DAOS: A High-Performance Scale-out Storage Stack

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

https://foundation.daos.io

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

THE LINUX FOUNDATION

daos

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

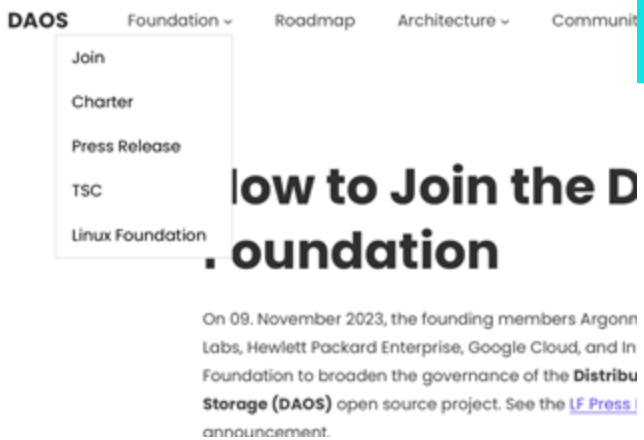
- Governing Board
  - Defines budget and approves expenses
  - Oversee efforts of other subcommittees
  - Approve roadmap provided by TSC
  - Vote on matters as needed

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

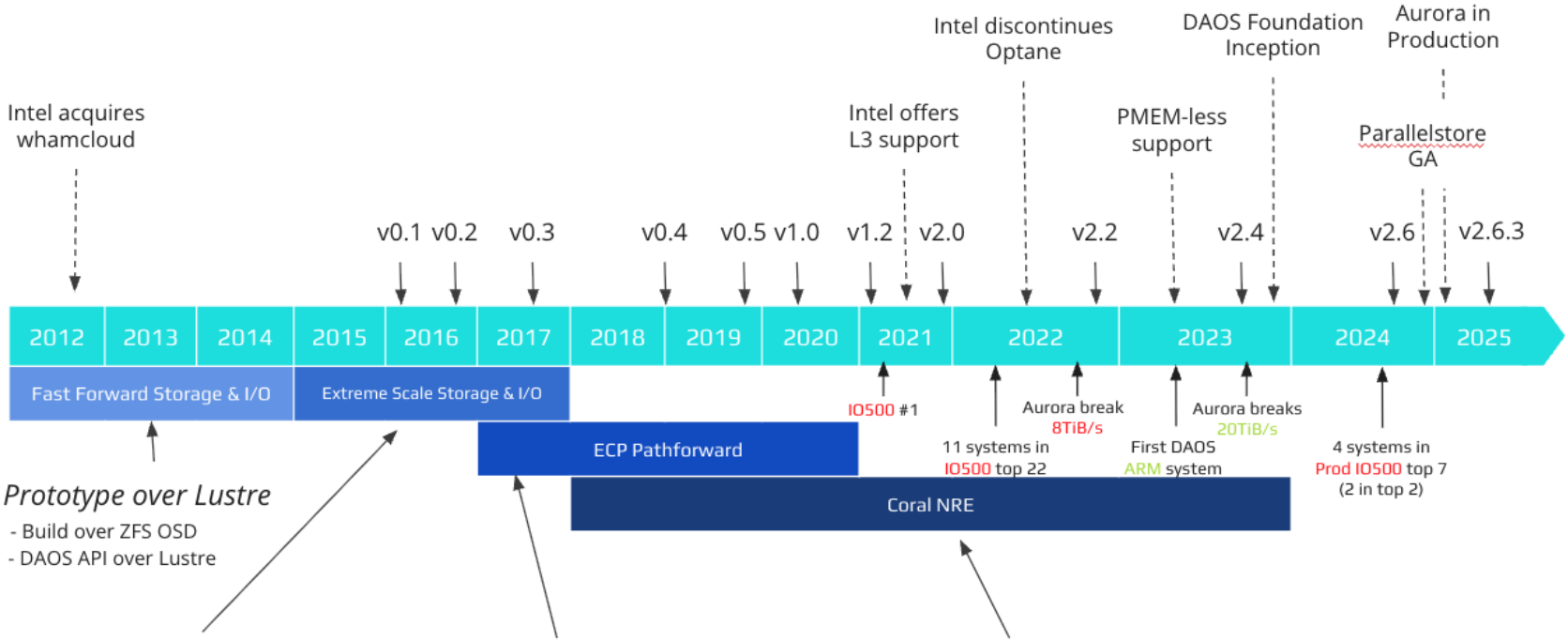


DAOS Foundation Membership Level	Annual Fees
Premier	25,000 USD
Premier for LF Associate Members	15,000 USD
General	15,000 USD
General for LF Associate Members	6,000 USD
Associate for LF Associate Members	0 USD

THE LINUX FOUNDATION

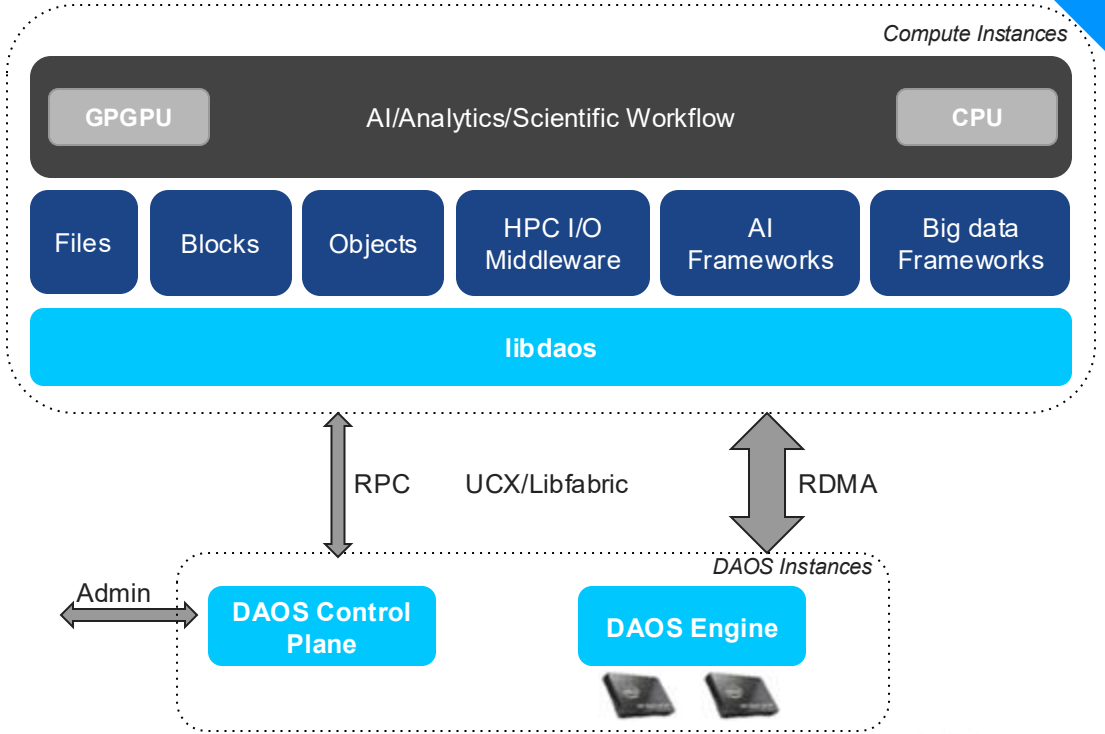
daos

DAOS History



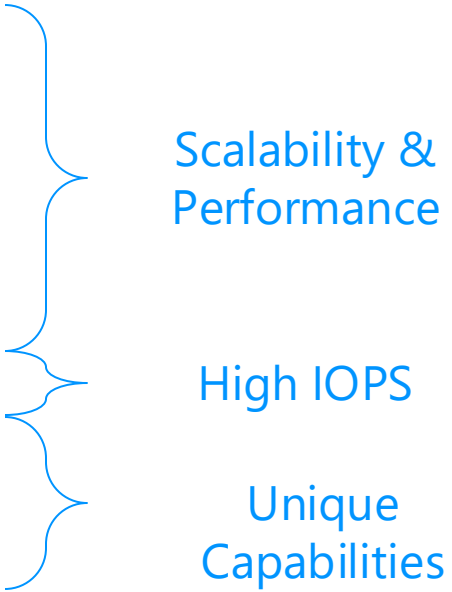
DAOS: Nextgen Open Storage Platform

- 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) µs

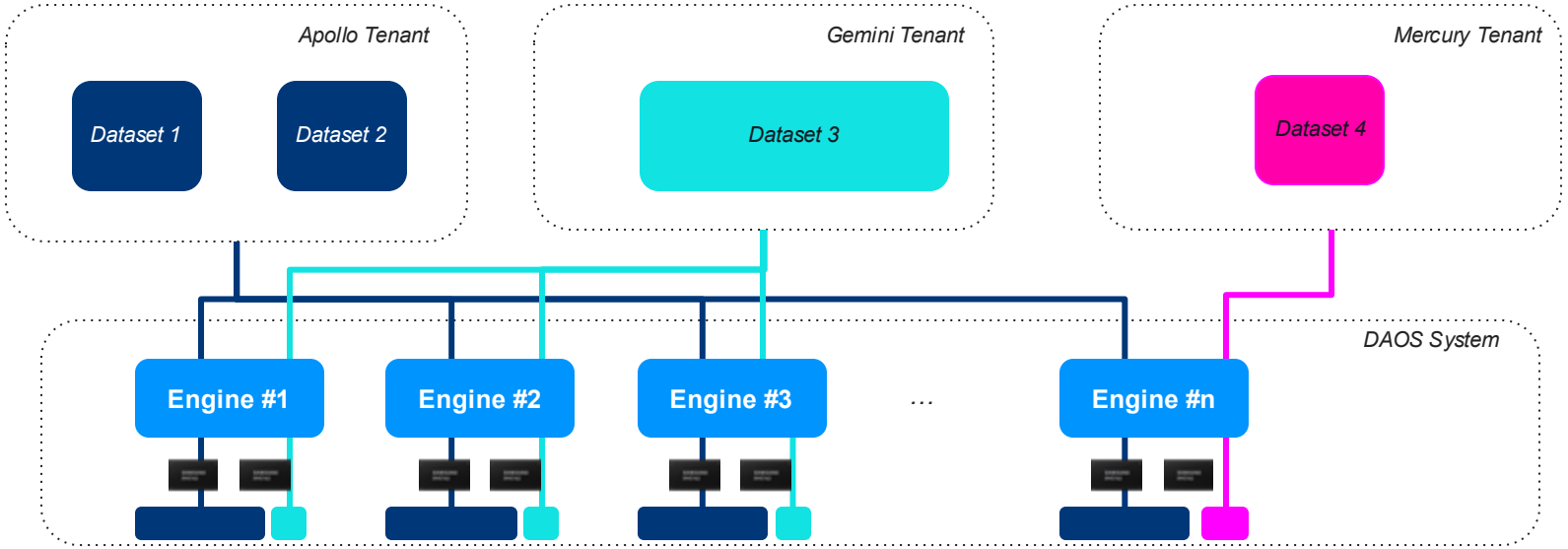


DAOS Design Fundamentals

- 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



Storage Pooling - Multi-tenancy

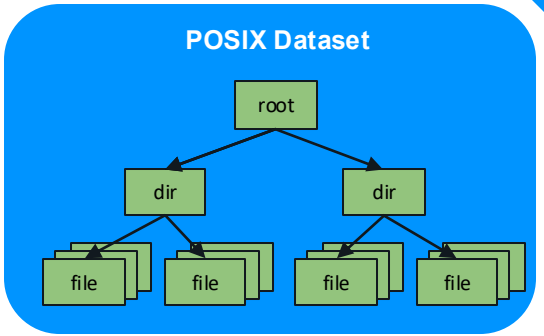


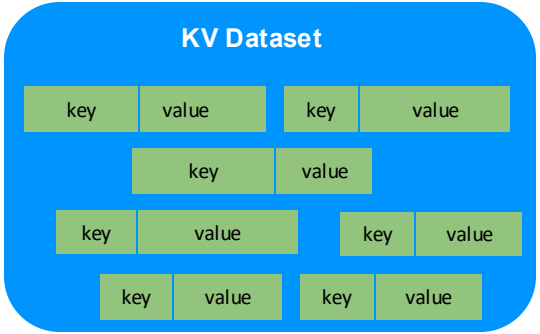
Pool 1	Apollo Tenant	100PB	20TB/s	200M IOPS
Pool 2	Gemini Tenant	10PB	2TB/s	20M IOPS
Pool 3	Mercury Tenant	30TB	80GB/s	2M IOPS

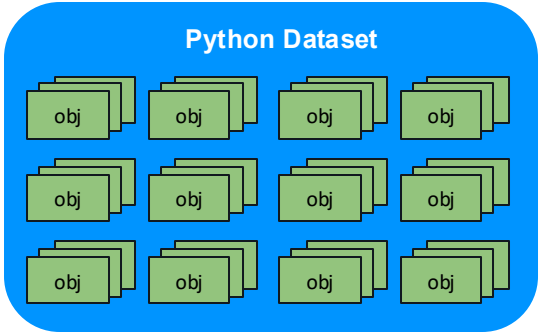
daos

Dataset Management

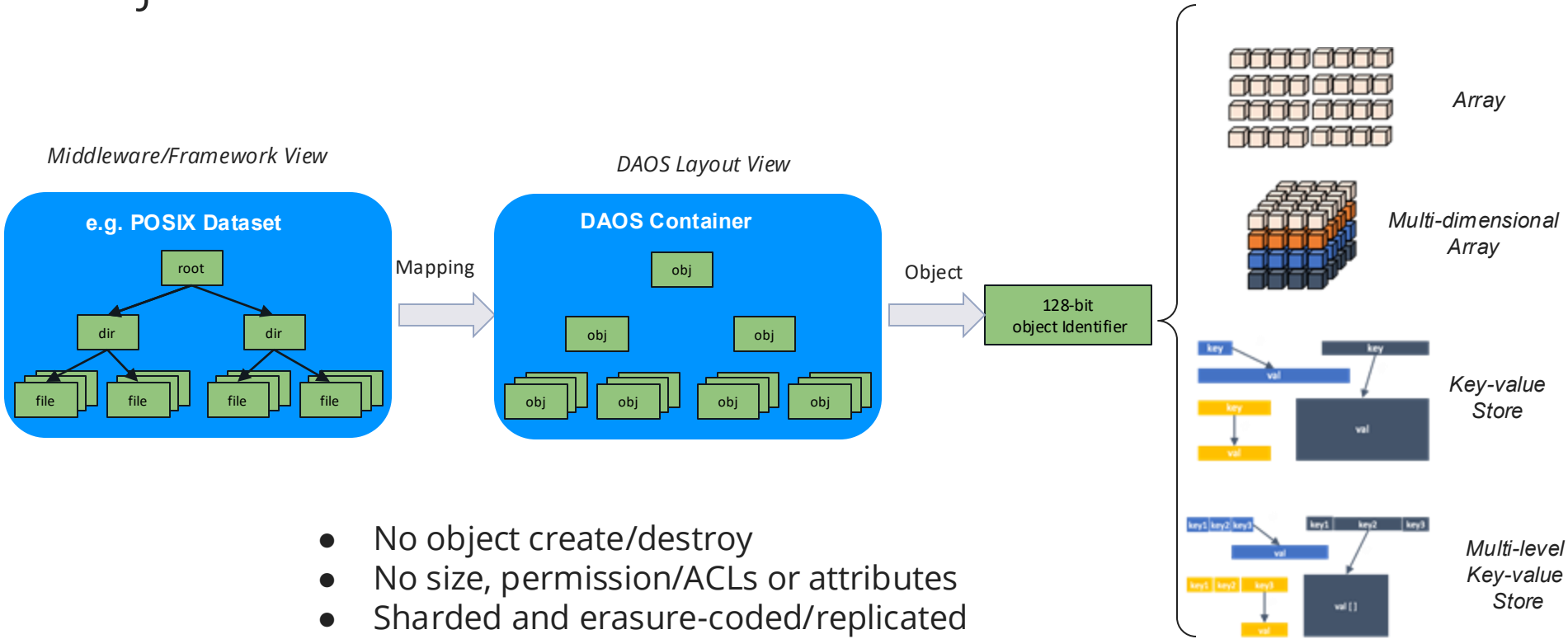
- New data movel to unwind 30+y of file-based management
- Introduce notion of dataset
- Basic unit of storage
- Datasets have a type
- POSIX datasets can include trillions of files/directories
- Advanced dataset query capabilities
- Unit of snapshots
- ACLs/IAM





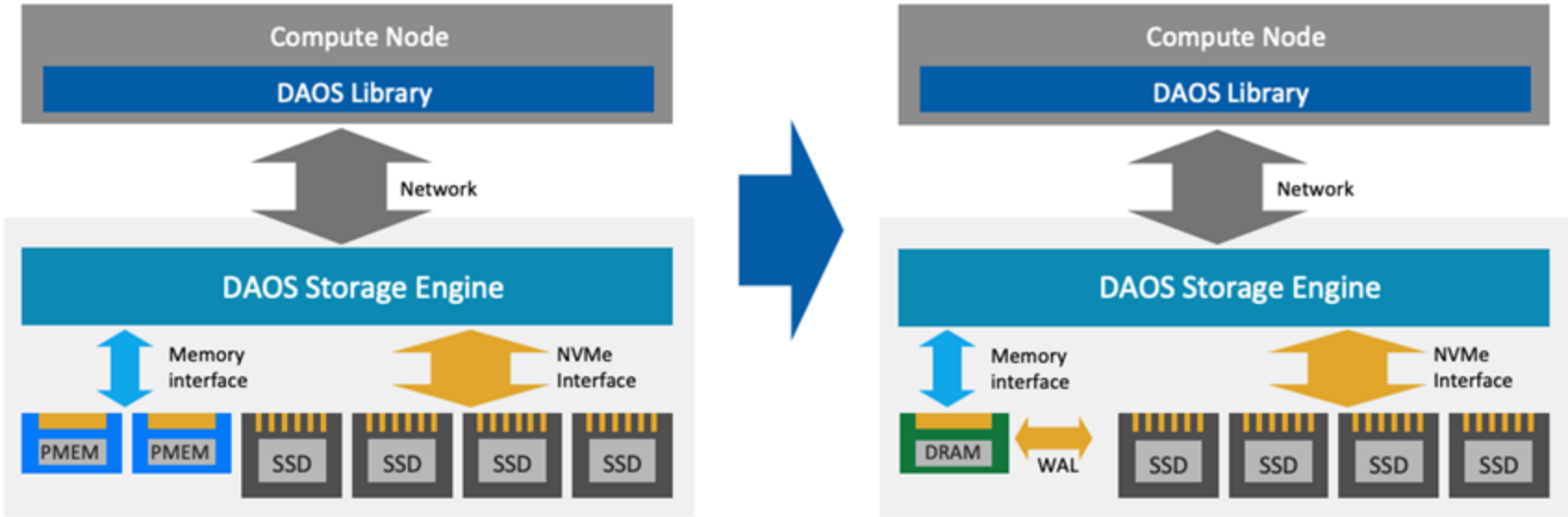


Object Interface

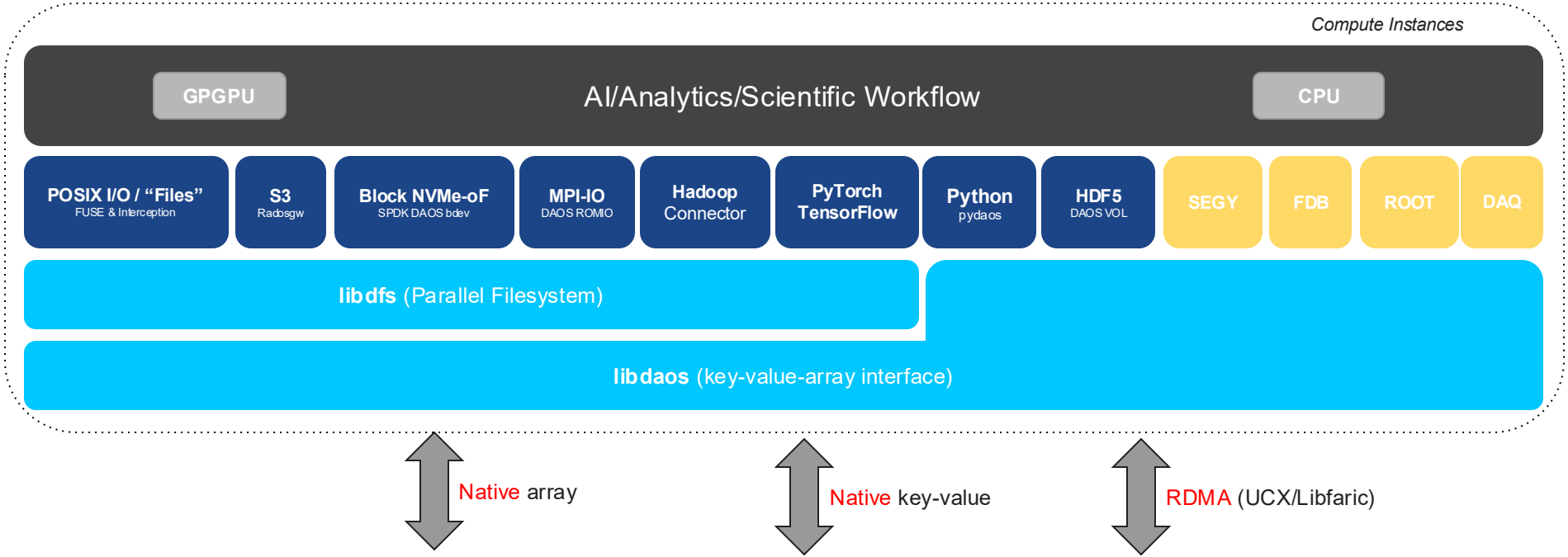


- No object create/destroy
- No size, permission/ACLs or attributes
- Sharded and erasure-coded/replicated
- Algorithmic object placement
- Very short Time To First Byte (TTFB)

DAOS Architecture Evolution



Software Ecosystem



Generic I/O Middleware/frameworks

Domain-specific data models under development in co-design with partners

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

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

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

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

System Interconnect
Slingshot 11; Dragonfly topology with adaptive routing

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

Peak Performance
≥ 2 Exaflop DP

High-Performance Storage
≥ 230 PB, ≥ 25 TB/s (DAOS)

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

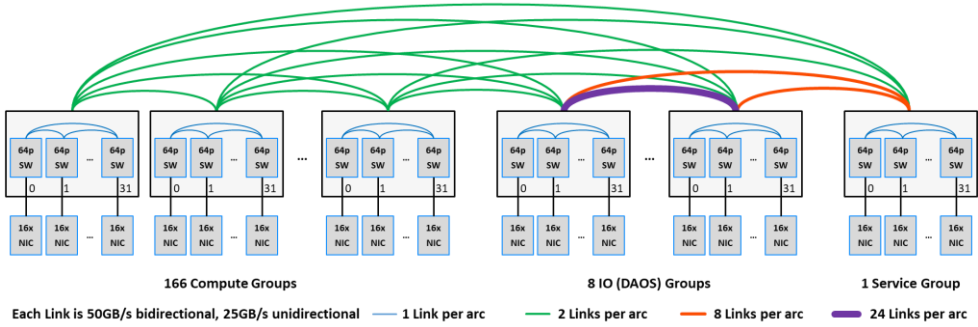
Platform
HPE Cray EX supercomputer

Aggregate System Memory
> 10 PB

Node Performance (TF)
> 130

System Size (# Nodes)
> 9,000

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

#	INFORMATION							IO500			
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW (GIB/S)	MD (KIOP/S)	REPRO.
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

THE LINUX FOUNDATION

daos

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

Source: https://io500.org/submissions/configuration/719
https://io500.org/submissions/view/683

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

Source: https://io500.org/submissions/configuration/719
https://io500.org/submissions/view/683

SuperMUC NG Performance Comparison

SuperMUC NG Phase 2 **DAOS**

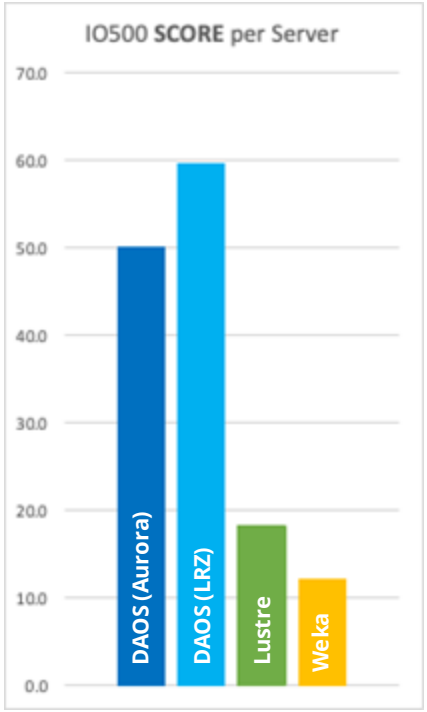
METADATA	
EASY WRITE	6,324.79 kIOP/s
EASY STAT	29,403.34 kIOP/s
EASY DELETE	3,442.67 kIOP/s
HARD WRITE	2,644.93 kIOP/s
HARD READ	17,023.13 kIOP/s
HARD STAT	23,242.01 kIOP/s
HARD DELETE	3,112.59 kIOP/s

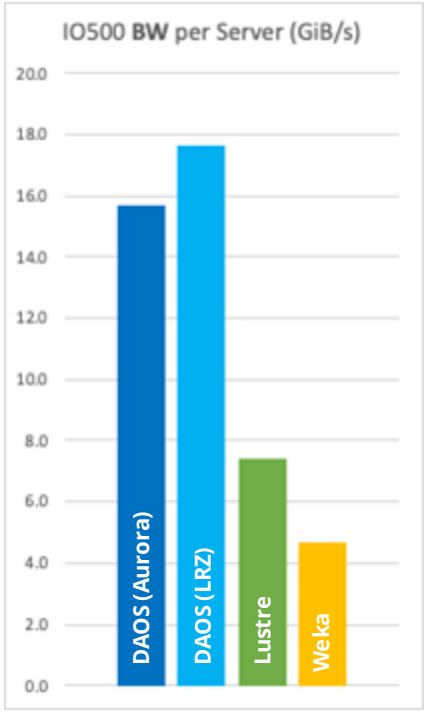
IRIS MSKCC **WekaIO**

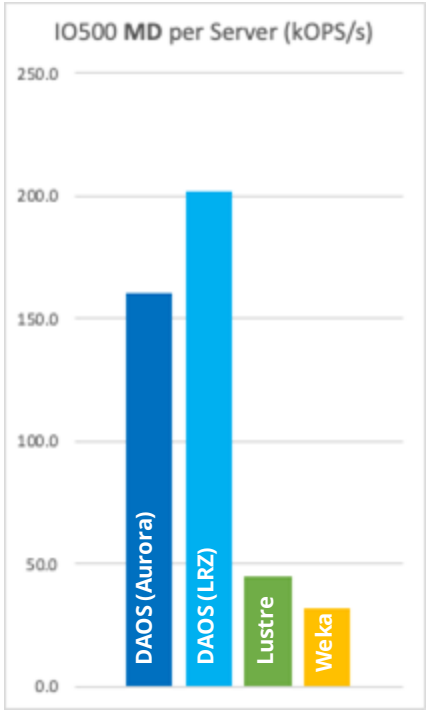
METADATA	
EASY WRITE	1,484.48 kIOP/s
EASY STAT	15,370.21 kIOP/s
EASY DELETE	1,693.76 kIOP/s
HARD WRITE	281.11 kIOP/s
HARD READ	6,806.84 kIOP/s
HARD STAT	8,791.83 kIOP/s
HARD DELETE	324.23 kIOP/s

Source: https://io500.org/submissions/configuration/719
https://io500.org/submissions/view/683

IO500 Per-server Performance (production list)



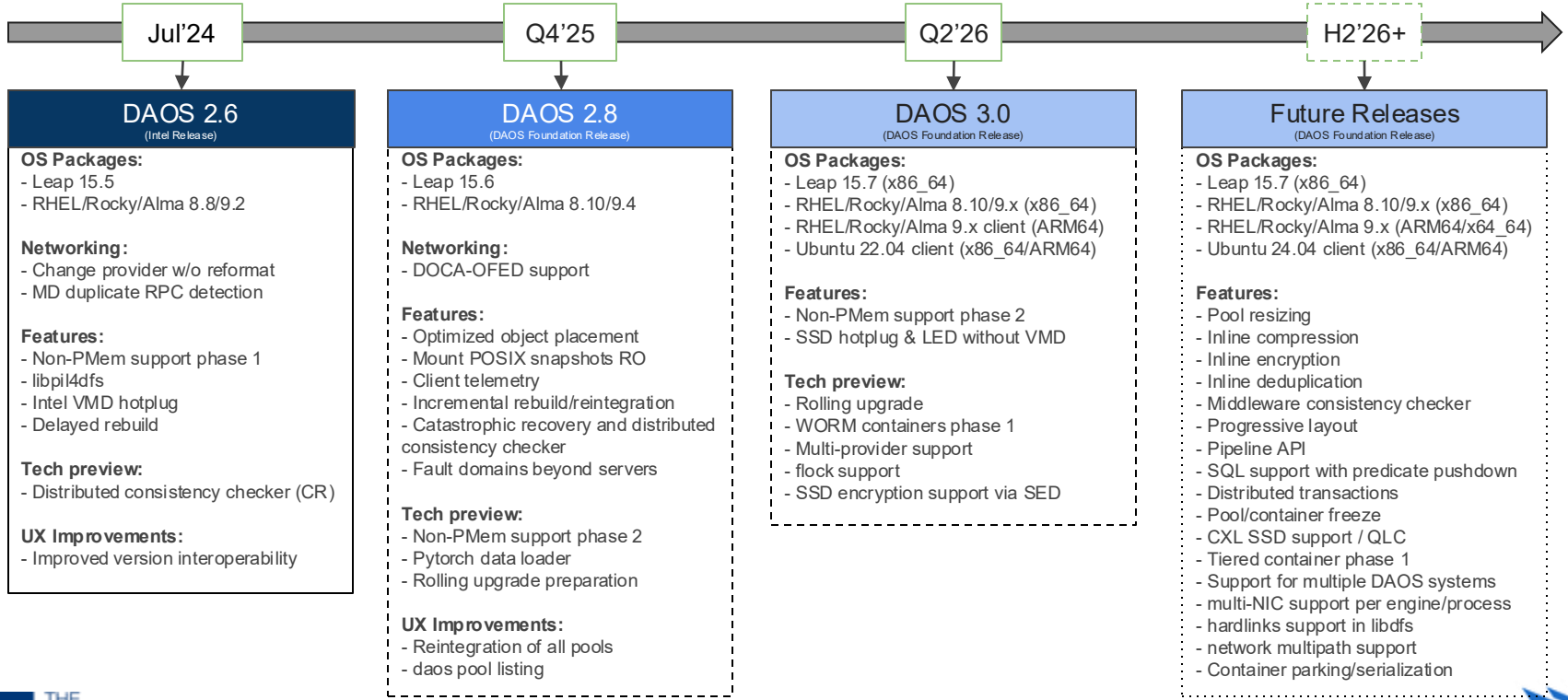




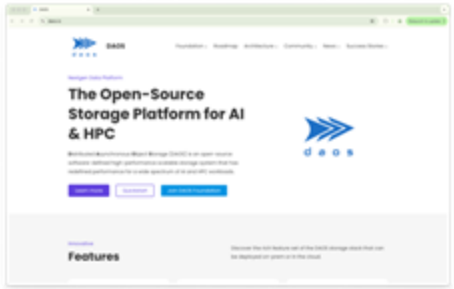
Source: https://io500.org

DAOS Community Release

Color coding schema:
- Committed (or released) release/features
- In-planning release/features
- Future possible release/features



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 https://daos.io/event/virtual-dug-25

THE LINUX FOUNDATION

daos