CERN Openlab Technical Workshop – March 2021

DAOS: Nextgen Storage Stack for AI, Big Data & Exascale HPC

Johann Lombardi, Senior Principal Engineer, CESG, Intel

intel®

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 .

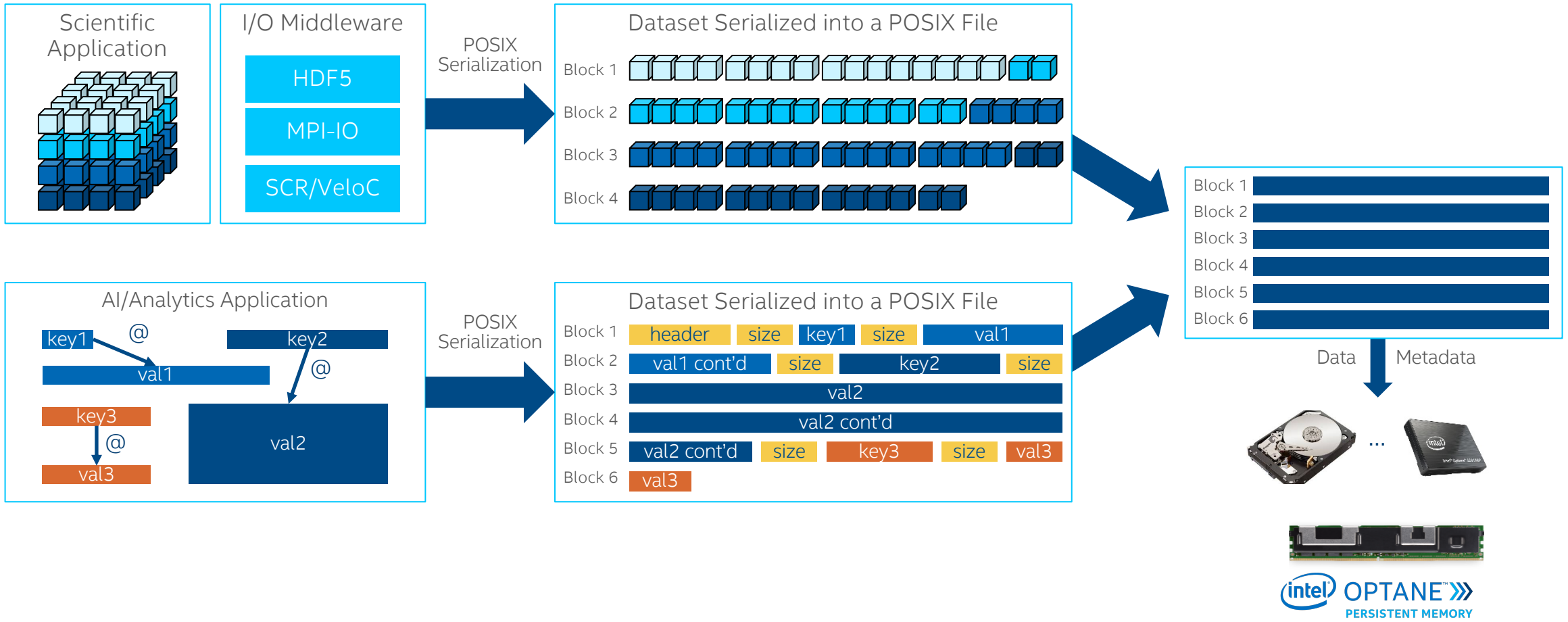
Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 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.

The Problem with POSIX & Blocks (& Objects)

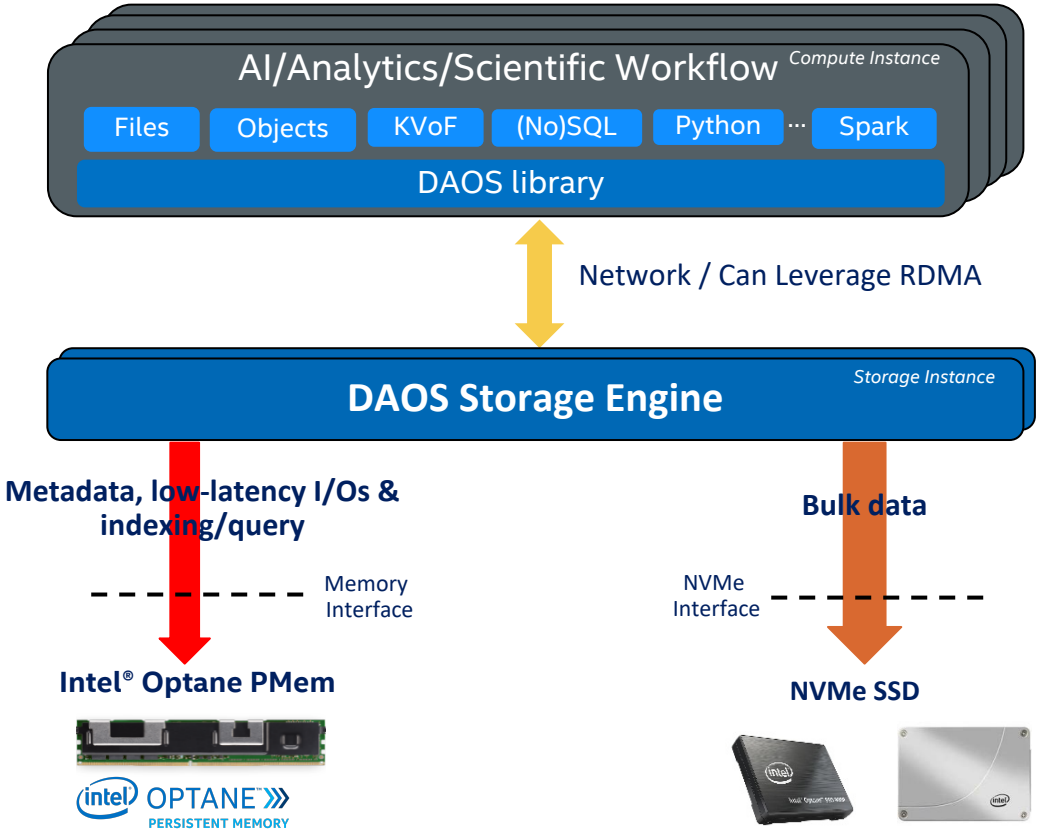


Nextgen Storage Stack Requirements

- Rich data semantics to support emerging use cases
  - Native support for structured, semi-structured & unstructured data models
  - Built-in producer/consumer workflow pipeline support
- Offload/storage acceleration capability
- Provide smooth migration path
- Maximize performance/utilization of hardware
- Elasticity & built-in storage management
- Multi-tenancy features
- Can run in cloud or on-premise
- Highly scalable over COTS Hardware

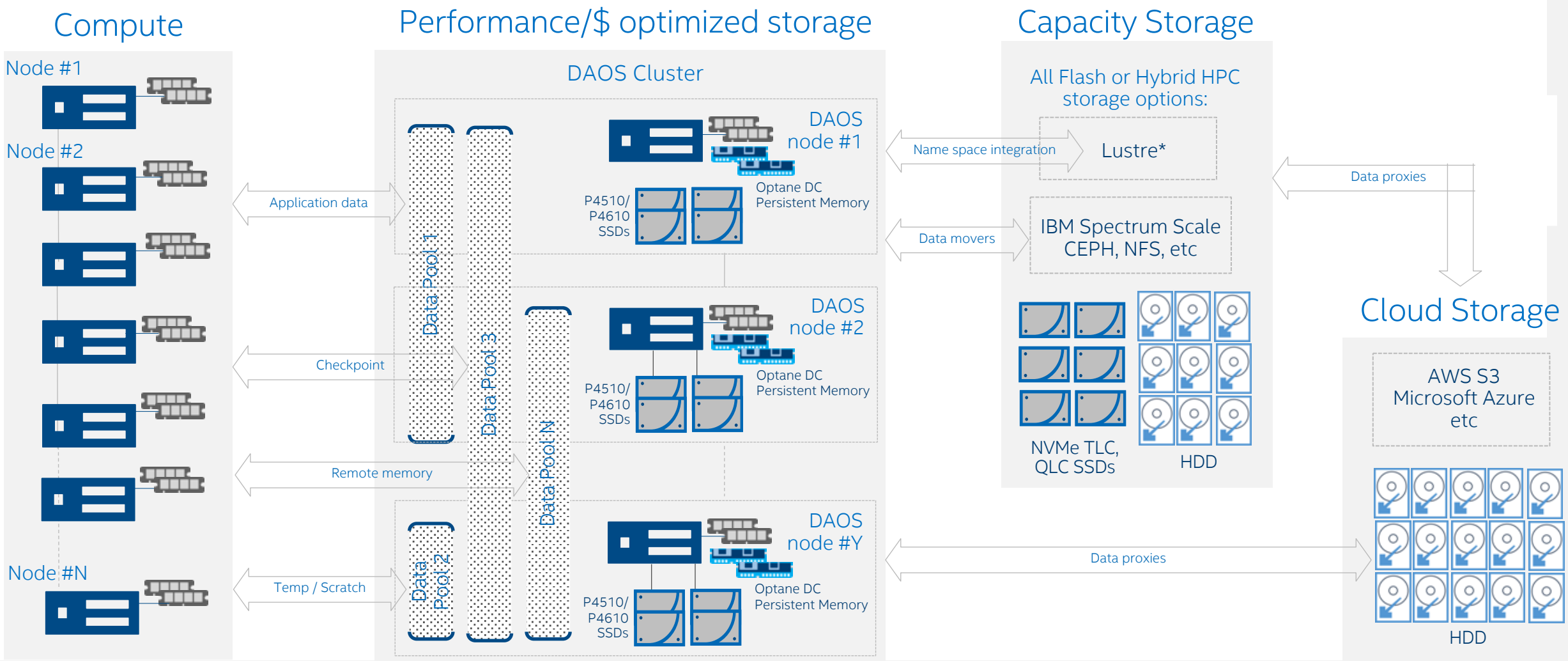
What is DAOS?

- A new, innovative scale-out storage-as-a-service stack based on Intel Optane Persistent Memory and NVMe SSDs
- Globally accessible from many nodes
- Delivers exceptionally high bandwidth and IOPS on commodity servers
- Can be utilized either as a standalone file system, or as a performance tier integrated with existing storage systems

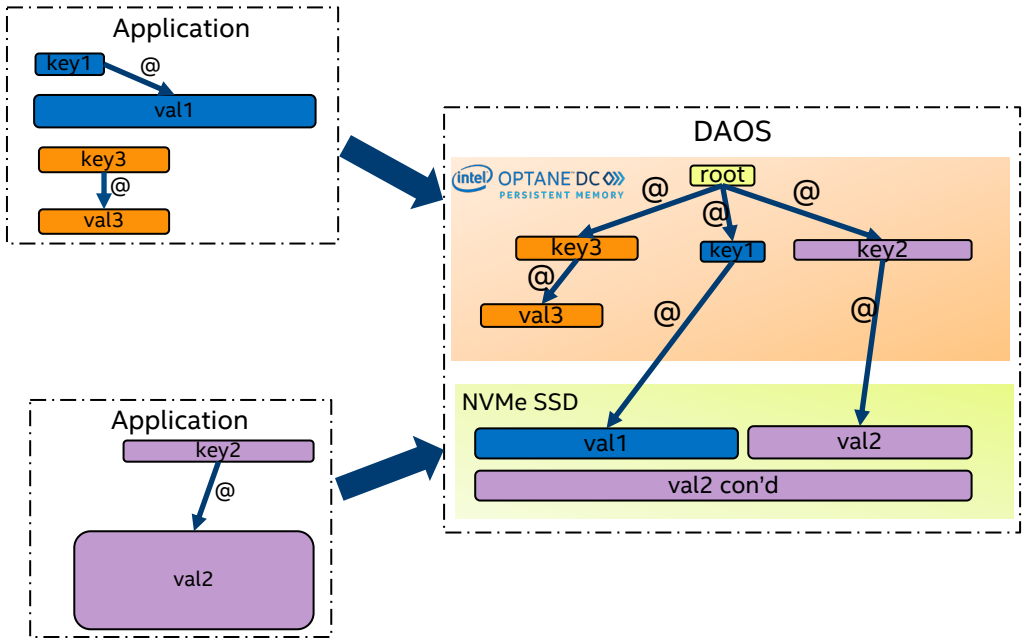


More IOPs and bandwidth per dollar

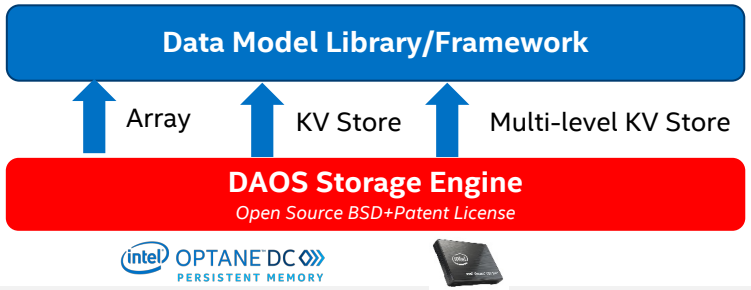
DAOS in the Overall Cluster Architecture



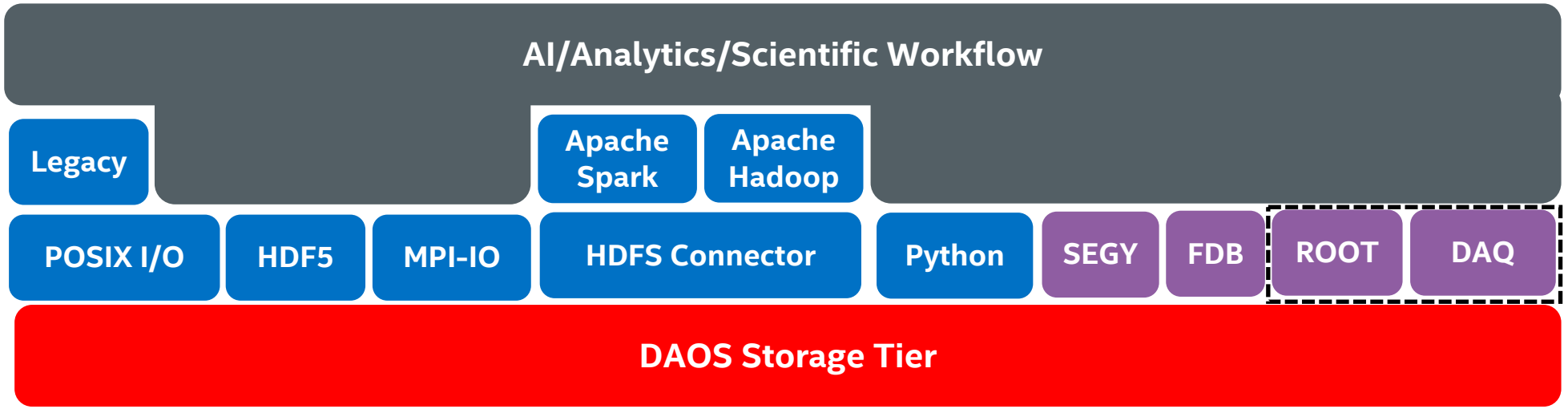
DAOS Data Model



- Native support for structured, semi-structured & unstructured data models
  - Built on top of DCPMM
  - Unconstrained by POSIX serialization
  - Custom attributes
  - Data access time orders of magnitude faster (µs)
  - Scalable concurrent updates & high IOPS
  - Non-blocking
  - Enable in-storage computing



DAOS Software Ecosystem

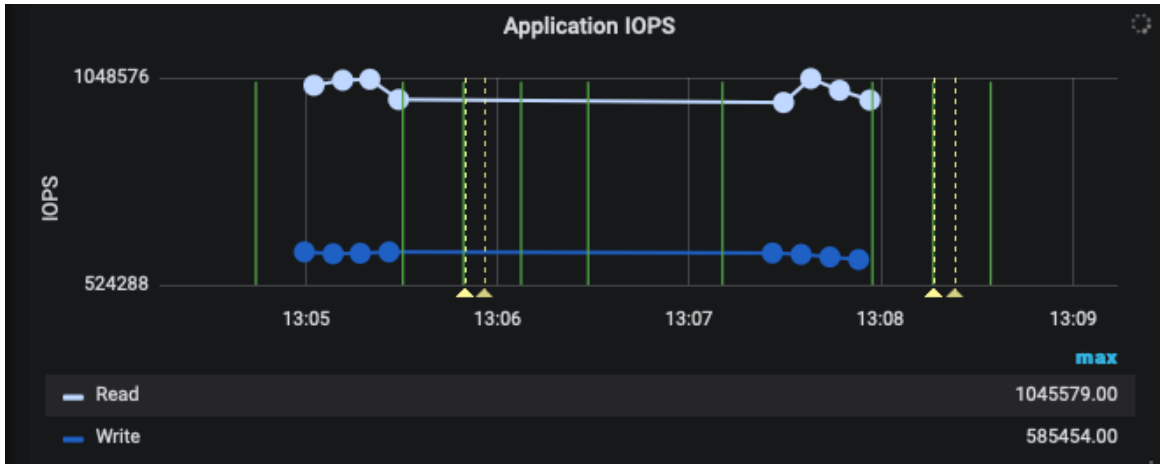


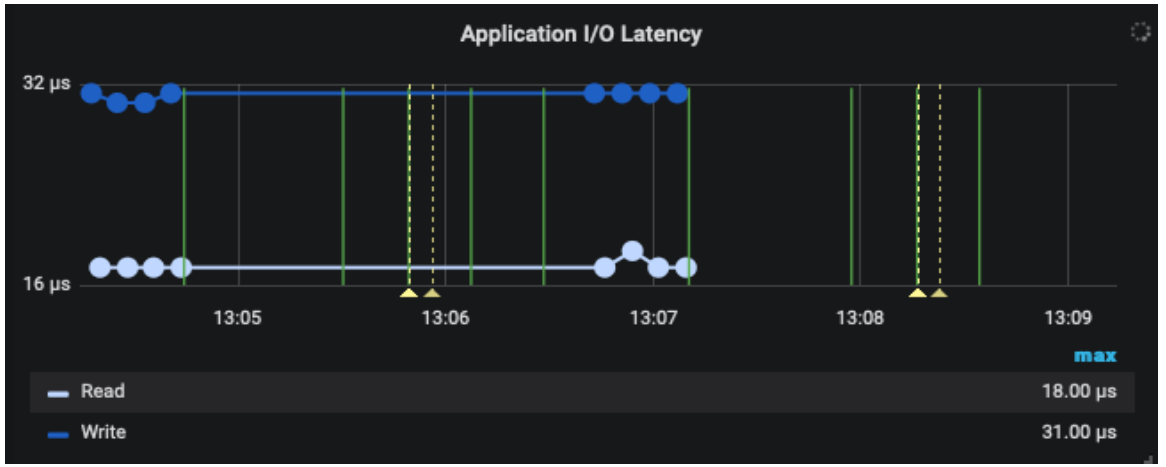
Generic I/O middleware supported today

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

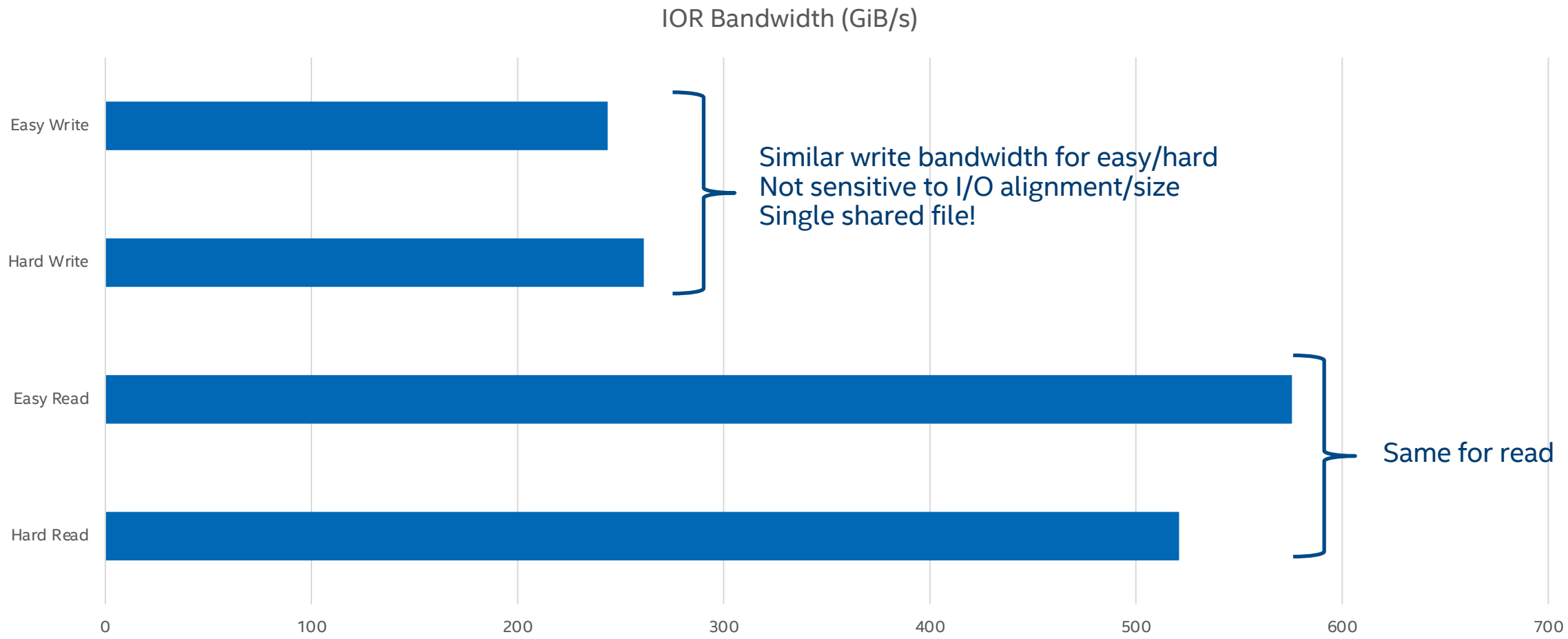
DAOS Performance (ISC'19)

- Deliver HW performance
  - Saturate SSD bandwidth
  - Latency/IOPS of persistent memory for metadata & small I/Os
  - Only need a few clients to reach max BW
    - One task enough to reach 10GB+/s
- Latency/IOPS demonstrated at ISC
  - ½U server
  - See:
    - https://www.youtube.com/watch?v=EMGBcvnftwQ
    - https://www.youtube.com/watch?v=e69Rgz2FMbE



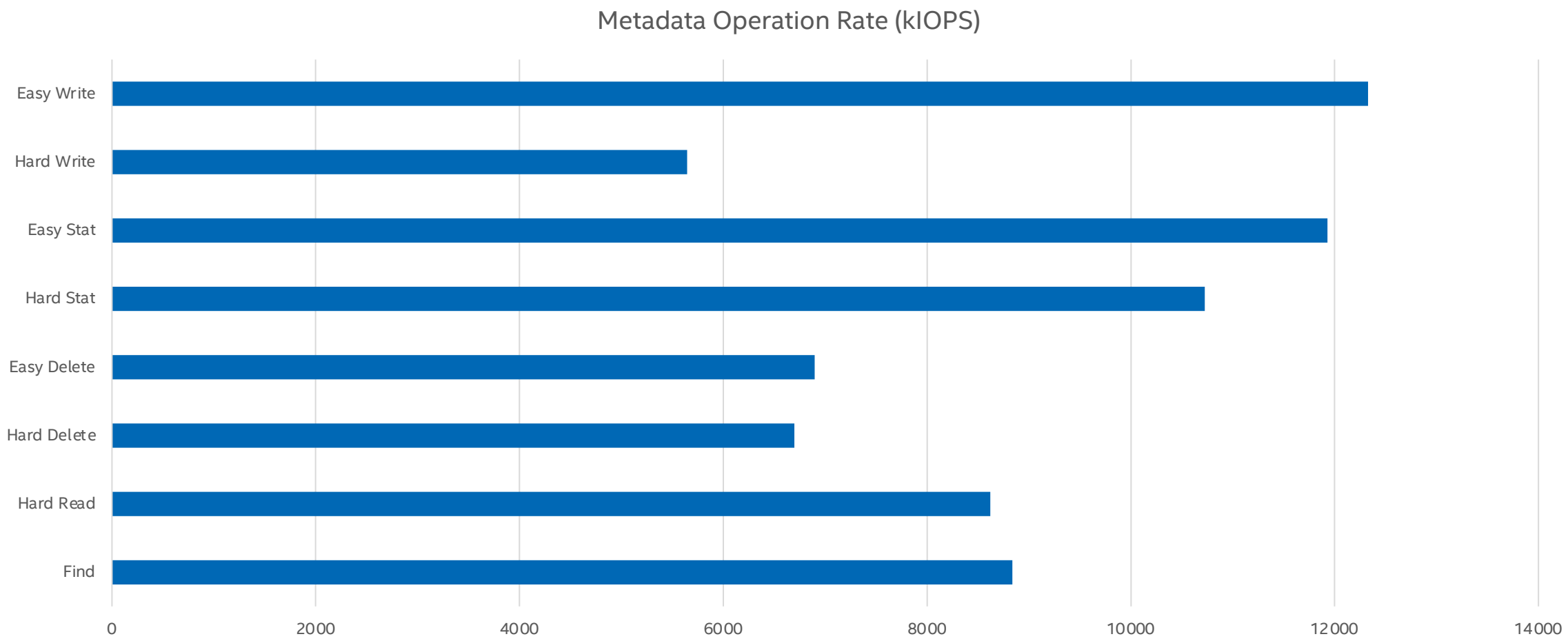


DAOS Bandwidth on IO500



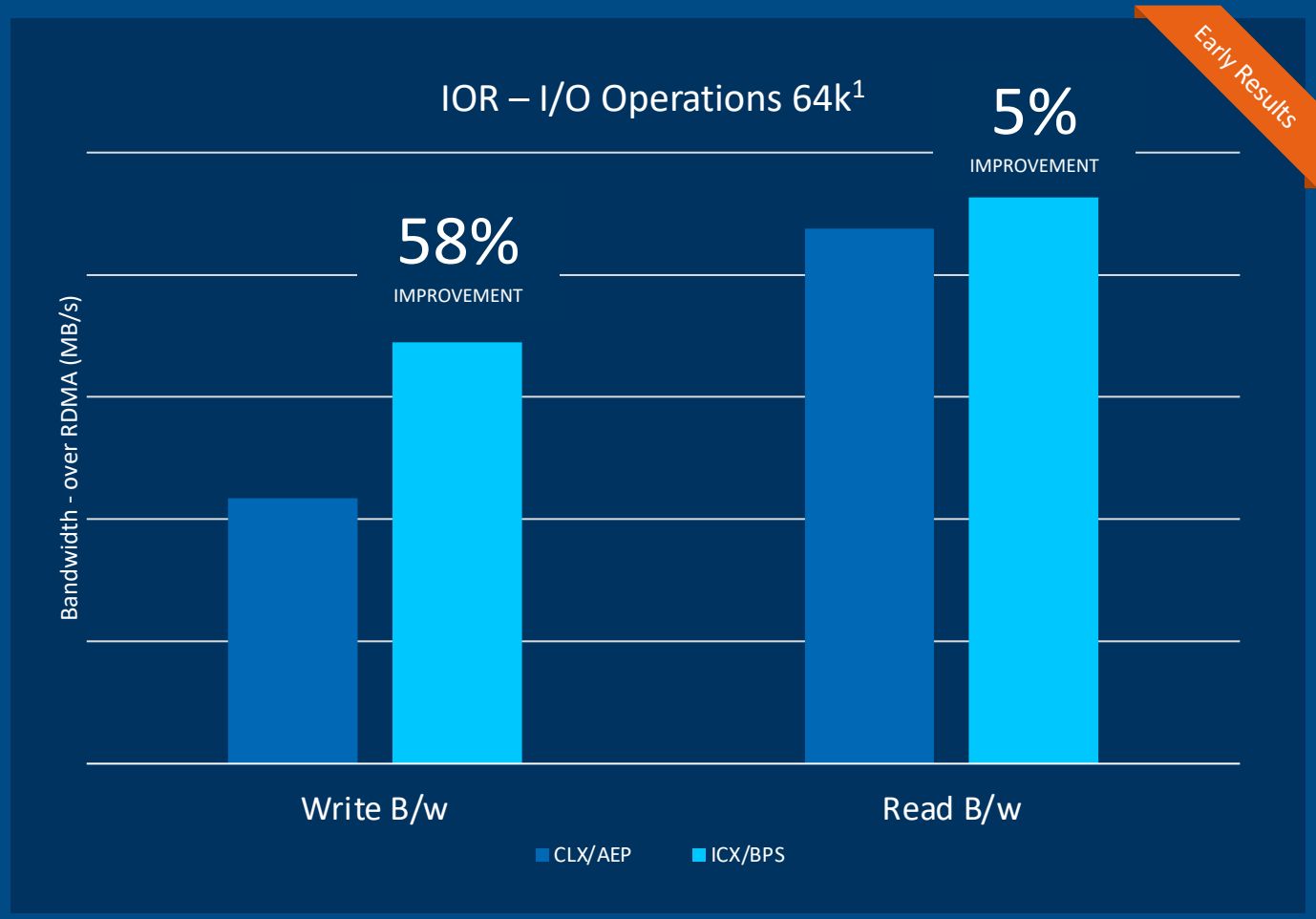
Source: https://io500.org/site/submissions/view/1

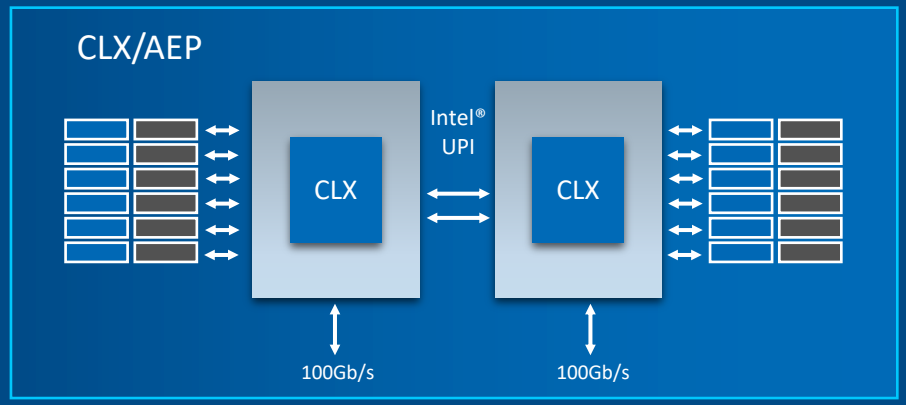
DAOS Metadata Performance on IO500

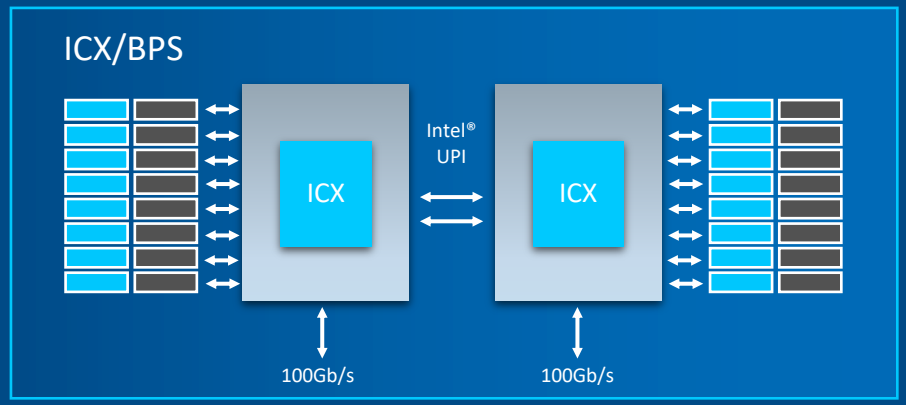


Source: https://io500.org/site/submissions/view/1

First Look: DAOS with ICX/BPS IOR Early Results



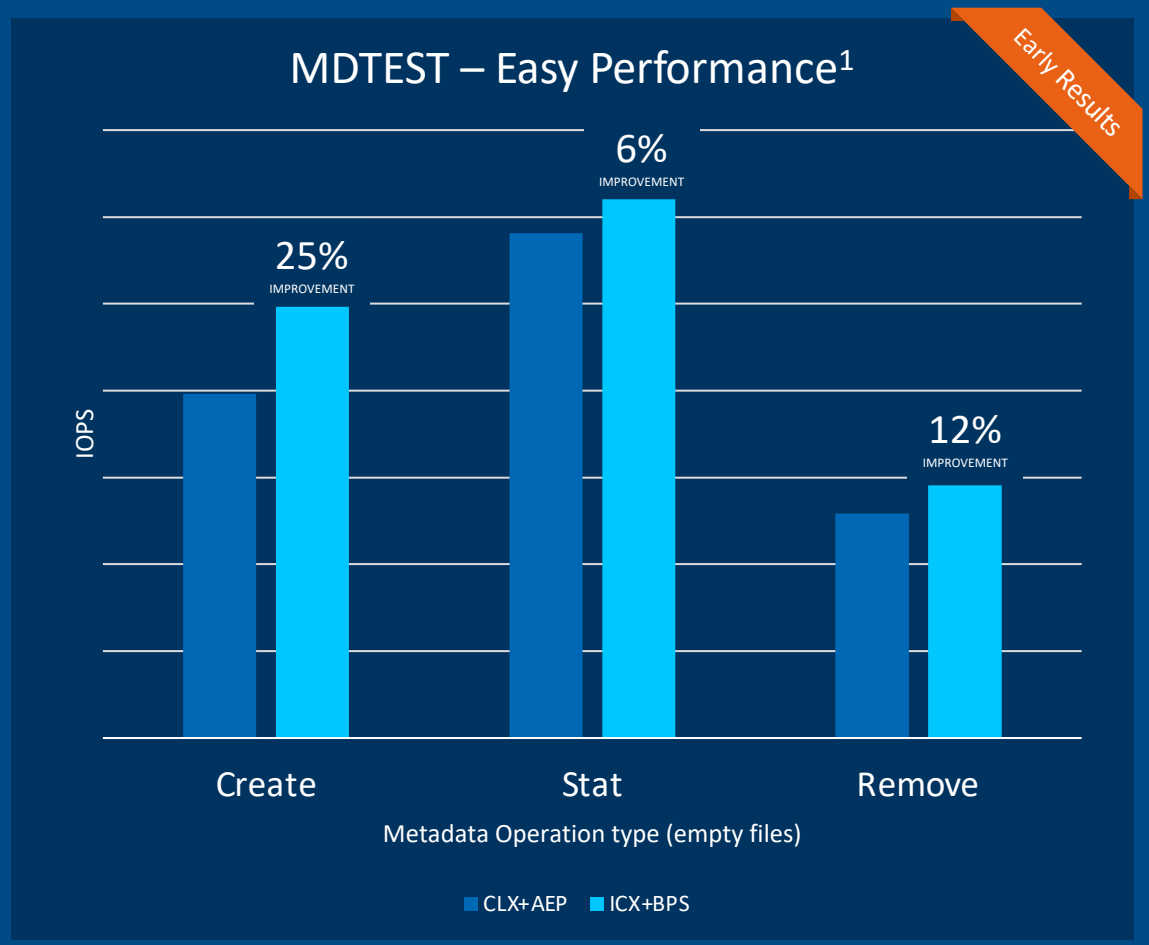


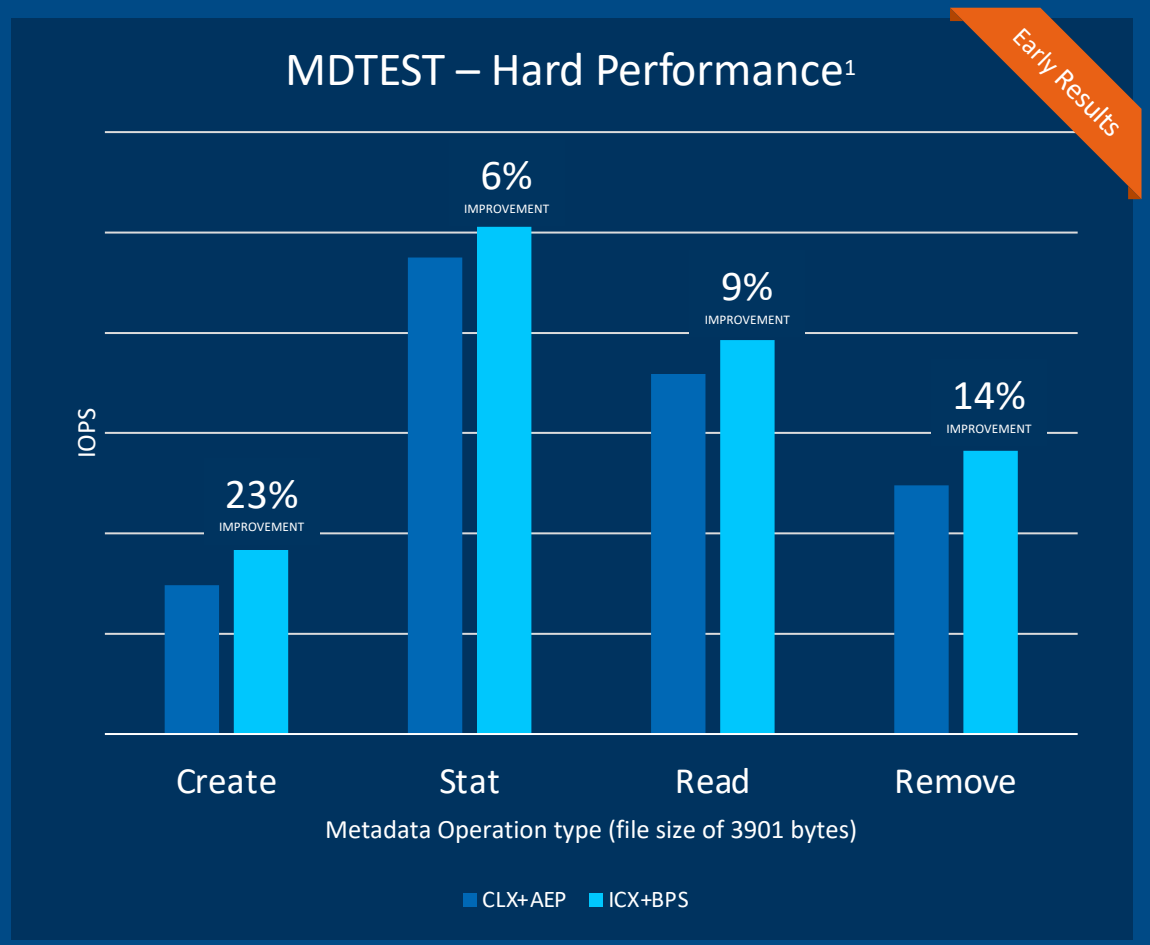


CLX = 2nd Gen Intel® Xeon® Scalable Processor (Codename: Cascade Lake) + AEP = Intel® Optane™ persistent Memory 100 Series (Codename: Apache Pass)
ICX = 3rd Gen Intel® Xeon® Scalable Processor (Codename: Ice Lake) + BPS = Intel® Optane™ persistent Memory 200 Series (Codename: Barlow Bass)

[1] Results have been estimated based on pre-production tests as of 10/15/2020. Performance varies by use, configuration and other factors, for details, see Slide 19. Learn more at www.Intel.com/PerformanceIndex.

First Look: DAOS with ICX/BPS Metadata Early Results





CLX = 2nd Gen Intel® Xeon® Scalable Processor (Codename: Cascade Lake) + AEP = Intel® Optane™ persistent Memory 100 Series (Codename: Apache Pass)
ICX = 3rd Gen Intel® Xeon® Scalable Processor (Codename: Ice Lake) + BPS = Intel® Optane™ persistent Memory 200 Series (Codename: Barlow Bass)

[1] Results have been estimated based on pre-production tests as of 10/15/2020. Performance varies by use, configuration and other factors, for details, see Slide 19. Learn more at www.Intel.com/PerformanceIndex.

DAOS: Primary Storage on Aurora

Aurora DAOS configuration

- Capacity: 230PB
- Bandwidth: >25TB/s

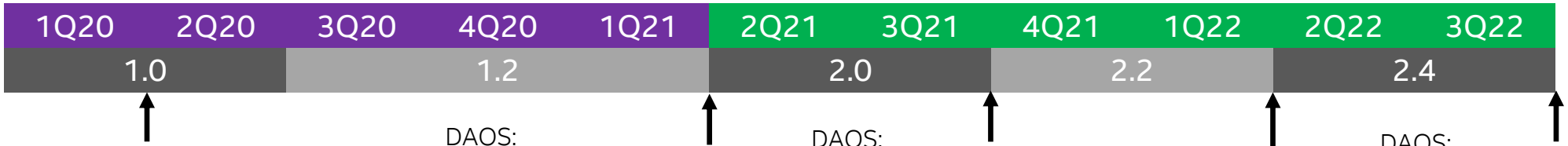
"Combined in Aurora, the Intel compute system, Cray Slingshot network, and the Intel DAOS storage open new possibilities for accelerating the scientific research needed to solve critical human challenges such as cancer and disease. DAOS enables the creation of new storage data models tailored specifically to applications like the Cancer Distributed Learning Environment (CANDLE) which provide a powerful platform to advance a wide array of scientific challenges using deep learning."

– Rick Stevens, Associate Laboratory Director for Computing, Environment and Life Sciences

"The Argonne Leadership Computing Facility is excited to be the first major production deployment of the DAOS storage system as part of Aurora, a US exascale system coming in 2021. As designed, it will provide us unprecedented levels of metadata operation rates and extremely high bandwidth for I/O intensive workloads."

– Susan Coghlan, ALCF-X Project Director/Exascale Computing Systems Deputy Director

DAOS Community Roadmap



Released on June 18

DAOS:
- NVMe & DCPMM support
- Per-pool ACL
- UNS in DAOS via dfuse
- Replication & self-healing (Preview)

Application Interface:
- MPI-IO Driver
- HDF5 DAOS Connector
- Basic POSIX I/O support

DAOS:
- End-to-end data integrity
- Per-container ACL
- Improved control plane
- Replication & self-healing
- Conditional updates
- Erasure Code (Preview)
- Online server addition
- Lustre/UNS integration

Application Interface:
- HDF5 vol w/o async
- POSIX I/O with conditional update support
- POSIX data mover
- Async HDF5 operations over DAOS
- Spark

DAOS:
- Erasure code
- Telemetry& per-job statistics
- Advanced control plane
- Distributed transactions
- Multi OFI provider support

Application Interface:
- POSIX I/O with distributed transaction support
- HDF5 data mover
- Container parking/serialization

DAOS:
- Progressive layout
- Placement optimizations
- Checksum scrubbing

DAOS:
- Catastrophic recovery tools

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

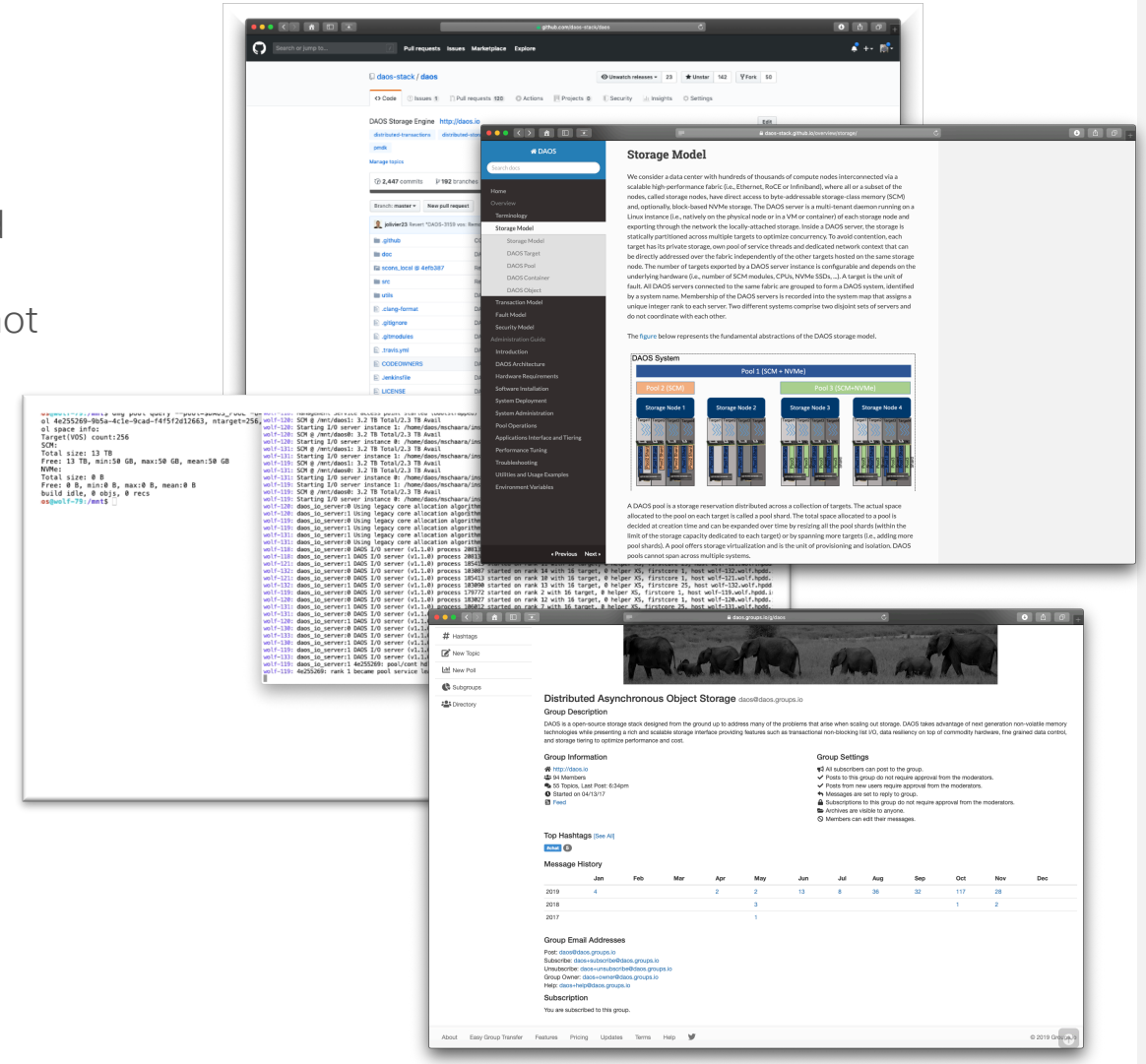
DAOS Technical Training Program

Program goal:

- Provide easy access to learn DAOS software stack.
- Firsthand DAOS experience to assist during common installation and configuration steps.
- Focused to educate Technical Sales (TSS) and customer previously not familiar with DAOS.
- Not a replacement to the documentation.

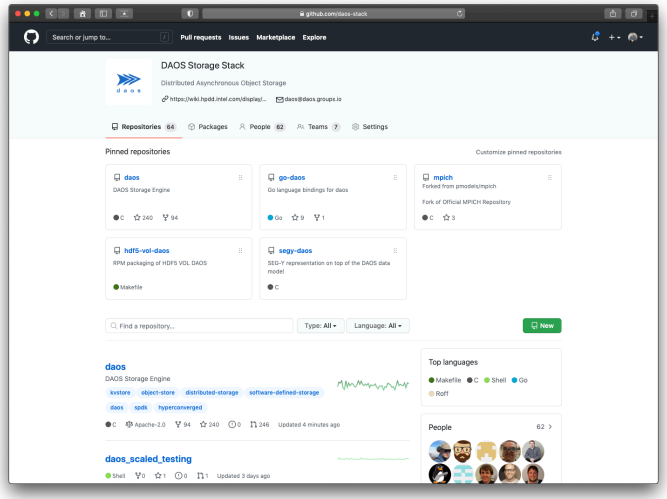
Distribution platform and training format:

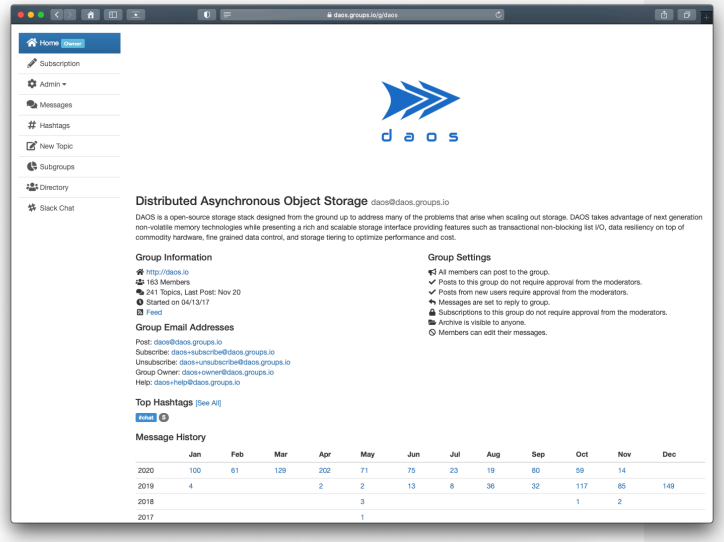
- Short 20-minute video recordings publicly available.
- Hands on demos and screen sharing sessions focused on:
  - Installation and configuration from RPMs
  - DAOS performance sizing for different configuration.
  - Control plane demo, storage configuration
  - Data redundancy and self healing
  - Middleware overview, what interfaces we have
  - POSIX interface
  - HDFS adapter
  - DAOS native API programming

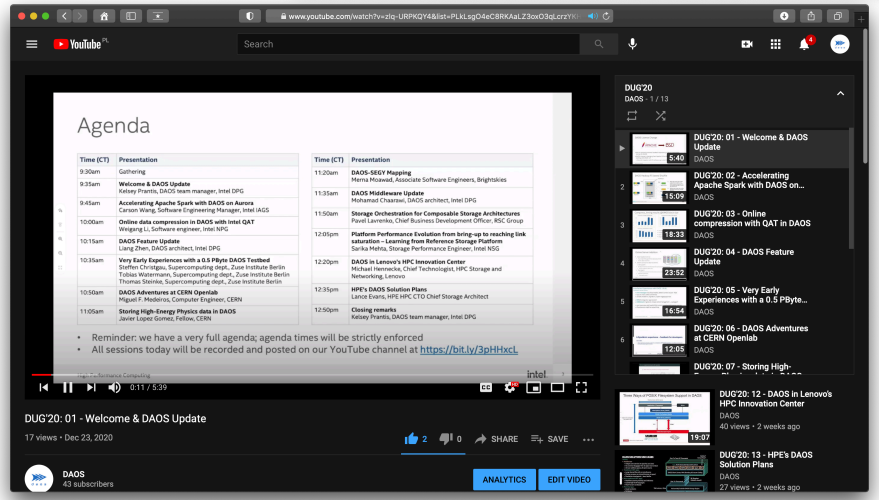


Resources

- Github
  - https://github.com/daos-stack/daos
- DAOS online documentation
  - http://daos.io
- Community mailing list
  - https://daos.groups.io
- DAOS User Group
  - https://wiki.hpdd.intel.com/display/DC/DUG20
  - https://www.youtube.com/playlist?list=PLkLsgO4eC8RKAaLZ3oxO3qLcrzYKHXNDm





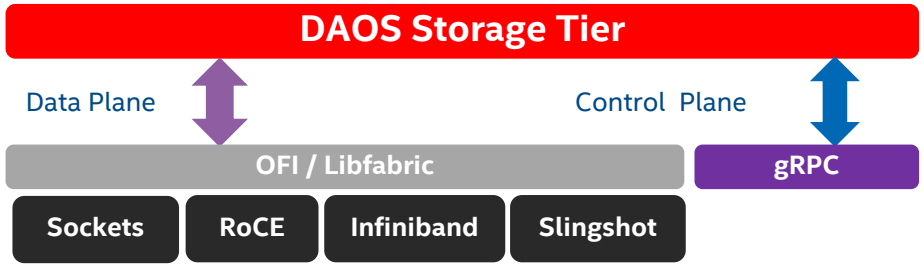



intel

ICX/BPS Comparison System Config

- CLX+AEP (Baseline) Config/Results – Tested by Intel on 10/15/2020. Platform; S2600WF0, 1 Node with 2x8260L Platinum Intel Xeon 2nd Gen Scalable CPUs, microcode 0x400002c, HT & Turbo On, Performance Mode, System BIOS SE5C620.86B.02.01.0008.031920191559, PMem Firmware 01.00.00.5127, System DRAM Config 12 slots / 16 GB / 2666 (192 GB Total Memory), System PMem Config 12 slots / 512 GB / 2666 (6 TB Total PMem), 1xIntel SATA SSD, 2xIntel OPA100 NIC, PCH Intel C621, OS openSUSE Leap 15.2, Kernel 5.3.18-lp152.44-default, Workload DAOS 1.1.0.
- ICX+BPS (New) Config/Results – Tested by Intel on 10/15/2020. Platform; WLYDCRB1, 1 Node with 2xICX-24C Intel Xeon 3rd Gen Scalable CPUs (Ice Lake pre-production), microcode 0x8b000260, HT & Turbo On, Performance Mode, System BIOS WLYDCRB1.SYS.0017.D75.2007020055, PMem Firmware 02.01.00.1110, System DRAM Config 16 slots / 16 GB / 3200 run at 2933 (256 GB Total Memory), System PMem Config 16 slots / 512 GB / 2933 (8 TB Total PMem), 1xIntel SATA SSD, 2xIntel OPA100 NIC, PCH Intel C621, OS openSUSE Leap 15.2, Kernel 5.3.18-lp152.44-default, Workload DAOS 1.1.0.

DAOS Feature Set



- Storage management
  - Integrated control plane
    - Deployment, firmware upgrade, …
    - Monitoring, telemetry & per-job stats
    - RAS events
  - Elastic storage
    - Storage node/SSD drain/reintegration
    - Online server addition
    - Online rebalancing
  - Security
    - Certificates & Access Control List (ACL)
  - DAOS-aware parallel data mover
- Networking
  - Native support for many interconnects
  - Optimized data placement
- Data protection & self-healing
  - Replication & Erasure code
  - End-to-end data integrity
  - Catastrophic recovery
- Advanced storage features
  - Dataset snapshot
  - Distributed serializable transactions