Advanced Storage and Memory Hierarchy in AI and HPC with DAOS

Andrey Kudryavtsev

DAOS Product Manager, Intel Corp.

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.

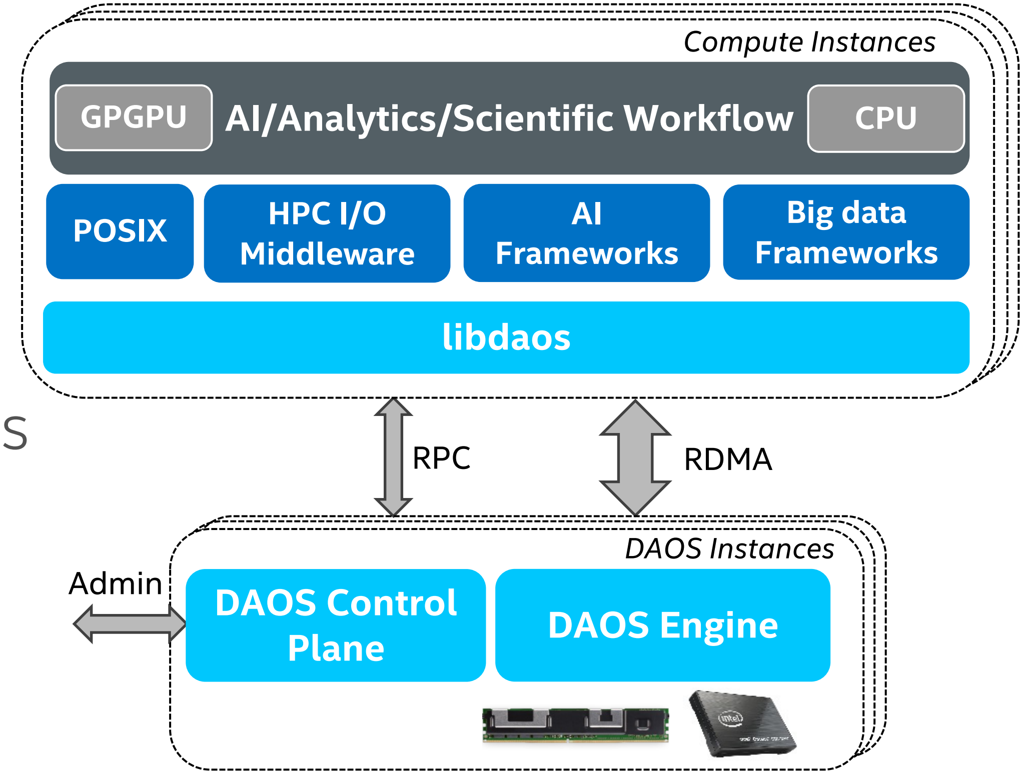
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality,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

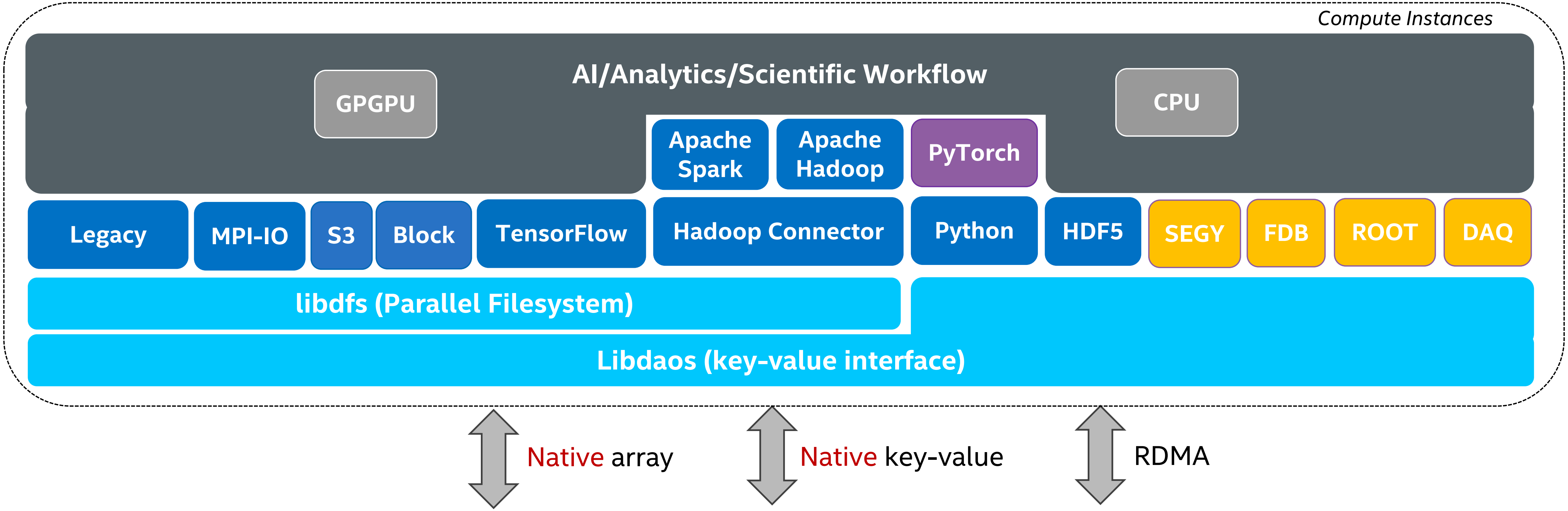
DAOS: Nextgen Open Storage Platform

- 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



Middleware Ecosystem

python PyTorch TensorFlow



Compute Instances

AI/Analytics/Scientific Workflow

GPGPU

CPU

Apache Spark

Apache Hadoop

PyTorch

Legacy

MPI-IO

S3

Block

TensorFlow

Hadoop Connector

Python

HDF5

SEGY

FDB

ROOT

DAQ

libdfs (Parallel Filesystem)

Libdaos (key-value interface)

Native array

Native key-value

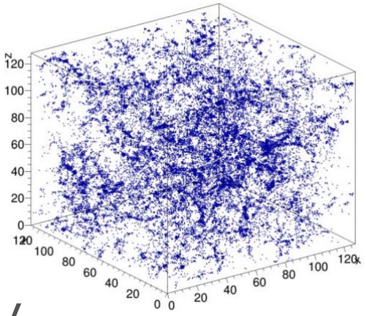
RDMA

🟦 Generic I/O middleware supported today

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

🟪 Enablement in progress

APP: CosmoFlow

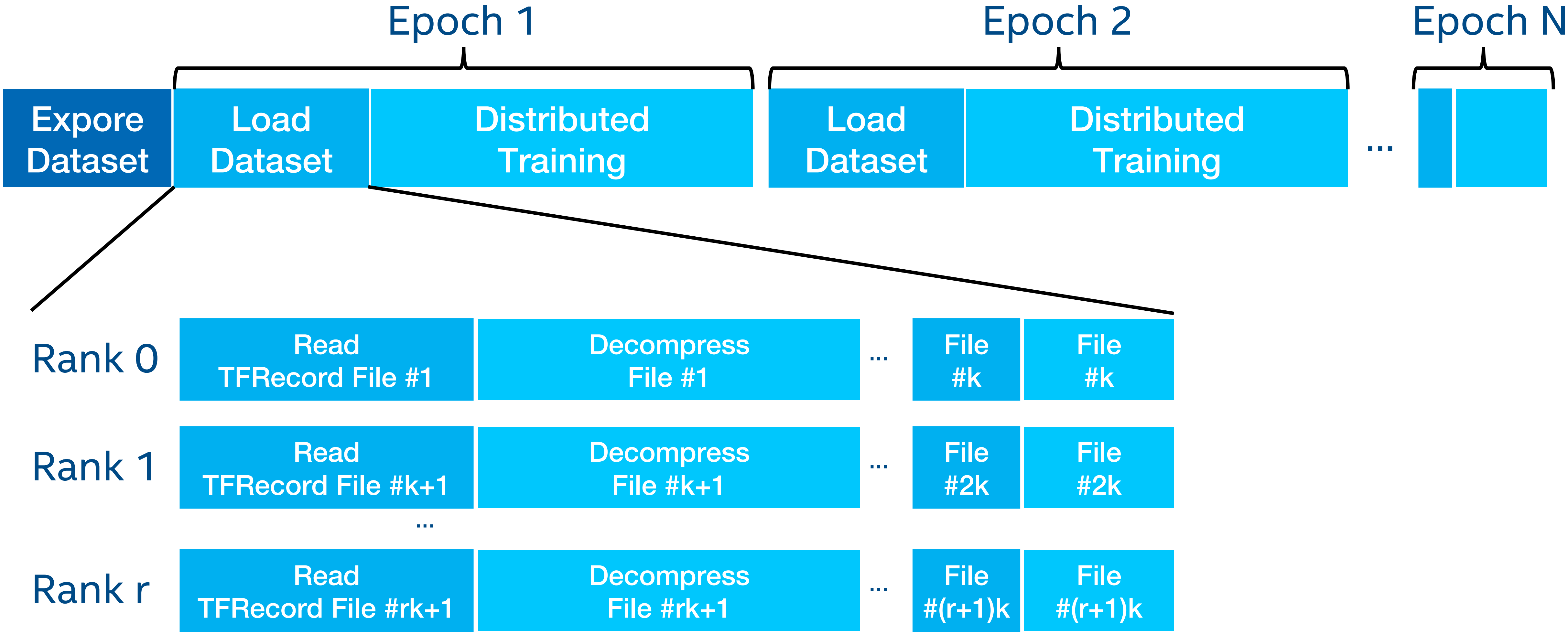


- A highly scalable Deep Learning application for Cosmology, using a 3D convolutional neural network trained on N-body cosmological simulation data to study dark matter in the universe
  - https://arxiv.org/pdf/1808.04728.pdf
  - Built over **TensorFlow**; uses **Horovod** for distributed training
- Community dataset: cosmoUniverse_2019_05_4parE_tf_v2
  - 1.7PiB
  - 524,288 samples for training
  - 65,536 samples for validation
  - Compressed TFRecord files

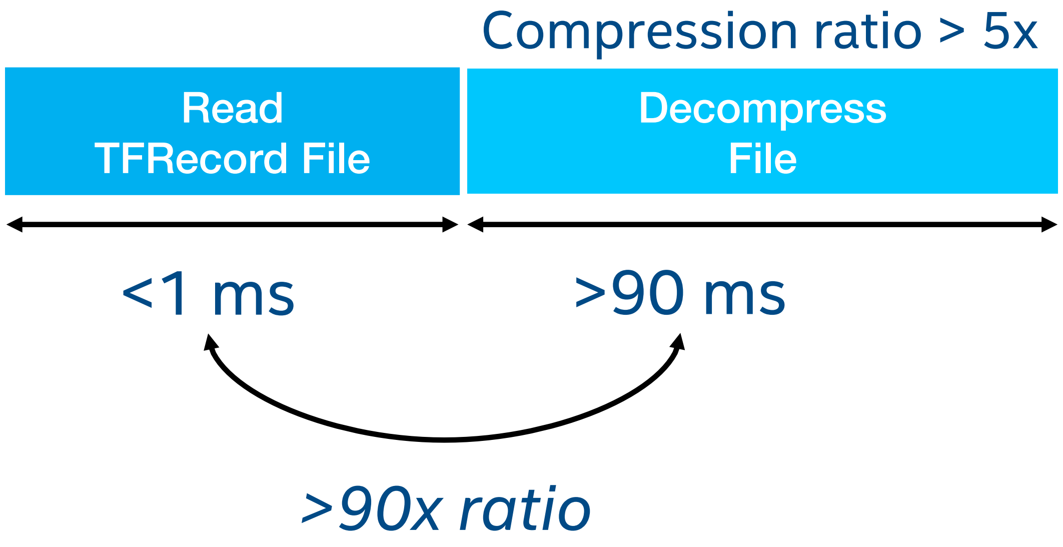
HOROVOD

TensorFlow

APP: Workflow



APP: TFRecord Loading



Compression ratio > 5x

Read TFRecord File

Decompress File

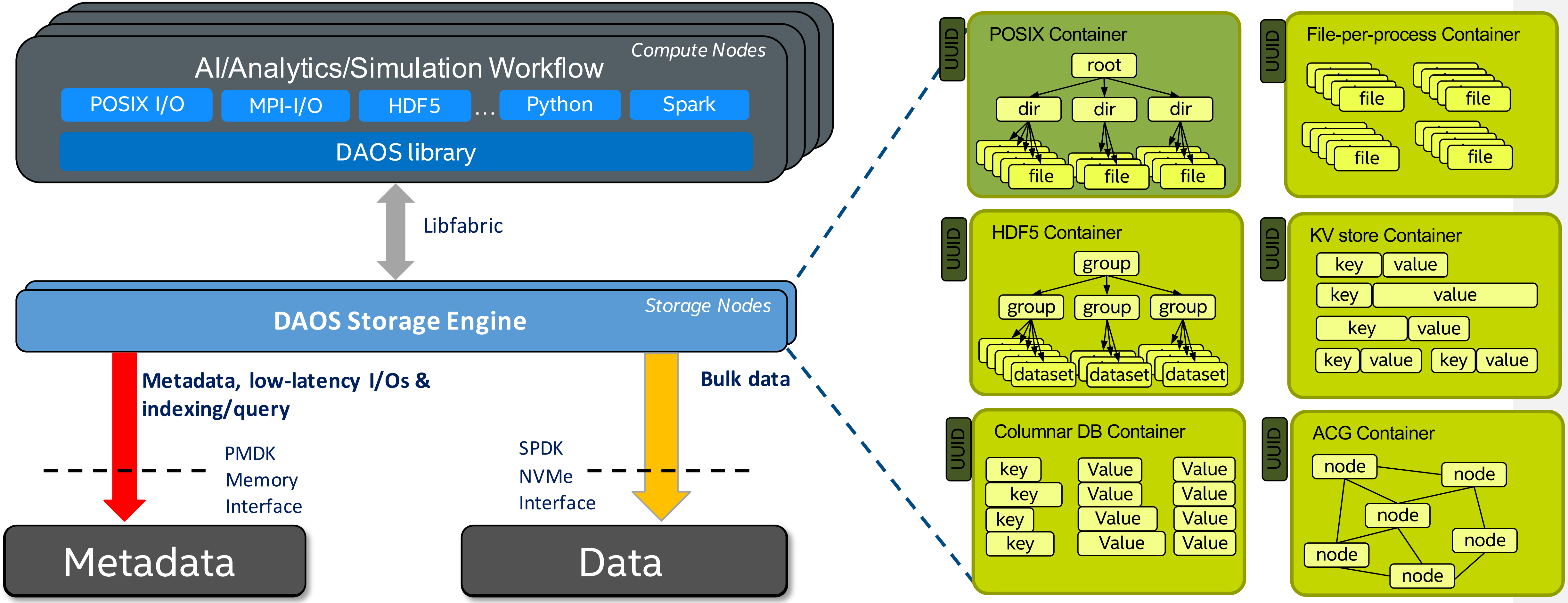
<1 ms

>90 ms

>90x ratio

Decompression coud be optimitized in future DAOS release with inline compression support with Intel® QAT

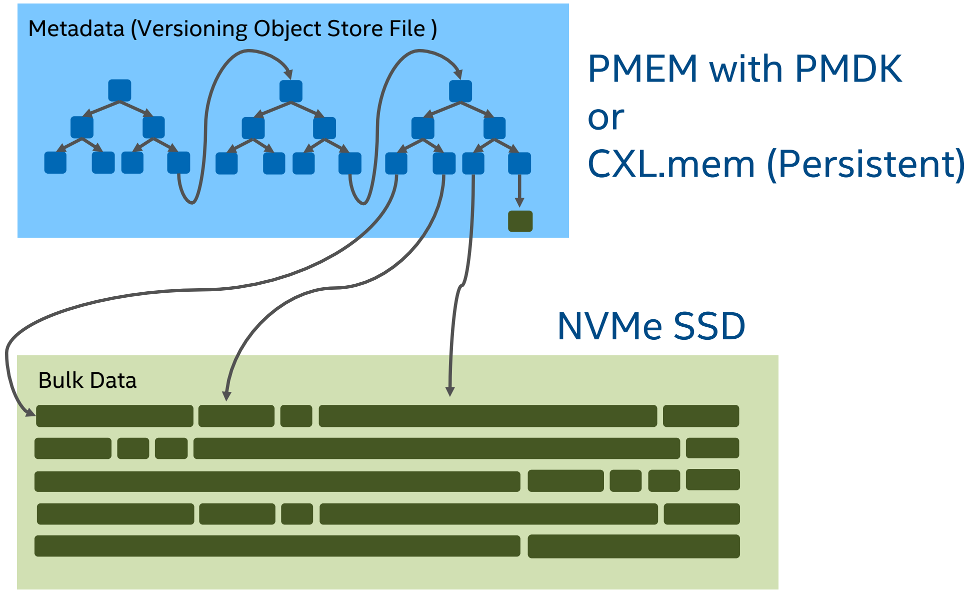
The DAOS Exascale Storage Stack – Software Architecture



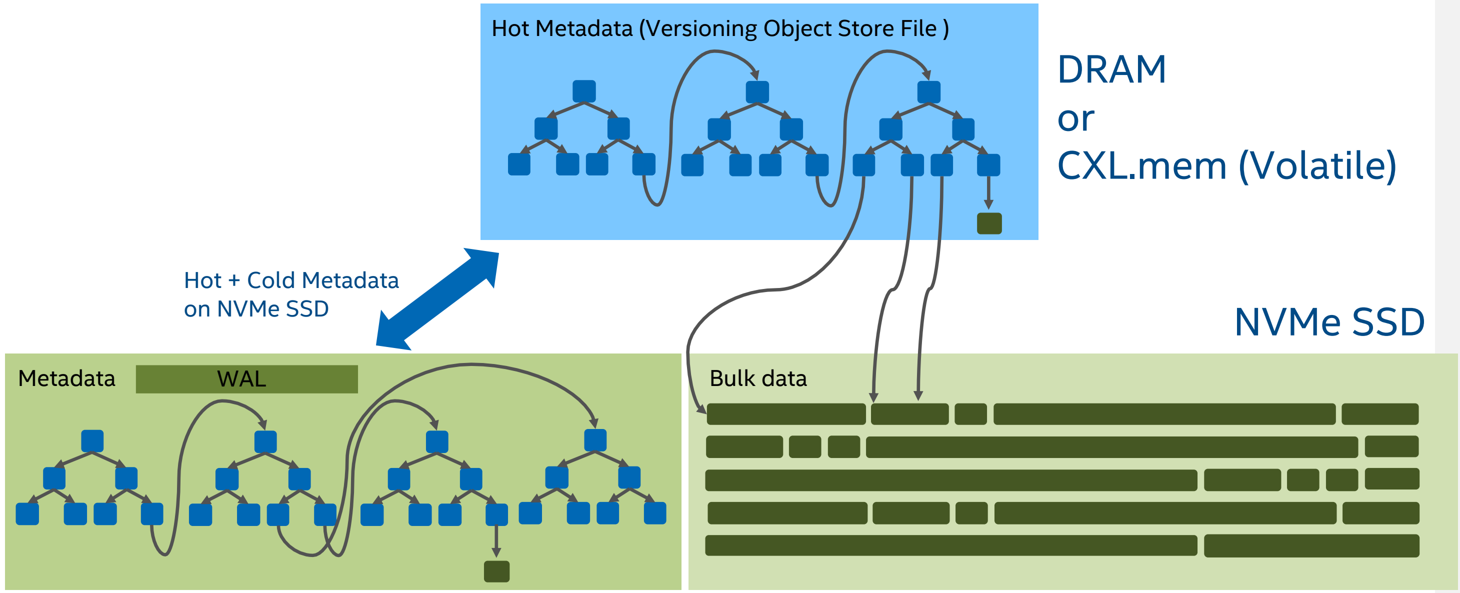
DAOS Program Updates

DAOS to introduce two development paths:

- Stay on PMEM with transitioning to CXL2.0 Persistent memory third party products
- Implement DRAM / CXL.mem (volatile) + NVMe staging for metadata



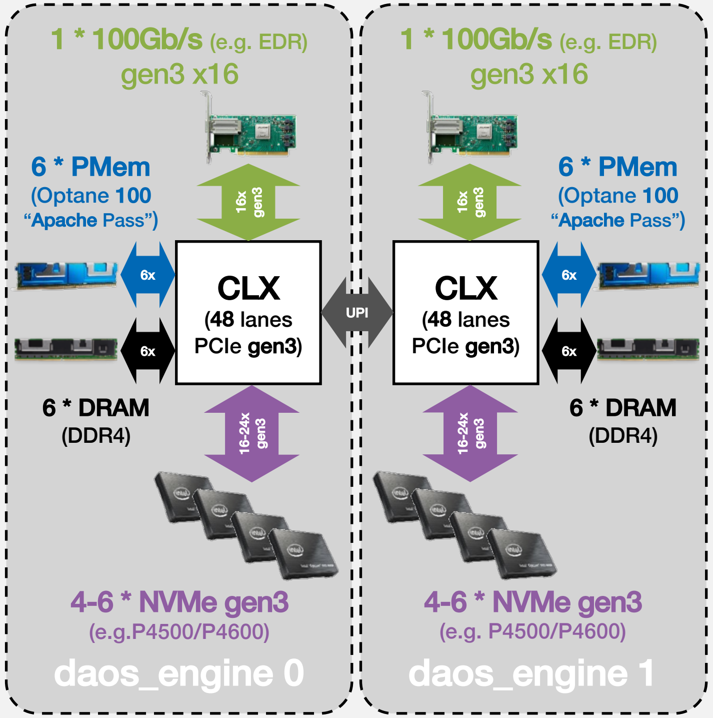
Persistent Device



Volatile Device

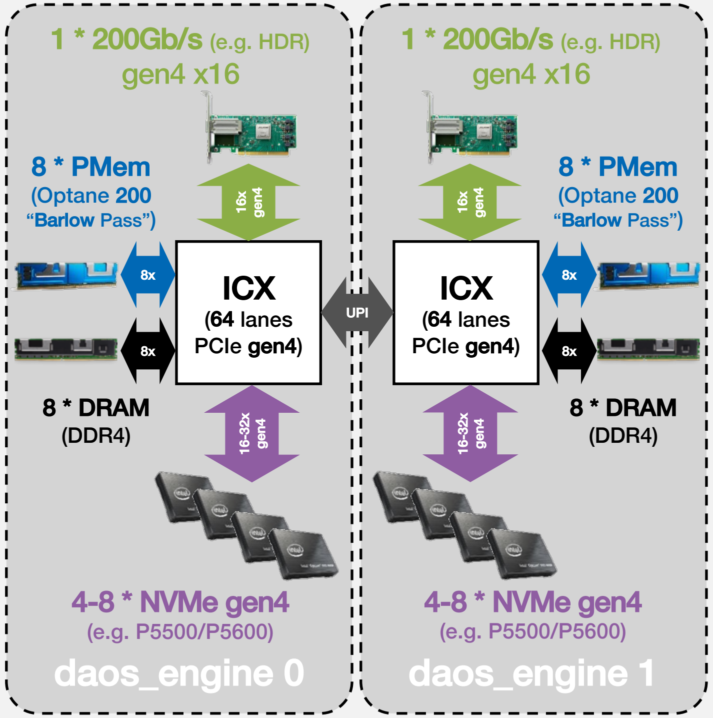
DAOS Servers on Intel Xeon Scalable Processors

2nd Gen Intel Xeon Scalable Processors



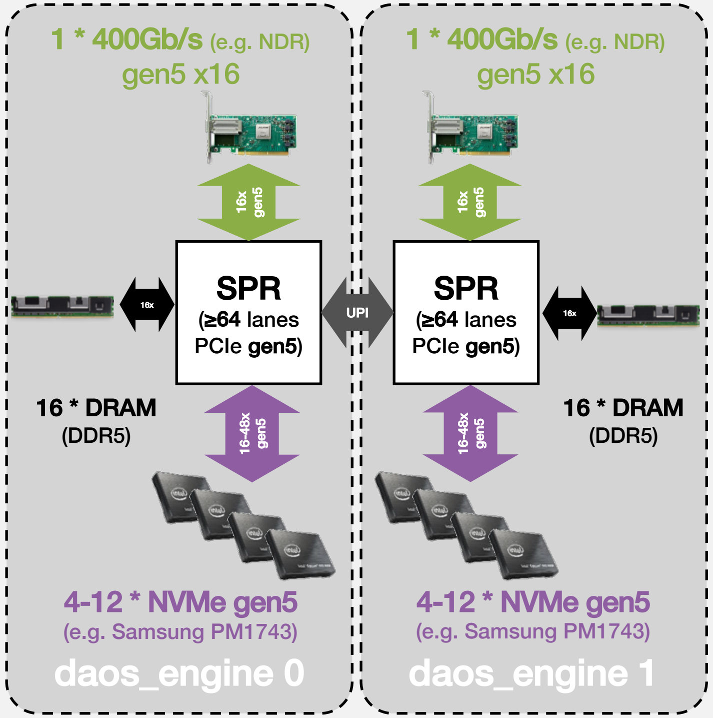
Bandwidth: **≤23 GiB/s** per server

3rd Gen Intel Xeon Scalable Processors



Bandwidth: **≤46 GiB/s** per server

4th Gen Intel Xeon Scalable Processors

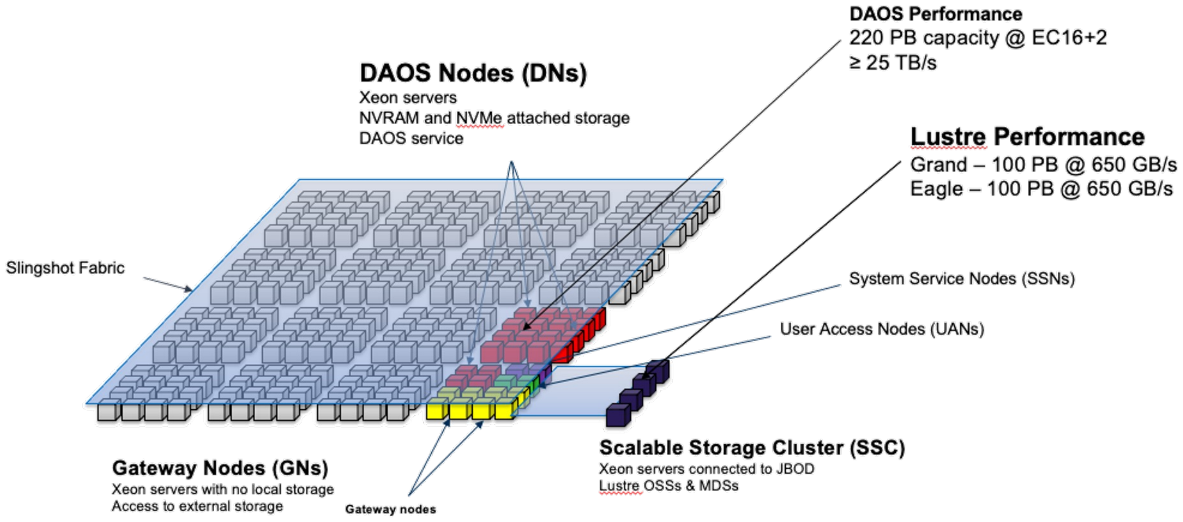


Bandwidth: **≤92 GiB/s** per server

DAOS on Aurora

- 1024x DAOS nodes, each with:
  - 2x Xeon 5320 CPUs
  - 512GB DRAM
  - 8TB Optane Persistent Memory 200
  - 244TB NVMe SSDs
  - 2x HPE Slingshot NIC
- Usable capacity
  - between 220PB and 249PB

    depending on redundancy level chosen



DAOS Performance
220 PB capacity @ EC16+2
≥ 25 TB/s

DAOS Nodes (DNs)
Xeon servers
NVRAM and NVMe attached storage
DAOS service

Lustre Performance
Grand – 100 PB @ 650 GB/s
Eagle – 100 PB @ 650 GB/s

Slingshot Fabric

System Service Nodes (SSNs)

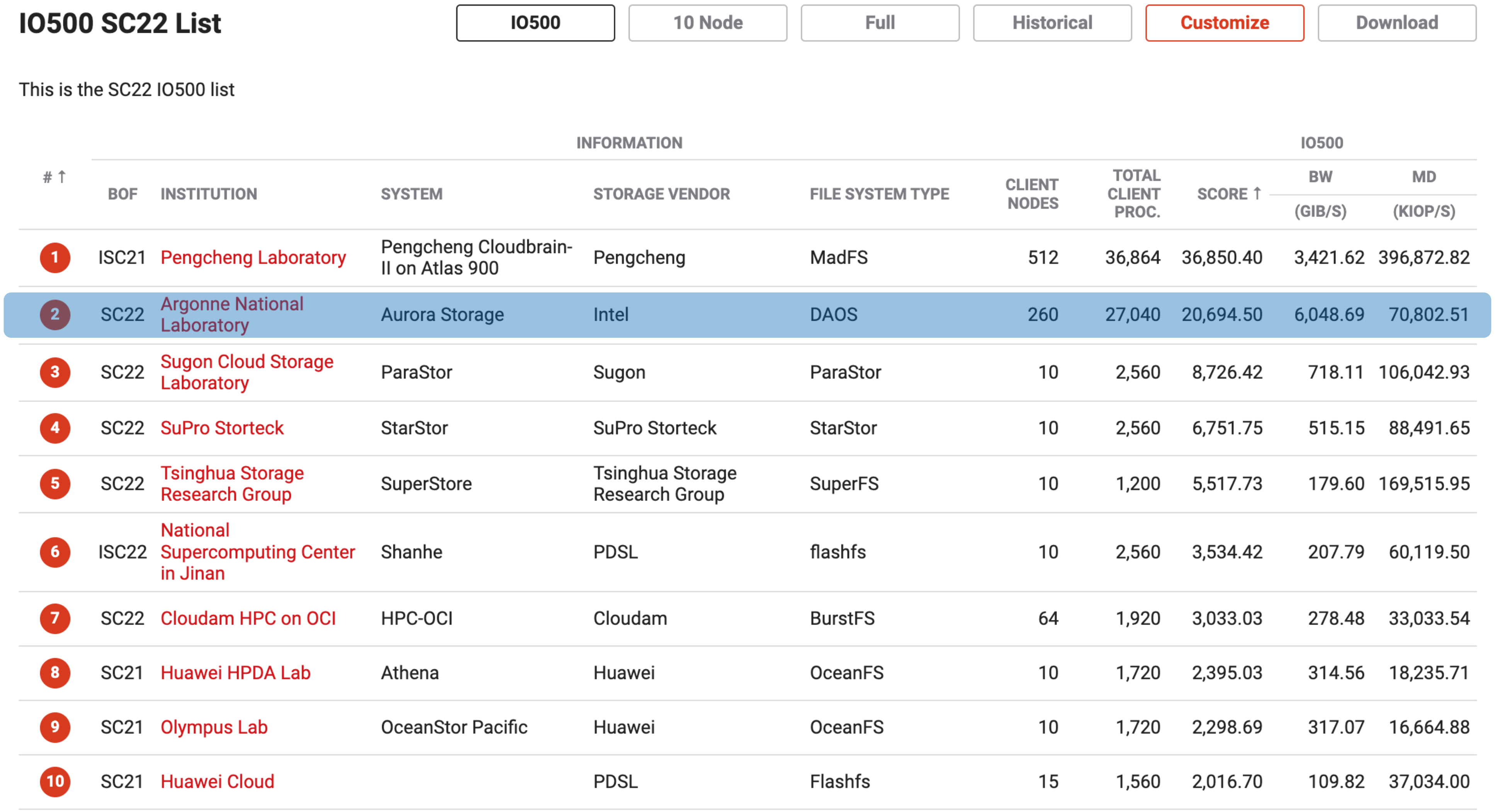
User Access Nodes (UANs)

Gateway Nodes (GNs)
Xeon servers with no local storage
Access to external storage

Gateway nodes

Scalable Storage Cluster (SSC)
Xeon servers connected to JBOD
Lustre OSSs & MDSs

IO500 SC22 Results



IO500 SC22 List

This is the SC22 IO500 list

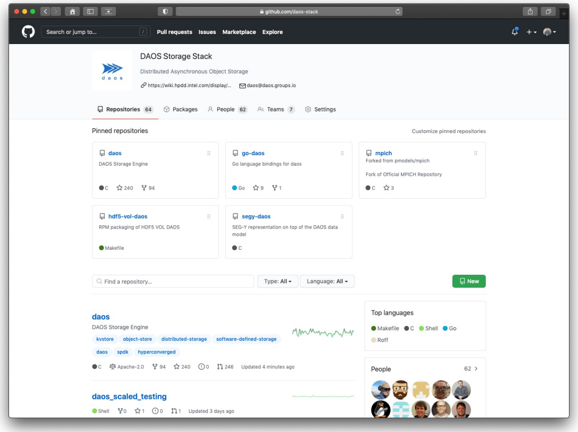
	INFORMATION								IO500	
# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	512	36,864	36,850.40	3,421.62	396,872.82
2	SC22	Argonne National Laboratory	Aurora Storage	Intel	DAOS	260	27,040	20,694.50	6,048.69	70,802.51
3	SC22	Sugon Cloud Storage Laboratory	ParaStor	Sugon	ParaStor	10	2,560	8,726.42	718.11	106,042.93
4	SC22	SuPro Storteck	StarStor	SuPro Storteck	StarStor	10	2,560	6,751.75	515.15	88,491.65
5	SC22	Tsinghua Storage Research Group	SuperStore	Tsinghua Storage Research Group	SuperFS	10	1,200	5,517.73	179.60	169,515.95
6	ISC22	National Supercomputing Center in Jinan	Shanhe	PDSL	flashfs	10	2,560	3,534.42	207.79	60,119.50
7	SC22	Cloudam HPC on OCI	HPC-OCI	Cloudam	BurstFS	64	1,920	3,033.03	278.48	33,033.54
8	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720	2,395.03	314.56	18,235.71
9	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720	2,298.69	317.07	16,664.88
10	SC21	Huawei Cloud		PDSL	Flashfs	15	1,560	2,016.70	109.82	37,034.00

DAOS Ecosystem

Hardware Partners	Hewlett Packard Enterprise Lenovo intel SUPERMICRO inspur Atos Orchestrating a brighter world NEC QCT
Reseller Partners	COLFAX Customized Solutions MEGWARE RSC E4 COMPUTER ENGINEERING 2crsi ScaleWorX
Software Development and 3rd party support	croit SEAGATE Brightskies Google Cloud Nettrix宁畅 H3C 数字化解决方案领导者
End Customers	JINR Argonne NATIONAL LABORATORY RUIJIN HOSPITAL SHANGHAI JIAOTONG UNIVERSITY SCHOOL OF MEDICINE 瑞金 UNIVERSITY OF CAMBRIDGE lrz

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 available at http://dug.daos.io
- Upcoming BoFs at SC22
- 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

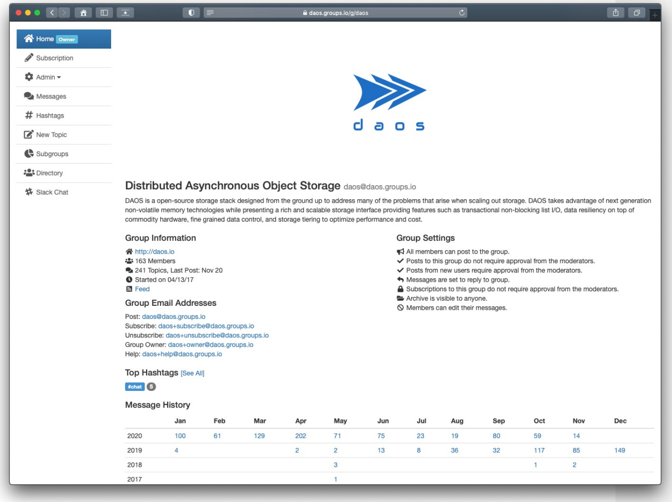
segy-daos — SEG-Y representation on top of the DAOS data model — ● C

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 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.

Group Email Addresses

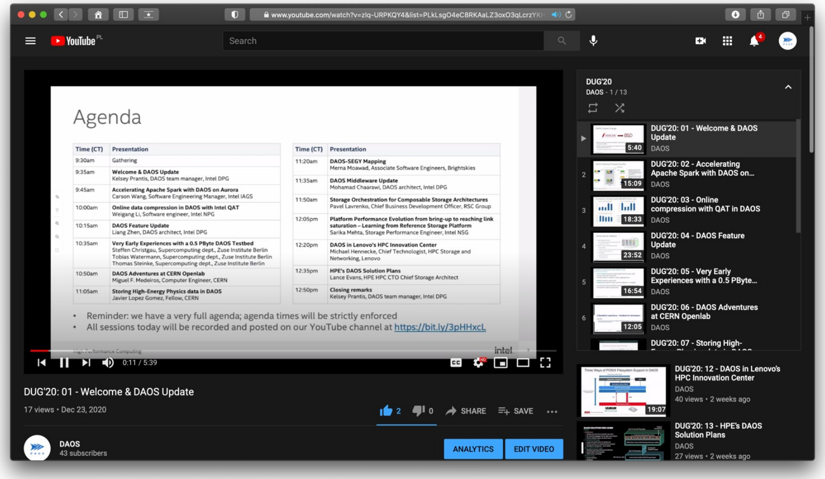
Post: daos@daos.groups.io
Subscribe: daos+subscribe@daos.groups.io
Unsubscribe: daos+unsubscribe@daos.groups.io
Group Owner: daos+owner@daos.groups.io
Help: daos+help@daos.groups.io

Top Hashtags [See All]

#chat 5

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
2
0
SHARE
SAVE
ANALYTICS
EDIT VIDEO
intel.

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

DUG'20: 01 - Welcome & DAOS Update

17 views • Dec 23, 2020

DAOS
43 subscribers

DUG'20
DAOS - 1 / 13

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
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