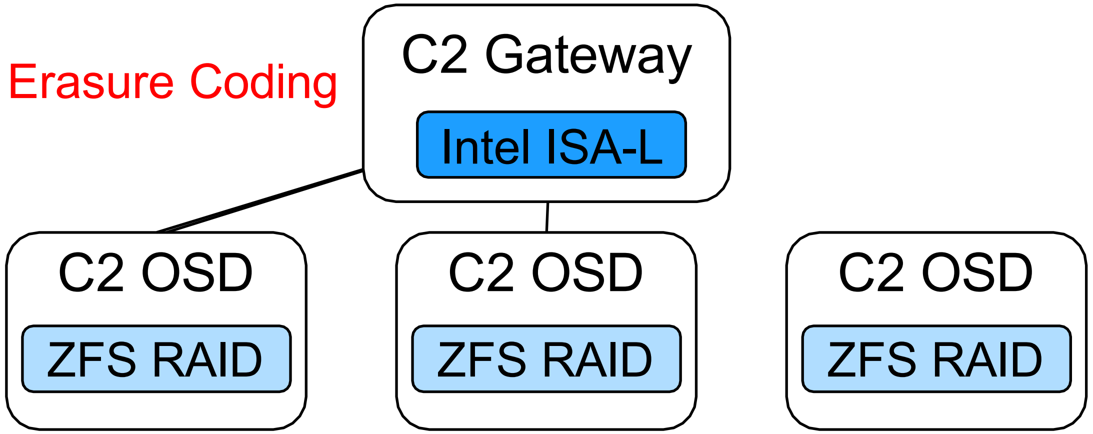
A collaboration with our excellent partners at Seagate

- Use standard analytics with things like Duckdb or Apache Drill and have the power of SQL and joins on columns and the simple indexes do massive reduction in parallel

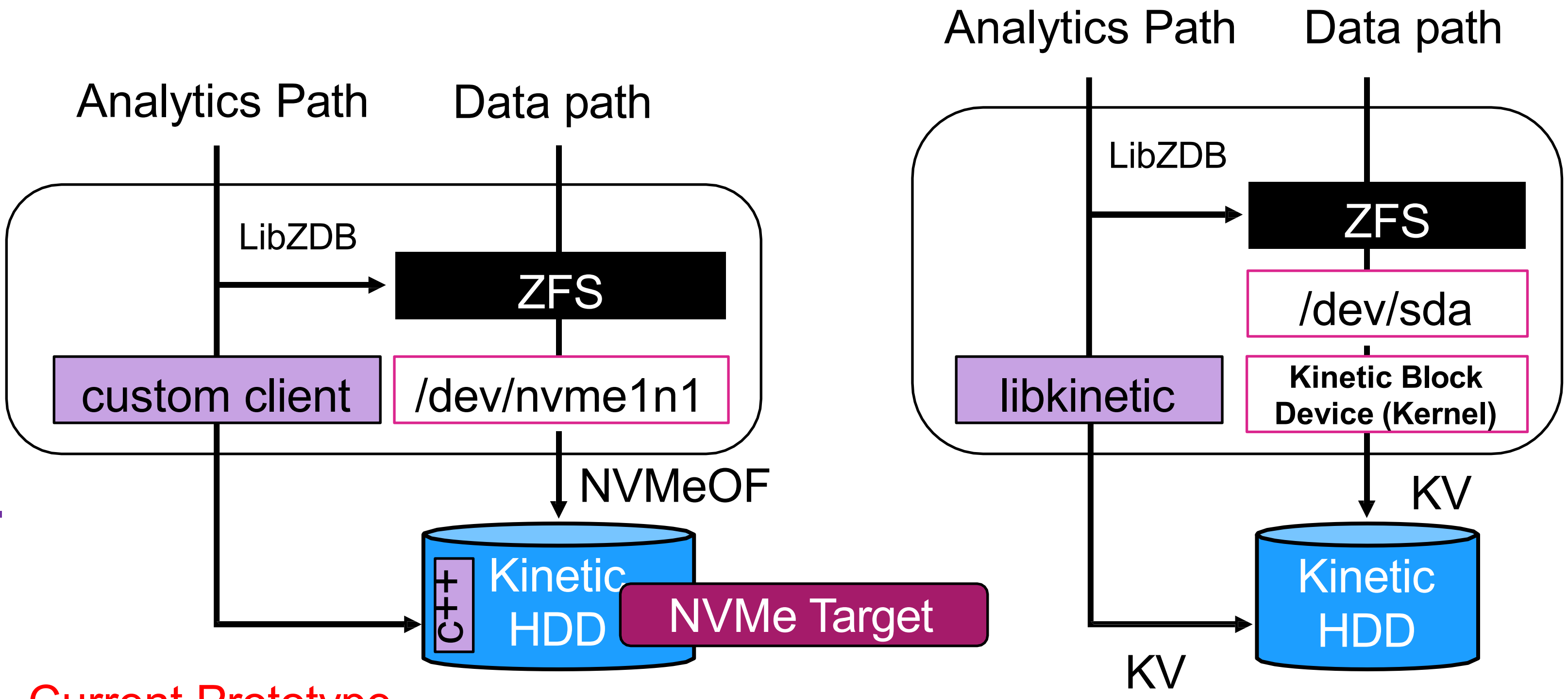
© SNIA. All Rights Reserved.

Prototype Implementation

Multi-level Erasure Coding



- LibZDB (LANL's stripped-down version of ZDB) for mapping ZFS filenames to disk LBAs
- Drive exposed as an NVMeOF block device to ZFS
- Custom C++ code for in-drive analytics



Current Prototype

Longer-term Design

 © SNIA. All Rights Reserved.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

Please take a moment to rate this session.

Your feedback is important to us.

 © SNIA. All Rights Reserved.