

DAOS IT Press Tour

Johann Lombardi, TSC Chair, DAOS Foundation
London, April 2025



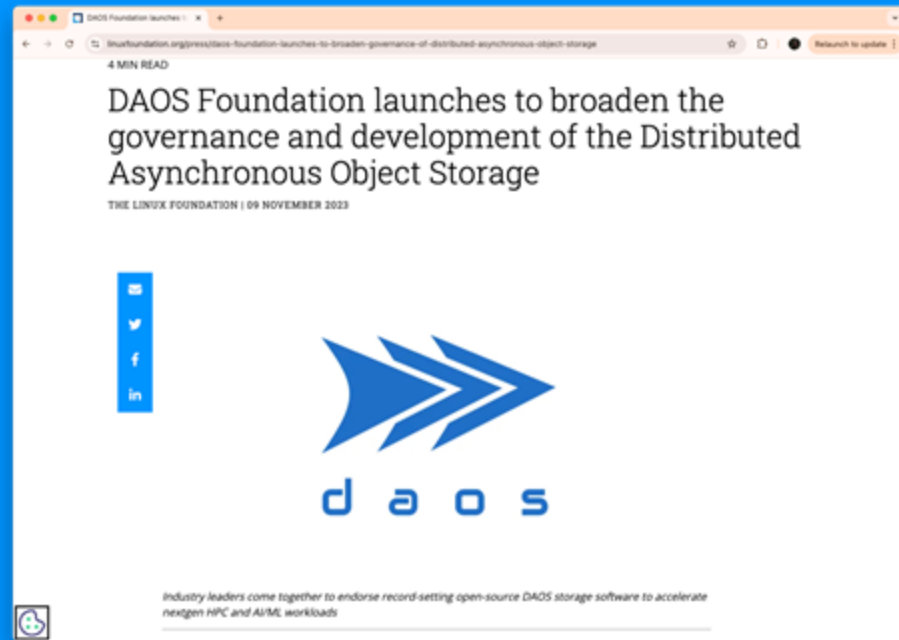
<https://foundation.daos.io>



Agenda

- DAOS Foundation
- Project History & Vision
- Technical Overview
- Software Ecosystem
- Roadmap
- Deployments & Performance

DAOS Foundation



DAOS Foundation Members



Board



Kevin Harms, Chair
Argonne National
Laboratory



Dean Hildebrand
Google



Lance Evans
HPE



Allison Goodman
Intel



Chris Girard
VDURA



Johann Lombardi
TSC Chair



Michael Hennecke
Outreach Committee Chair

Mission

- The DAOS Foundation exists to
 - Maintain DAOS as an open source project independent of any one organization
 - Foster the developer and user communities around DAOS
 - Guide the direction of the overall DAOS project
 - Promote the use of DAOS
- Governing Board
 - Defines budget and approves expenses
 - Oversee efforts of other subcommittees
 - Approve roadmap provided by TSC
 - Vote on matters as needed

Meetings

- Governing Board
 - Weekly meeting on Wednesday
 - Currently open only to Board members
- Technical Steering Committee
 - Weekly on rotating schedule
 - Monday
 - Wednesday
 - Working Groups - rotating schedule

How to Join

- Two step process for any organization
 - Join the Linux Foundation (at any level)
 - Join the DAOS Foundation
- <https://daos.io/how-to-join-the-daos-foundation>
- DAOS Foundation
 - 3 levels with 5 fees



On 09. November 2023, the founding members Argonne Labs, Hewlett Packard Enterprise, Google Cloud, and Intel Foundation to broaden the governance of the **Distributed Storage (DAOS)** open source project. See the [LF Press Release](#) announcement.

DAOS Foundation Membership Level	Annual Fees
Premier	25,000 USD
Premier for LF Associate Members	15,000 USD
General	15,000 USD
General for LF Associate Members	6,000 USD
Associate for LF Associate Members	0 USD

DAOS Foundation Levels

- Premier Membership
 - Each Premier Member can appoint a voting member to the DAOS Foundation's Governing Board, its Outreach Committee, and to any other committee that the DAOS Foundation may establish (including the TSC).
- General Membership
 - The group of all General Members annually elect up to three voting representatives to the DAOS Foundation's Governing Board (depending on the number of General Members).
 - Each General Member can appoint a non-voting member to the DAOS Foundation's Outreach Committee.
- Associate Membership
 - The Associate Members can participate in the activities of the DAOS Foundation, but have no seat on the Governing Board and no voting rights.

2024 Expense Summary

Area	Budget (USD)	Actual Spend	Description
Community Engagement	27,500	0	DUG Event(s) and press releases
Legal	11,000	0	Trademarks and filings
Board Operations	23,750	23,750	LF project management (prorated)
Development	18,400	3,200	Cloud/Hosting/Tools, Community travel, CI/CD
General & Administrative	8,100	10,350	LF fee on membership revenue (9%)

2024 Achievements and 2025 Goals

2024

- Added VDURA to Foundation
- Completed transfer of DAOS assets from Intel to Foundation
- Completed charters for foundation and TSC
- Regular TSC meetings including collaboration to align v2.6
- DUG'24!

2025

- Recruiting new members
- Update website and promotional materials
- Complete trademark of DAOS
- Release DAOS v2.8
 - First community release
- Event Planning
 - In-person DUG event
 - Virtual DUG event
 - Continued presence at conferences

TSC Structure

- Voting Members

- *Argonne*: Kevin Harms
- *Google*: corwin
- *HPE*: Lance Evans
- *Intel*: Allison Goodman
- *Vdura*: Brian Mueller
- *TSC Chair*: Johann Lombardi

- Meet weekly (public) with rotating schedule

- Members distributed across US, EU, China and Australia



TSC Scope

- Define community roadmap (2.8+)
 - Gather contributions from all community members
 - Publish roadmap on <https://daos.io>
- Produce community releases (2.8+)
 - Track progress, review jira tickets & test results
 - Tag release and sign/distribute packages
 - Provide docker images
- Organize DAOS development
 - Simplify contributions
 - Organize gatekeeping (members, responsibilities, process)
 - Document contribution process

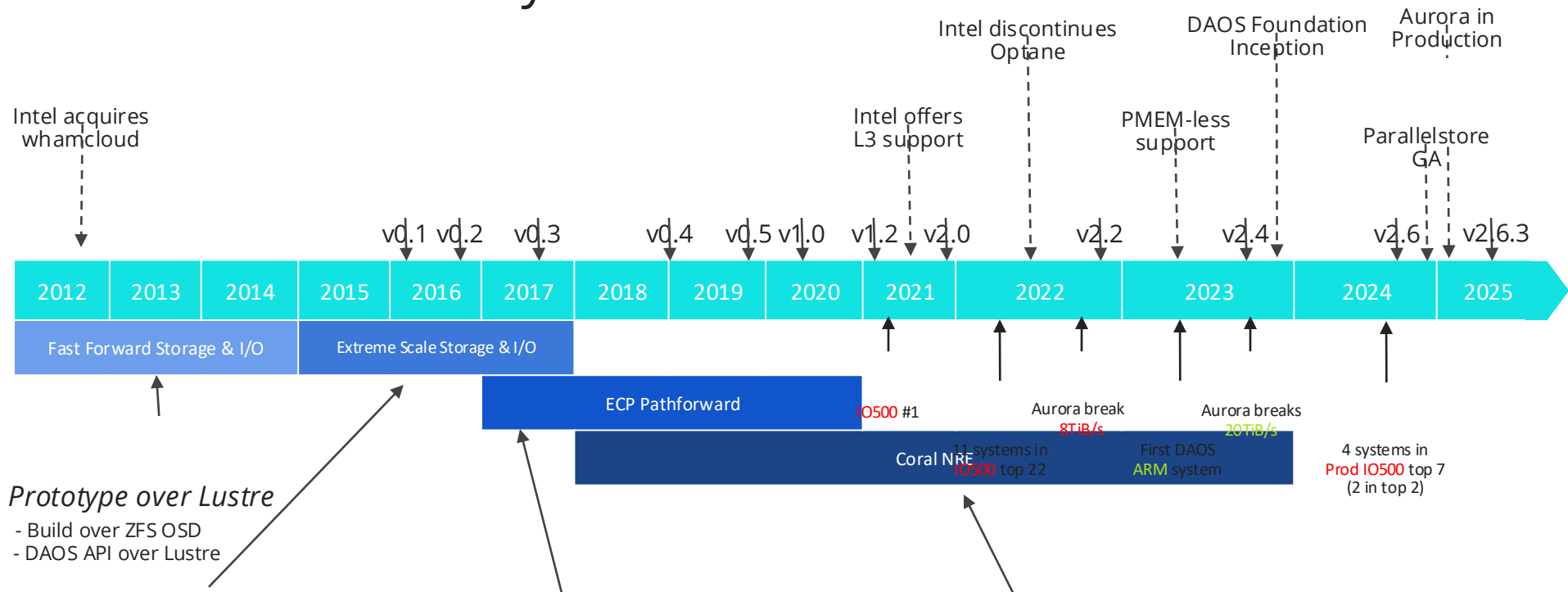
TSC Scope

- Community test infrastructure
 - Goal: artifacts and logs available to all contributors
 - Expand coverage
 - ARM/AMD
 - More fabrics
 - More linux distributions
 - Cloud environments
 - Focus on pmem-less mode
- Working groups
 - Open to anyone
 - Forums for DAOS users/administrators/contributors to exchange
 - Rotating schedule

Project History & Vision



DAOS History



Standalone prototype

- OS-bypass
- Persistent memory via PMDK
- Replication & self healing

DAOS embedded on FPGA

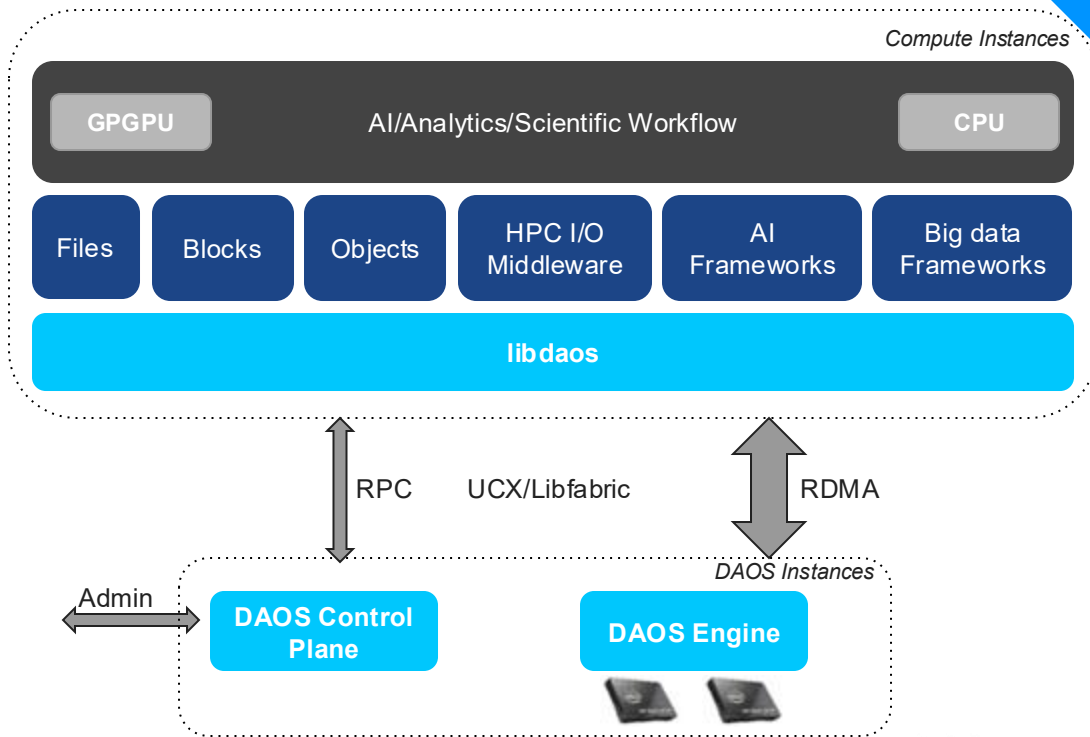
- Disaggregated I/O
- Monitoring
- NVMe SSD support via SPDK

DAOS Productization for Aurora

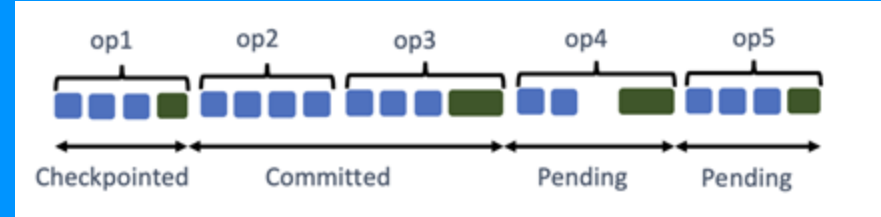
- Hardening
- 10+ new features
- Support for extra AI/Big data frameworks

DAOS: Nextgen Open Storage Platform

- Platform for innovation
- Files, blocks, objects and more
- Full end-to-end userspace
- Flexible built-in data protection
 - EC/replication with self-healing
- Flexible network layer
- Efficient single server
 - O(100)GB/s and O(1M) IOPS per server
- Highly scalable
 - TB/s and billions IOPS of aggregated performance
 - O(1M) client processes
- Time to first byte in O(10) μ s



Technical Overview



DAOS Design Fundamentals

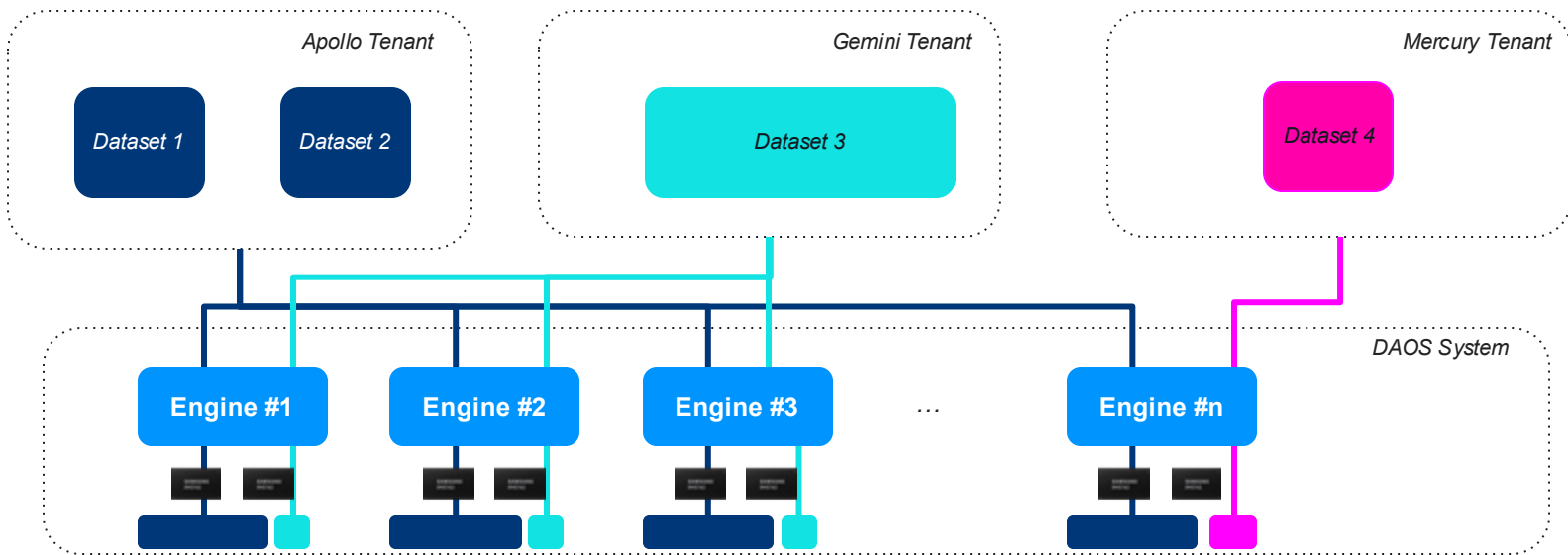
- No read-modify-write on I/O path (use versioning)
- No locking/DLM (use MVCC)
- No client tracking or client recovery
- No centralized (meta)data server
- No global object table
- Non-blocking I/O processing (futures & promises)
- Serializable distributed transactions
- Built-in multi-tenancy
- User snapshot




Scalability &
Performance

High IOPS

Unique
Capabilities

Storage Pooling - Multi-tenancy

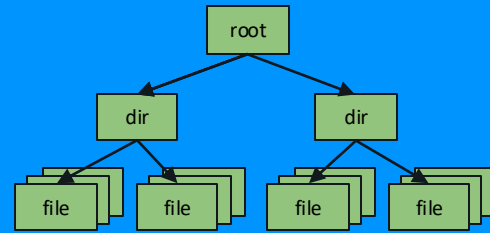


Pool 1		Apollo Tenant	100PB	20TB/s	200M IOPS
Pool 2		Gemini Tenant	10PB	2TB/s	20M IOPS
Pool 3		Mercury Tenant	30TB	80GB/s	2M IOPS

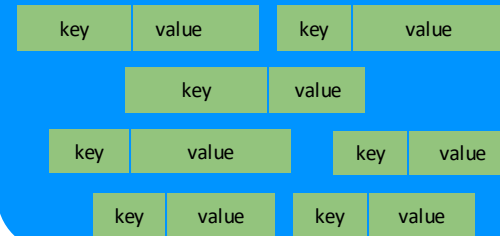
Dataset Management

- New data model to unwind 30+ years of file-based management
- Introduce notion of dataset
- Basic unit of storage
- Datasets have a type
- POSIX datasets can include trillions of files/directories
- Advanced dataset query capabilities
- Unit of snapshots
- ACLs/IAM

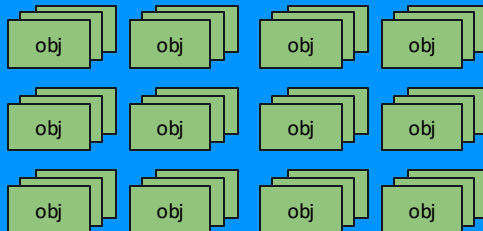
POSIX Dataset



KV Dataset



Python Dataset



Object Interface

Middleware/Framework View

DAOS Layout View

Mapping

Object

128-bit
object Identifier

Array

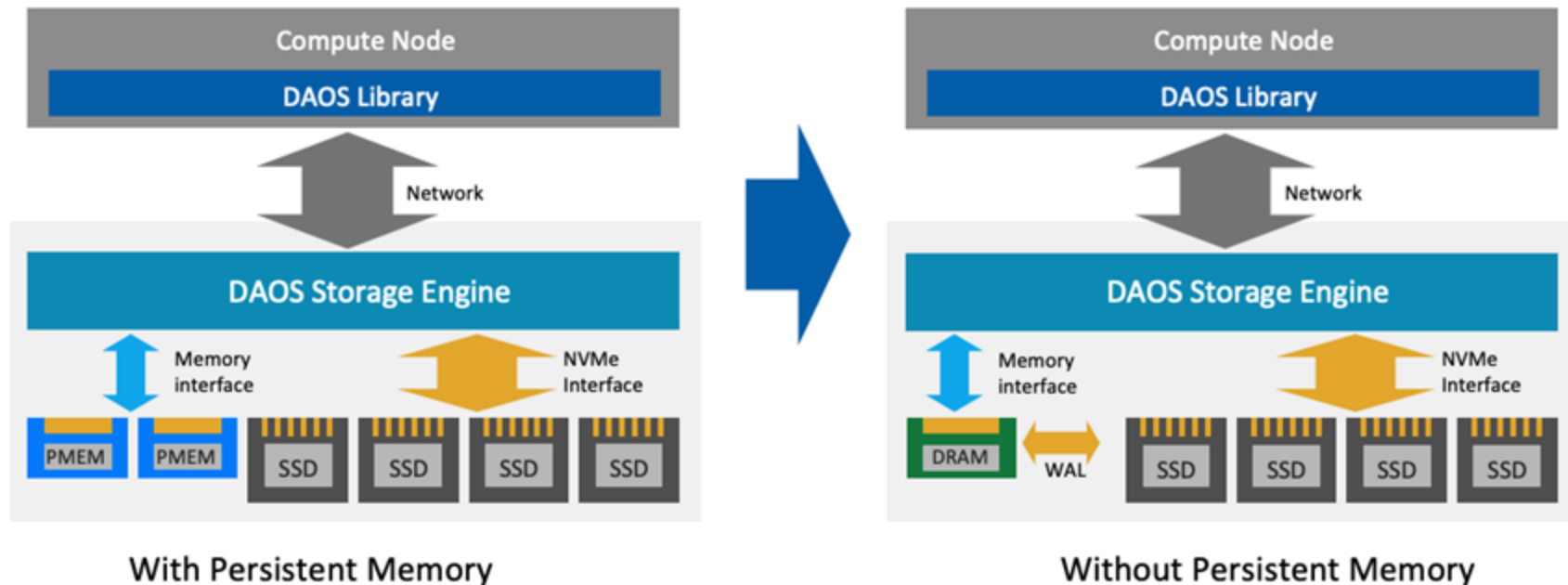
Multi-dimensional
Array

Key-value
Store

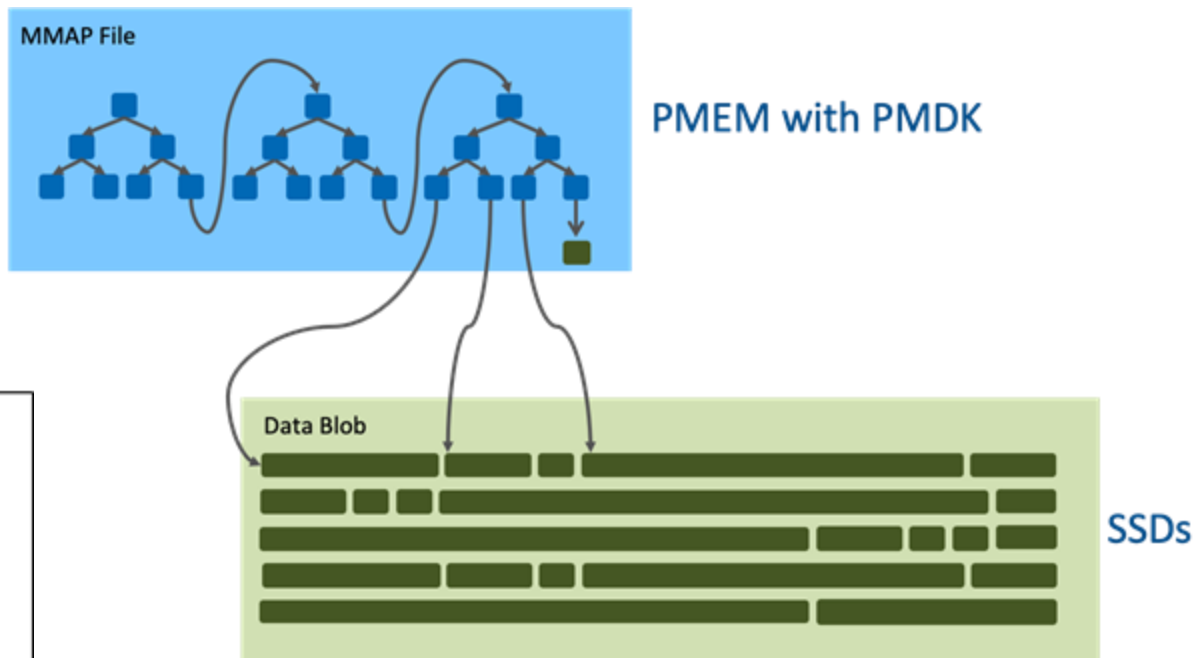
Multi-level
Key-value
Store

- No object create/destroy
- No size, permission/ACLs or attributes
- Sharded and erasure-coded/replicated
- Algorithmic object placement
- Very short Time To First Byte (TTFB)

DAOS Architecture Evolution

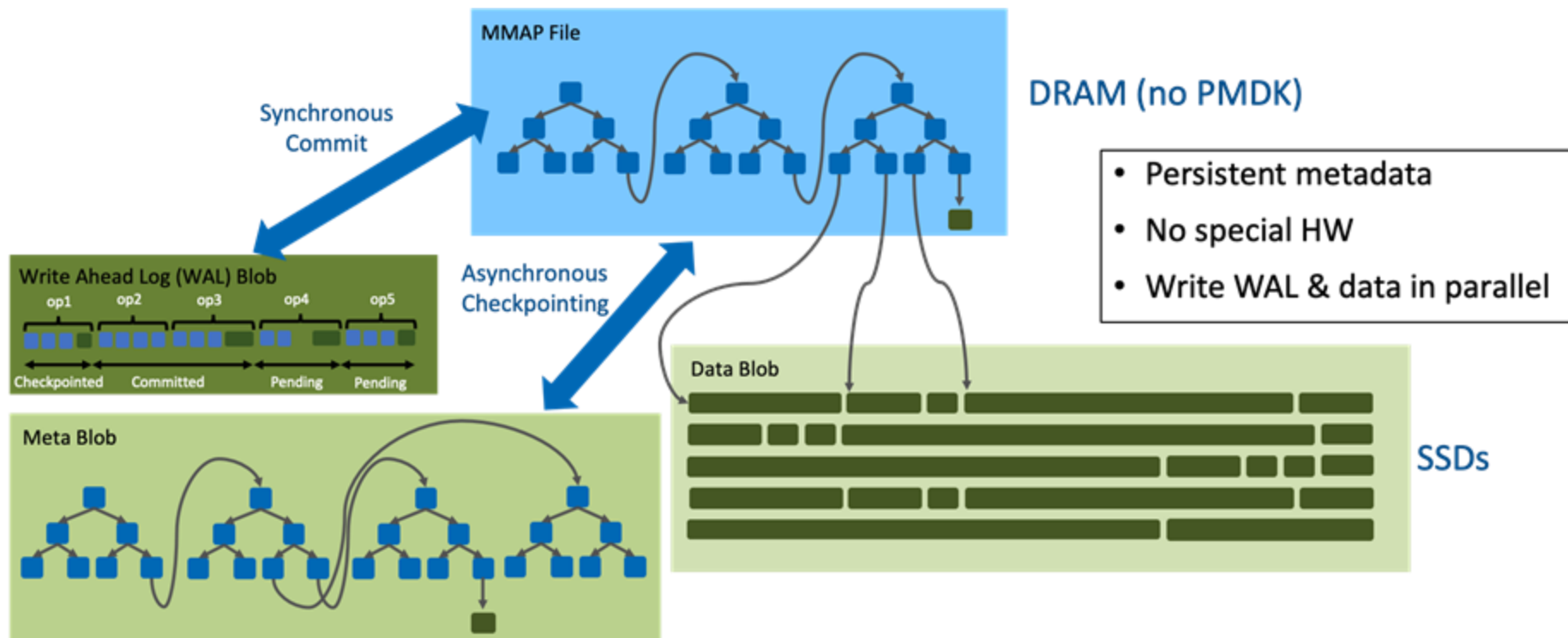


Pmem Mode



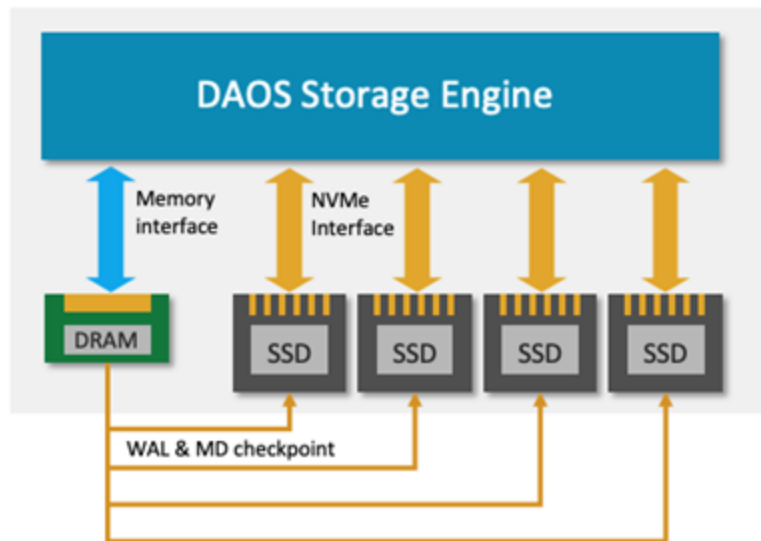
- Persistent metadata
- Require Intel Optane PMEM (or NVDIMM-N)
- App Direct mode
- Mode used on Aurora

Pmem-less Mode

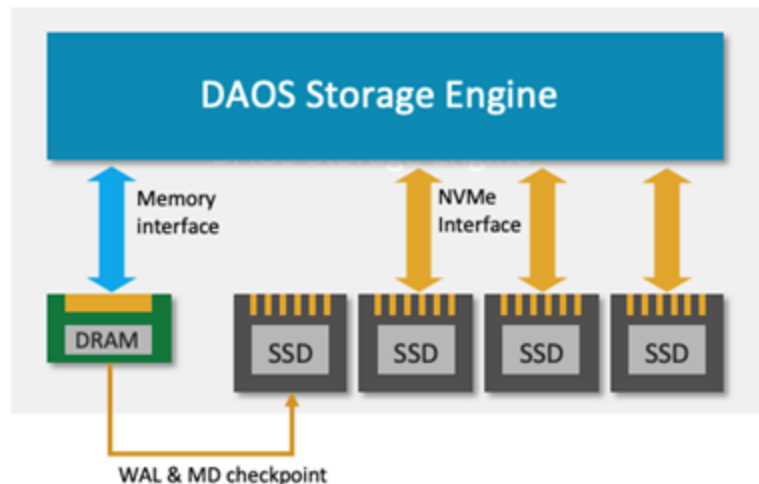


Pmem-less Configuration

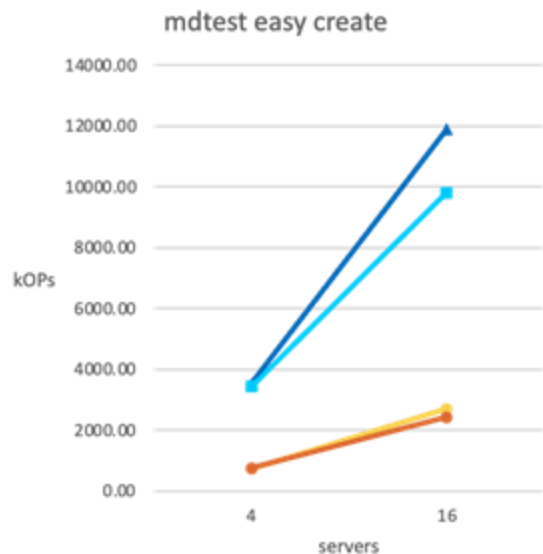
Mixed mode



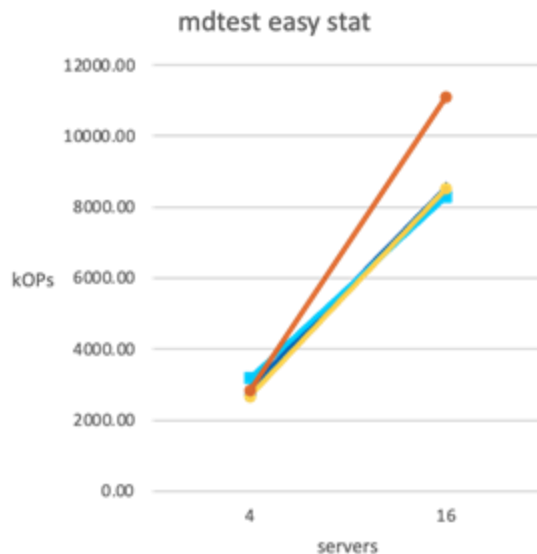
Dedicated SSD for MD/WAL



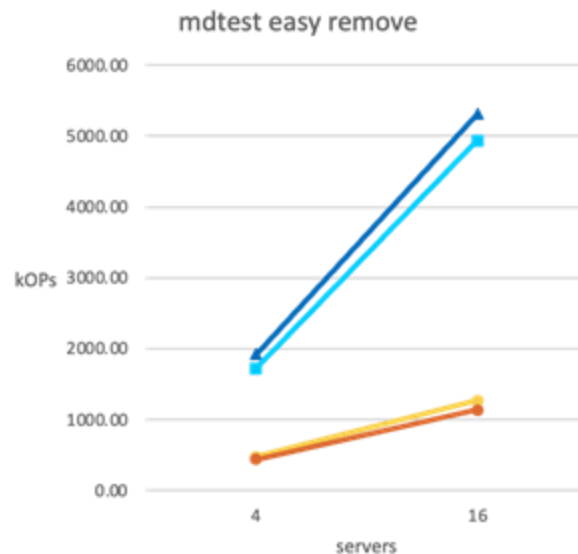
Pmem vs Pmem-less Performance



pmem_create_s1 mdssd_create_s1
pmem_create_rp3g1 mdssd_create_rp3g1

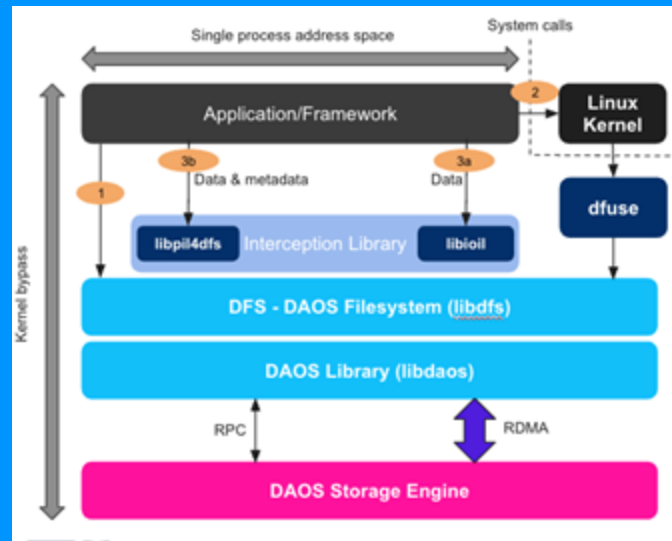


pmem_stat_s1 mdssd_stat_s1
pmem_stat_rp3g1 mdssd_stat_rp3g1



pmem_remove_s1 mdssd_remove_s1
pmem_remove_rp3g1 mdssd_remove_rp3g1

Software Ecosystem





python

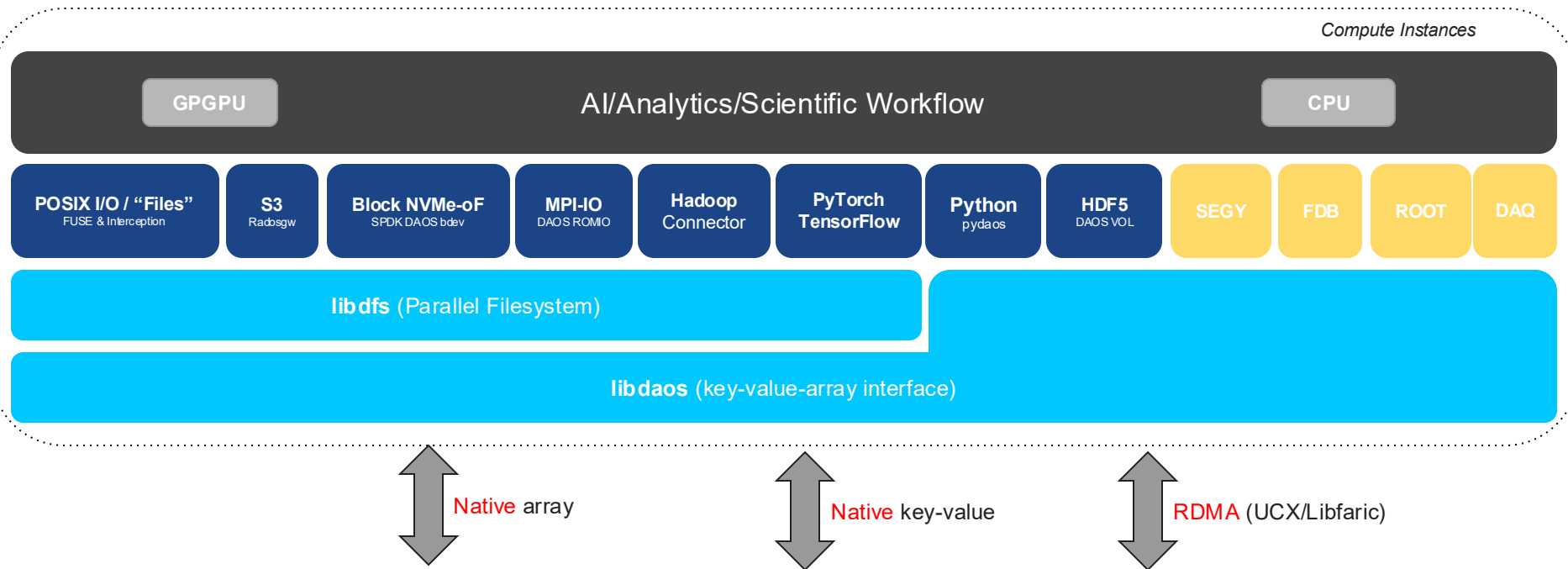
PyTorch



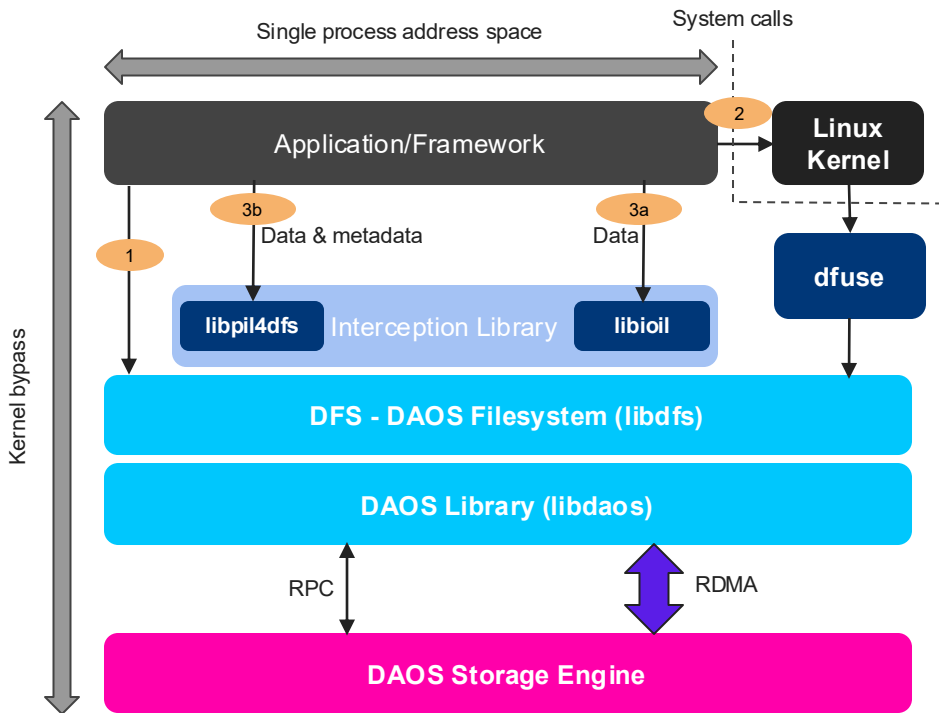
TensorFlow



Software Ecosystem



POSIX Support & Interception



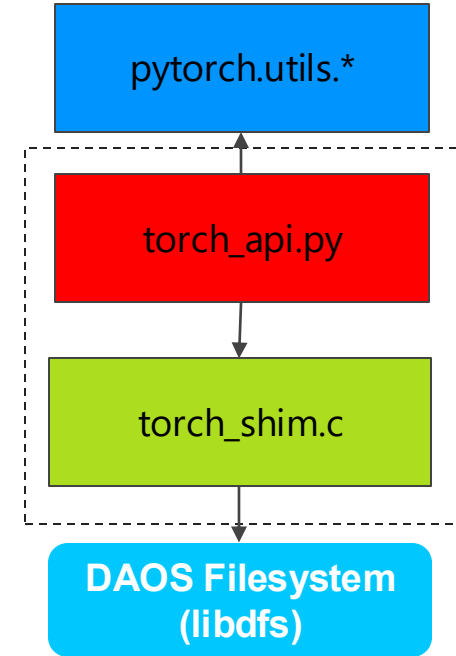
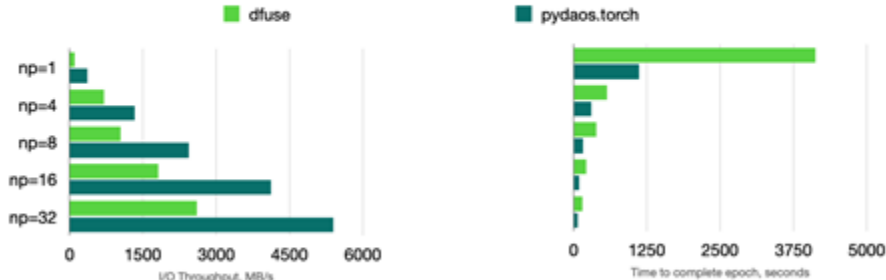
1. Userspace DFS library with API like POSIX
 - **Require** application changes
 - Low latency & high concurrency
 - No caching
2. DFUSE daemon to support POSIX API
 - **No** application changes
 - VFS mount point & high latency
 - Caching by Linux kernel
3. DFUSE + Interception library
 - **No** application changes
 - 2 flavors using LD_PRELOAD
- 3a. libioil
 - (f)read/write interception
 - Metadata via dfuse
- 3b. libpil4dfs
 - Data & metadata interception
 - Aim at delivering same performance as #1 w/o any application change
 - Mmap & binary execution via fuse

PyTorch DAOS Modules

- Collaboration between Enakta Labs and Google
- DataLoader and Checkpoint modules
 - Support for both iterable and map-style datasets
 - High parallelism using several DAOS event queues
 - Parallel namespace scanning using dfs anchor API

Test results

Large IO: 100,000 of 4MB samples, batch size=32, 1 reader



	Time to scan 1.1M Files
Regular scan	291s
Optimized scan	32s

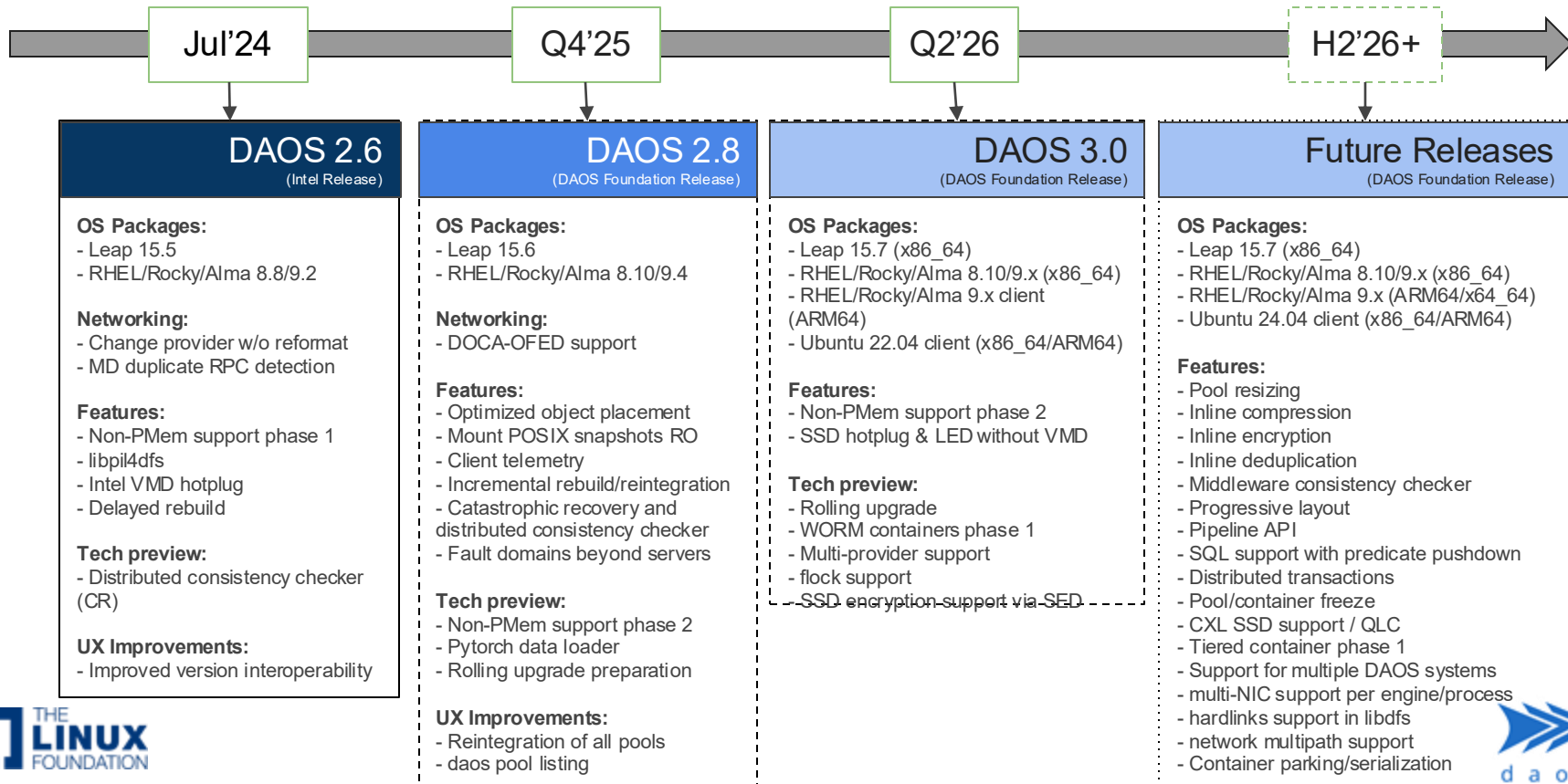
Roadmap



DAOS Community Roadmap

Color coding schema:

- Committed (or released) release/features
- In-planning release/features
- Future possible release/features



Deployments & Performance



Aurora Overview



Aurora System Specifications

Compute Node

2 Intel Xeon scalable "Sapphire Rapids" processors;
6 Xe arch-based GPUs; Unified Memory
Architecture; 8 fabric endpoints; RAMBO

CPU-GPU Interconnect

CPU-GPU: PCIe; GPU-GPU: Xe Link

Peak Performance

≈ 2 Exaflop DP

Platform

HPE Cray EX supercomputer

System Size (# Nodes)

> 9,000

Software Stack

HPE Cray EX supercomputer software stack + Intel
enhancements + data and learning

System Interconnect

Slingshot 11; Dragonfly topology with adaptive
routing

High-Performance Storage

≈ 230 PB, ≈ 25 TB/s (DAOS)

Aggregate System Memory

> 10 PB

GPU Architecture

Xe arch-based "Ponte Vecchio" GPU; Tile-based
chiplets, HBM stack, Foveros 3D integration, 7nm

Network Switch

25.6 Tb/s per switch, from 64–200 Gbs ports (25
GB/s per direction)

Programming Models

Intel oneAPI, MPI, OpenMP, C/C++, Fortran,
SYCL/DPC++

Node Performance (TF)

> 130

Aurora DAOS System

- 1024x DAOS Storage nodes
 - 2x Xeon 5320 CPUs (ICX)
 - 512GB DRAM
 - 8TB Optane Persistent Memory 200
 - 244TB NVMe SSDs
 - 2x HPE Slingshot NICs
- Supported data protection schemes
 - No data protection
 - All EC flavors: 2+1, 2+2, 4+1, 4+2, 8+1, 8+2, 16+1 and 16+2
 - N-way replication
- Usable DAOS capacity
 - between 220PB and 249PB depending on redundancy level chosen



DAOS Performance - SC'24 Production List

# 1	INFORMATION								IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE 1	BW	MD	REPRO.
									(GiB/s)	(KiOP/s)	
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	✓
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	✓
3	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	✓
4	SC24	MSKCC	IRIS	WekaIO	WekaIO	261	27,144	665.49	252.54	1,753.69	✓
5	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	✓

IOR & FIND

EASY WRITE	20,693.63 GiB/s
EASY READ	12,122.87 GiB/s
HARD WRITE	4,216.34 GiB/s
HARD READ	9,706.55 GiB/s
FIND	229,672.10 KiOP/s

METADATA

EASY WRITE	60,985.13 KiOP/s
EASY STAT	225,295.35 KiOP/s
EASY DELETE	57,648.44 KiOP/s
HARD WRITE	33,827.19 KiOP/s
HARD READ	141,467.16 KiOP/s
HARD STAT	230,086.03 KiOP/s
HARD DELETE	62,196.78 KiOP/s

Aurora IO500 Run

Features	Values
Number of MPI tasks/processes	63k
Number of DAOS servers	642
Number of DAOS engines	1284
Largest Pool	160PiB
Largest file	8.5PiB
Total number of files	177 Billions
Number of files in a single directory	33 Billions

SuperMUC NG System

SuperMUC NG Phase 2 **DAOS**

- 42x Lenovo Storage nodes
 - 2x Xeon 8352Y CPUs (ICX)
 - 512GB DRAM
 - 8x 3.84TB NVMe SSDs
 - 2x HDR IB NICs
 - 2TB Optane Persistent Memory 200
- 90x Client nodes



SuperMUC NG System Comparison

SuperMUC NG Phase 2 **DAOS**

- 42x Lenovo Storage nodes
 - 2x Xeon 8352Y CPUs (ICX)
 - 512GB DRAM
 - 8x 3.84TB NVMe SSDs
 - 2x HDR IB NICs
 - 2TB Optane Persistent Memory 200
- 90x Client nodes



IRIS MSKCC **WekaIO**

- 54x Dell Storage nodes
 - 2x Xeon 5317 CPUs (ICX)
 - 256GB DRAM
 - 8x 15TB NVMe SSDs
 - 2x HDR IB NICs
- 261x Client nodes

SuperMUC NG Performance Comparison

SuperMUC NG Phase 2 **DAOS**

IOR & FIND	
EASY WRITE	896.71 GiB/s
EASY READ	1,872.09 GiB/s
HARD WRITE	252.43 GiB/s
HARD READ	718.81 GiB/s
FIND	12,733.44 kIOP/s

IRIS MSKCC **WekaIO**

IOR & FIND	
EASY WRITE	383.77 GiB/s
EASY READ	1,076.53 GiB/s
HARD WRITE	51.68 GiB/s
HARD READ	190.49 GiB/s
FIND	424.40 kIOP/s

SuperMUC NG Performance Comparison

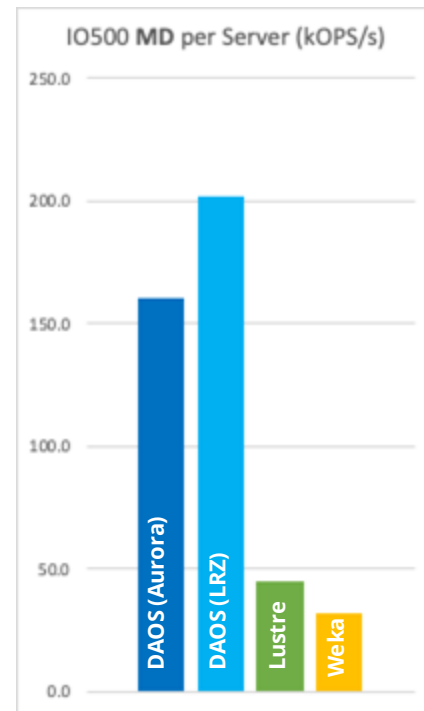
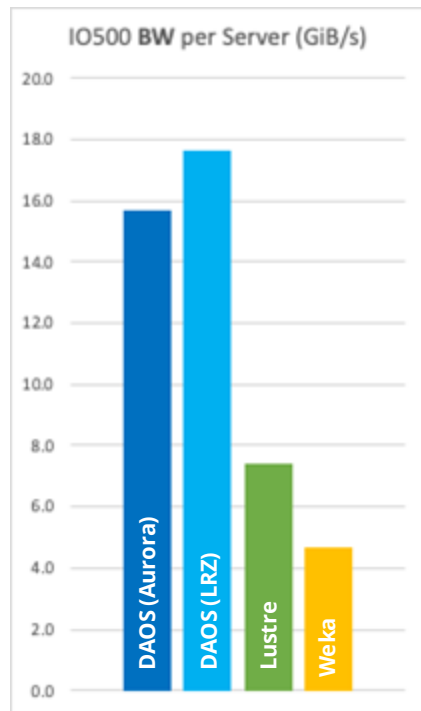
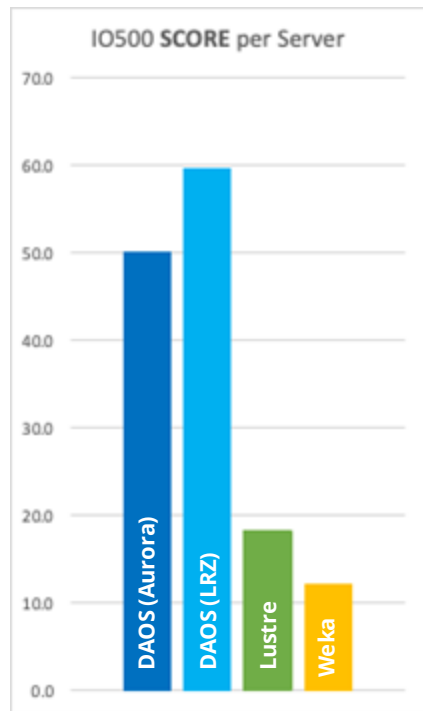
SuperMUC NG Phase 2 **DAOS**

METADATA	
EASY WRITE	6,324.79 kIOP/s
EASY STAT	29,403.34 kIOP/s
EASY DELETE	3,442.67 kIOP/s
HARD WRITE	2,644.93 kIOP/s
HARD READ	17,023.13 kIOP/s
HARD STAT	23,242.01 kIOP/s
HARD DELETE	3,112.59 kIOP/s

IRIS MSKCC **WekaIO**

METADATA	
EASY WRITE	1,484.48 kIOP/s
EASY STAT	15,370.21 kIOP/s
EASY DELETE	1,693.76 kIOP/s
HARD WRITE	281.11 kIOP/s
HARD READ	6,806.84 kIOP/s
HARD STAT	8,791.83 kIOP/s
HARD DELETE	324.23 kIOP/s

IO500 Per-server Performance (production list)



Google Parallelstore Performance

Performance

Expected performance from Parallelstore is shown in the following table.

Metric	Result
Write Throughput	0.5 GiBps per TiB
Read throughput	1.15 GiBps per TiB
Read IOPS	30k IOPS per TiB
Write IOPS	10k IOPS per TiB
4K Read Latency	0.3 ms
Number of client processes supported	4000
Transfer speed (Parallelstore <-> Cloud Storage)	Maximum transfer rate of 20 GiBps or 5000 files per second
Mean time to data loss (MTTDL)	100 TiB capacity: 2 months 48 TiB capacity: 4 months 12 TiB capacity: 16 months

These numbers are measured using 256 client connections to a single instance. Latency is measured from a single client. Directory and file striping settings are optimized for each metric.

Resources

- Foundation website: <https://daos.io/>
- Github: <https://github.com/daos-stack/daos>
- Online doc: <https://docs.daos.io>
- Mailing list & slack: <https://daos.groups.io>
- YouTube channel: <http://video.daos.io>
- Virtual DAOS User Group on May 22, 2025: <https://daos.io/event/virtual-dug-25>

