SC22

Dallas, TX | hpc accelerates.

Overview of the DAOS Object Store

Johann Lombardi, Senior Principal Engineer, AXG, Intel

PMDK & DAOS Tutorial, 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.

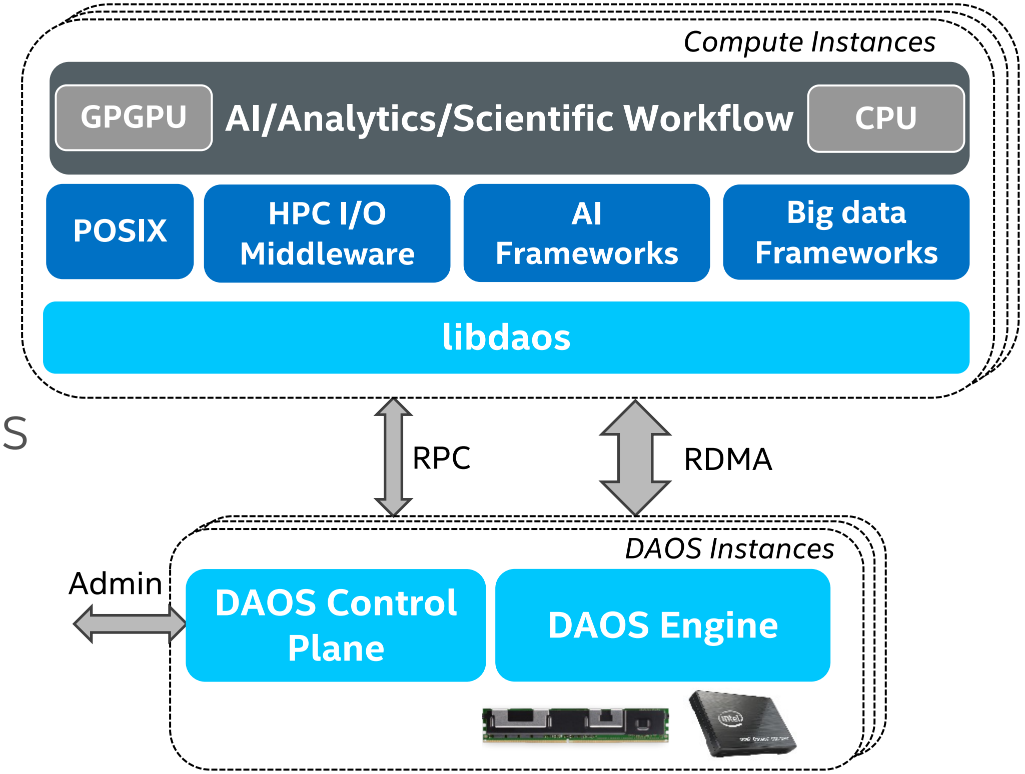
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, 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.

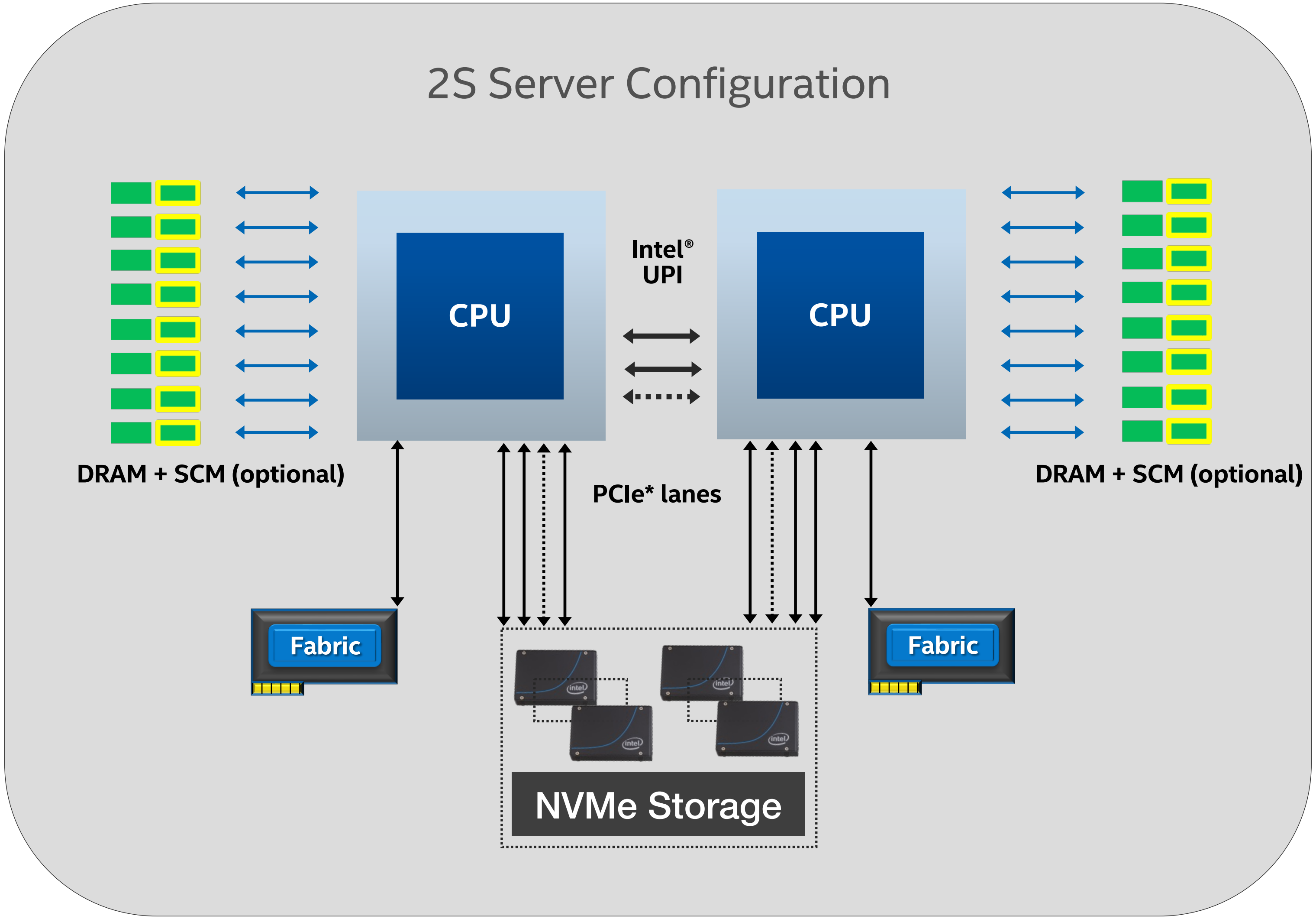
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

DAOS: Nextgen Open Storage Platform

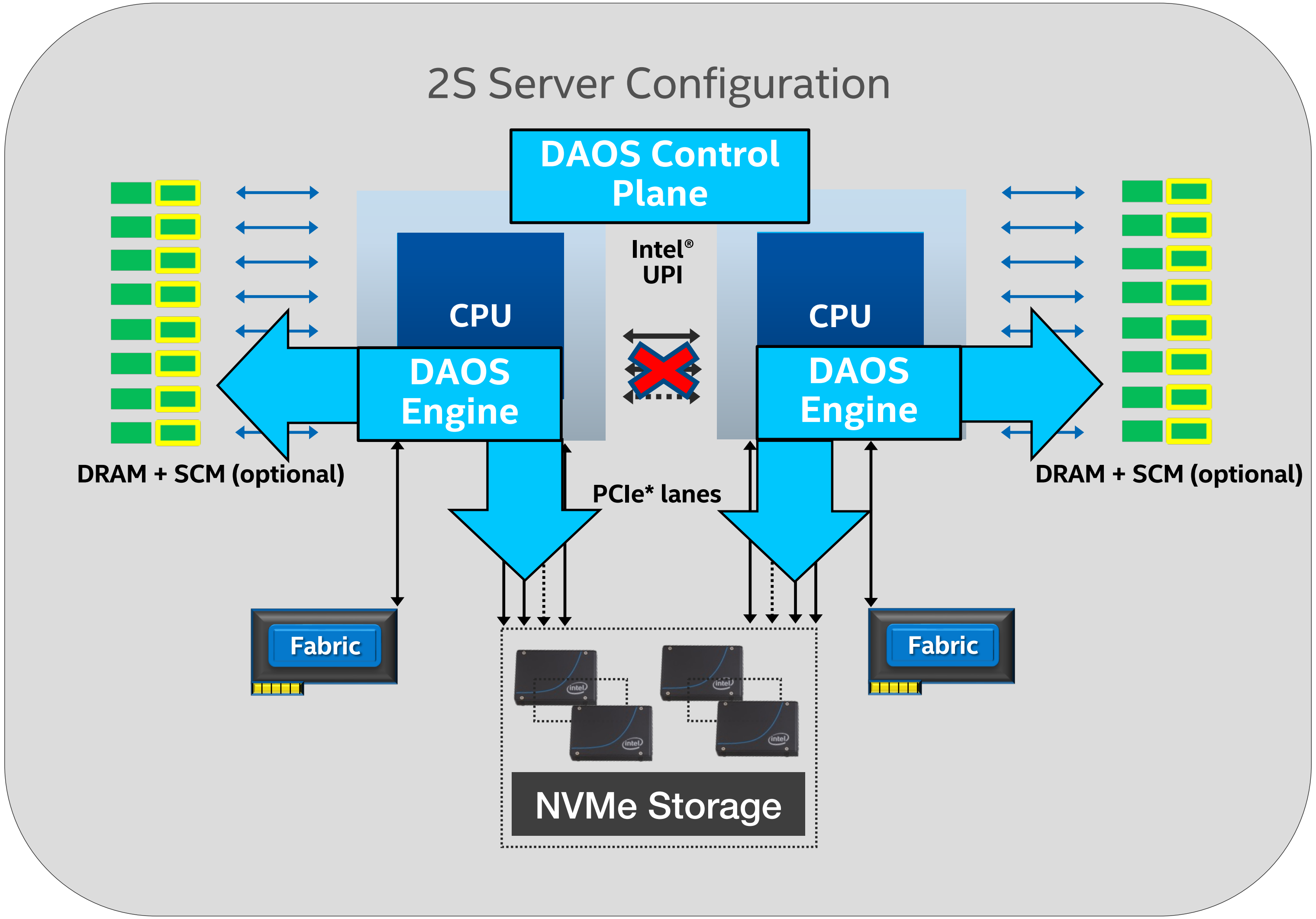
- Fully Distributed multi-tenant global namespace
- Platform for innovation
  - Modular API and layering
  - Can leverage latest HW & SW technology
- Built for high performance
  - 10’s µs latency, billions of IOPS, TB/s to PB/s
- Full userspace model
  - Run on-prem or in the cloud
- Growing open-source community



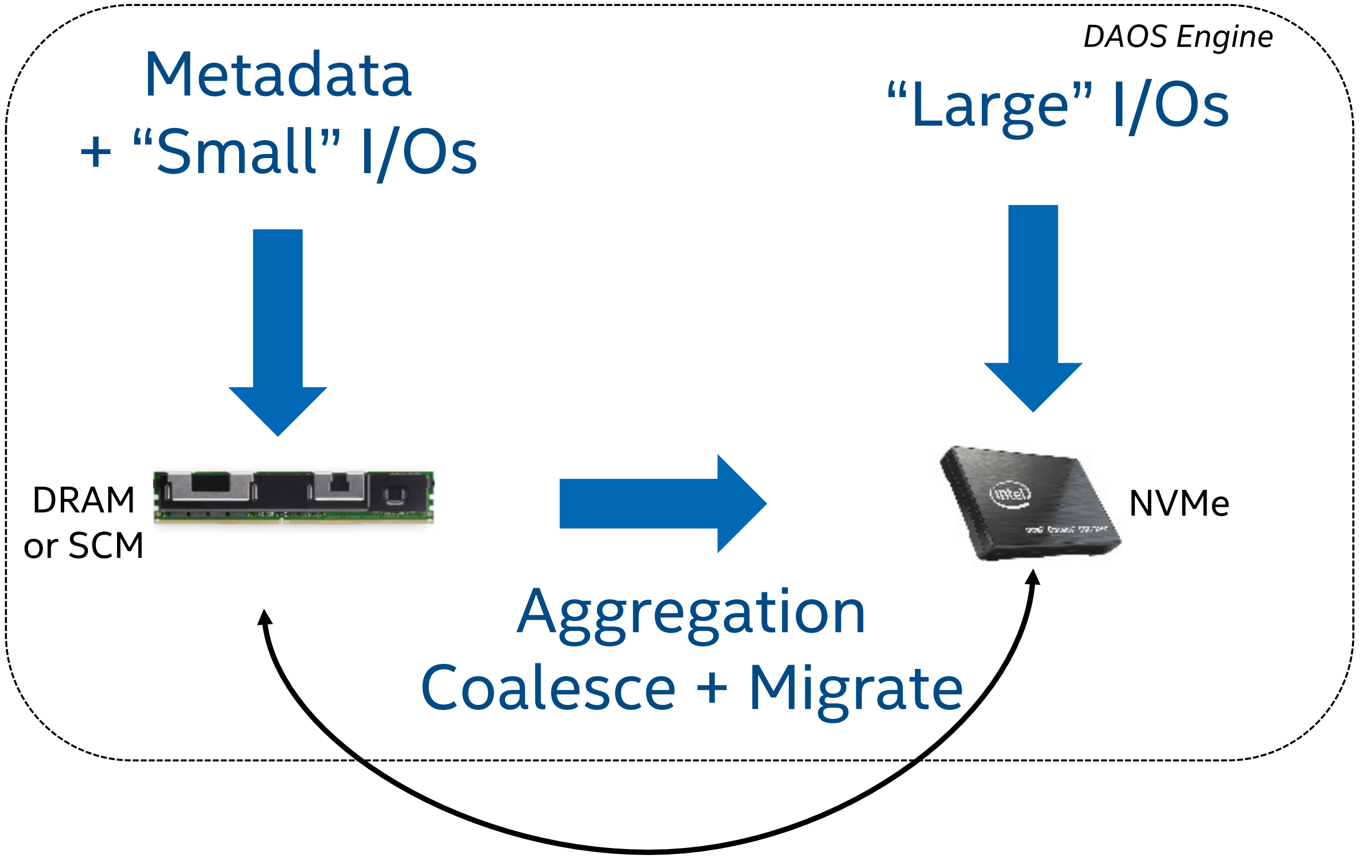
DAOS Node Design



DAOS Node Design



Engine: Media Management


DAOS Engine

Metadata
+ "Small" I/Os

"Large" I/Os

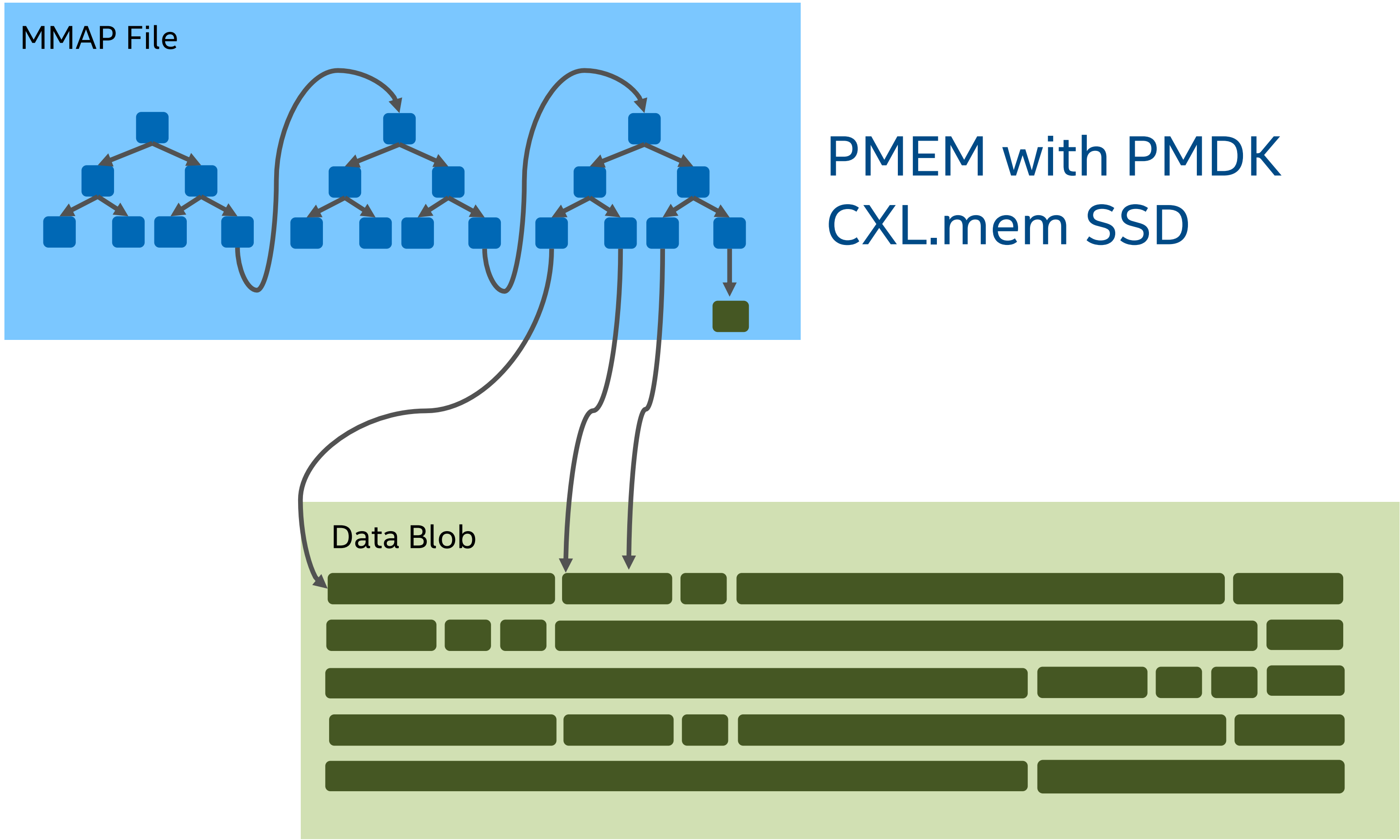
DRAM
or SCM

NVMe

Aggregation
Coalesce + Migrate

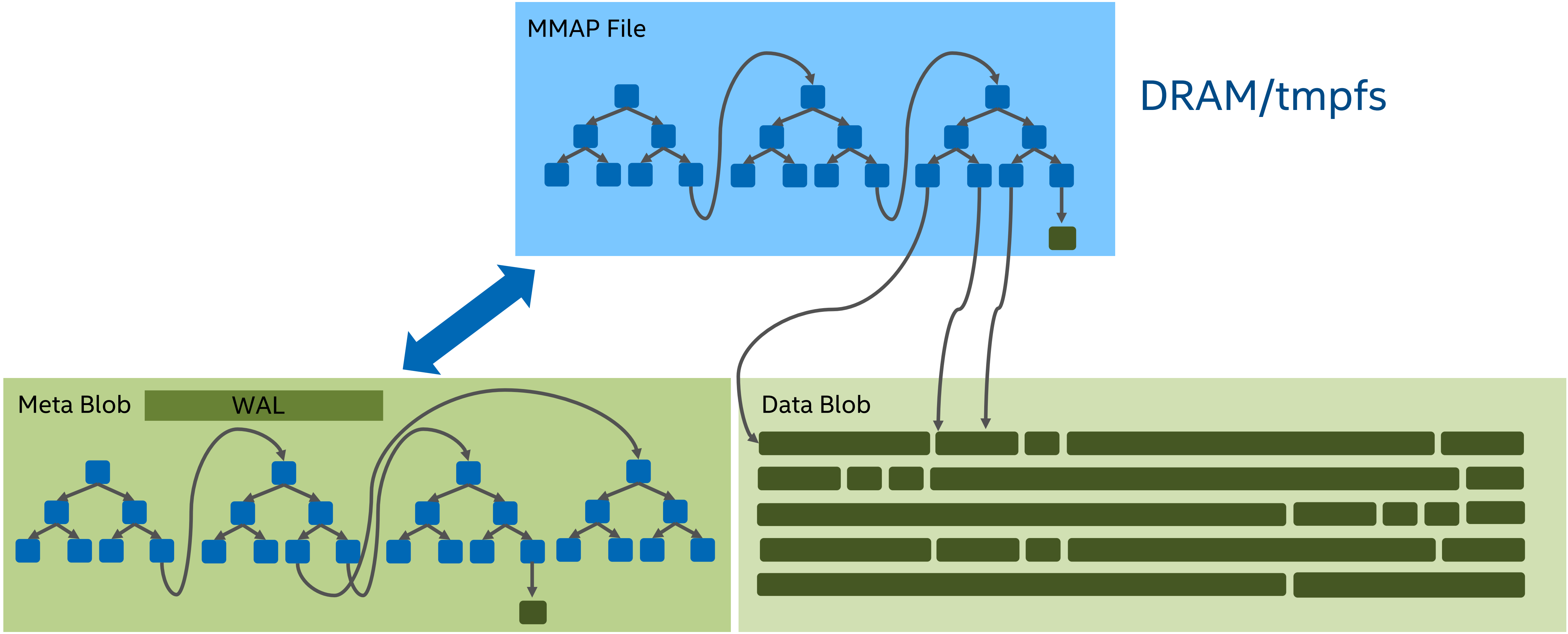
1-6% Capacity Ratio

Metadata: Persistent Device

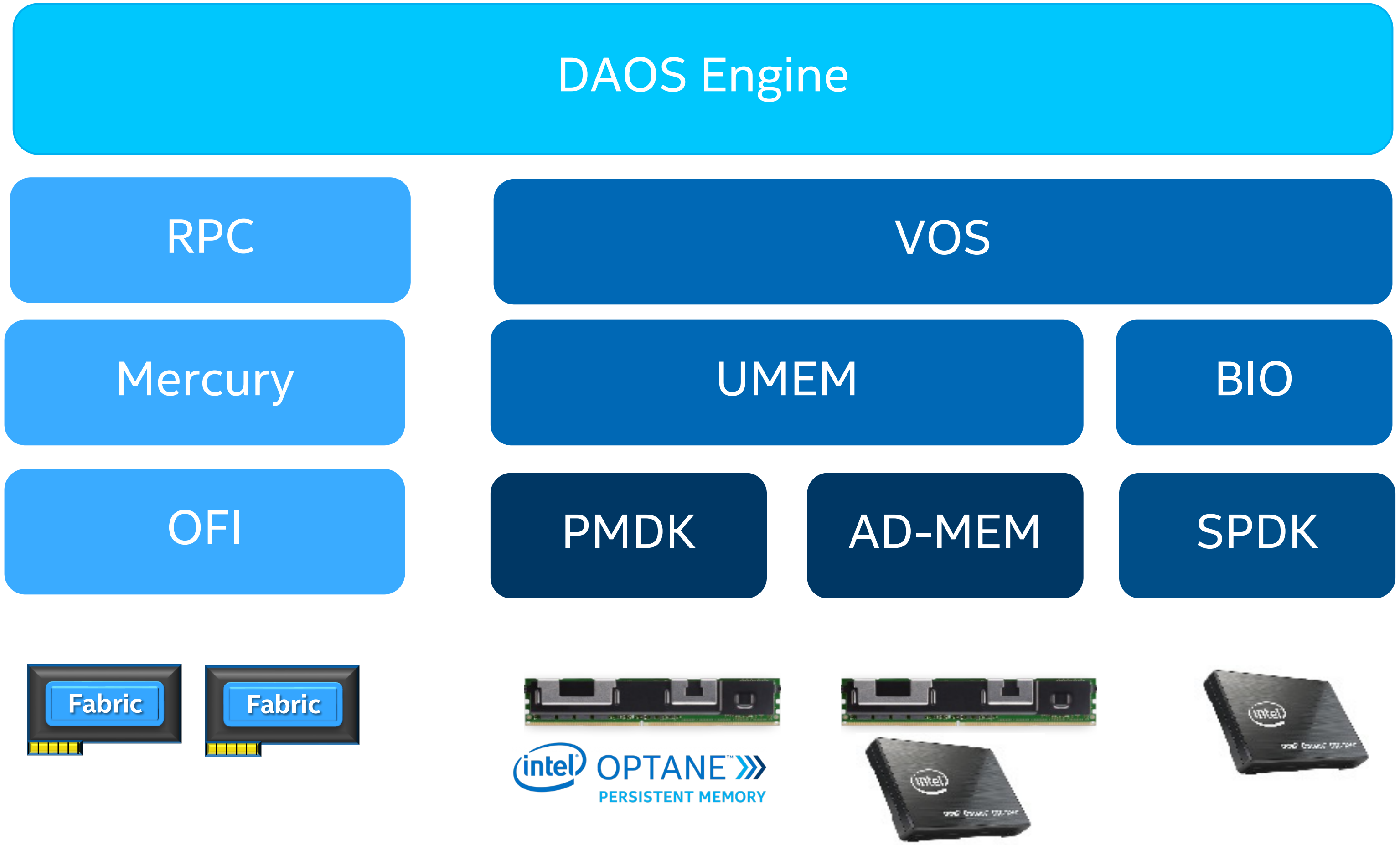


PMEM with PMDK
CXL.mem SSD

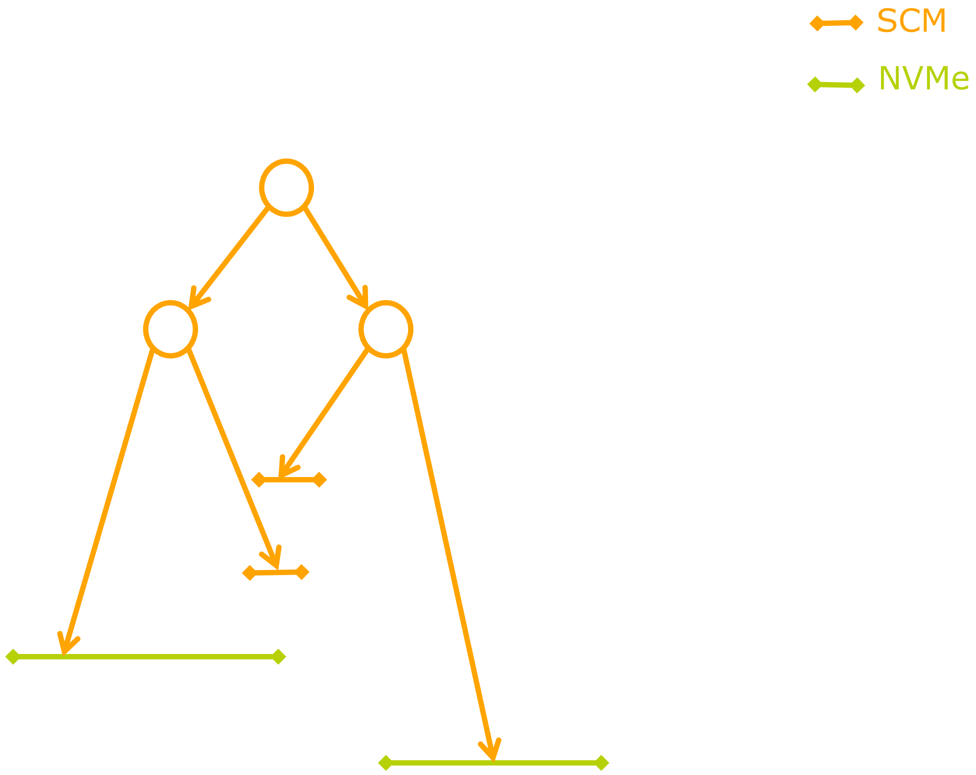
Metadata: Volatile Device



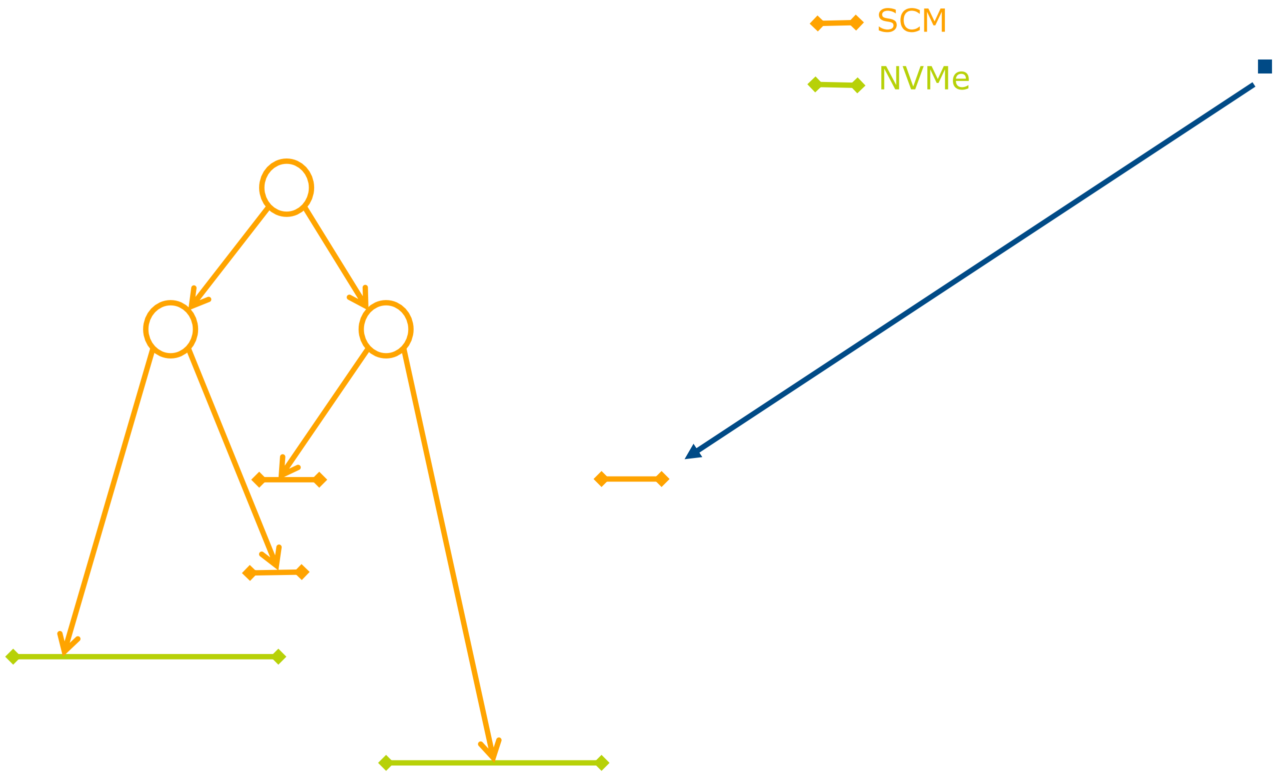
Engine Software Stack



I/O Operation Flow



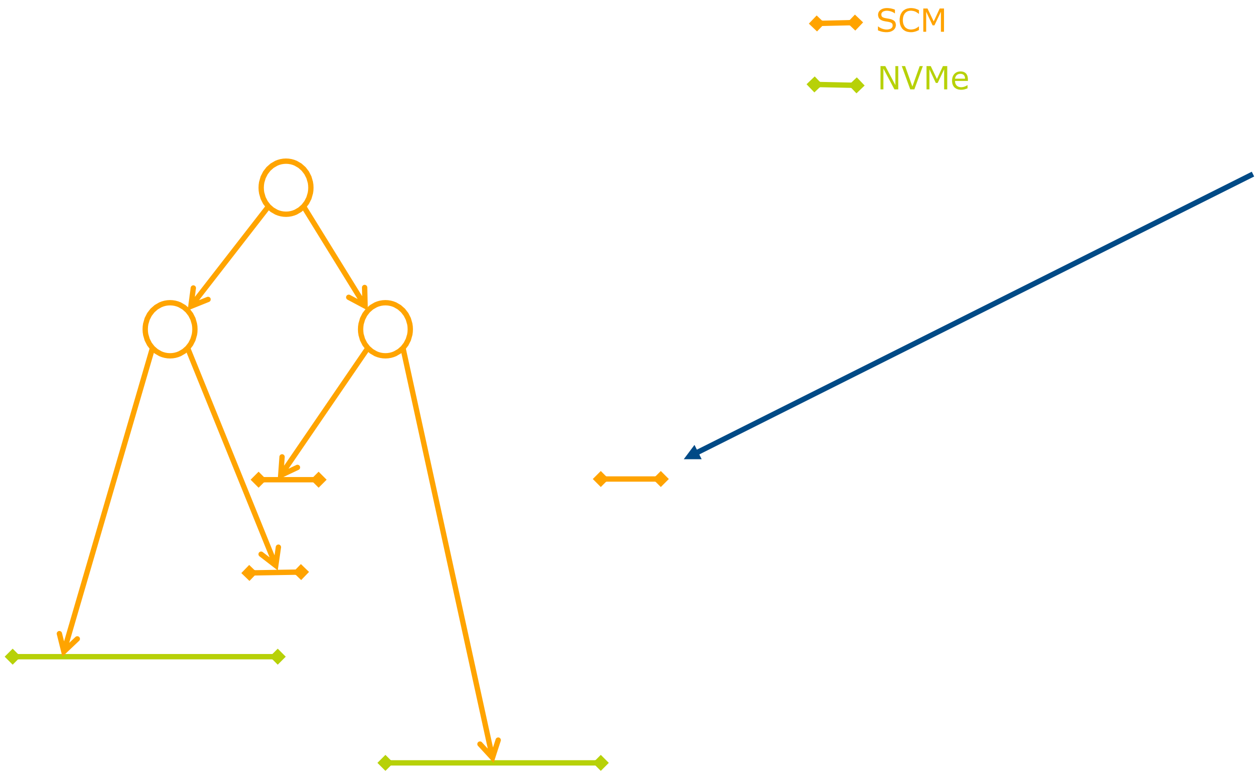
I/O Operation Flow



DAOS XStream

- pmemobj_reserve() new buffer

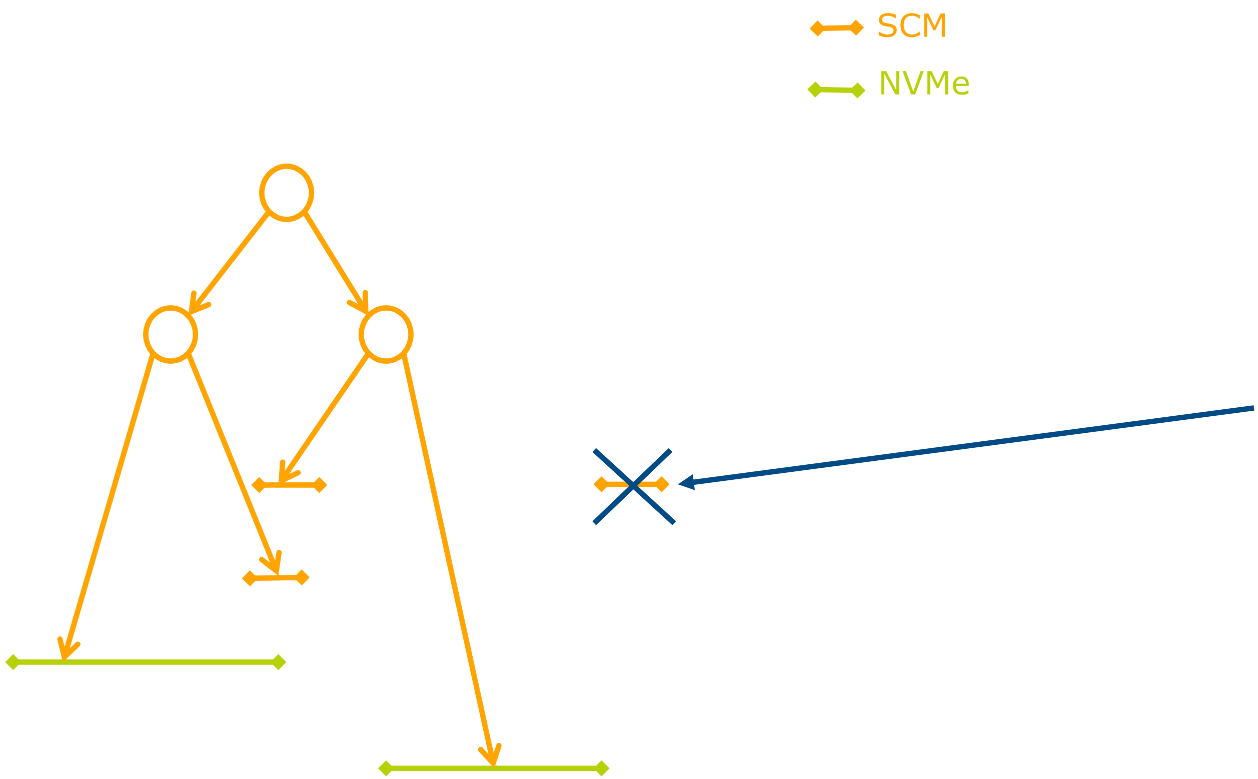
I/O Operation Flow



DAOS XStream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported

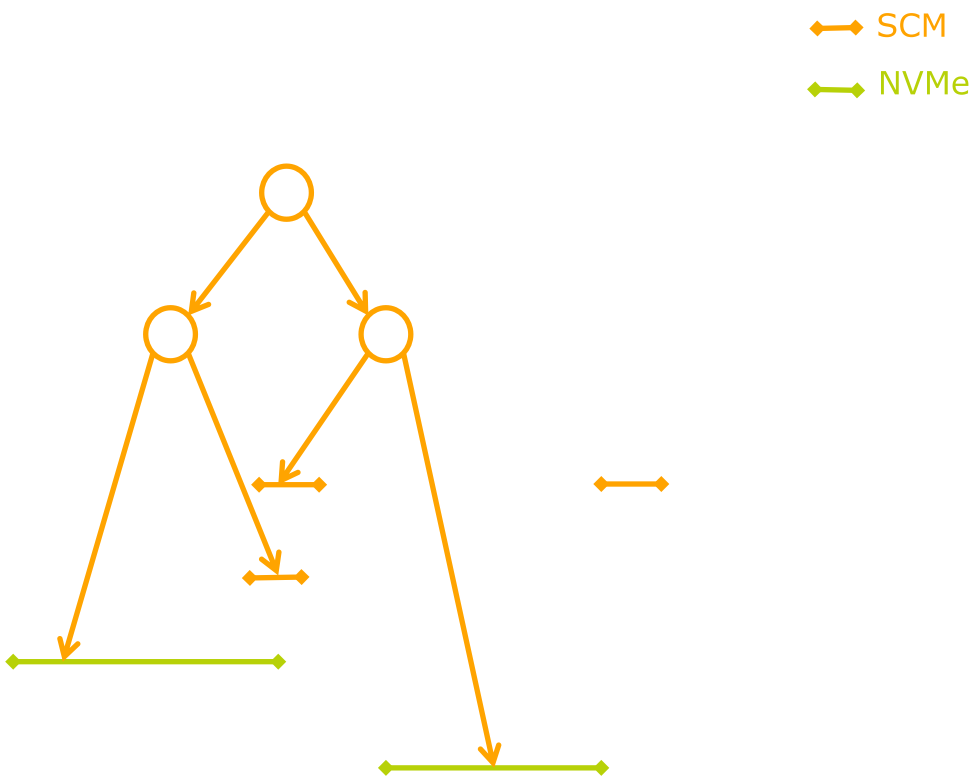
I/O Operation Flow



DAOS XStream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported
- If RDMA transfer failed, free buffer with pmemobj_cancel()

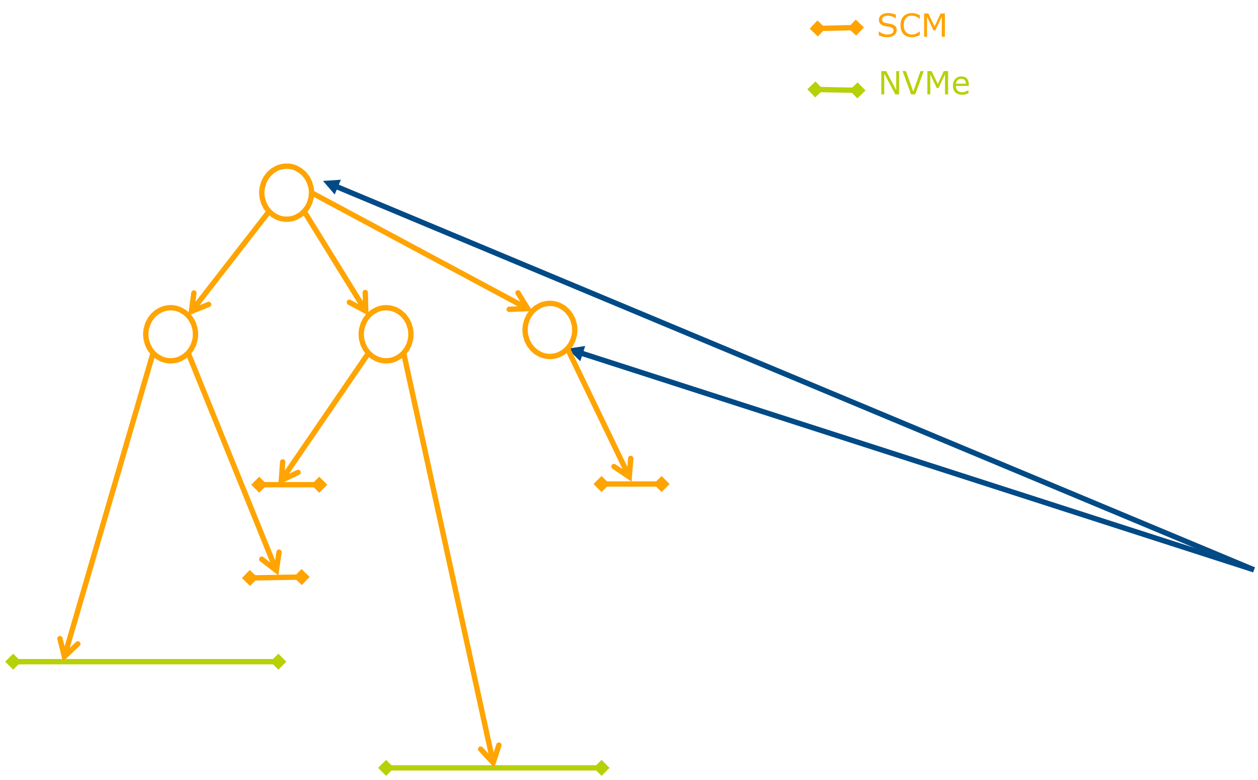
I/O Operation Flow



DAOS XStream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported
- If RDMA transfer failed, free buffer with pmemobj_cancel()
- Otherwise, start pmemobj transaction

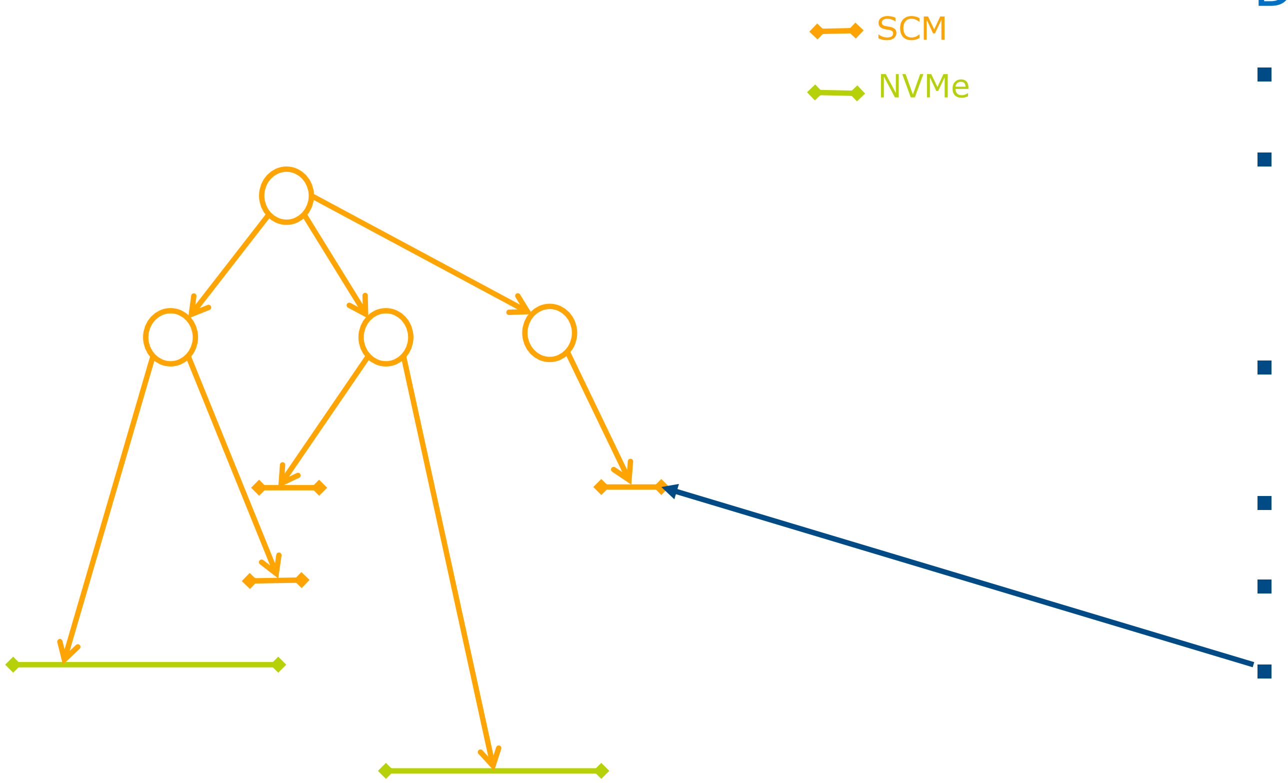
I/O Operation Flow



DAOS XStream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported
- If RDMA transfer failed, free buffer with pmemobj_cancel()
- Otherwise, start pmemobj transaction
- Modify index to insert new extent

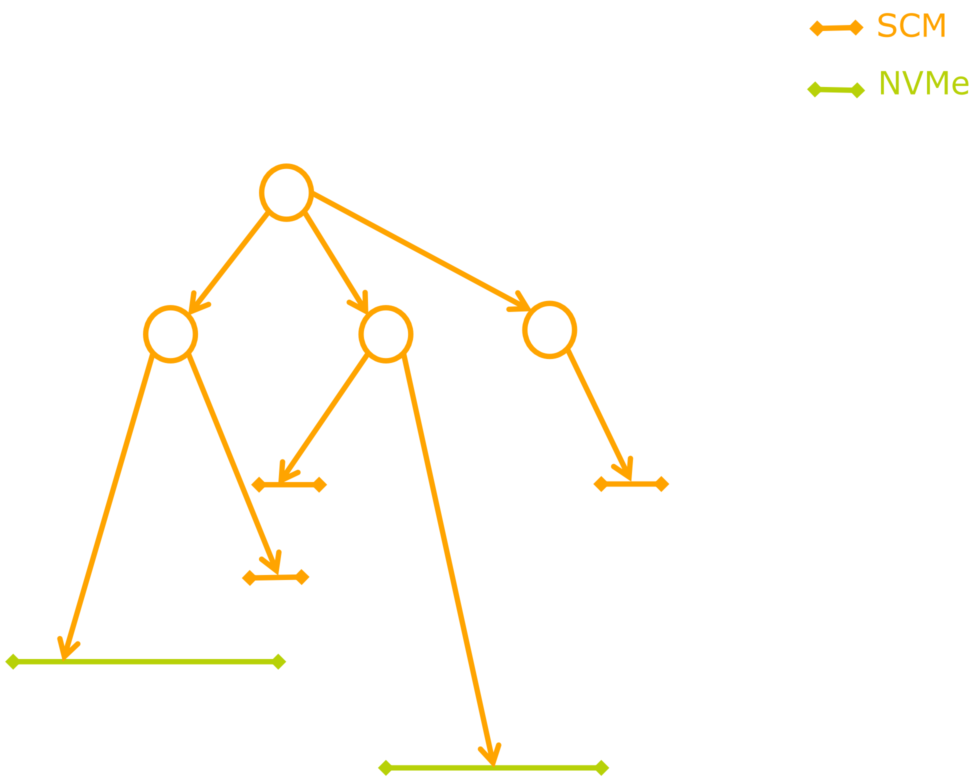
I/O Operation Flow



DAOS Xstream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported
- If RDMA transfer failed, free buffer with pmemobj_cancel()
- Otherwise, start pmemobj transaction
- Modify index to insert new extent
- pmemobj_tx_publish()

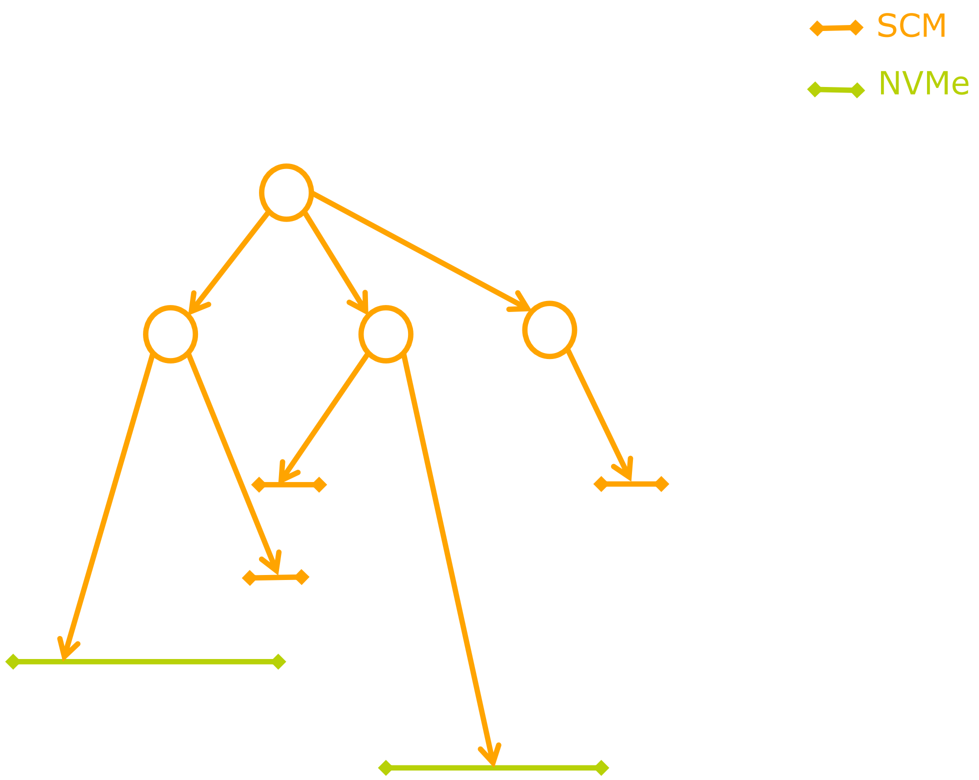
I/O Operation Flow



DAOS Xstream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported
- If RDMA transfer failed, free buffer with pmemobj_cancel()
- Otherwise, start pmemobj transaction
- Modify index to insert new extent
- pmemobj_tx_publish()
- Commit pmemobj transaction

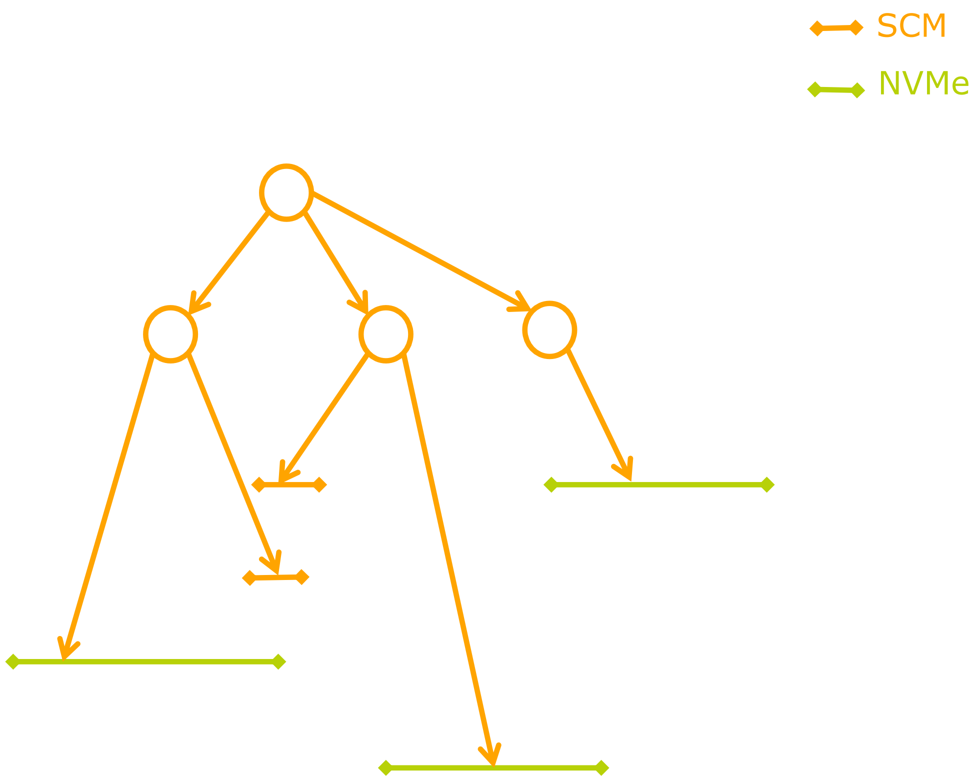
I/O Operation Flow



DAOS Xstream

- pmemobj_reserve() new buffer
- Start RDMA transfer to newly allocated buffer
  - Switch to other ULT until completion is reported
- If RDMA transfer failed, free buffer with pmemobj_cancel()
- Otherwise, start pmemobj transaction
- Modify index to insert new extent
- pmemobj_tx_publish()
- Commit pmemobj transaction
- Reply to the client

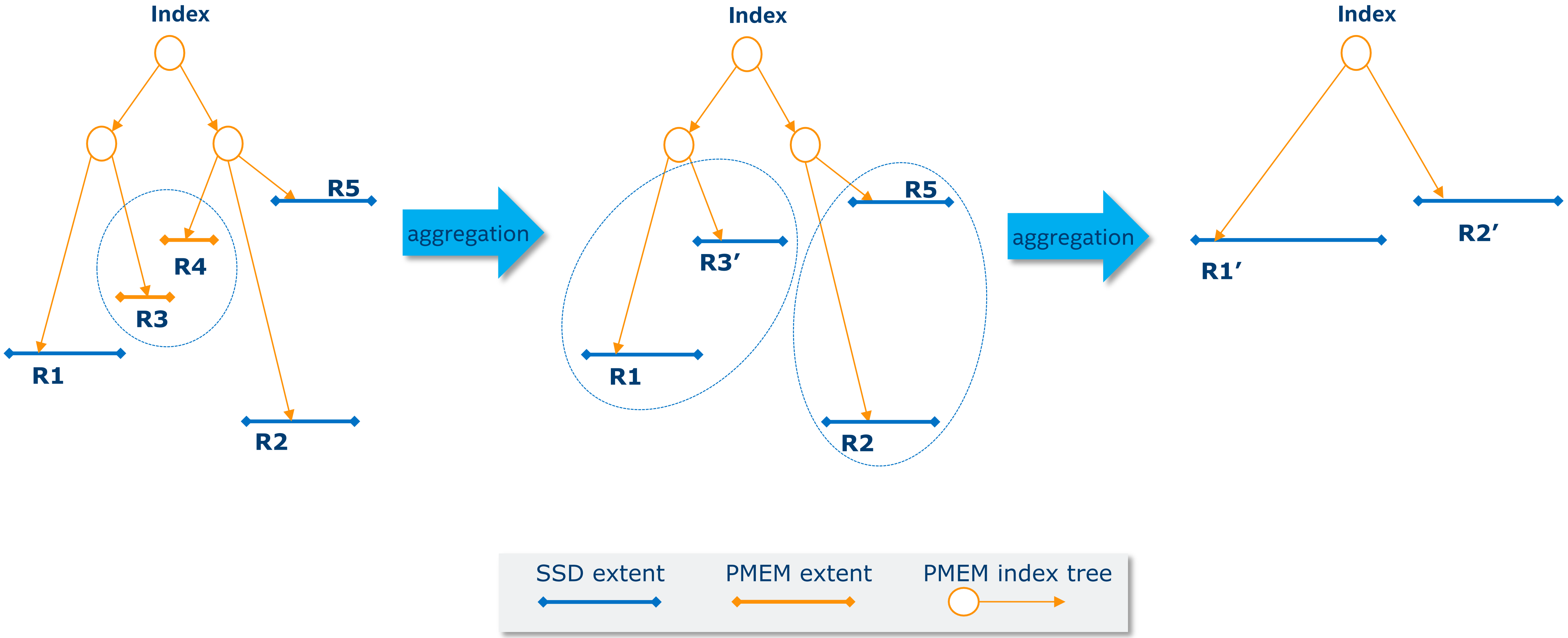
I/O Operation Flow



Extent in NVMe

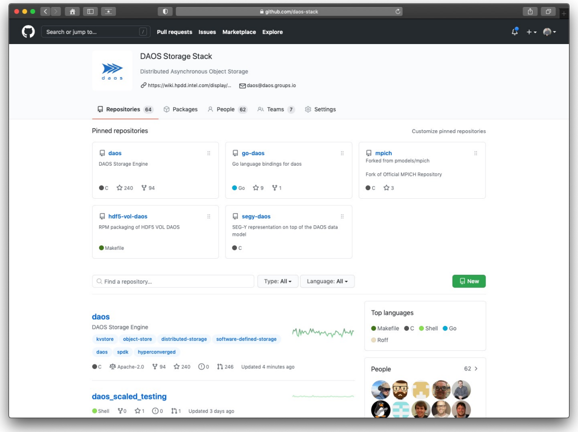
- Track NVMe allocated/free space in data structures maintained in pmemobj pool
- Same processing flow

Data Aggregation



Resources

- 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



DAOS Storage Stack

Distributed Asynchronous Object Storage

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

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

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

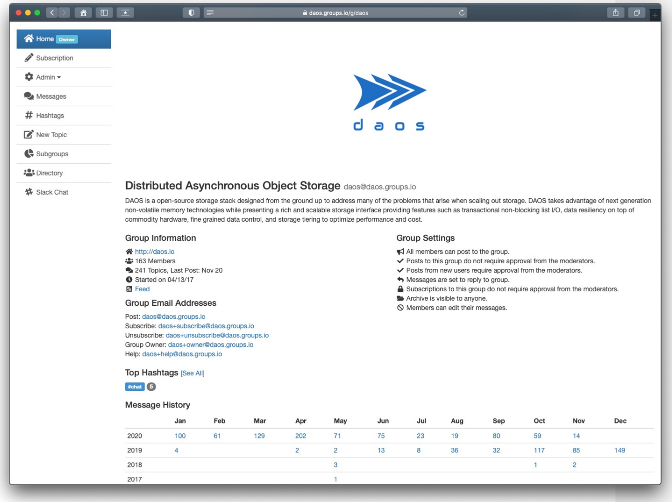
daos_scaled_testing

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

Group Settings

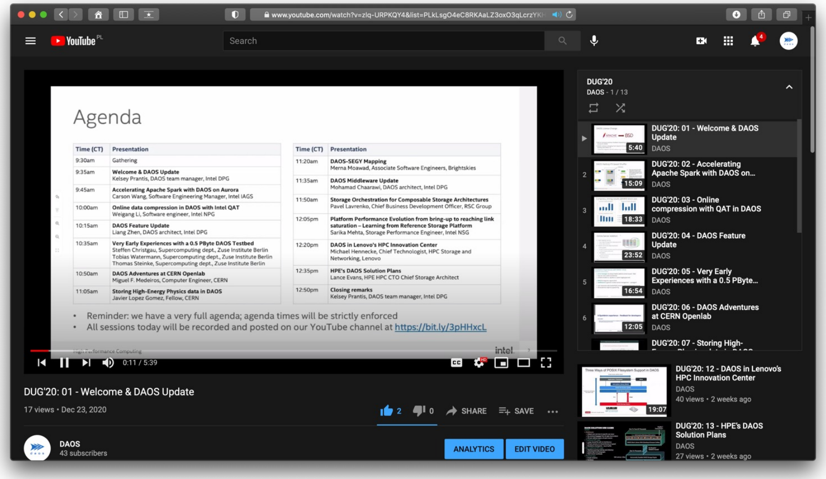
All members can post to the group.
✓ Posts to this group do not require approval from the moderators.
✓ Posts from new users require approval from the moderators.
Messages are set to reply to group.
Subscriptions to this group do not require approval from the moderators.
Archive is visible to anyone.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


www.youtube.com/watch?v=zlq-URPKQY4&list=PLkLsgO4eC8RKAaLZ3oxO3qLcrzYKH
YouTube PL
Search
0:11 / 5:39
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

High Performance Computing intel

DUG'20: 01 - Welcome & DAOS Update

17 views • Dec 23, 2020 👍 2 👎 0

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