Hewlett Packard
Enterprise

DAOS – New Horizons for High Performance Storage

Michael Hennecke and Jerome Soumagne, HPE

May 8, 2025

Introduction

- Monday Tutorial
  - Exploring High Performance Storage with DAOS (hands-on)

- Today’s Presentations
  - DAOS – New Horizons for High Performance Storage (overview)
  - Enhancing RPC on Slingshot for Aurora’s DAOS Storage System (networking aspects)

DAOS Foundation – A Project of the Linux Foundation

Learn more about how to join the DAOS Foundation here: https://daos.io/how-to-join-the-daos-foundation.

Argonne
NATIONAL LABORATORY

Google

Hewlett Packard
Enterprise

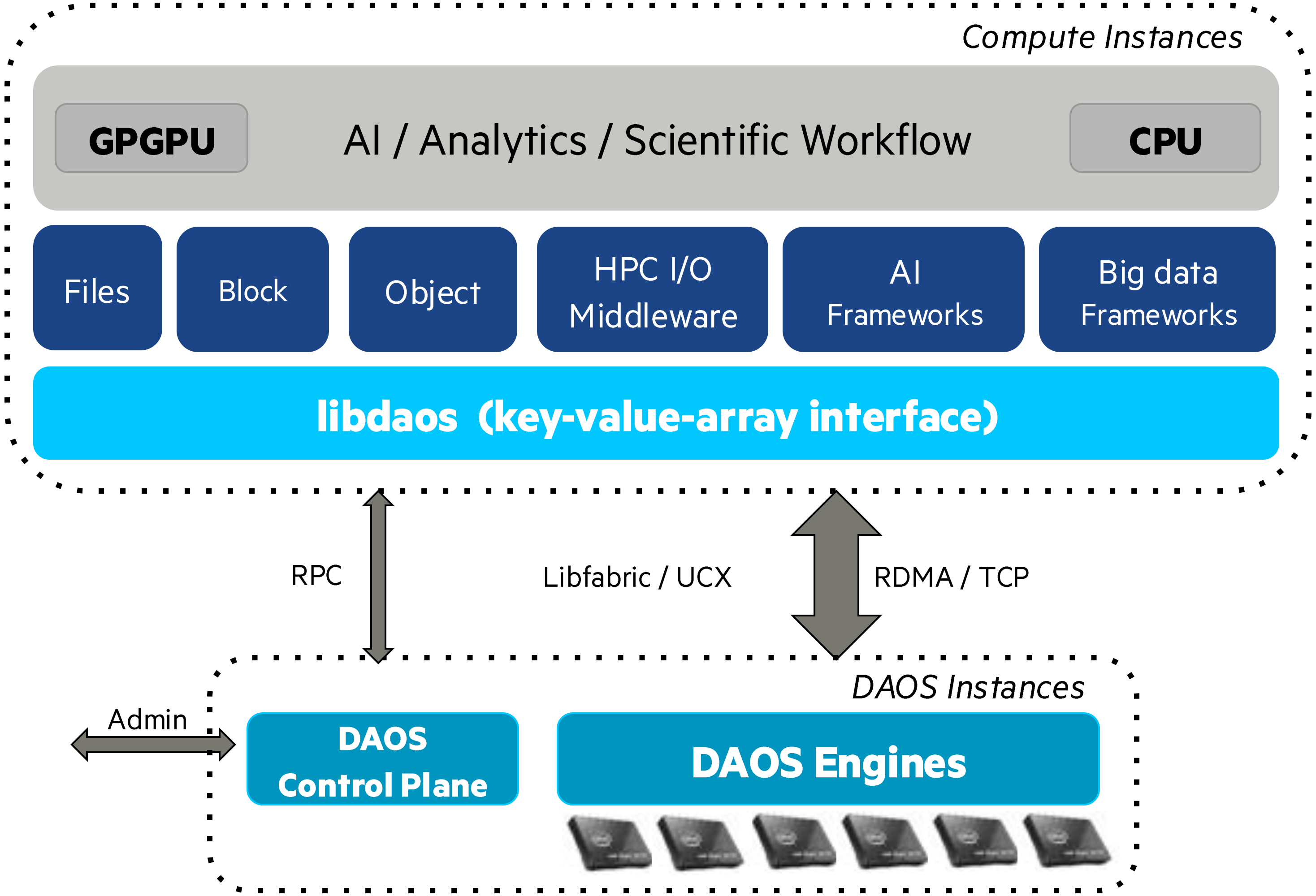
intel®

VDURA

DAOS: Nextgen Open Source Storage Platform

Distributed Asynchronous Object Storage

- Platform for innovation (beyond POSIX)
- 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
  - Tens of TB/s, billions of 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

Scalability & Performance

- Non-blocking I/O processing (not limited by threads)

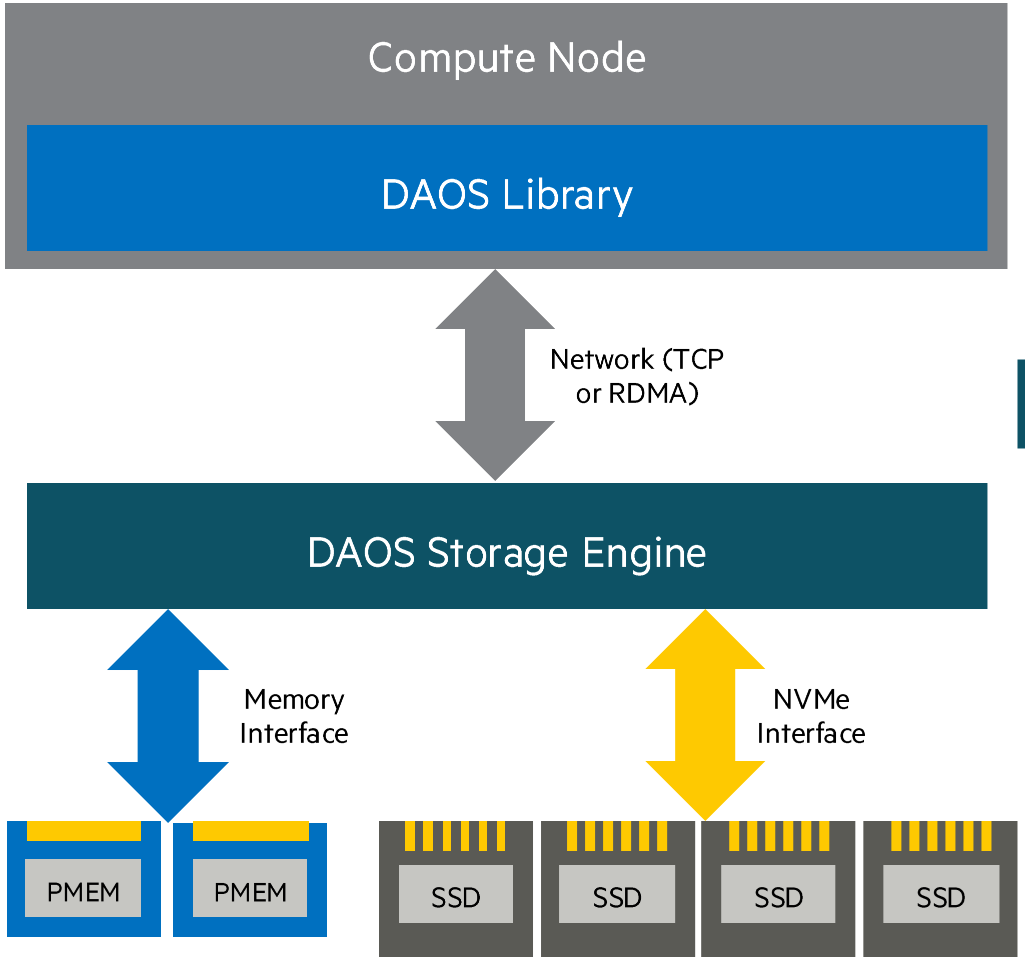
High IOPS

- Serializable distributed transactions
- Built-in multi-tenancy
- User snapshots (rollback exposed to user)

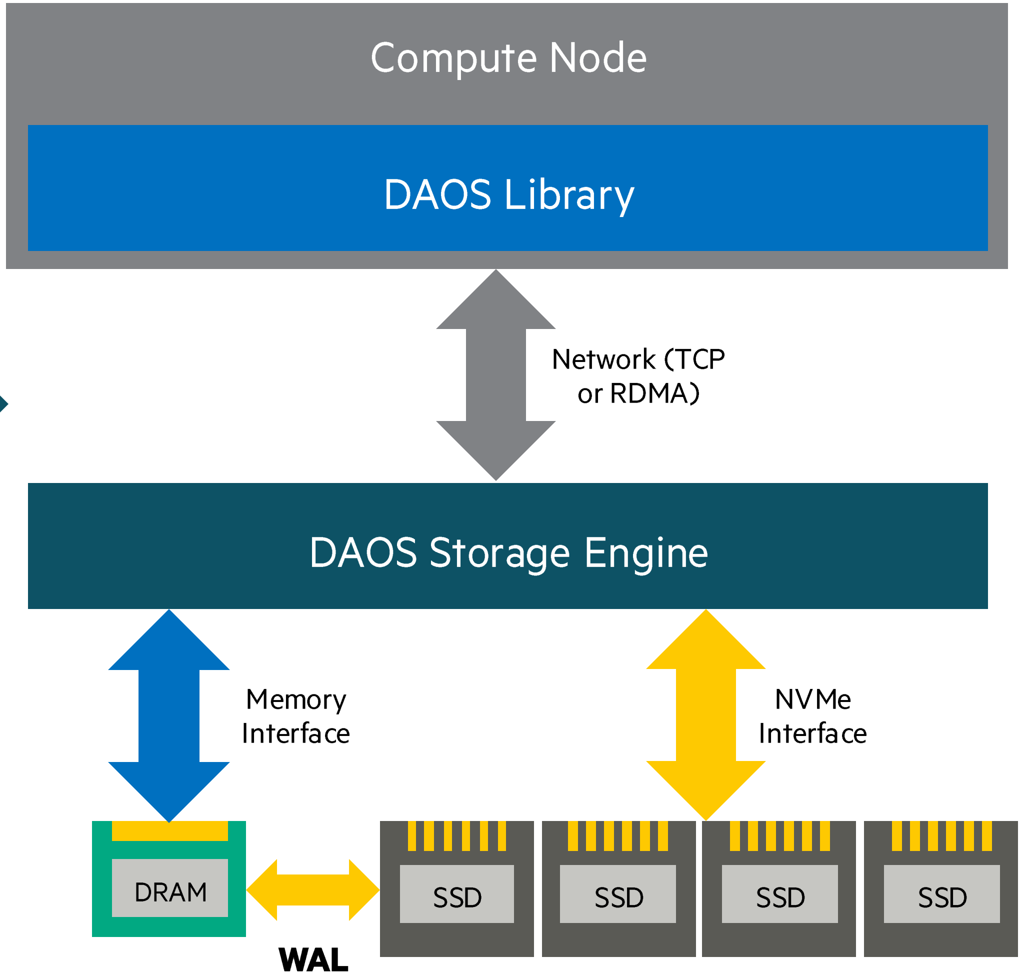
Unique Capabilities

DAOS Architecture Evolution

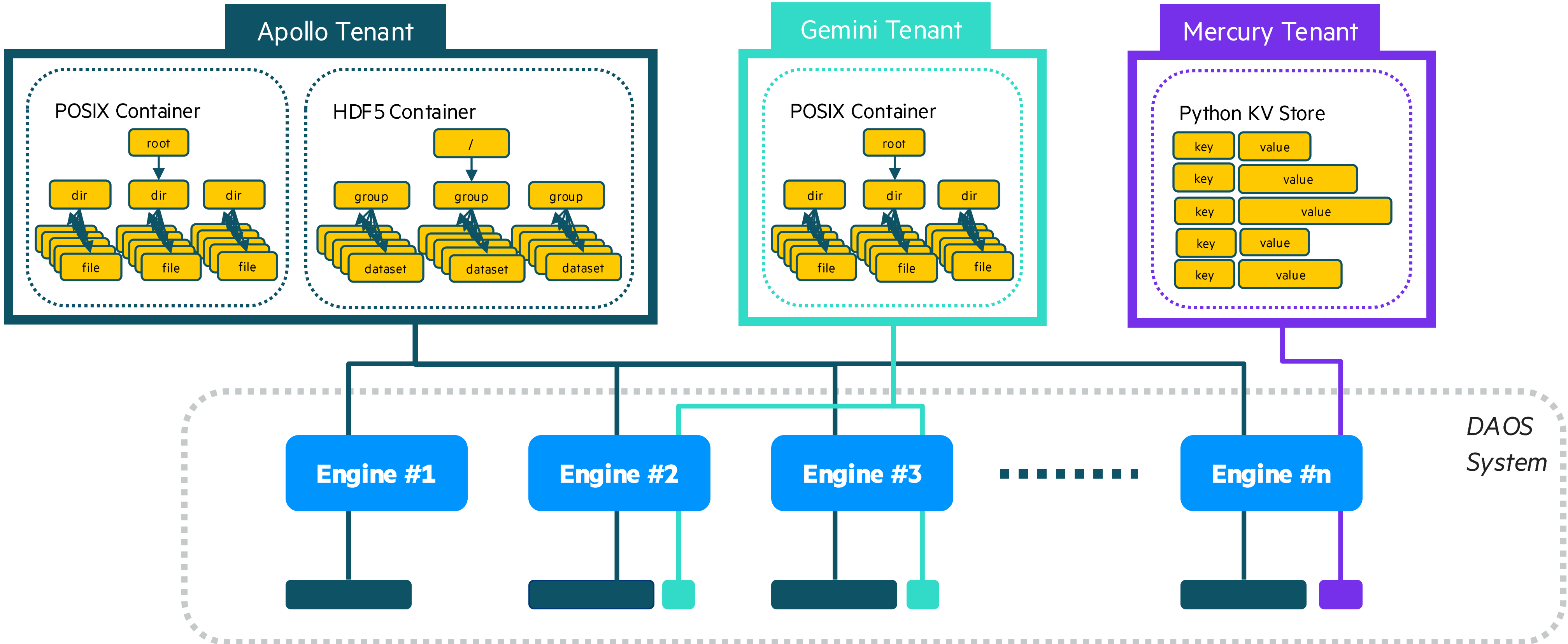
With Persistent Memory



Without Persistent Memory

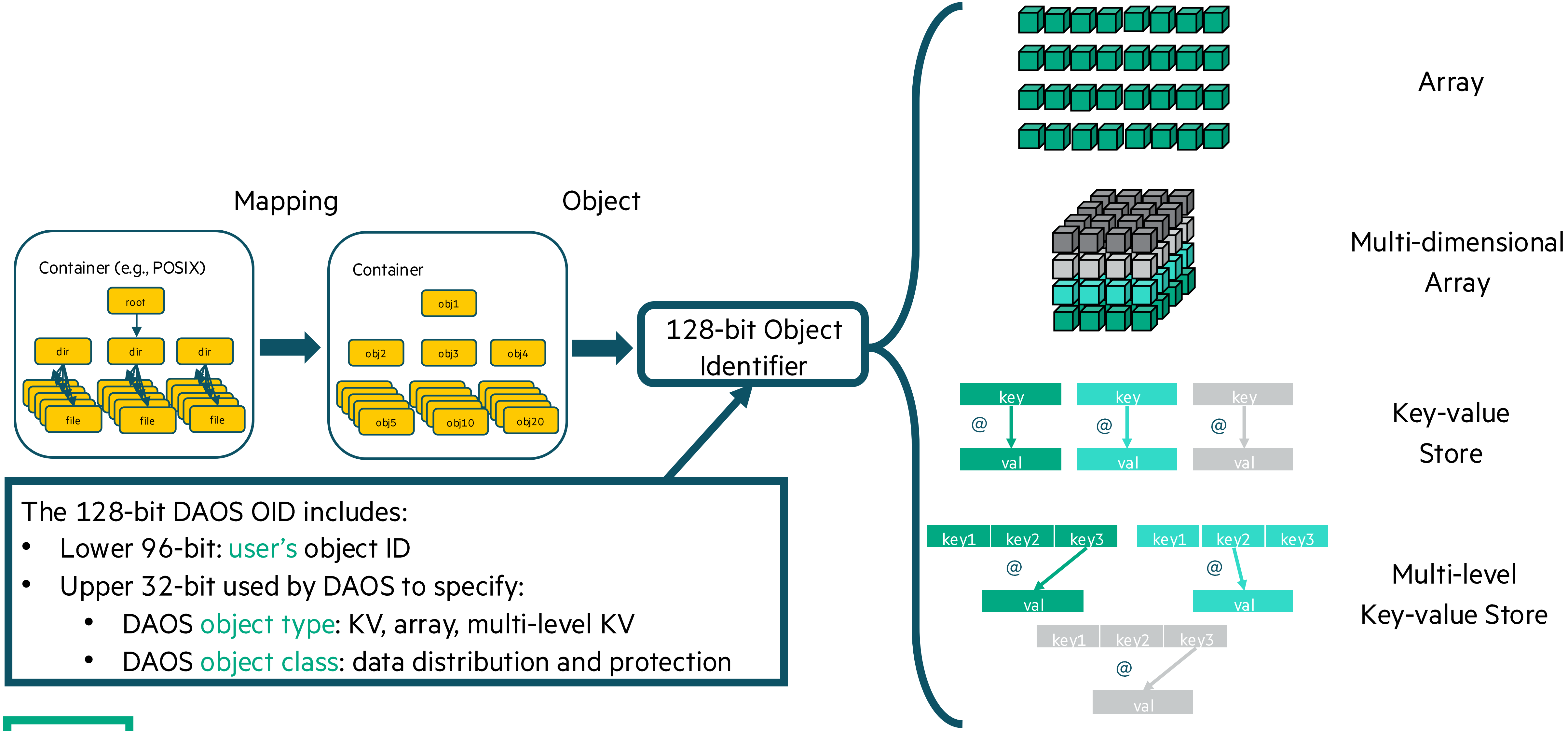


DAOS Storage Pooling and Multi-Tenancy



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

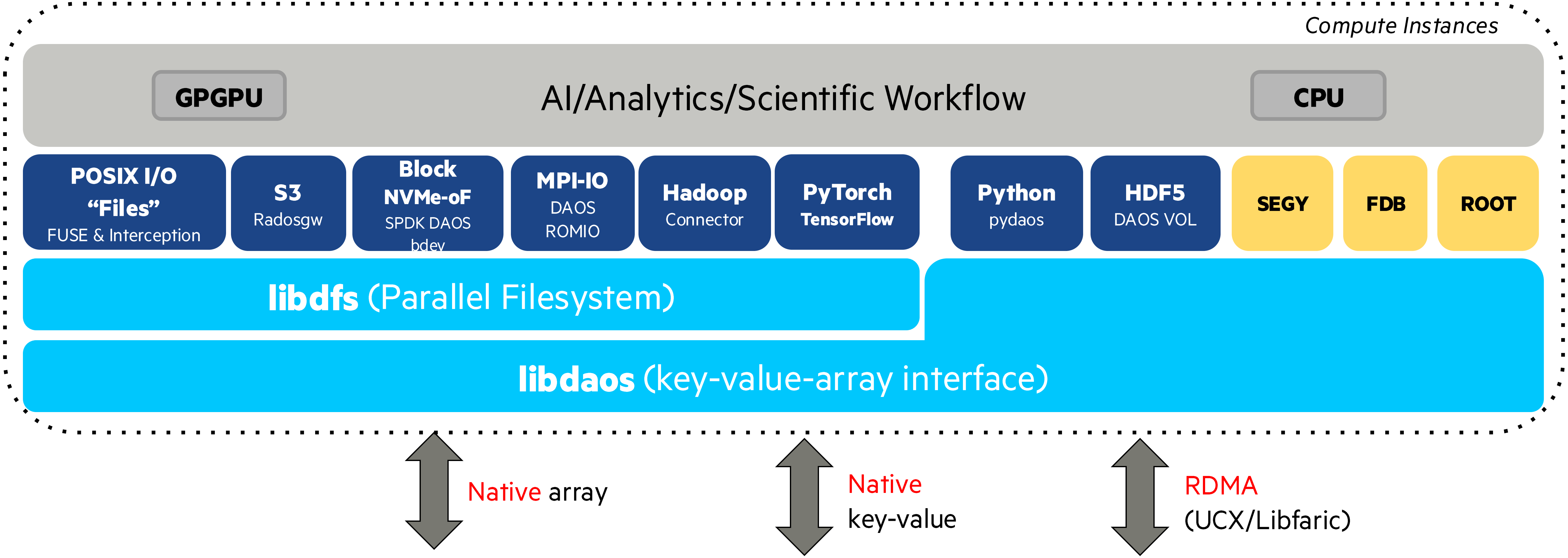
DAOS Object Model



The 128-bit DAOS OID includes:

- Lower 96-bit: user’s object ID
- Upper 32-bit used by DAOS to specify:
  - DAOS object type: KV, array, multi-level KV
  - DAOS object class: data distribution and protection

DAOS Software Ecosystem

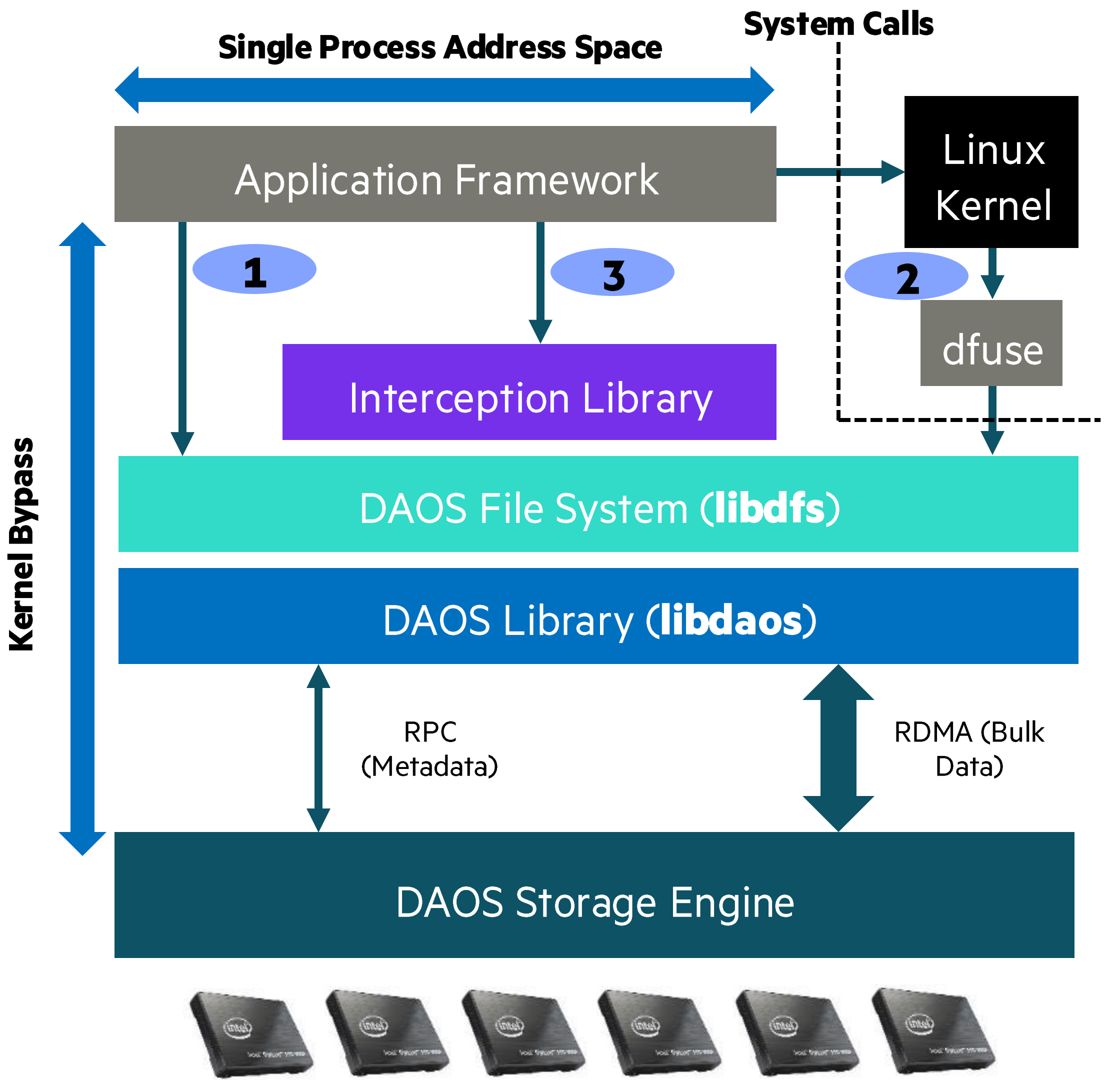


🟦 Generic I/O Middleware/frameworks

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

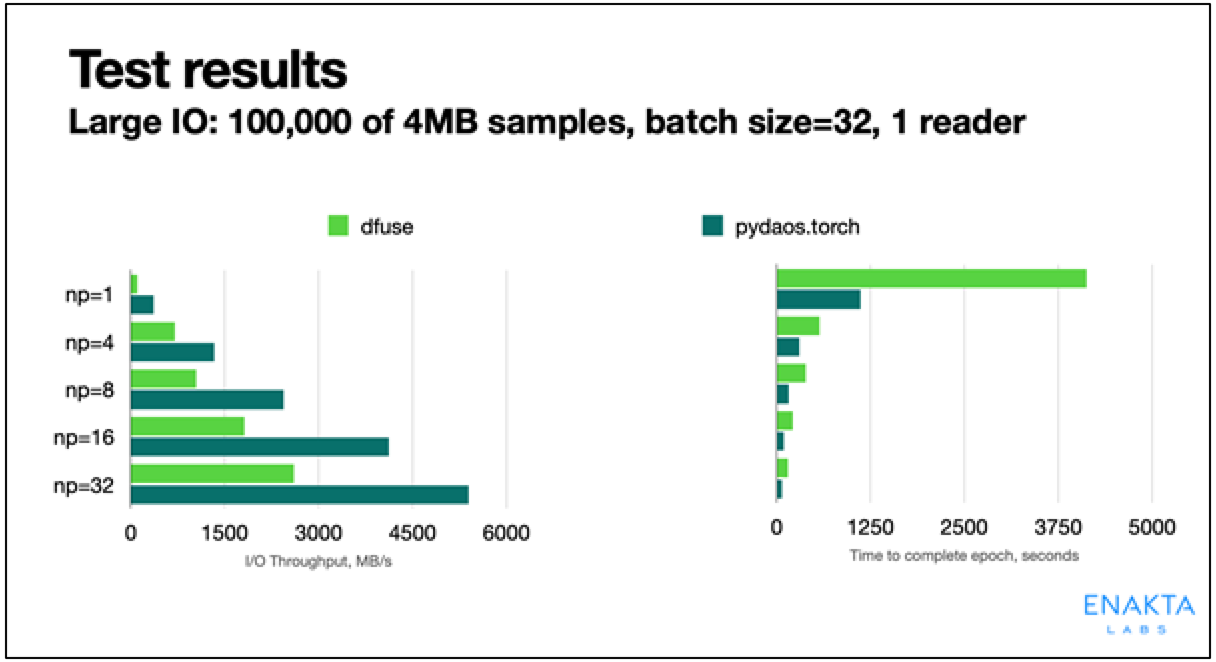
POSIX I/O Support

- DAOS File System (**libdfs**) 1
  - Implements POSIX namespace in a container
  - Userspace **`libdfs.so`** library (requires application changes)
  - Full OS bypass, asynchronous I/O; no caching
- FUSE Daemon (**dfuse**) 2
  - Transparent POSIX access (no application changes)
  - VFS mountpoint through dfuse; high latency
  - Caching by Linux kernel
- I/O interception library (**libpil4dfs**) 3
  - Combined with dfuse; data & metadata interception
  - **`LD_PRELOAD=/usr/lib64/libpil4dfs.so (or libioil.so)`**
  - mmap and binary execution via dfuse



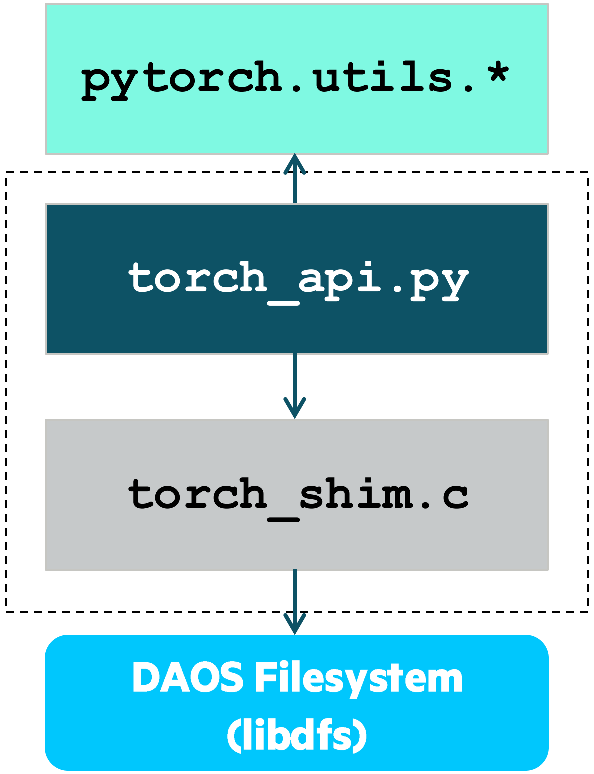
Pytorch Data Modules

- Collaboration between Enakta Labs and Google
- DataLoader and Checkpoint modules (uses DFS)
  - 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



pytorch.utils.*

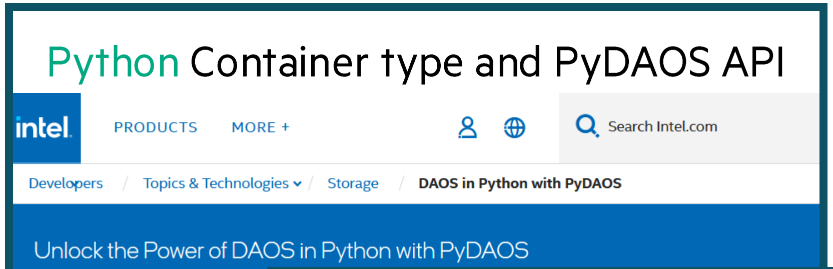
torch_api.py

torch_shim.c

DAOS Filesystem (libdfs)

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

DAOS Beyond POSIX: Storage Innovations



Python Container type and PyDAOS API

Developers / Topics & Technologies / Storage / DAOS in Python with PyDAOS

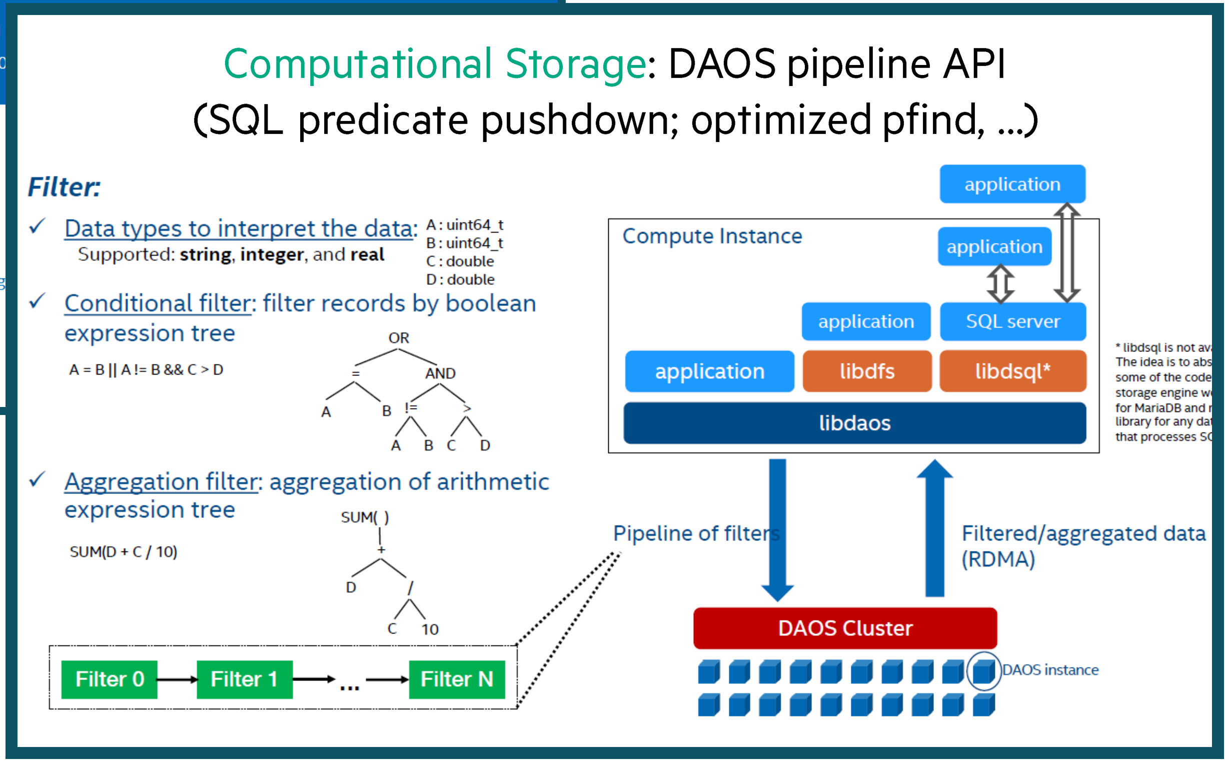
Unlock the Power of DAOS in Python with PyDAOS

ID 685683

Updated 11/17/20

Introduction
Pools and Containers
PyDAOS Step by Step
A complete Example
An Additional Example Using JSON Files
Summary

By Eduardo Berrocal Garcia De Carellan, Eeheet Hayer

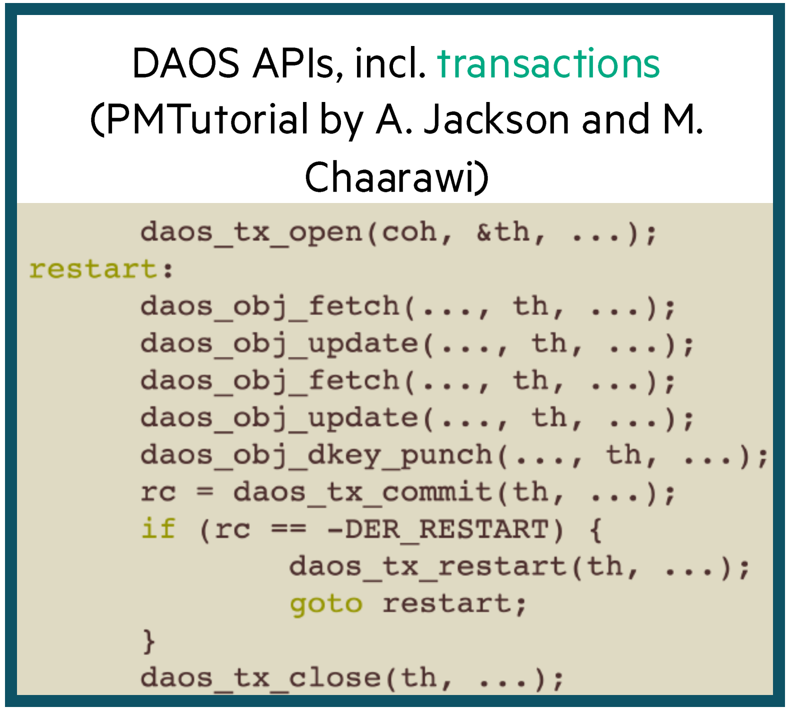


Computational Storage: DAOS pipeline API
(SQL predicate pushdown; optimized pfind, …)

Filter:

- ✓ Data types to interpret the data: A : uint64_t, B : uint64_t, C : double, D : double
  Supported: **string**, **integer**, and **real**
- ✓ Conditional filter: filter records by boolean expression tree
  A = B || A != B && C > D
- ✓ Aggregation filter: aggregation of arithmetic expression tree
  SUM(D + C / 10)

* libdsql is not ava
The idea is to abs
some of the code
storage engine w
for MariaDB and r
library for any dat
that processes SQ

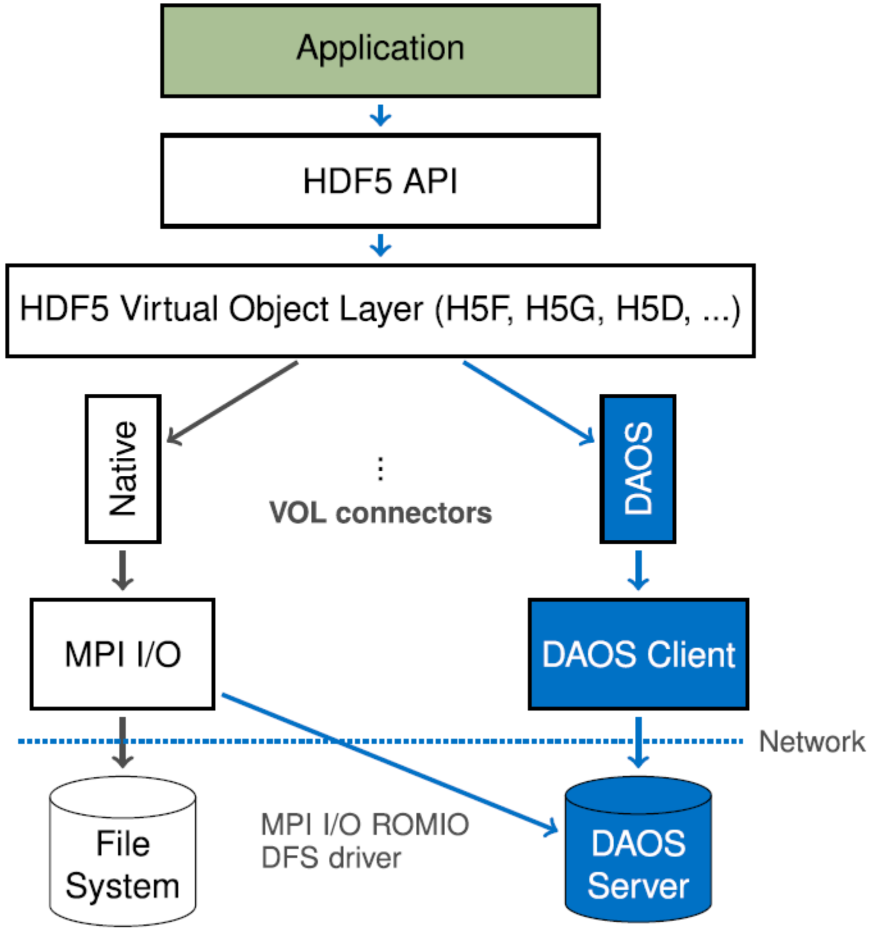


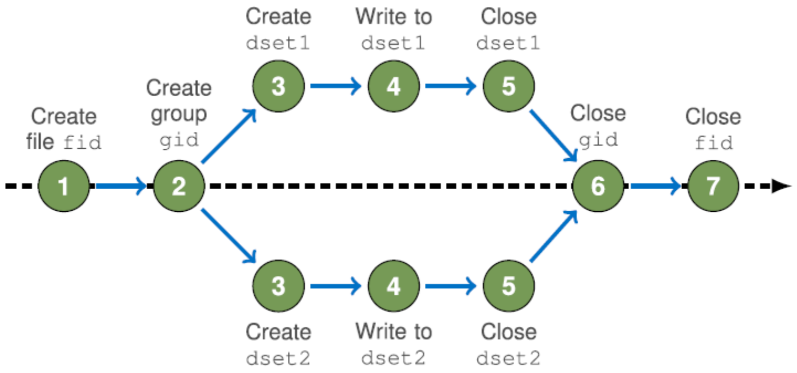
DAOS APIs, incl. transactions
(PMTutorial by A. Jackson and M. Chaarawi)

```
        daos_tx_open(coh, &th, ...);
restart:
        daos_obj_fetch(..., th, ...);
        daos_obj_update(..., th, ...);
        daos_obj_fetch(..., th, ...);
        daos_obj_update(..., th, ...);
        daos_obj_dkey_punch(..., th, ...);
        rc = daos_tx_commit(th, ...);
        if (rc == -DER_RESTART) {
                daos_tx_restart(th, ...);
                goto restart;
        }
        daos_tx_close(th, ...);
```

DAOS Beyond POSIX: Storage Innovations – HDF5

- Remove inherent POSIX limitations from HDF5
- New functionality provided in DAOS VOL plugin:
  - Independent object creation
    - H5daos_set_all_ind_metadata_ops
    - Avoids collective synchronisation across ranks
  - Maps: application-defined K/V store
    - H5Mput/H5Mget
  - Asynchronous I/O





```
es_id = H5EScreate();
fid = H5Fcreate_async(..., es_id);
gid = H5Gcreate_async(fid, ..., es_id);
did1 = H5Dcreate_async(gid, ..., es_id);
H5Dwrite_async(did1, ..., es_id);
did2 = H5Dcreate_async(gid, ..., es_id);
H5Dwrite_async(did2, ..., es_id);
H5Dclose_async(did1, es_id);
H5Dclose_async(did2, es_id);
H5Gclose_async(gid);
H5Fclose_async(fid);
H5ESwait(es_id, ...);
```

DAOS Deployments at ALCF (Aurora) and LRZ (SNG2)

Argonne NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
intel
Hewlett Packard Enterprise
Aurora

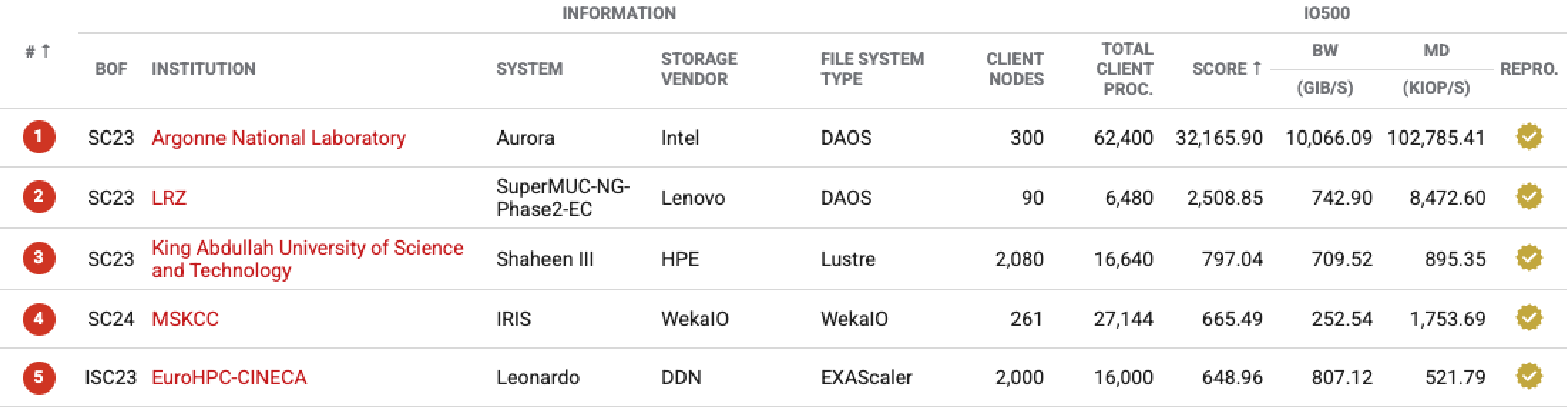
Compute Nodes:
2x Intel SPR+**HBM**, **6x** Intel Xe "PVC" GPUs, **8x** HPE Slingshot

1024 DAOS Servers (Intel M50CYP):
2x Xeon 5320 26core 2.2GHz CPUs
16x 32GB DDR4 DRAM
16x **512GB** Intel Optane 200 PMem
16x Samsung PM1733 **15.36TB** NVMe (gen4)
2x HPE Slingshot (200Gbps)
→ 16k NVMe (250PB), 16k PMem (8PB), 2k engines

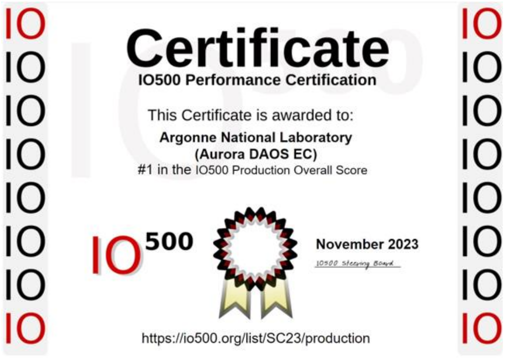
Compute Nodes:
2x Intel SPR, **4x** Intel Xe "PVC" GPUs, **2x** NVIDIA HDR

42 DAOS Servers (Lenovo SR630v2):
2x Xeon 8352Y 32core 2.2GHz CPUs
16x 32GB DDR4 DRAM
16x **128GB** Intel Optane 200 PMem
8x Intel P5500 **3.84TB** NVMe (gen4)
2x NVIDIA HDR InfiniBand (200Gbps)
→ 336 NVMe (1.3PB), 672 PMem (84TB), 84 engines

DAOS Performance – SC’24 IO500 Production List



# ↑	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	✓



Certificate

IO500 Performance Certification

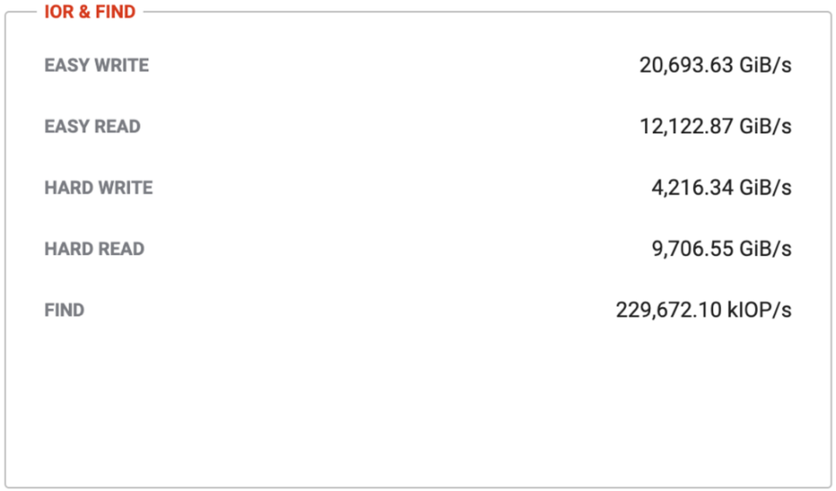
This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)
#1 in the IO500 Production Overall Score

November 2023

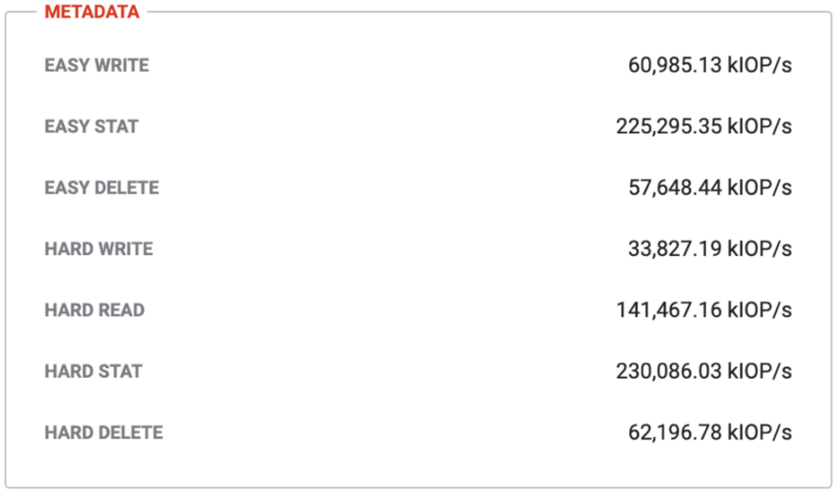
IO500 Steering Board

https://io500.org/list/SC23/production



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

DAOS Performance – 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

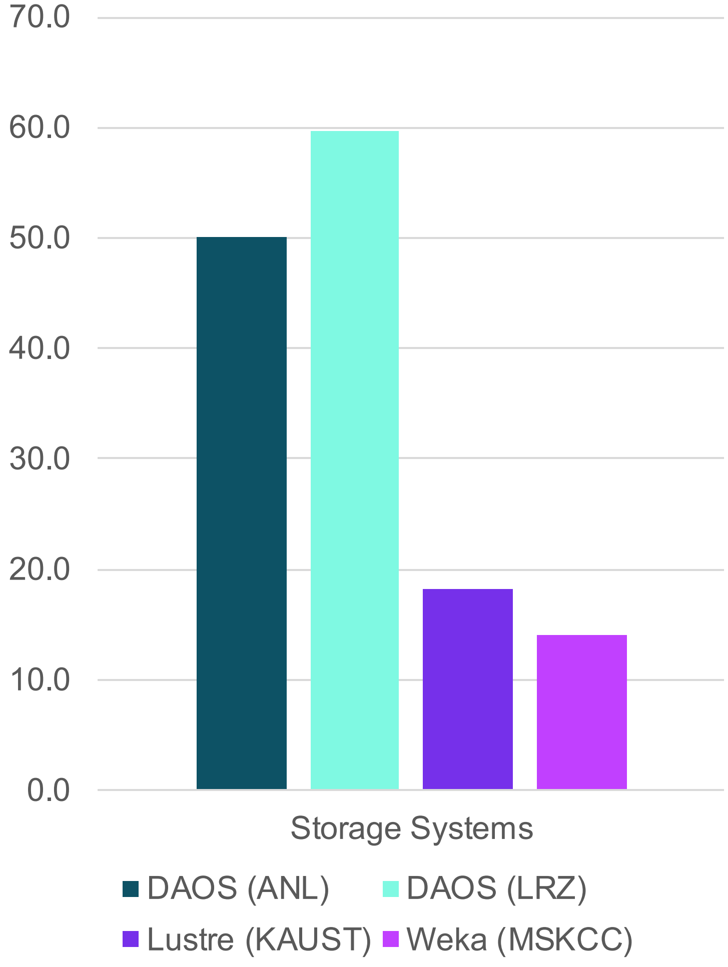
IRIS MSKCC WekaIO

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

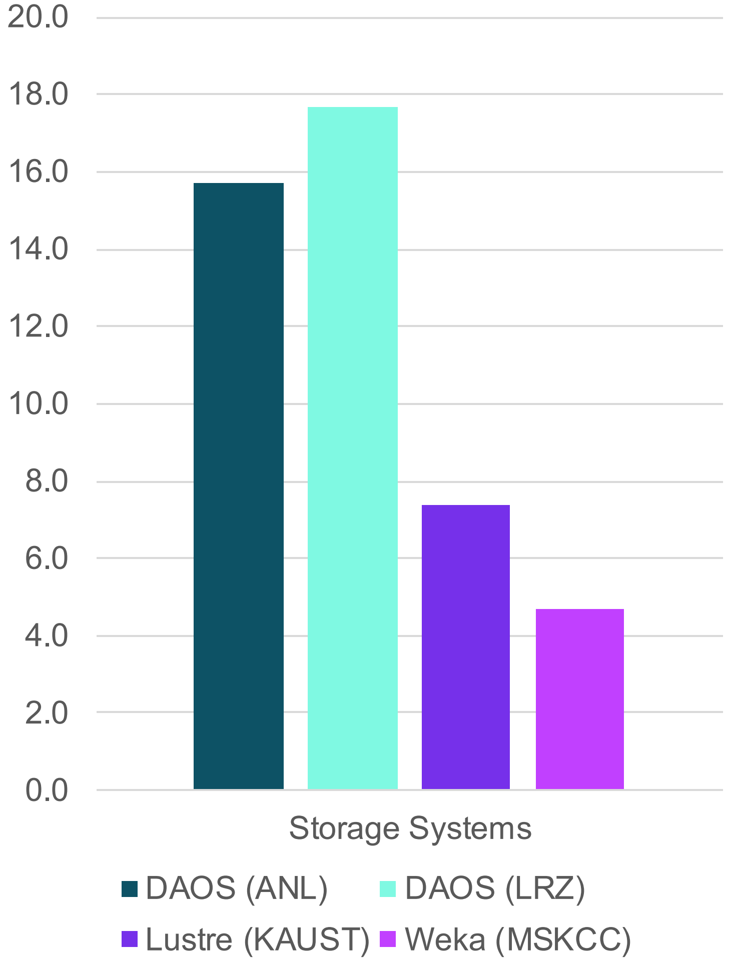
- 261x Client nodes

DAOS Performance – SC'24 IO500 Production List (bigger is better)

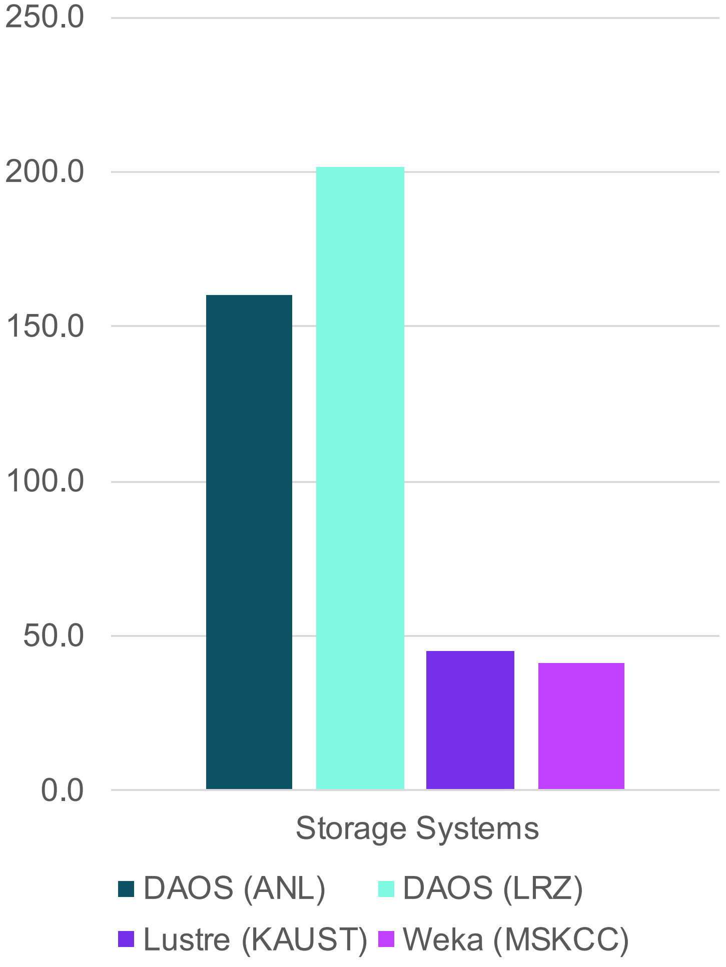
IO500 Score per Server



IO500 BW per Server (GiB/s)



IO500 MD per Server (kOps/s)



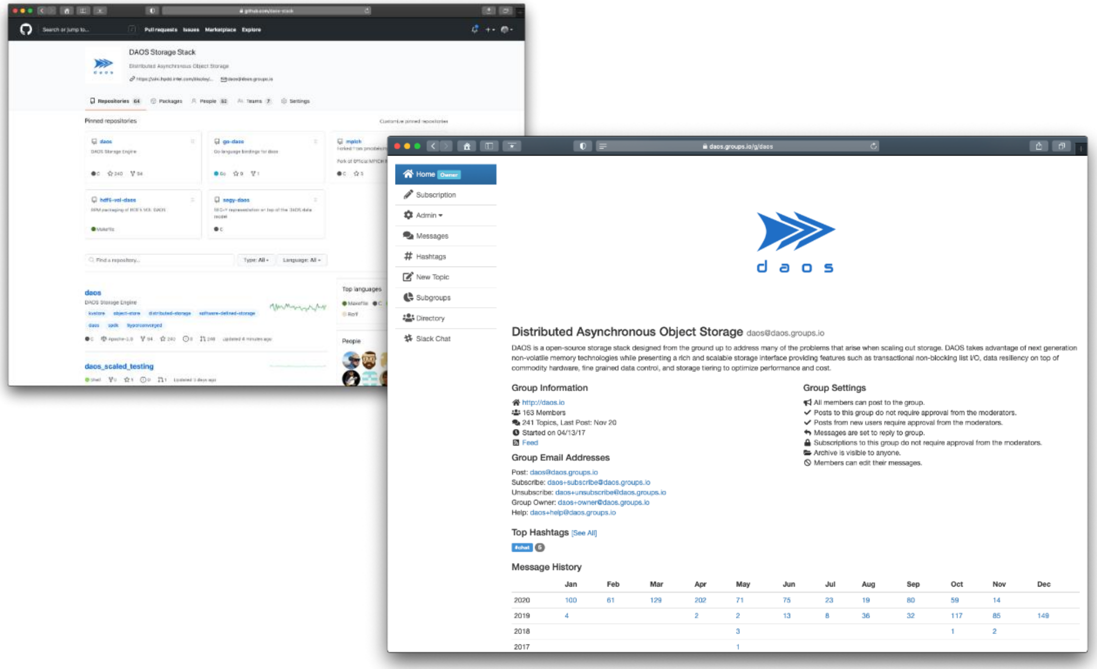
Source: https://io500.org

DAOS Resources

- DAOS Foundation landing page
  - https://daos.io/

- Community Resources
  - Github: https://github.com/daos-stack/daos
  - Online doc: https://docs.daos.io/
  - Mailing list & slack: https://daos.groups.io/
  - YouTube channel: https://video.daos.io/

- Virtual DAOS User Group on May 22, 2025:
  - https://daos.io/event/virtual-dug-25



Distributed Asynchronous Object Storage daos@daos.groups.io

DAOS is a open-source storage stack designed from the ground up to address many of the problems that arise when scaling out storage. DAOS takes advantage of next generation non-volatile memory technologies while presenting a rich and scalable storage interface providing features such as transactional non-blocking list I/O, data resiliency on top of commodity hardware, fine grained data control, and storage tiering to optimize performance and cost.

Group Information

- http://daos.io
- 163 Members
- 241 Topics, Last Post: Nov 20
- Started on 04/13/17
- Feed

Group Email Addresses

Post: daos@daos.groups.io
Subscribe: daos+subscribe@daos.groups.io
Unsubscribe: daos+unsubscribe@daos.groups.io
Group Owner: daos+owner@daos.groups.io
Help: daos+help@daos.groups.io

Group Settings

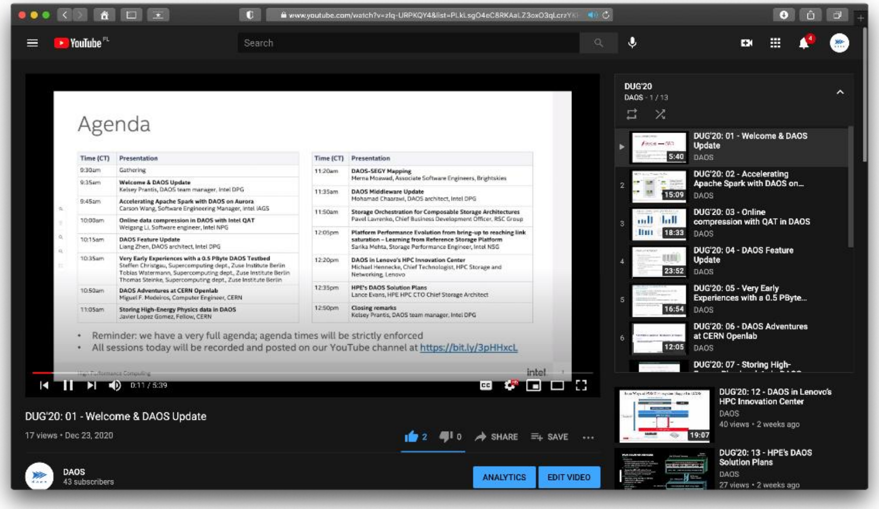
- All members can post to the group.
- ✓ Posts to this group do not require approval from the moderators.
- ✓ Posts from new users require approval from the moderators.
- Messages are set to reply to group.
- Subscriptions to this group do not require approval from the moderators.
- Archive is visible to anyone.
- Members can edit their messages.

Top Hashtags [See All]

#chat 5

Message History

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2020	100	61	129	202	71	75	23	19	80	59	14	
2019	4			2	2	13	8	36	32	117	85	149
2018					3					1	2	
2017					1							



Agenda

Time (CT)	Presentation
9:30am	Gathering
9:35am	**Welcome & DAOS Update** Kelsey Prantis, DAOS team manager, Intel DPG
9:45am	**Accelerating Apache Spark with DAOS on Aurora** Carson Wang, Software Engineering Manager, Intel IAGS
10:00am	**Online data compression in DAOS with Intel QAT** Weigang Li, Software engineer, Intel NPG
10:15am	**DAOS Feature Update** Liang Zhen, DAOS architect, Intel DPG
10:35am	**Very Early Experiences with a 0.5 PByte DAOS Testbed** Steffen Christgau, Supercomputing dept., Zuse Institute Berlin Tobias Watermann, Supercomputing dept., Zuse Institute Berlin Thomas Steinke, Supercomputing dept., Zuse Institute Berlin
10:50am	**DAOS Adventures at CERN Openlab** Miguel F. Medeiros, Computer Engineer, CERN
11:05am	**Storing High-Energy Physics data in DAOS** Javier Lopez Gomez, Fellow, CERN

Time (CT)	Presentation
11:20am	**DAOS-SEGY Mapping** Merna Moawad, Associate Software Engineers, Brightskies
11:35am	**DAOS Middleware Update** Mohamad Chaarawi, DAOS architect, Intel DPG
11:50am	**Storage Orchestration for Composable Storage Architectures** Pavel Lavrenko, Chief Business Development Officer, RSC Group
12:05pm	**Platform Performance Evolution from bring-up to reaching link saturation – Learning from Reference Storage Platform** Sarika Mehta, Storage Performance Engineer, Intel NSG
12:20pm	**DAOS in Lenovo's HPC Innovation Center** Michael Hennecke, Chief Technologist, HPC Storage and Networking, Lenovo
12:35pm	**HPE's DAOS Solution Plans** Lance Evans, HPE HPC CTO Chief Storage Architect
12:50pm	**Closing remarks** Kelsey Prantis, DAOS team manager, Intel DPG

- Reminder: we have a very full agenda; agenda times will be strictly enforced
- All sessions today will be recorded and posted on our YouTube channel at https://bit.ly/3pHHxcL

DUG'20: 01 - Welcome & DAOS Update

17 views • Dec 23, 2020

DAOS
43 subscribers

DUG'20
DAOS - 1 / 13

1. DUG'20: 01 - Welcome & DAOS Update — DAOS — 5:40
2. DUG'20: 02 - Accelerating Apache Spark with DAOS on... — DAOS — 15:09
3. DUG'20: 03 - Online compression with QAT in DAOS — DAOS — 18:33
4. DUG'20: 04 - DAOS Feature Update — DAOS — 23:52
5. DUG'20: 05 - Very Early Experiences with a 0.5 PByte... — DAOS — 16:54
6. DUG'20: 06 - DAOS Adventures at CERN Openlab — DAOS — 12:05
7. DUG'20: 07 - Storing High-

DUG'20: 12 - DAOS in Lenovo's HPC Innovation Center
DAOS
40 views • 2 weeks ago
19:07

DUG'20: 13 - HPE's DAOS Solution Plans
DAOS
27 views • 2 weeks ago

Thank you

jerome.soumagne@hpe.com
michael.hennecke@hpe.com