DAOS 2.4 and Beyond

Johann Lombardi, Senior Principal Engineer, AXG, Intel

6th DAOS User Group, Dallas, Nov 22, 2022

Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks .

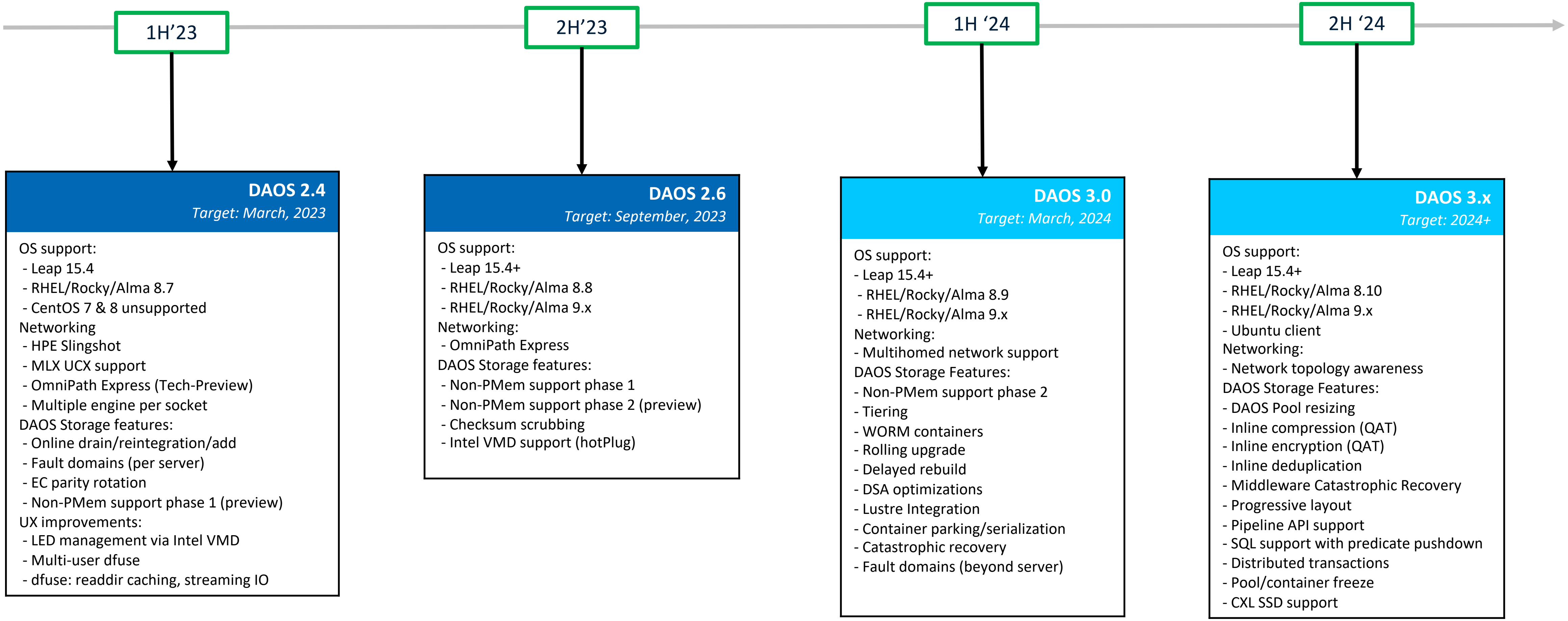
Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

DAOS Upcoming Community Releases



DAOS 2.4

Target: March, 2023

OS support:
- Leap 15.4
- RHEL/Rocky/Alma 8.7
- CentOS 7 & 8 unsupported

Networking
- HPE Slingshot
- MLX UCX support
- OmniPath Express (Tech-Preview)
- Multiple engine per socket

DAOS Storage features:
- Online drain/reintegration/add
- Fault domains (per server)
- EC parity rotation
- Non-PMem support phase 1 (preview)

UX improvements:
- LED management via Intel VMD
- Multi-user dfuse
- dfuse: readdir caching, streaming IO

DAOS 2.6

Target: September, 2023

OS support:
- Leap 15.4+
- RHEL/Rocky/Alma 8.8
- RHEL/Rocky/Alma 9.x

Networking:
- OmniPath Express

DAOS Storage features:
- Non-PMem support phase 1
- Non-PMem support phase 2 (preview)
- Checksum scrubbing
- Intel VMD support (hotPlug)

DAOS 3.0

Target: March, 2024

OS support:
- Leap 15.4+
- RHEL/Rocky/Alma 8.9
- RHEL/Rocky/Alma 9.x

Networking:
- Multihomed network support

DAOS Storage Features:
- Non-PMem support phase 2
- Tiering
- WORM containers
- Rolling upgrade
- Delayed rebuild
- DSA optimizations
- Lustre Integration
- Container parking/serialization
- Catastrophic recovery
- Fault domains (beyond server)

DAOS 3.x

Target: 2024+

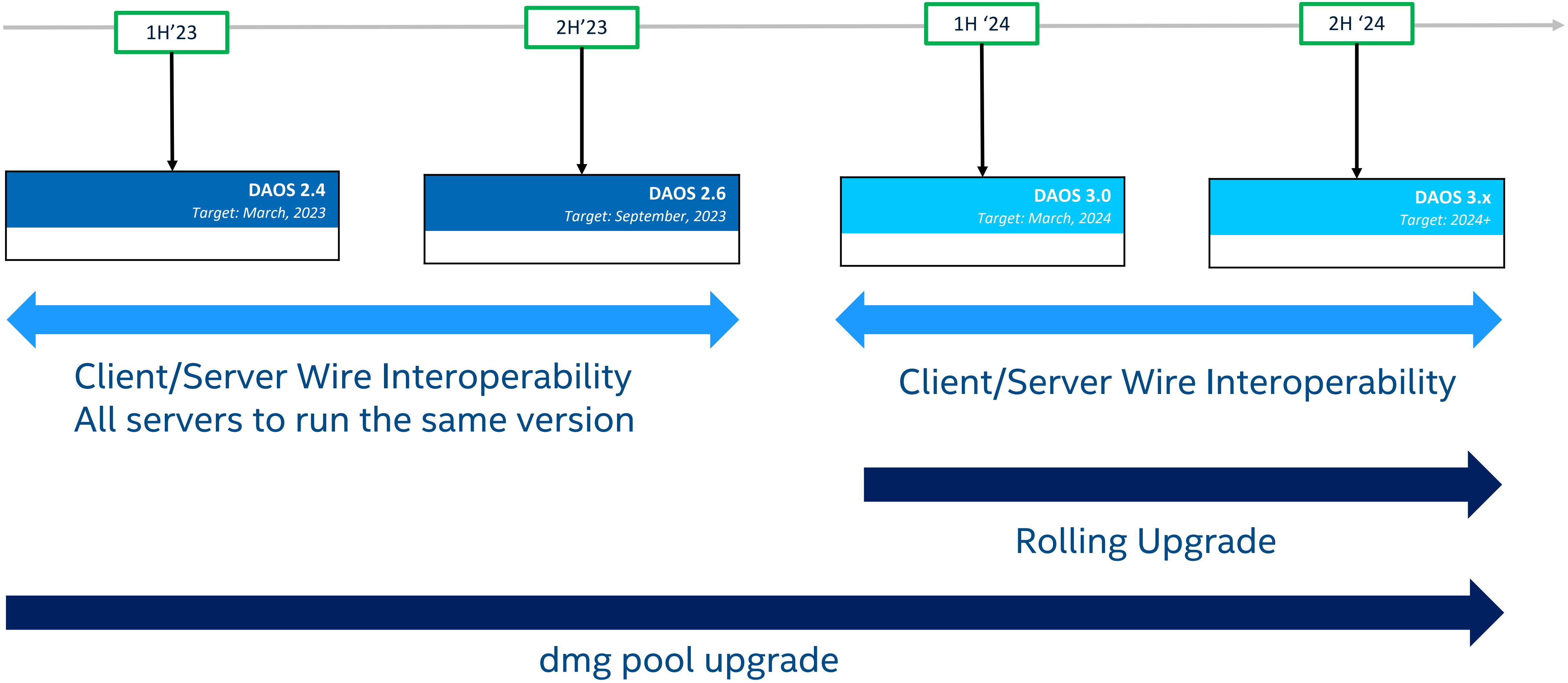
OS support:
- Leap 15.4+
- RHEL/Rocky/Alma 8.10
- RHEL/Rocky/Alma 9.x
- Ubuntu client

Networking:
- Network topology awareness

DAOS Storage Features:
- DAOS Pool resizing
- Inline compression (QAT)
- Inline encryption (QAT)
- Inline deduplication
- Middleware Catastrophic Recovery
- Progressive layout
- Pipeline API support
- SQL support with predicate pushdown
- Distributed transactions
- Pool/container freeze
- CXL SSD support

NOTE: All information provided in this roadmap is subject to change without notice.

Interoperability / Upgradability



1H'23

2H'23

1H '24

2H '24

DAOS 2.4
Target: March, 2023

DAOS 2.6
Target: September, 2023

DAOS 3.0
Target: March, 2024

DAOS 3.x
Target: 2024+

Client/Server Wire Interoperability
All servers to run the same version

Client/Server Wire Interoperability

Rolling Upgrade

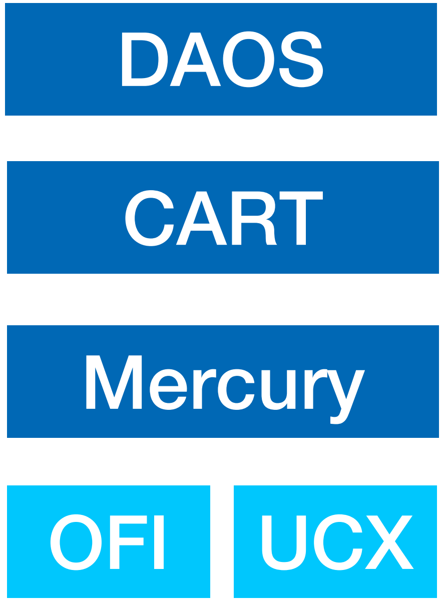
dmg pool upgrade

Network: Buffer Mgmt Improvements (2.4)

- Enable MR (memory registration) cache
  - Required for CXI support
  - Dfuse changes to be MR cache-friendly in progress
- OFI MULTI_RECV support
  - Faced issue with buffer exhaustion at larger scale
  - Change Mercury to support OFI multi-receive API
    - Post large buffers where incoming messages are appended
    - Protocol changes since no tag support
  - Supported by most providers

Network: UCX Support (2.4)

- Mellanox UCX library
- Future path for MLX HW support:
  - Shipped as part of MOFED packages
  - Better scalability on MLX fabric via DC & UD support
  - Native multi-rail support
  - GDS support on MLX/NVIDIA HW
- Shoud eventually replace OFI verbs provider
- Feature preview in 2.2 and official support in 2.4
- More work to be done to demonstrate scalability



DAOS

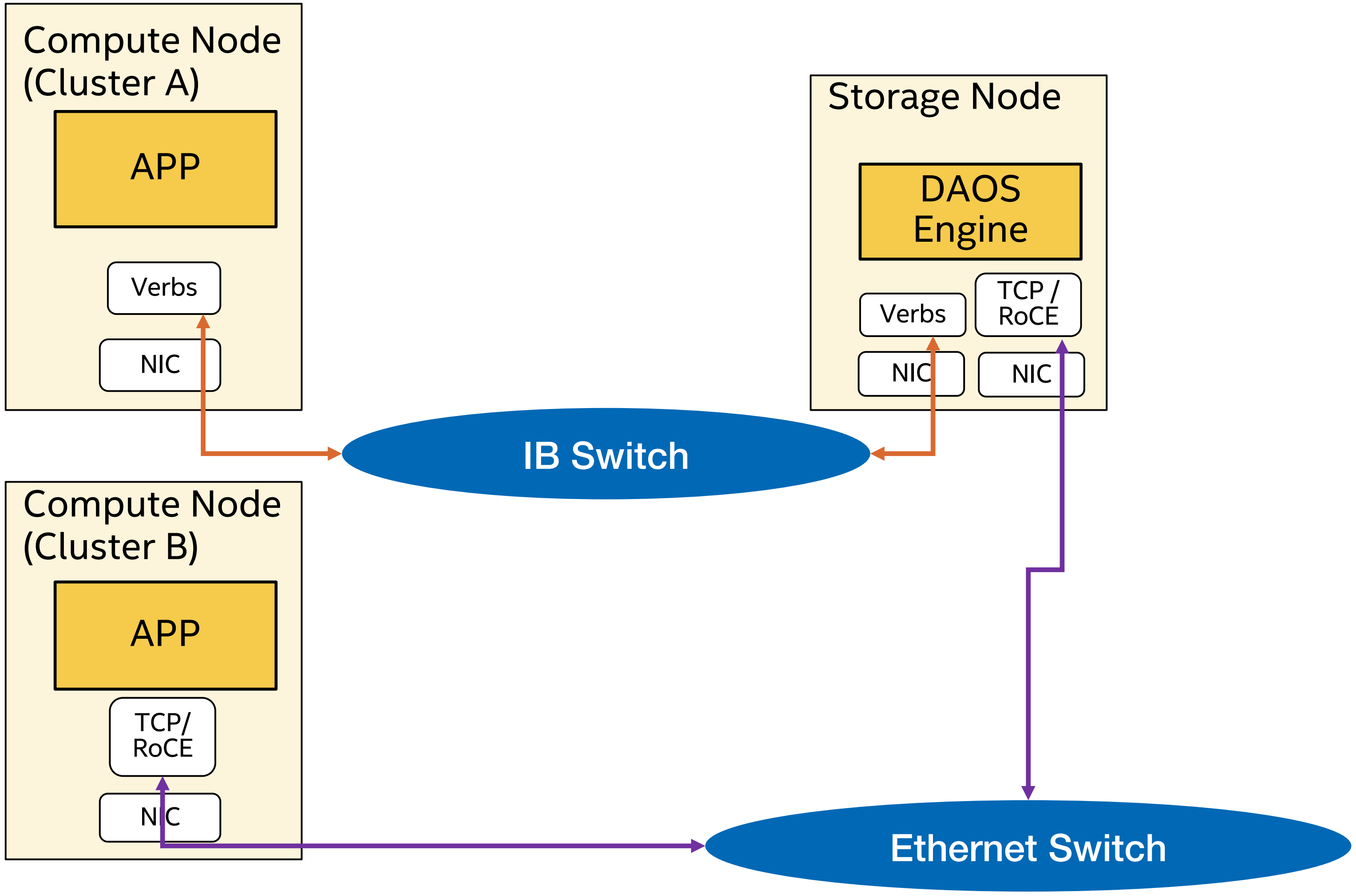
CART

Mercury

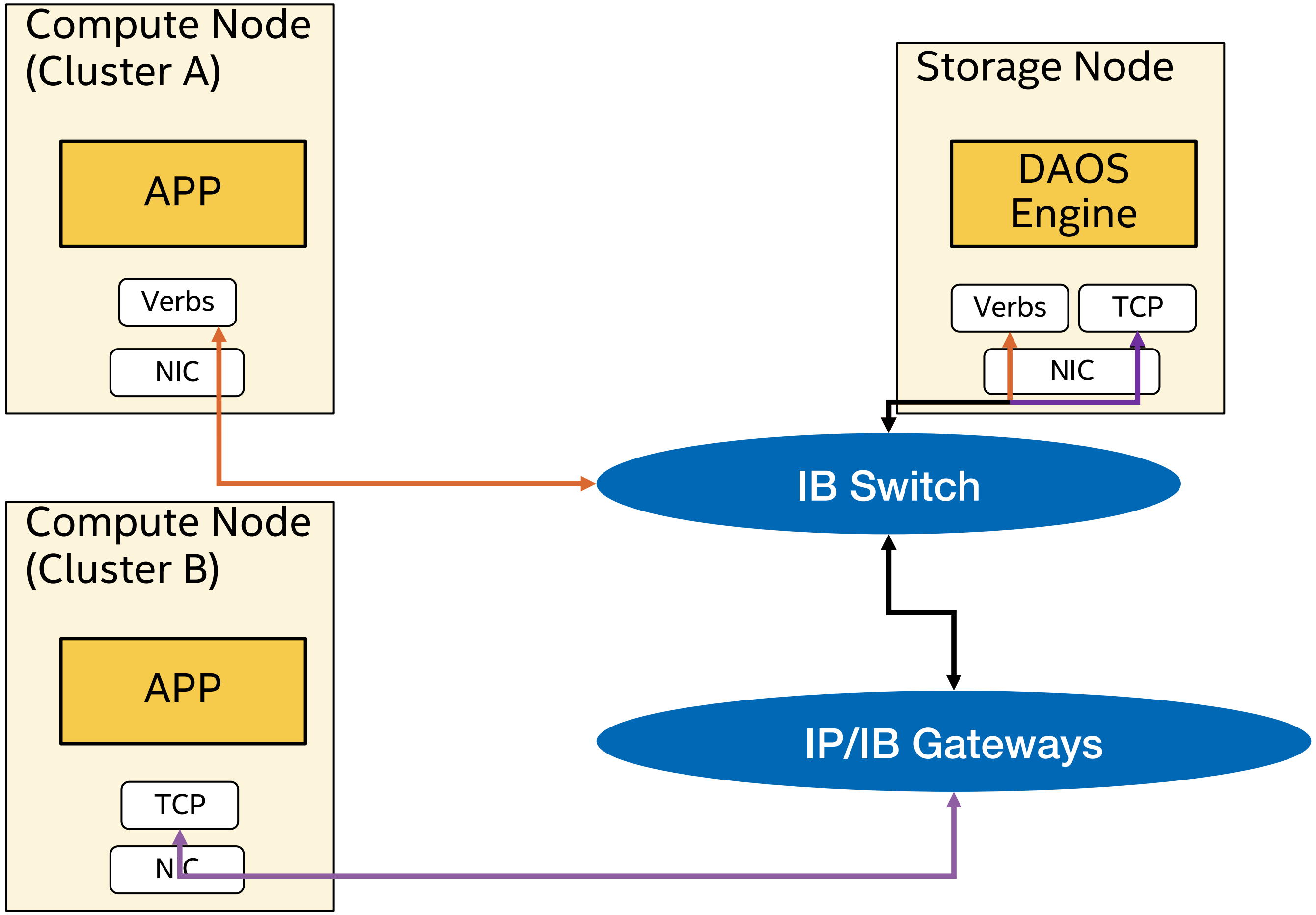
OFI

UCX

Network: Multi-Homed Network (3.0)



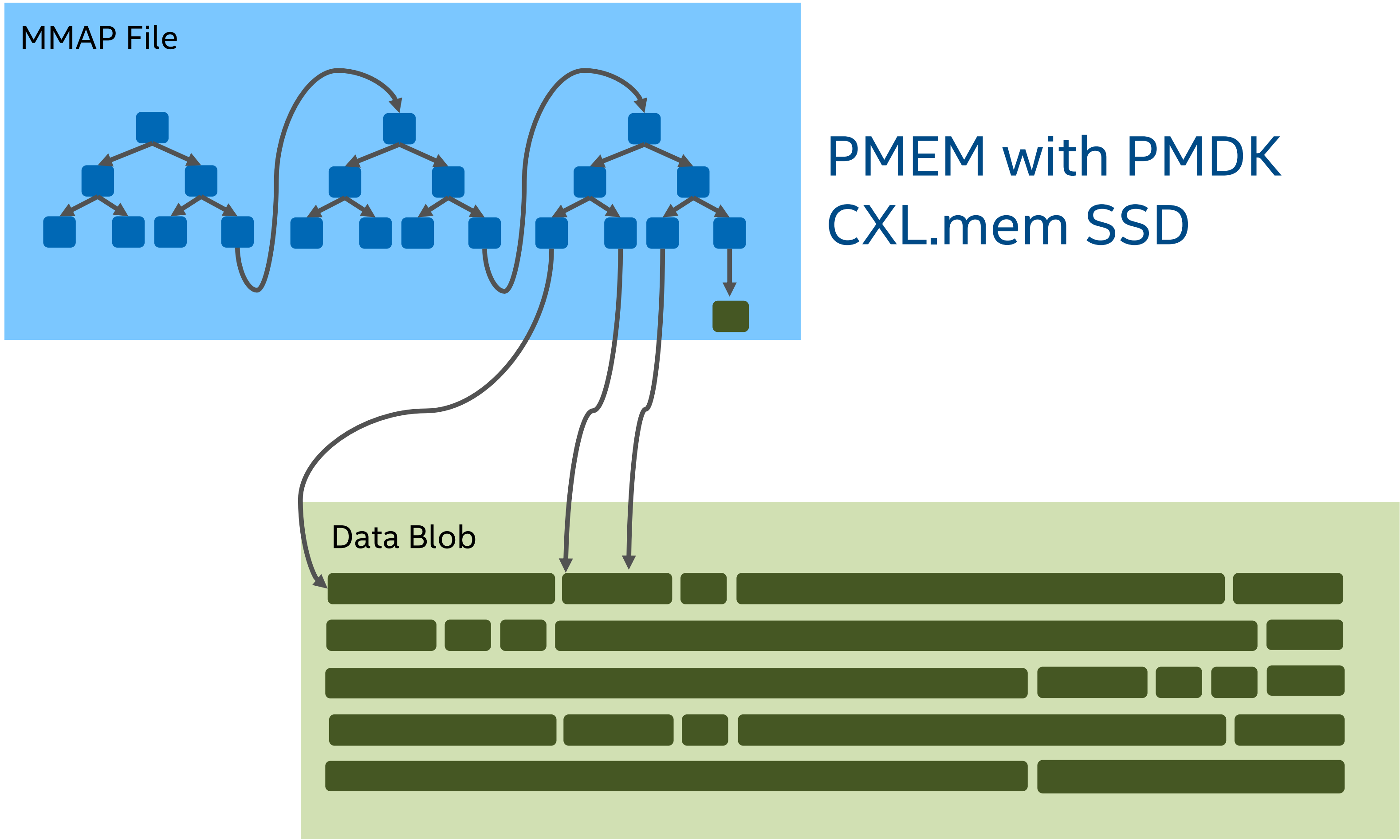
Network: Multi-Homed Network (3.0)



Backend: Metadata on SSD Requirements (2.4+)

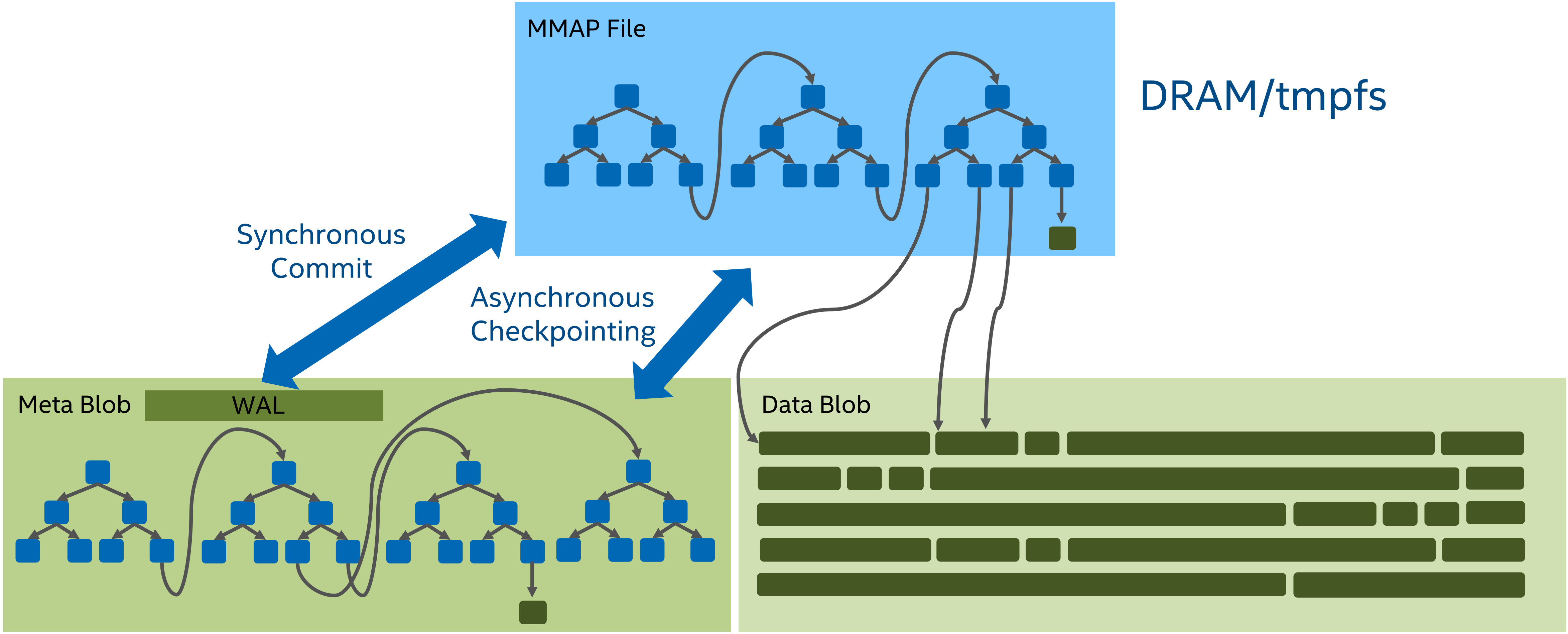
- Incremental changes
  - Deliver incremental functionality w/o waiting for another 4y
- Offer continuity to our current and future customers
- Maintain performance leadership
- Turn this into an opportunity to:
  - Strenghten the DAOS open-source community
  - Run DAOS on a wider range of hardware
  - Broaden the DAOS market
- https://daosio.atlassian.net/wiki/spaces/DC/pages/11196923911/Metadata+on+SSDs

Backend: Persistent Memory

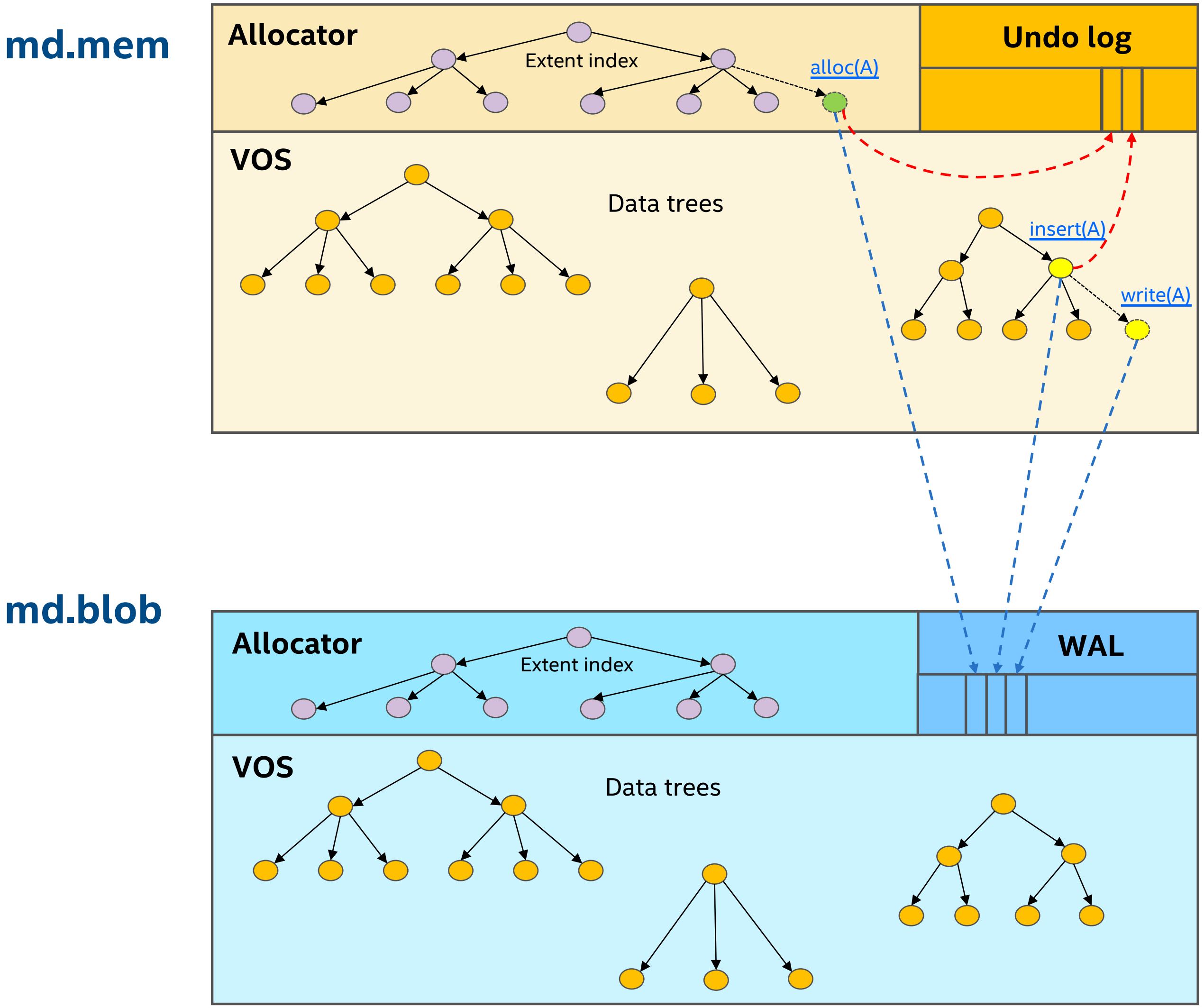


PMEM with PMDK
CXL.mem SSD

Backend: Volatile Memory



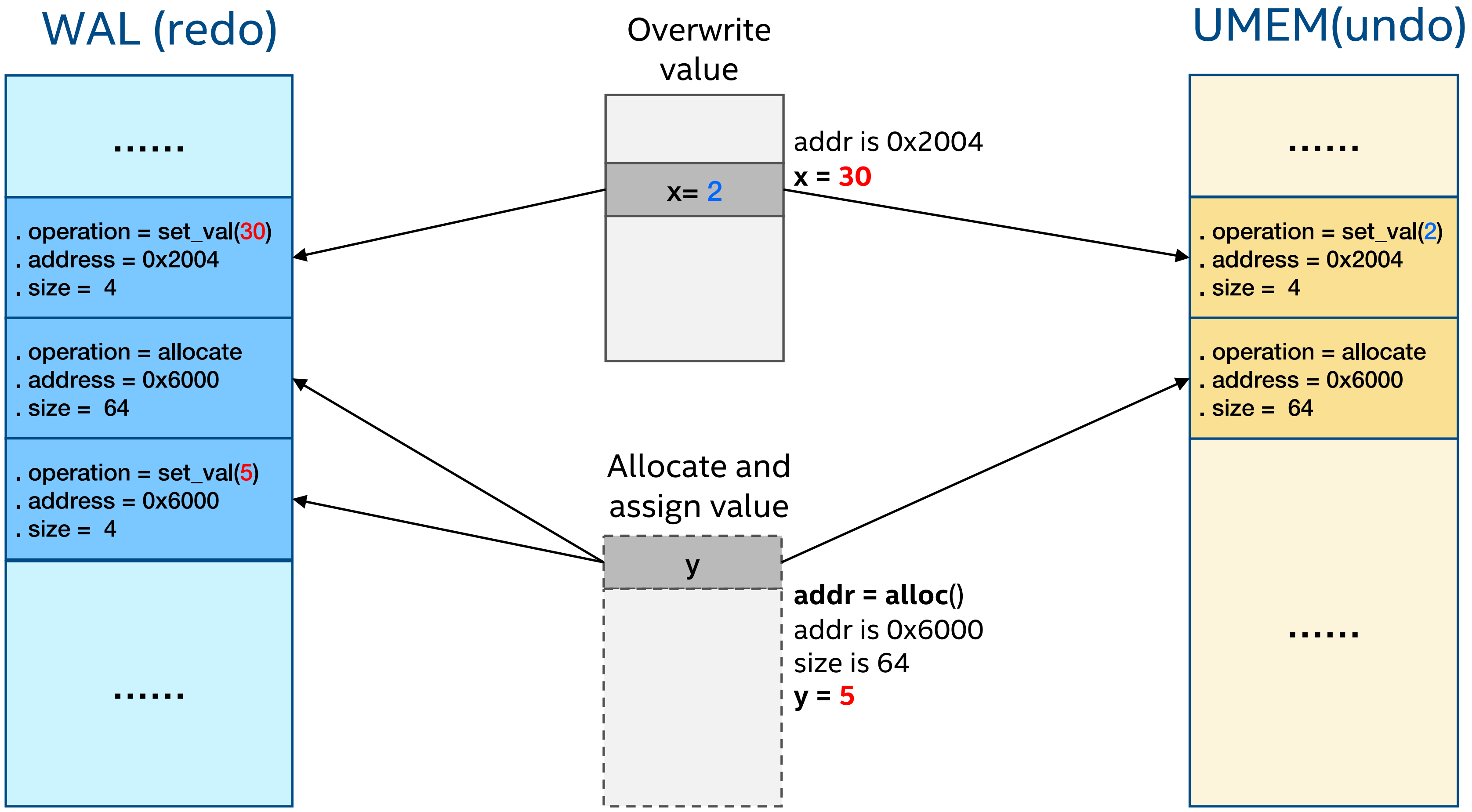
Backend: New (Meta)Data Structures



Undo log is **off-heap**, it has no persistent mirror

WAL is on-heap **(or in a separate blo**b), it has no ephemeral mirror.

Backend: Undo vs Redo



WAL (redo)

……

. operation = set_val(30)
. address = 0x2004
. size = 4

. operation = allocate
. address = 0x6000
. size = 64

. operation = set_val(5)
. address = 0x6000
. size = 4

……

Overwrite value

x= 2

addr is 0x2004
x = 30

Allocate and assign value

y

addr = alloc()
addr is 0x6000
size is 64
y = 5

UMEM(undo)

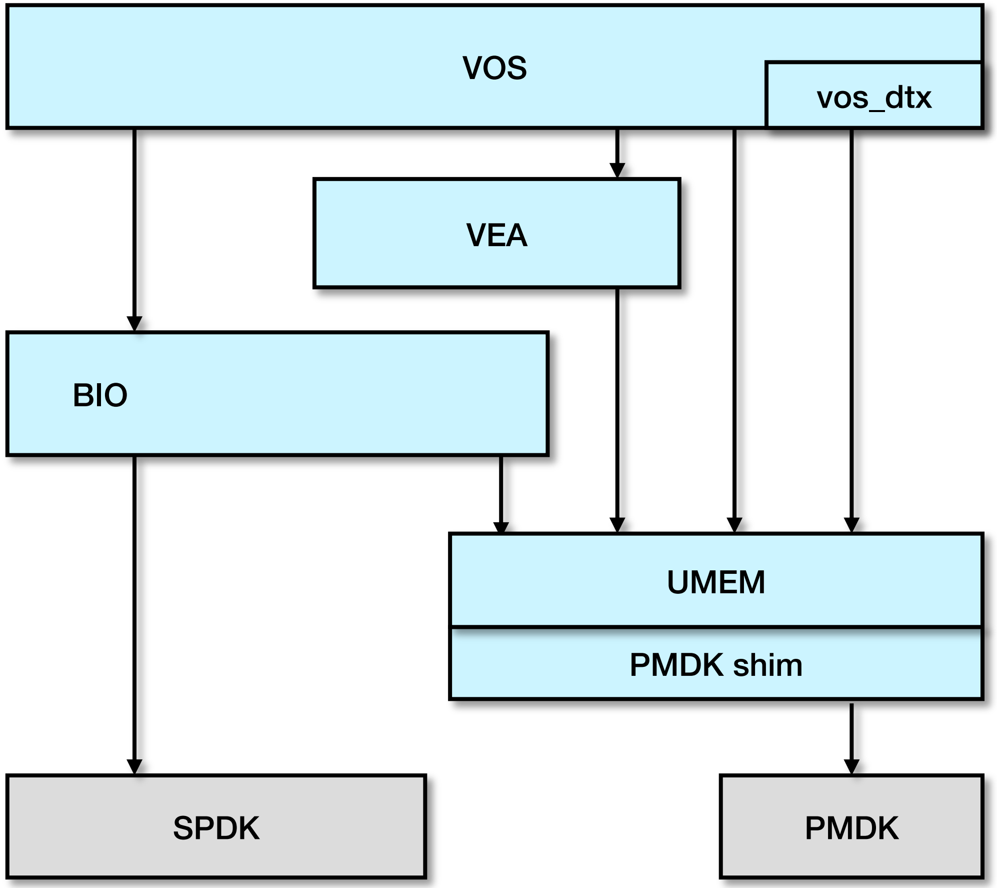
……

. operation = set_val(2)
. address = 0x2004
. size = 4

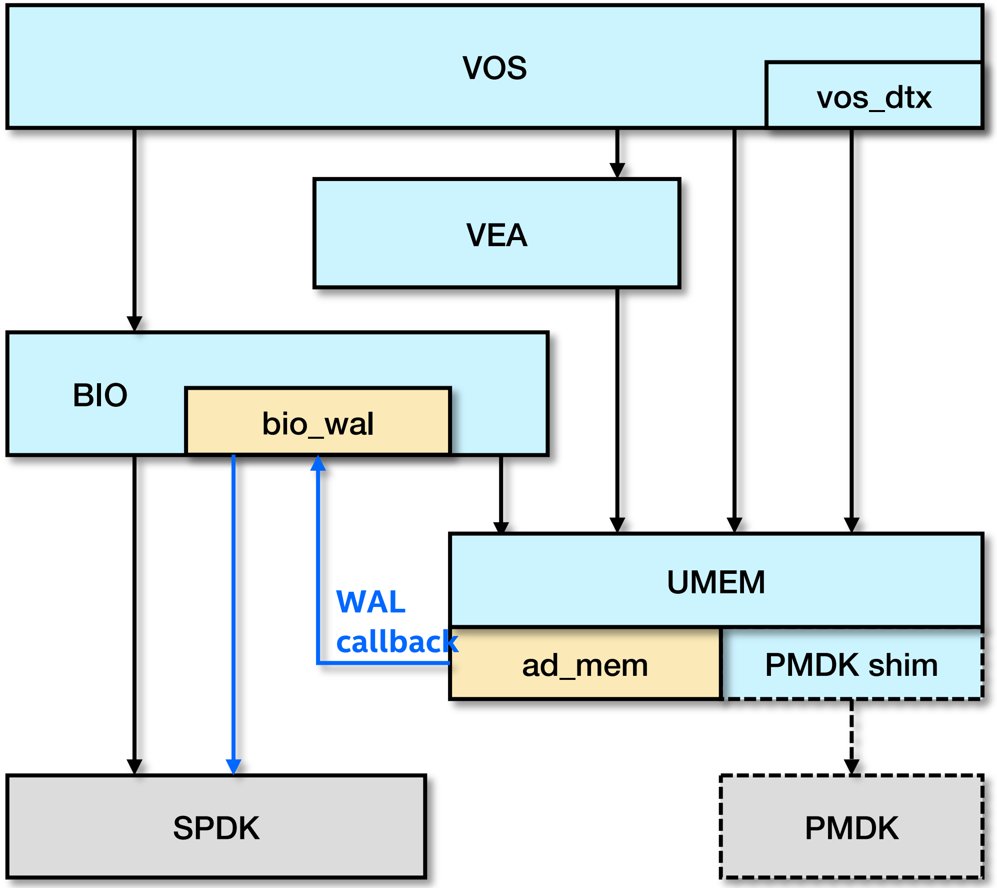
. operation = allocate
. address = 0x6000
. size = 64

……

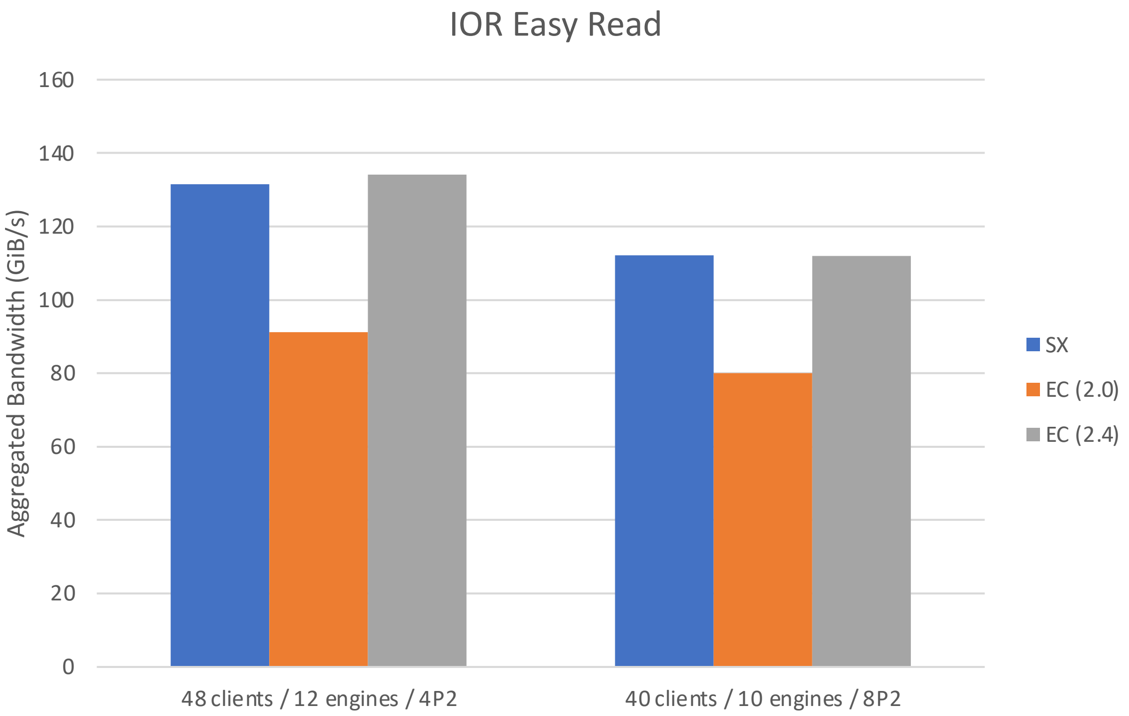
Backend: Stack Layering (2.2)

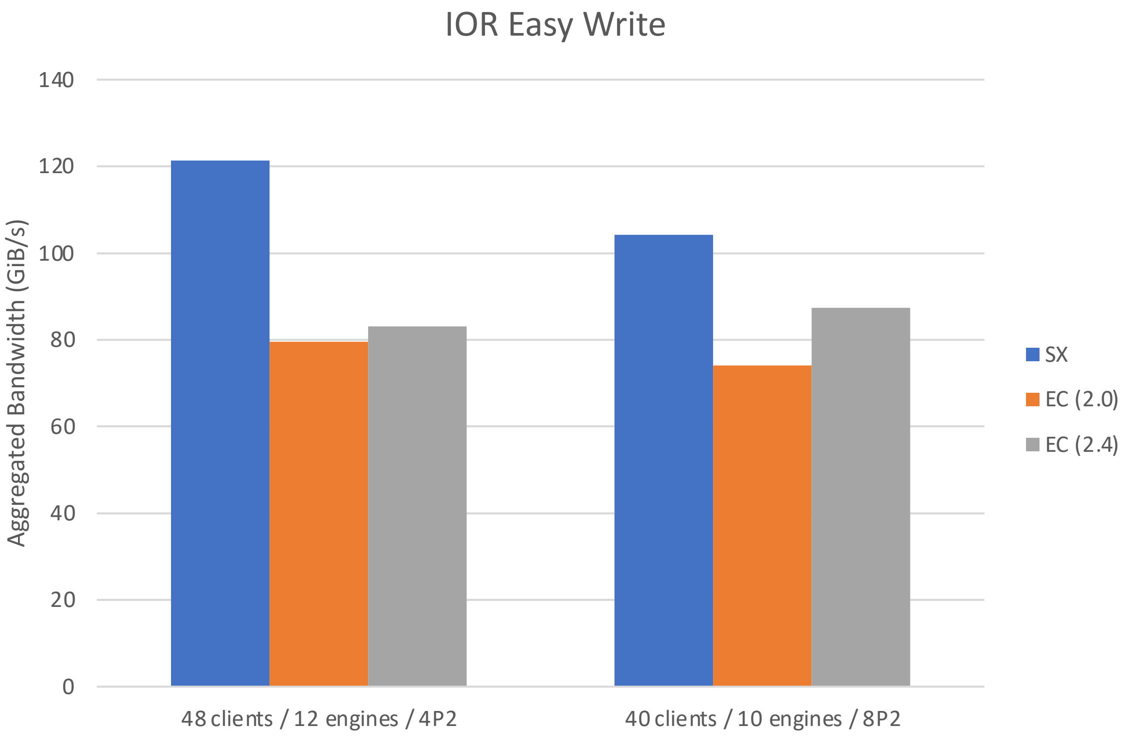


Backend: Stack Layering (2.4+)

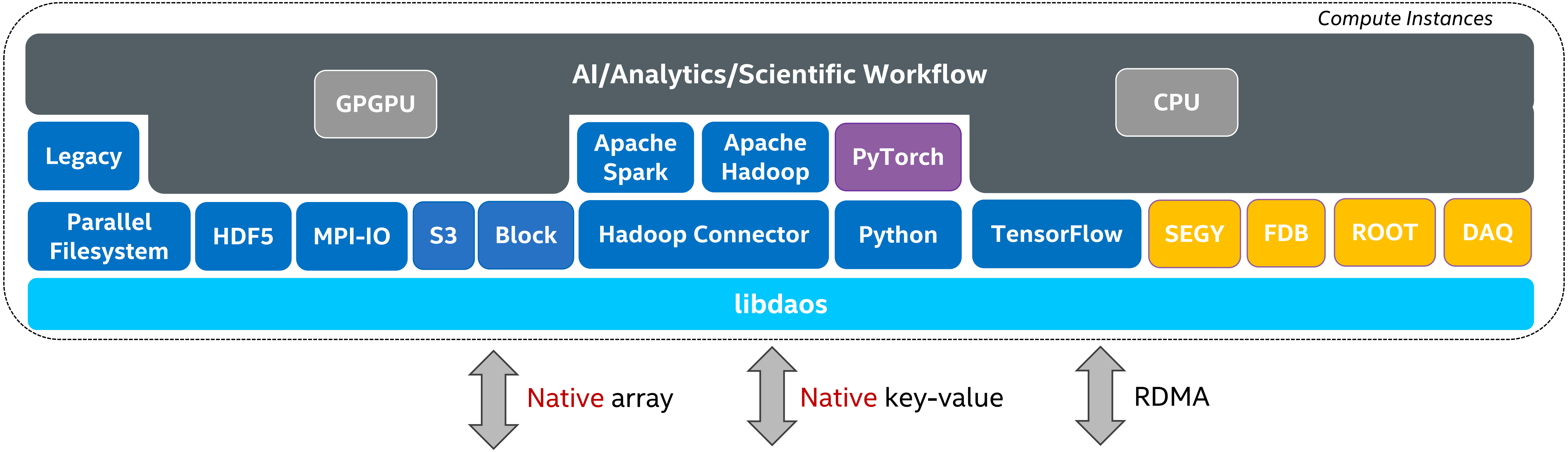


Erasure Code: Parity Rotation (2.4)





Middleware: State of Affairs



🟦 Generic I/O middleware supported today

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

🟪 Enablement in progress

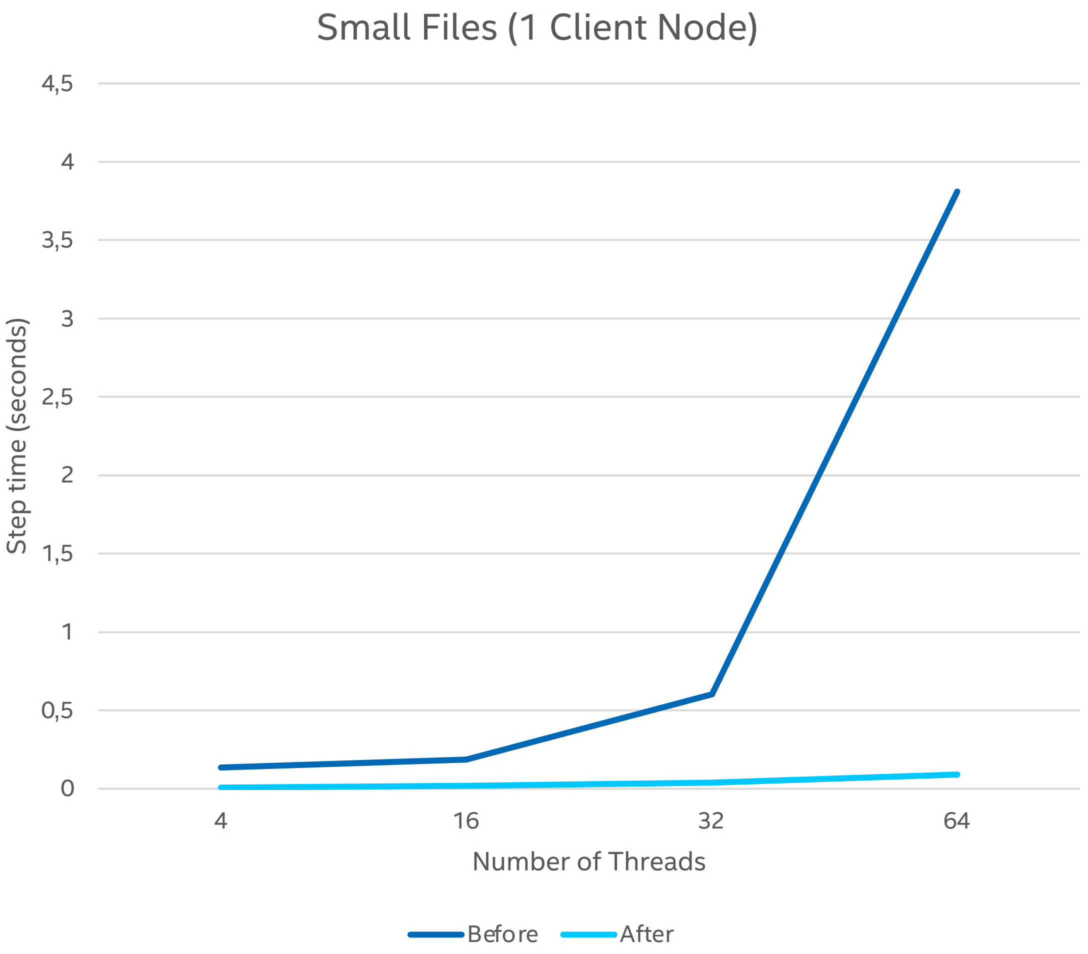
Middleware: dfuse & IL Improvements (2.4)

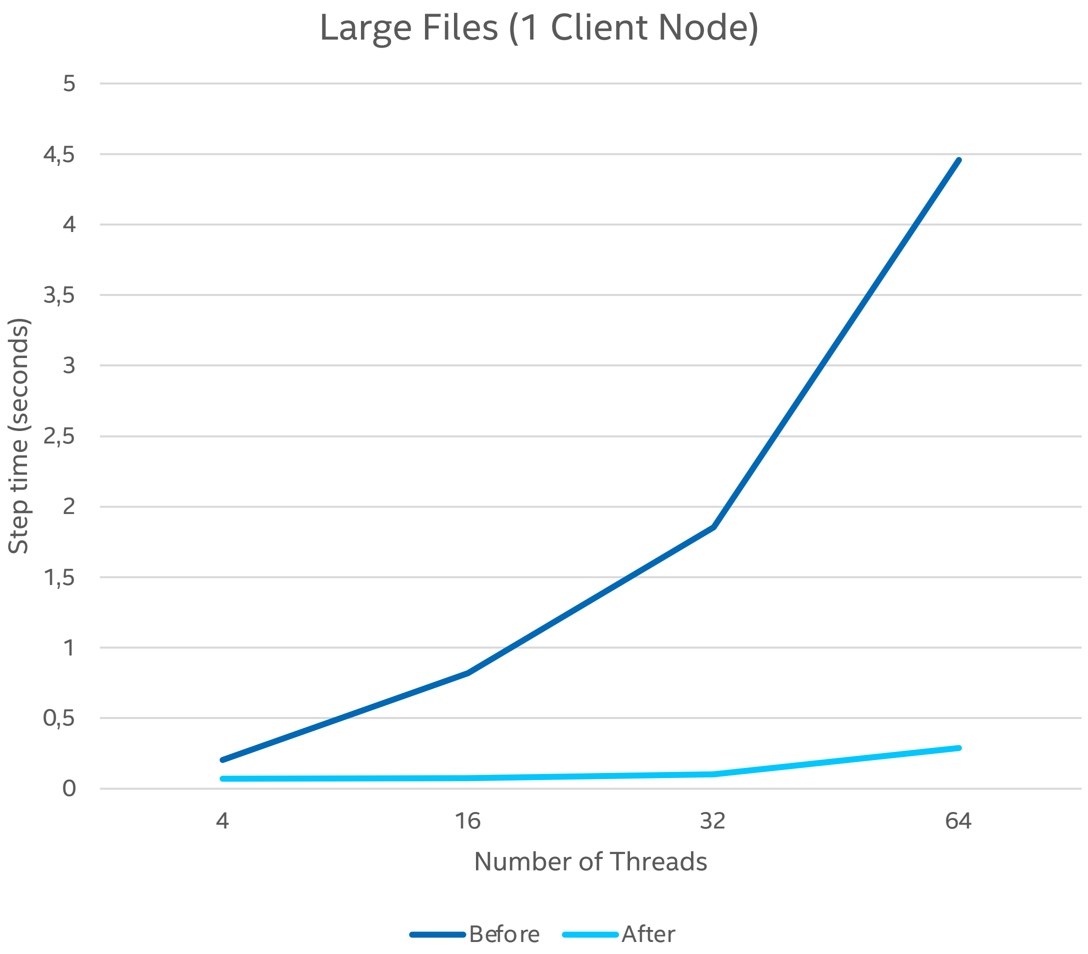
- Multi-user dfuse
  - Mountpoint owned by root and accessible by all users
  - Unix permissions apply
  - Can be used with IL, but no global handle
- Interception of streaming functions
  - f{open,read,write,close,…}
  - No caching
- Aggressive dfuse caching
  - Readdir and other operations

Middleware: dfuse Caching AI Benchmark

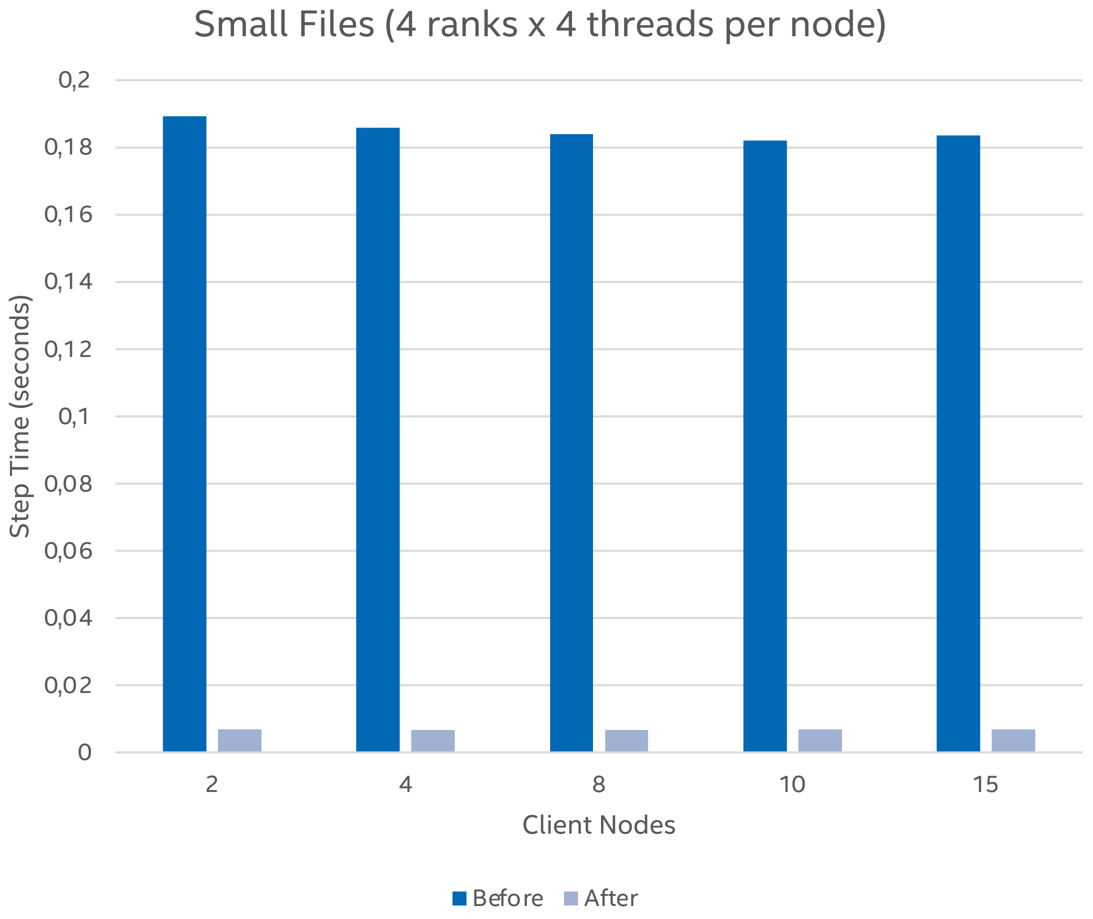
- Benchmark to simulate some AI / ML workloads
  - 2 Datasets: large number of 4k files, small number of 4m files
  - x MPI ranks, each having y threads, all reading the dataset at the same time
    - Stat, open, read, close
- Measure on wolf using dfuse:
  - Before results: all default options
  - After results: do all the optimizations for aggressive caching
- Benchmark from LRZ: Durillo Barrionuevo, Juan; Hammer, Nicolay

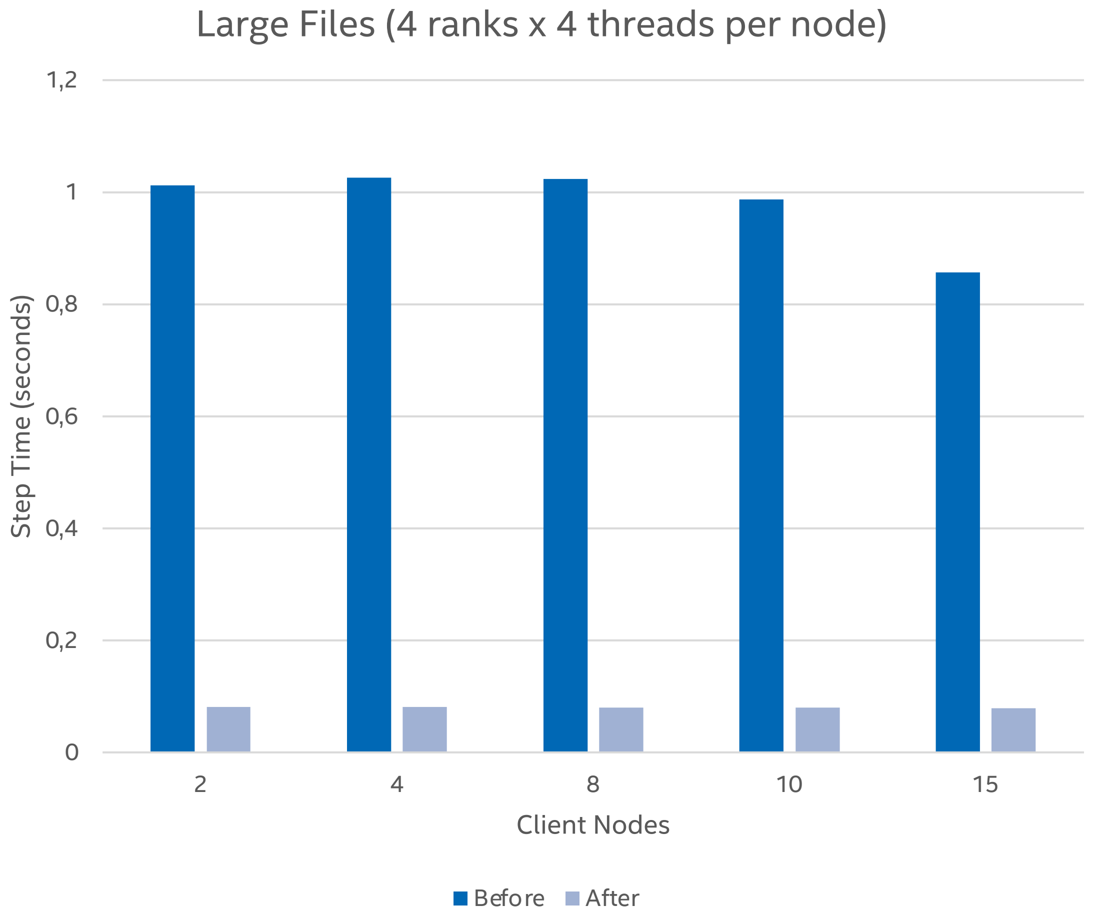
Middleware: dfuse Caching 1 Client





Middleware: dfuse Caching Many Clients





Middleware: WORM Containers (3.0)

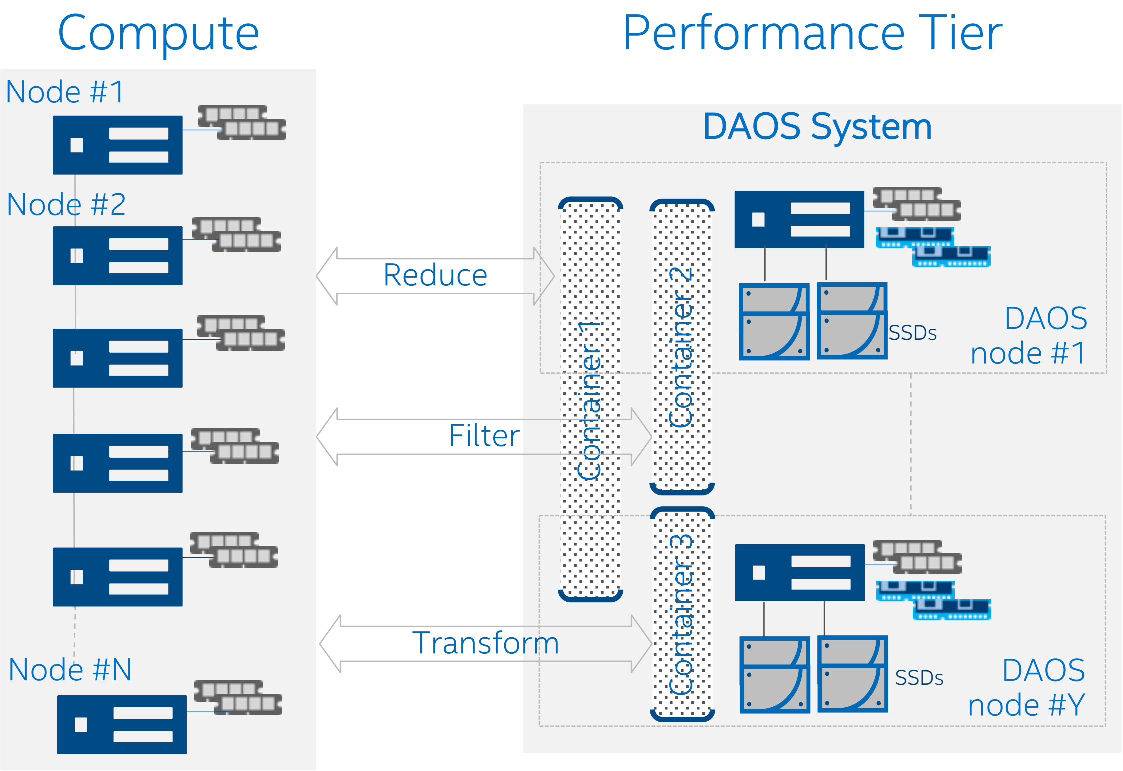
- Write Once Read Many Containers
- Optimizations for read-only datasets
  - Aggressive caching
  - Read-optimized layout
  - Size optimizations for immutable files
  - Indexing
- Use cases
  - Training/verification datasets

Direction: Full Userspace POSIX Support

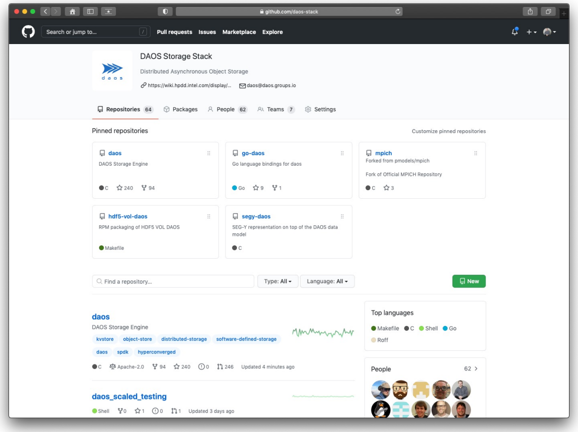
- Intercept all POSIX calls, including metadata operations
  - File open/create/stat/close and derivative
  - mmap via userfaultfd
- Collaborative caching
  - Use shared memory for data & metadata caching
  - Accessible by all ranks running on the same node
  - Size of cache configurable

Direction: Scale-out Active Storage

- DAOS pipeline API
  - Offload data-intensive processing to storage
  - Pre-defined or user-defined (ubpf)
- Leverage HW acceleration
  - computational storage devices
  - accelerators/smartNICs
- Many use cases
  - POSIX find(1)
  - SQL query / MariaDB prototype using predicate push-down
  - In-place data filtering/pre-processing/transformation for AI frameworks
  - Calculate max/min/sum/avg … or searching for specific pattern/value on metadata/data



Resources



DAOS Storage Stack

Distributed Asynchronous Object Storage

https://wiki.hpdd.intel.com/display/... daos@daos.groups.io

Repositories 64 · Packages · People 62 · Teams 7 · Settings

Pinned repositories — Customize pinned repositories

- **daos** — DAOS Storage Engine — C ☆ 240 · 94
- **go-daos** — Go language bindings for daos — Go ☆ 9 · 1
- **mpich** — Forked from pmodels/mpich — Fork of Official MPICH Repository — C ☆ 3
- **hdf5-vol-daos** — RPM packaging of HDF5 VOL DAOS — Makefile
- **segy-daos** — SEG-Y representation on top of the DAOS data model — C

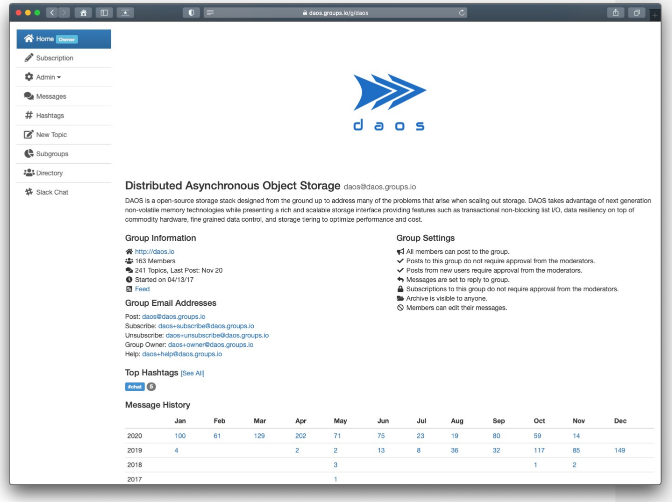
Find a repository... Type: All ▾ Language: All ▾ New

daos
DAOS Storage Engine
kvstore · object-store · distributed-storage · software-defined-storage · daos · spdk · hyperconverged
C · Apache-2.0 · 94 · ☆ 240 · 0 · 246 · Updated 4 minutes ago

daos_scaled_testing
Shell · 0 · ☆ 1 · 0 · 1 · Updated 3 days ago

Top languages: ● Makefile ● C ● Shell ● Go ● Roff

People 62



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

Top Hashtags [See All]

#chat 5

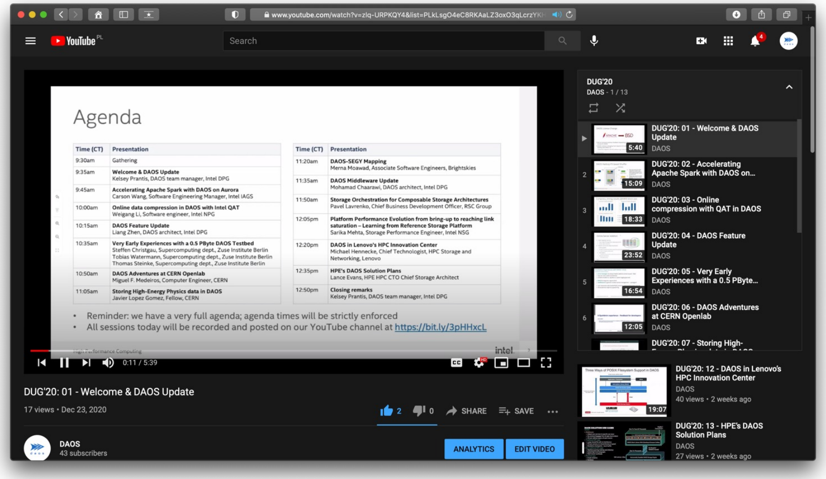
Group Settings

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.

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							

- Open-source Community
  - Github: https://github.com/daos-stack/daos
  - Online doc: http://daos.io
  - Mailing list & slack: https://daos.groups.io
  - YouTube channel: http://video.daos.io
- 6th DAOS User Group (DUG'22)
  - Recordings will be available at http://dug.daos.io
- DAOS BoF Community at SC'22
- Intel landing page
  - https://www.intel.com/content/www/us/en/high-performance-computing/daos.html


www.youtube.com/watch?v=zlq-URPKQY4&list=PLkLsgO4eC8RKAaLZ3oxO3qLcrzYKH
YouTube PL
Search
0:11 / 5:39
2
0
SHARE
SAVE
ANALYTICS
EDIT VIDEO

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

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

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