

# DAOS: A High-Performance Scale-out Storage Stack

Andrey Kudryavtsev, DAOS Foundation FMS, High-Performance Storage for Data Center Session





### DAOS Foundation













### 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
  - O Maintain DAOS as an open source project independent of any one organization
  - O Foster the developer and user communities around DAOS
  - O Guide the direction of the overall DAOS project
  - O Promote the use of DAOS

### Governing Board

- O Defines budget and approves expenses
- O Oversee efforts of other subcommittees
- Approve roadmap provided by TSC
- Vote on matters as needed





## Meetings

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





### How to Join

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

Join		
Charter		
Press Release		
TSC	low to Join the D	
Linux Foundation	oundation	
	Odiladioli	

On 09. November 2023, the founding members Argonn Labs, Hewlett Packard Enterprise, Google Cloud, and In Foundation to broaden the governance of the **Distribu Storage (DAOS)** open source project. See the <u>LF Pressign Parameters</u>

Architecture ~

Communit

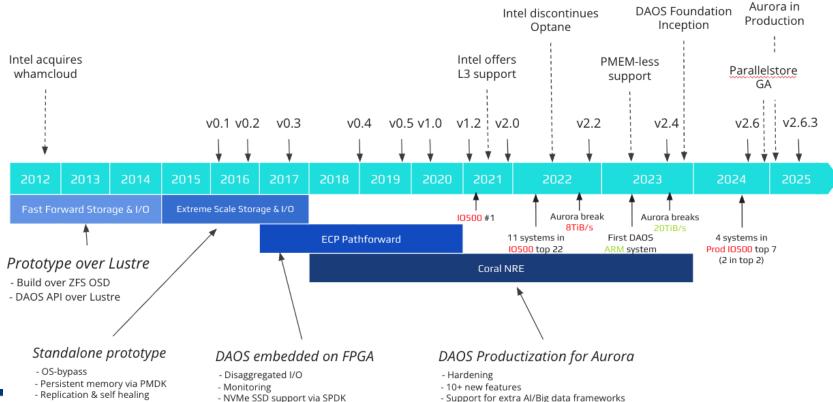
DAOS Foundation Membership Level	Annual Fees
Premier	25,000 USD
Premier for LF Associate Members	15,000 USD
General	15,000 USD
General for LF Associate Members	6,000 USD
Associate for LF Associate Members	0 USD

DAOS





## **DAOS** History

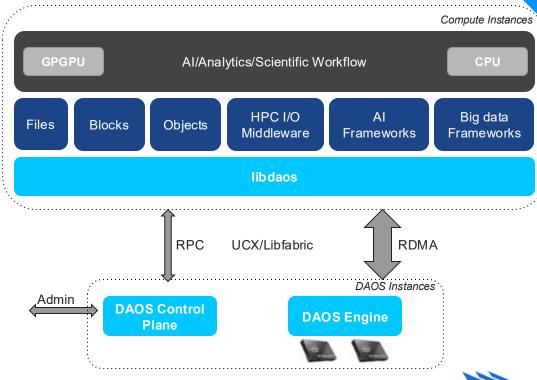




FOUNDATION

# DAOS: Nextgen Open Storage Platform

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







# 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

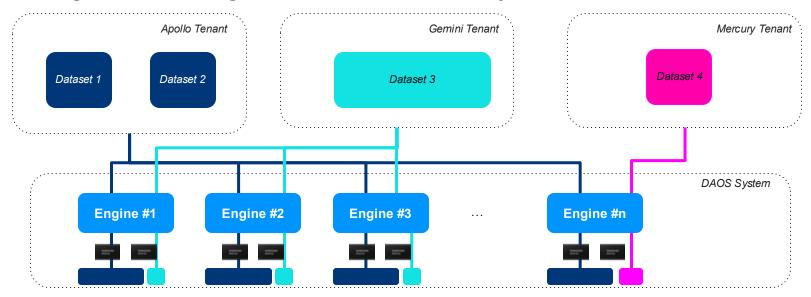
High IOPS

Unique Capabilities





# Storage Pooling - Multi-tenancy



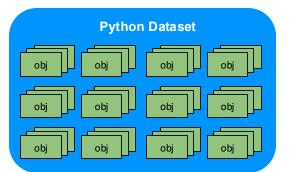
Pool 1	Apollo Tenant	100PB	20TB/s	200M IOPS	
Pool 2	Gemini Tenant	10PB	2TB/s	20M IOPS	
Pool 3	Mercury Tenant	30TB	80GB/s	2M IOPS	

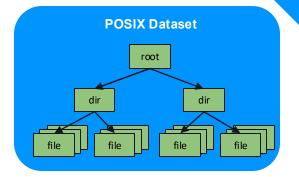


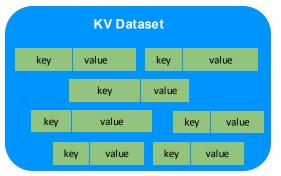


## Dataset Management

- New data movel to unwind 30+y 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



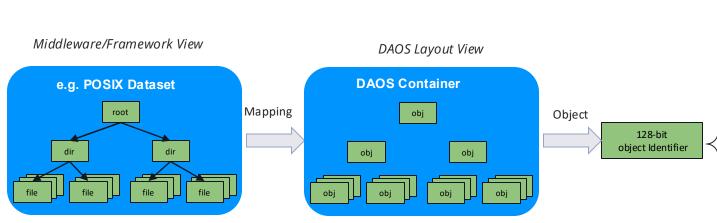




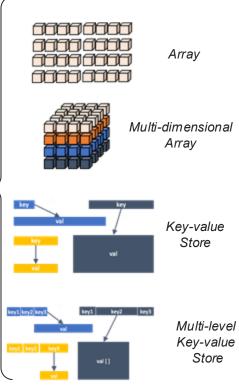




# Object Interface



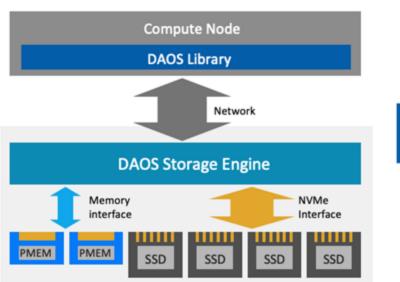
- No object create/destroy
- No size, permission/ACLs or attributes
- Sharded and erasure-coded/replicated
- Algorithmic object placement
- Very short Time To First Byte (TTFB)



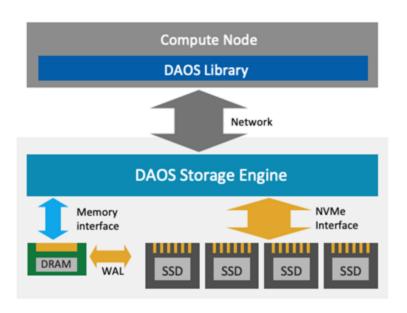




### DAOS Architecture Evolution







With Persistent Memory

Without Persistent Memory





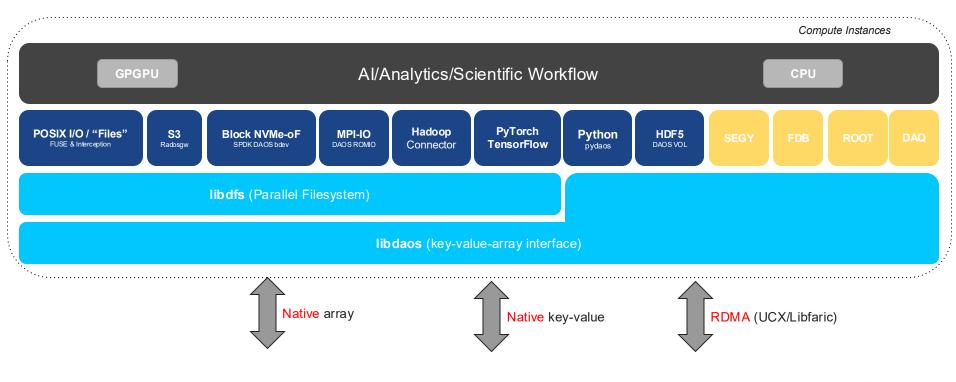






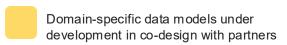


# Software Ecosystem











### 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

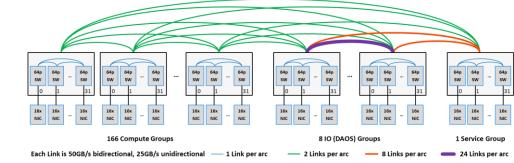


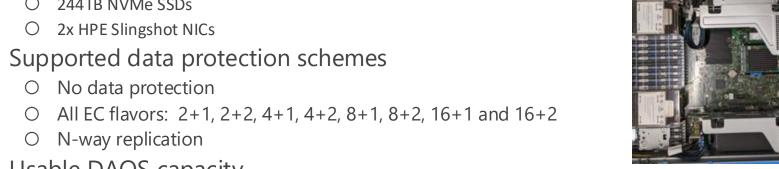


# Aurora DAOS System

- 1024x DAOS Storage nodes
  - 2x Xeon 5320 CPUs (ICX)
  - 512GB DRAM
  - 8TB Optane Persistent Memory 200
  - 244TB NVMe SSDs

- Usable DAOS capacity
  - between 220PB and 249PB depending on redundancy level chosen











### DAOS Performance - SC'24 Production List

			INFORMAT	TON					10500			
# T	BOF	INSTITUTION	SYSTEM	STORAGE	FILE SYSTEM	CLIENT	TOTAL	SCORE 1	BW	MD	REPRO.	
	BOF	P INSTITUTION	STSTEM VE	VENDOR	TYPE	NODES	PROC.	SCORE	(GIB/S)	(KIOP/S)	REPRO.	
0	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	0	
3	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	٥	
0	SC24	MSKCC	IRIS	WekalO	WekalO	261	27,144	665.49	252.54	1,753.69	٥	
5	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	0	



10

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







# SuperMUC NG System Comparison

### SuperMUC NG Phase 2 DAOS

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



### IRIS MSKCC WekalO

- 54x Dell Storage nodes
  - O 2x Xeon 5317 CPUs (ICX)
  - O 256GB DRAM
  - O 8x 15TB NVMe SSDs
  - O 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 WekalO

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

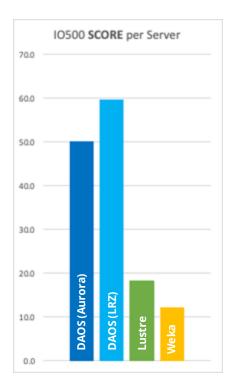
### IRIS MSKCC WekalO

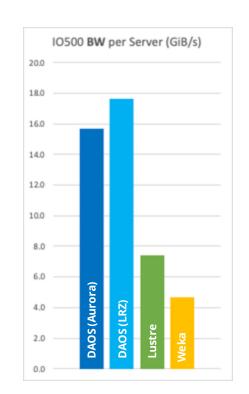
METADATA	
me rapara	
EASY WRITE	1,484.48 kIOP/s
EASY STAT	15,370.21 kIOP/s
EASY DELETE	1,693.76 kIOP/s
HARD WRITE	281.11 kIOP/s
HARD READ	6,806.84 kIOP/s
HARD STAT	8,791.83 kIOP/s
HARD DELETE	324.23 kIOP/s

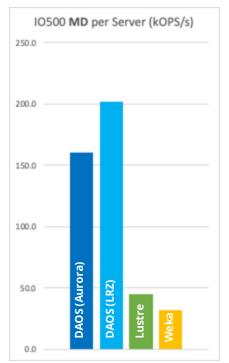




# IO500 Per-server Performance (production list)











### Color coding schema:

## DAOS Community Release

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

Jul'24 Q4'25 Q2'26 H2'26+

### DAOS 2.6

(Intel Release)

### OS Packages:

- Leap 15.5
- RHEL/Rocky/Alma 8.8/9.2

#### Networking:

- Change provider w/o reformat
- MD duplicate RPC detection

#### Features:

- Non-PMem support phase 1
- libpi14dfs
- Intel VMD hotplug
- Delayed rebuild

### Tech preview:

- Distributed consistency checker (CR)

### UX Improvements:

- Improved version interoperability

### (DAOS Foundation Release) OS Packages:

- Leap 15.6
- RHEL/Rocky/Alma 8.10/9.4

**DAOS 2.8** 

#### Networking:

- DOCA-OFED support

#### Features:

- Optimized object placement
- Mount POSIX snapshots RO
- Client telemetry
- Incremental rebuild/reintegration
- Catastrophic recovery and distributed consistency checker
- Fault domains beyond servers

### Tech preview:

- Non-PMem support phase 2
- Pytorch data loader
- Rolling upgrade preparation

### UX Improvements:

- Reintegration of all pools
- daos pool listing

### OS Packages:

- Leap 15.7 (x86 64)
- RHEL/Rocky/Alma 8.10/9.x (x86 64)
- RHEL/Rocky/Alma 9.x client (ARM64)

**DAOS 3.0** 

(DAOS Foundation Release)

- Ubuntu 22.04 client (x86 64/ARM64)

#### Features:

- Non-PMem support phase 2
- SSD hotplug & LED without VMD

#### Tech preview:

- Rolling upgrade
- WORM containers phase 1
- Multi-provider support
- flock support
- SSD encryption support via SED

### **Future Releases**

(DAOS Foundation Release)

### OS Packages:

- Leap 15.7 (x86\_64)
- RHEL/Rocky/Alma 8.10/9.x (x86\_64)
- RHEL/Rocky/Alma 9.x (ARM64/x64 64)
- Ubuntu 24.04 client (x86 64/ARM64)

#### Features:

- Pool resizing
- Inline compression
- Inline encryption
- Inline deduplication
- Middleware consistency checker
- Progressive layout
- Pipeline API
- SQL support with predicate pushdown
- Distributed transactions
- Pool/container freeze
- CXL SSD support / QLC
- Tiered container phase 1
- Support for multiple DAOS systems
- multi-NIC support per engine/process
- hardlinks support in libdfs
- network multipath support
- Container parking/serialization





### Resources



- Foundation website: <a href="https://daos.io/">https://daos.io/</a>
- Github: <a href="https://github.com/daos-stack/daos">https://github.com/daos-stack/daos</a>
- Online doc: <a href="https://docs.daos.io">https://docs.daos.io</a>
- Mailing list & slack: <a href="https://daos.groups.io">https://daos.groups.io</a>
- YouTube channel: <a href="http://video.daos.io">http://video.daos.io</a>
- Virtual DAOS User Group <a href="https://daos.io/event/virtual-dug-25">https://daos.io/event/virtual-dug-25</a>











