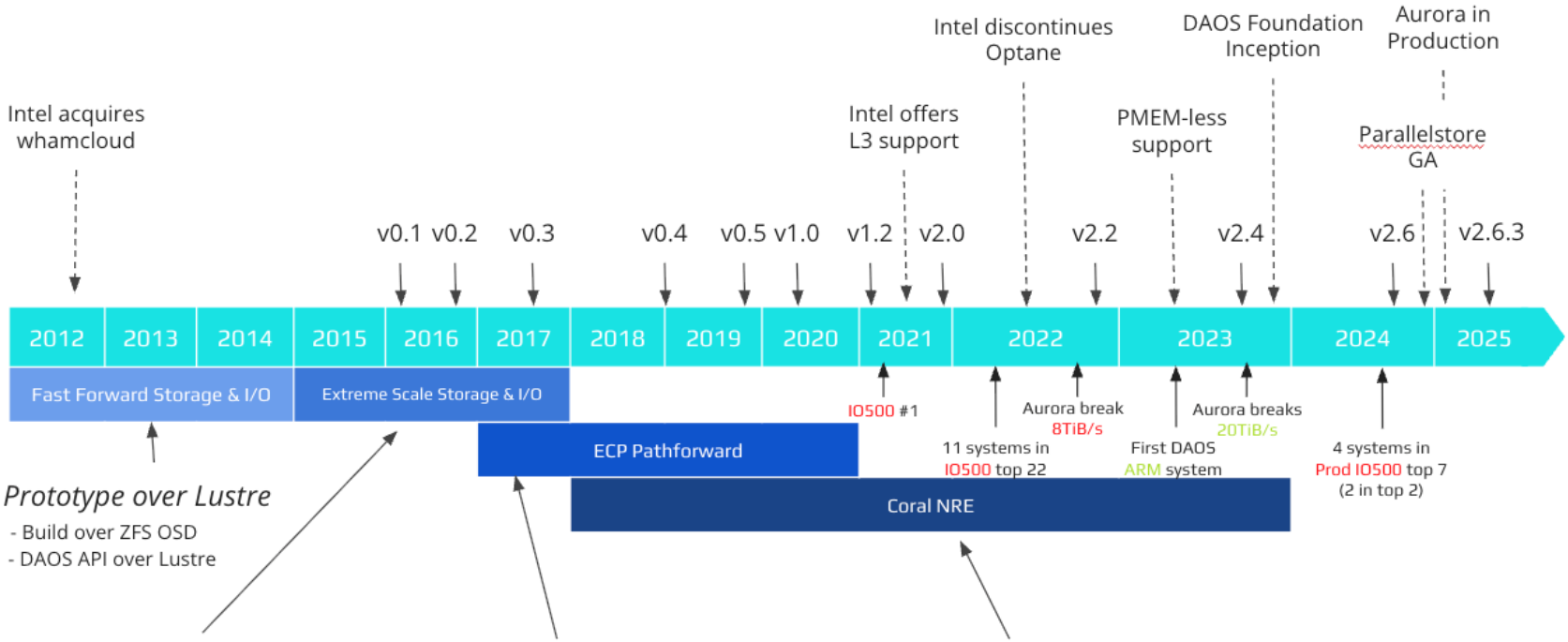
DAOS: A High-Performance Scale-out Storage Stack

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

https://foundation.daos.io

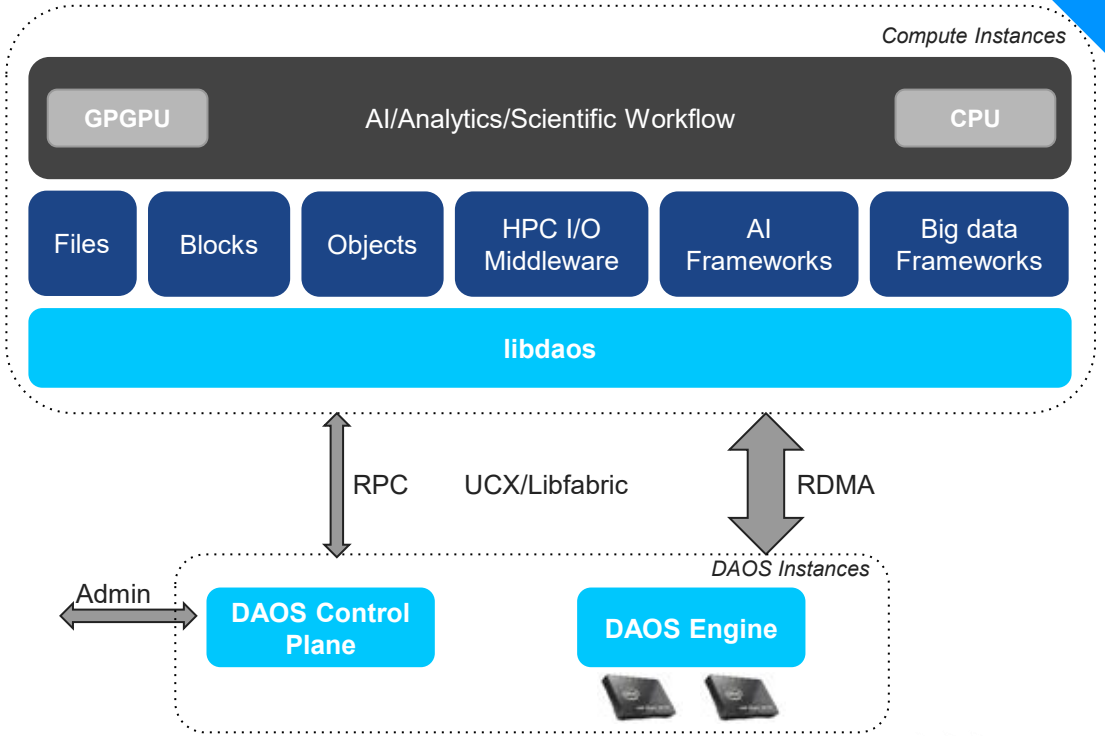
DAOS Foundation

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

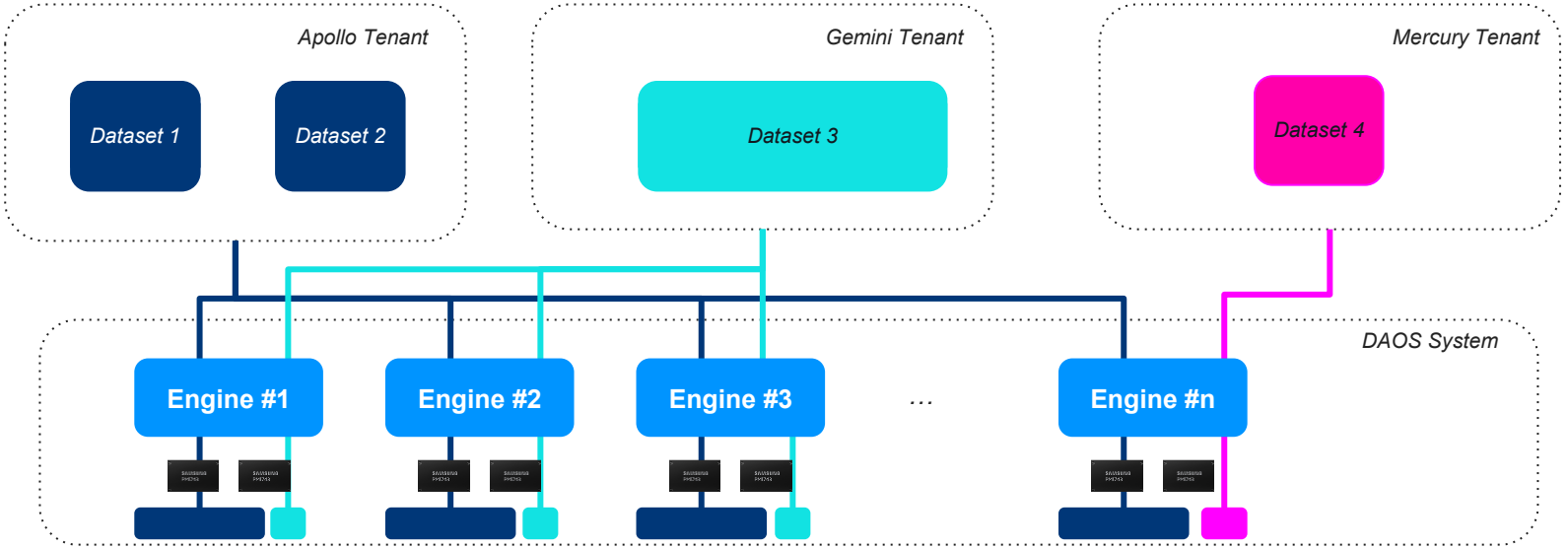
- No read-modify-write on I/O path (use versioning)
- No locking/DLM (use MVCC)
- No client tracking or client recovery
- No centralized (meta)data server
- No global object table
- Non-blocking I/O processing (futures & promises)
- Serializable distributed transactions
- Built-in multi-tenancy
- User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

Scalability & Performance

High IOPS

Unique Capabilities

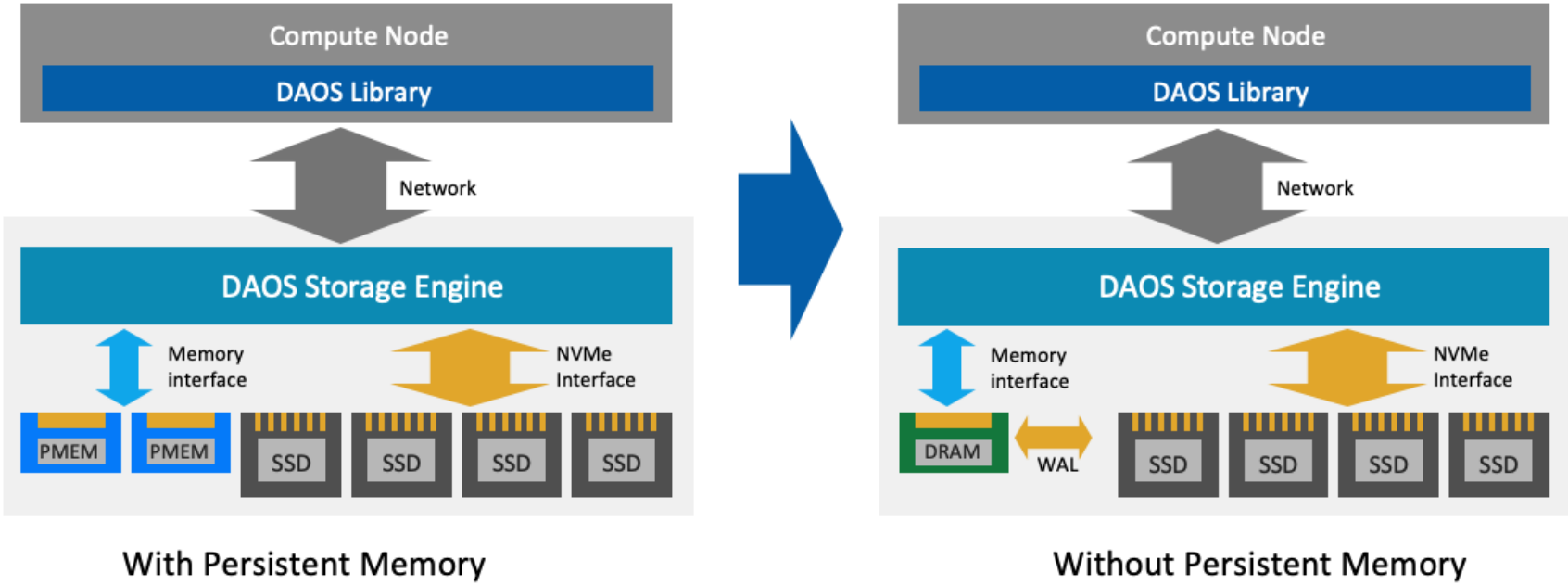
Storage Pooling - Multi-tenancy



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

daos

DAOS Architecture Evolution



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++

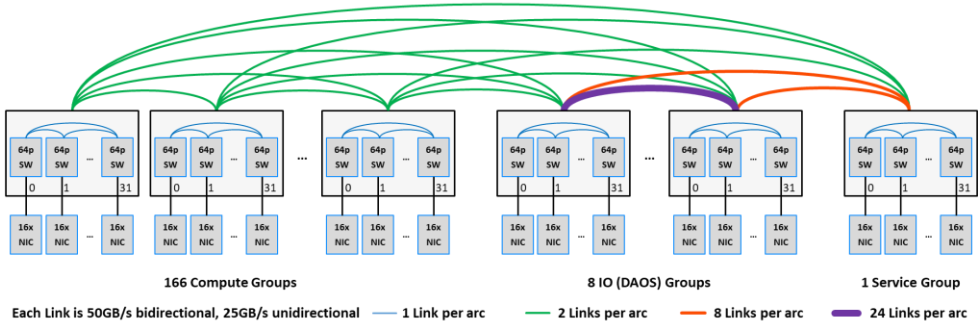
Platform
HPE Cray EX supercomputer

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

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**

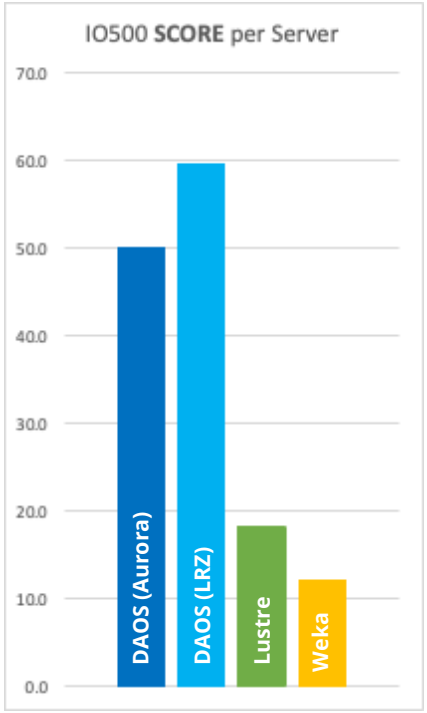
- 42x Lenovo Storage nodes
  - 2x Xeon 8352Y CPUs (ICX)
  - 512GB DRAM
  - 8x 3.84TB NVMe SSDs
  - 2x HDR IB NICs
  - 2TB Optane Persistent Memory 200
- 90x Client nodes

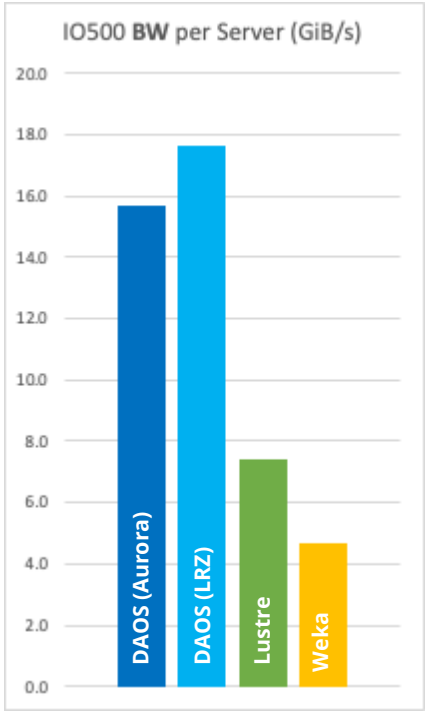
IRIS MSKCC **WekaIO**

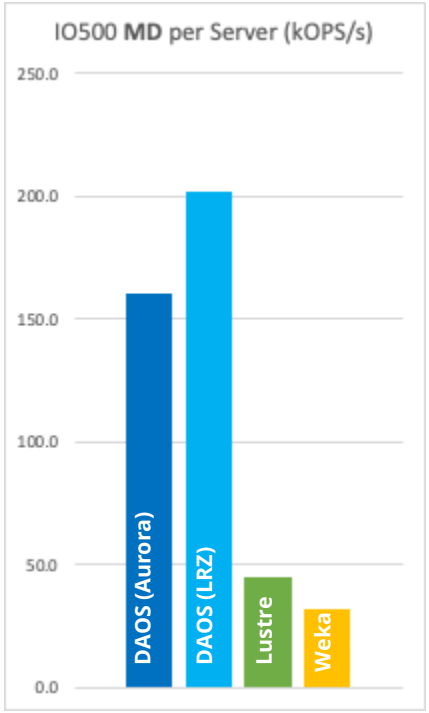
- 54x Dell Storage nodes
  - 2x Xeon 5317 CPUs (ICX)
  - 256GB DRAM
  - 8x 15TB NVMe SSDs
  - 2x HDR IB NICs
- 261x Client nodes

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

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

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
- libpil4dfs
- Intel VMD hotplug
- Delayed rebuild

Tech preview:
- Distributed consistency checker (CR)

UX Improvements:
- Improved version interoperability

DAOS 2.8
(DAOS Foundation Release)

OS Packages:
- Leap 15.6
- RHEL/Rocky/Alma 8.10/9.4

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

DAOS 3.0
(DAOS Foundation Release)

OS Packages:
- Leap 15.7 (x86_64)
- RHEL/Rocky/Alma 8.10/9.x (x86_64)
- RHEL/Rocky/Alma 9.x client (ARM64)
- 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

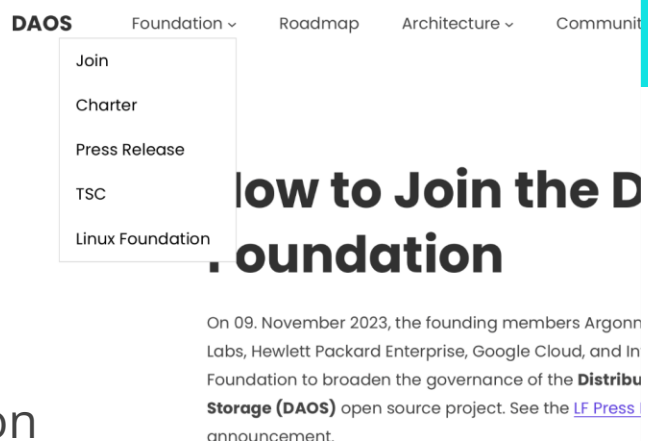
Future Releases
(DAOS Foundation Release)

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

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

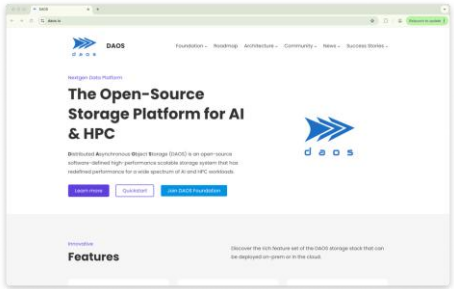


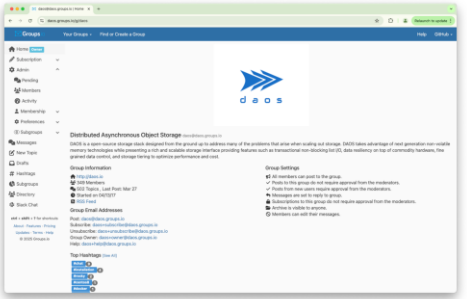
DAOS Foundation Membership Level	Annual Fees
Premier	25,000 USD
Premier for LF Associate Members	15,000 USD
General	15,000 USD
General for LF Associate Members	6,000 USD
Associate for LF Associate Members	0 USD

THE LINUX FOUNDATION

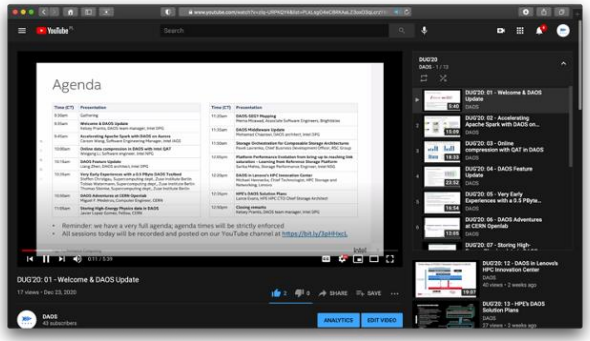
daos

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