Accelerating Storage with Optane and DAOS

Johann Lombardi, Senior Principal Engineer, Intel
Nicolau Manubens, Analyst, ECMWF

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 .

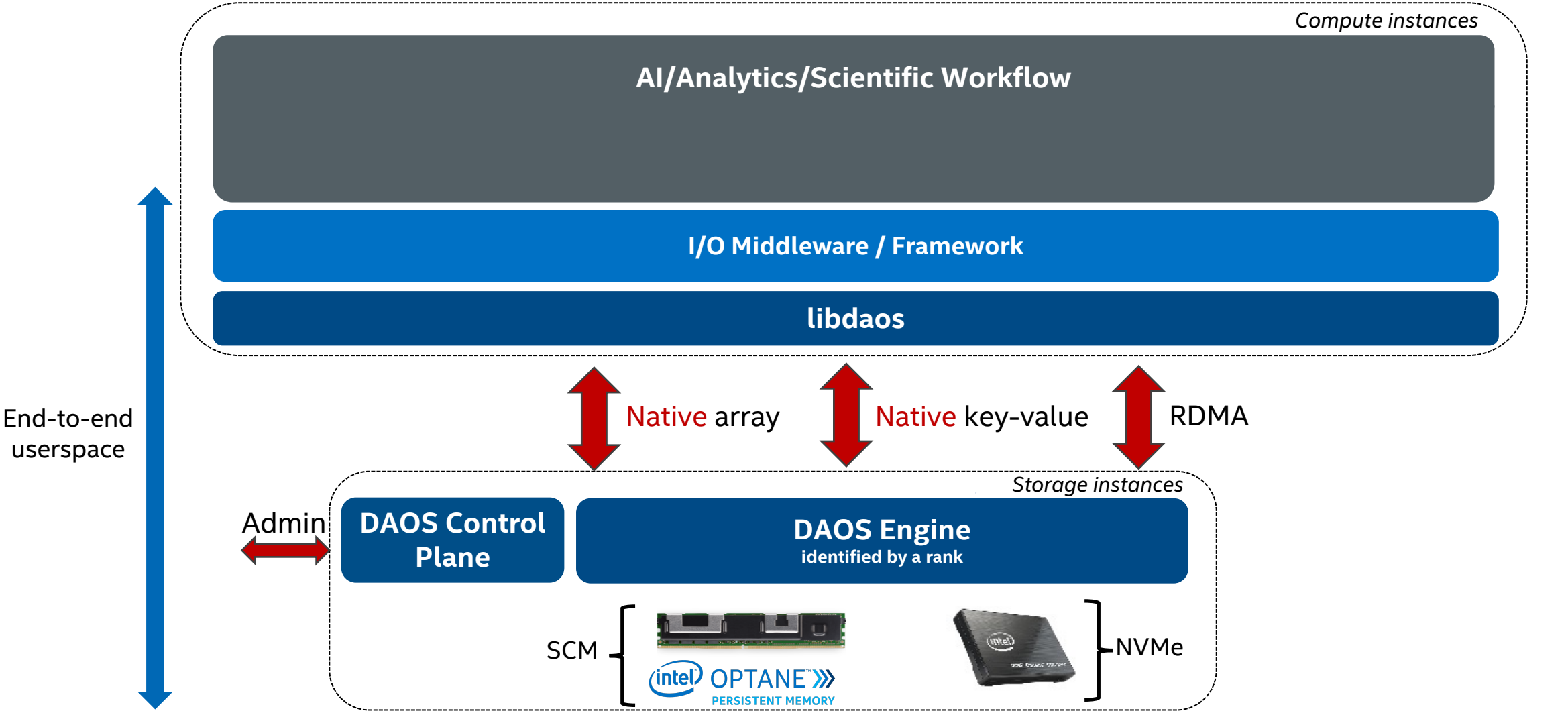
Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology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

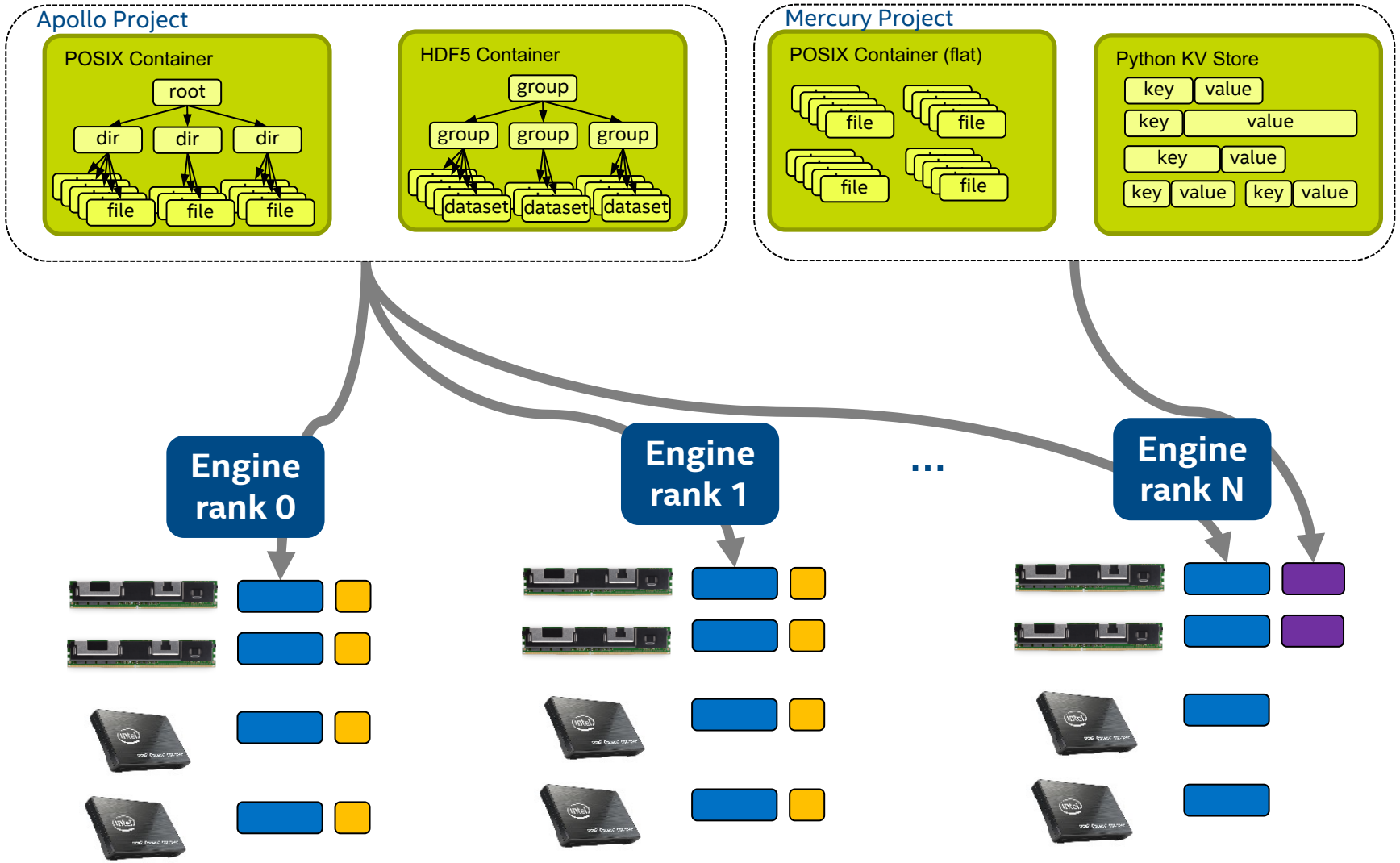
Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

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.

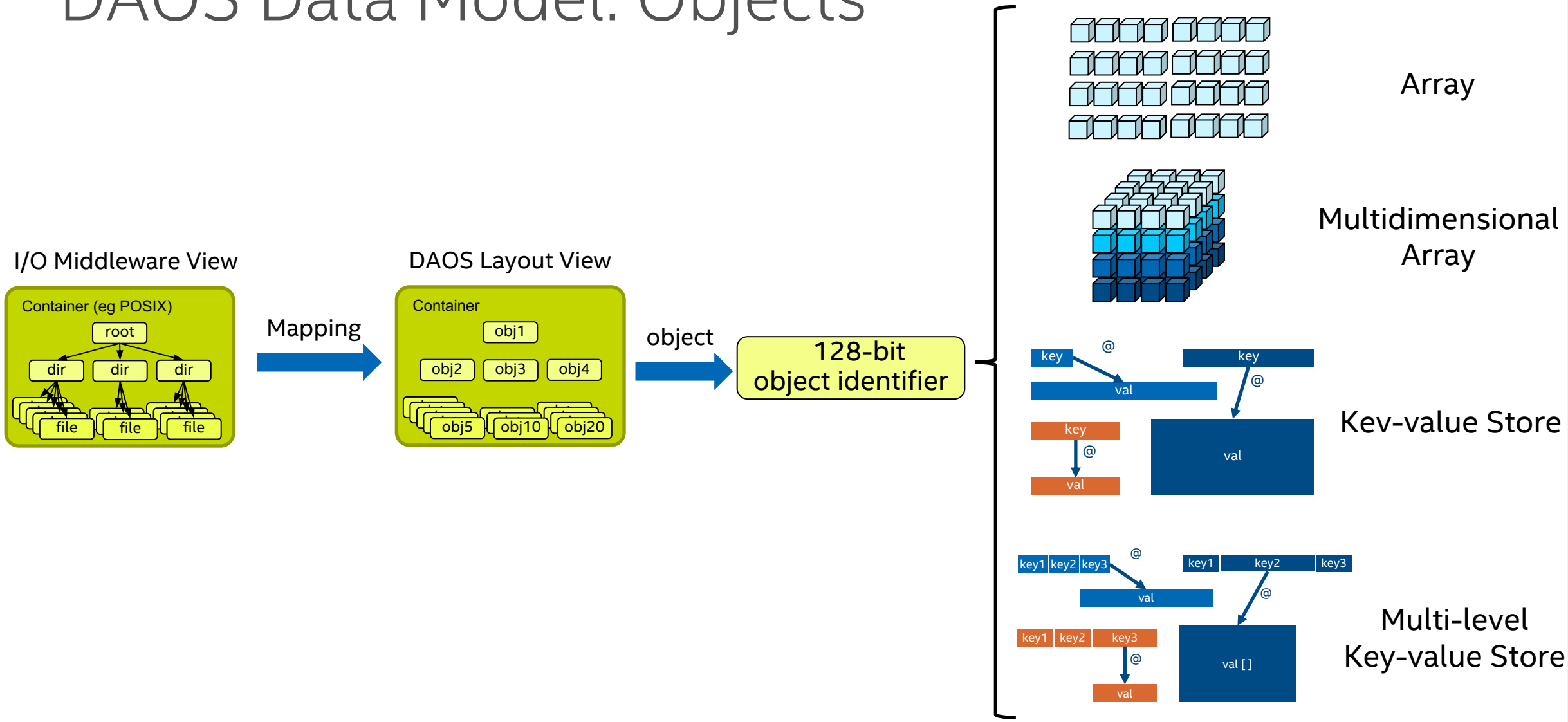
DAOS: Nextgen Storage Stack



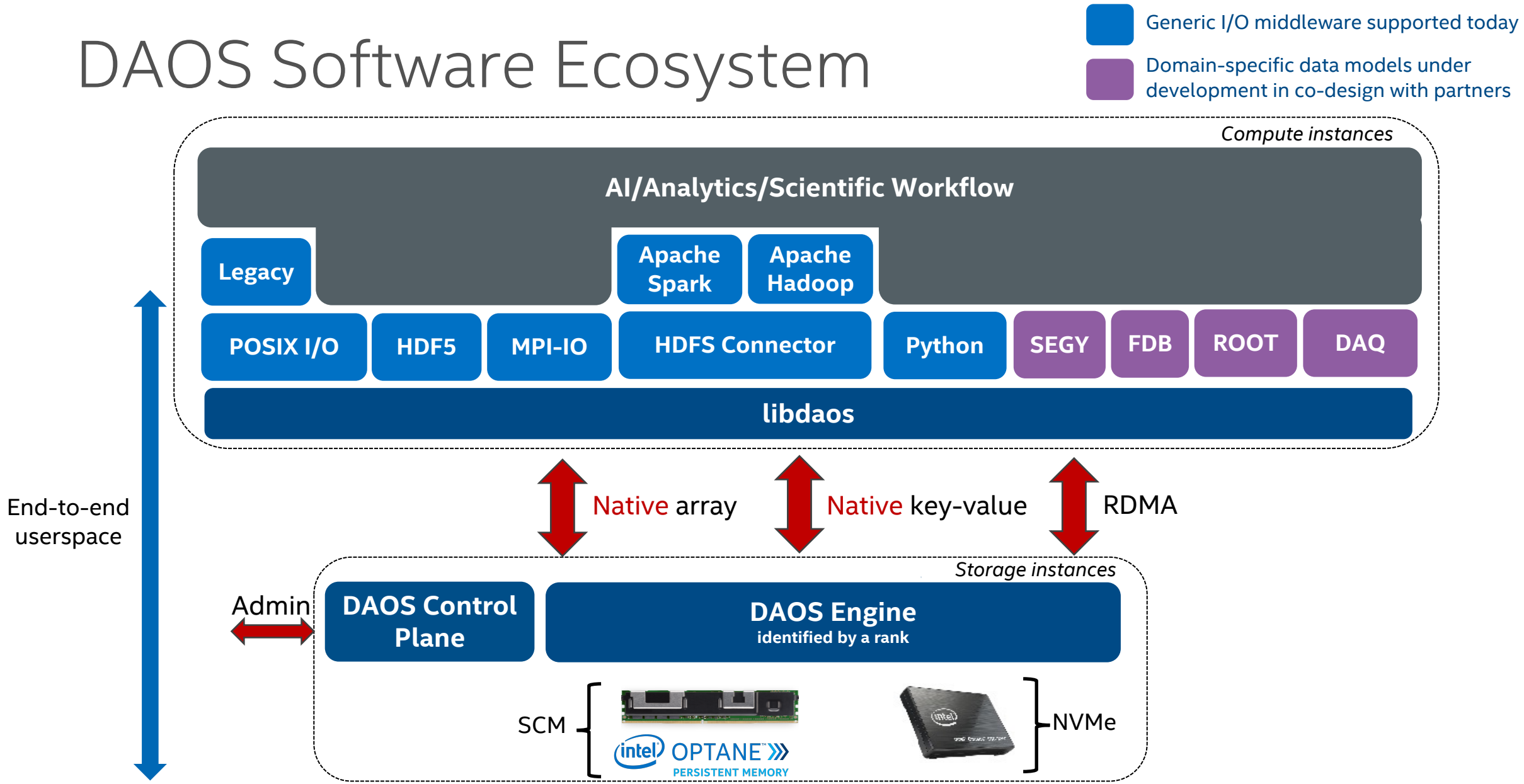
DAOS Data Model: Container



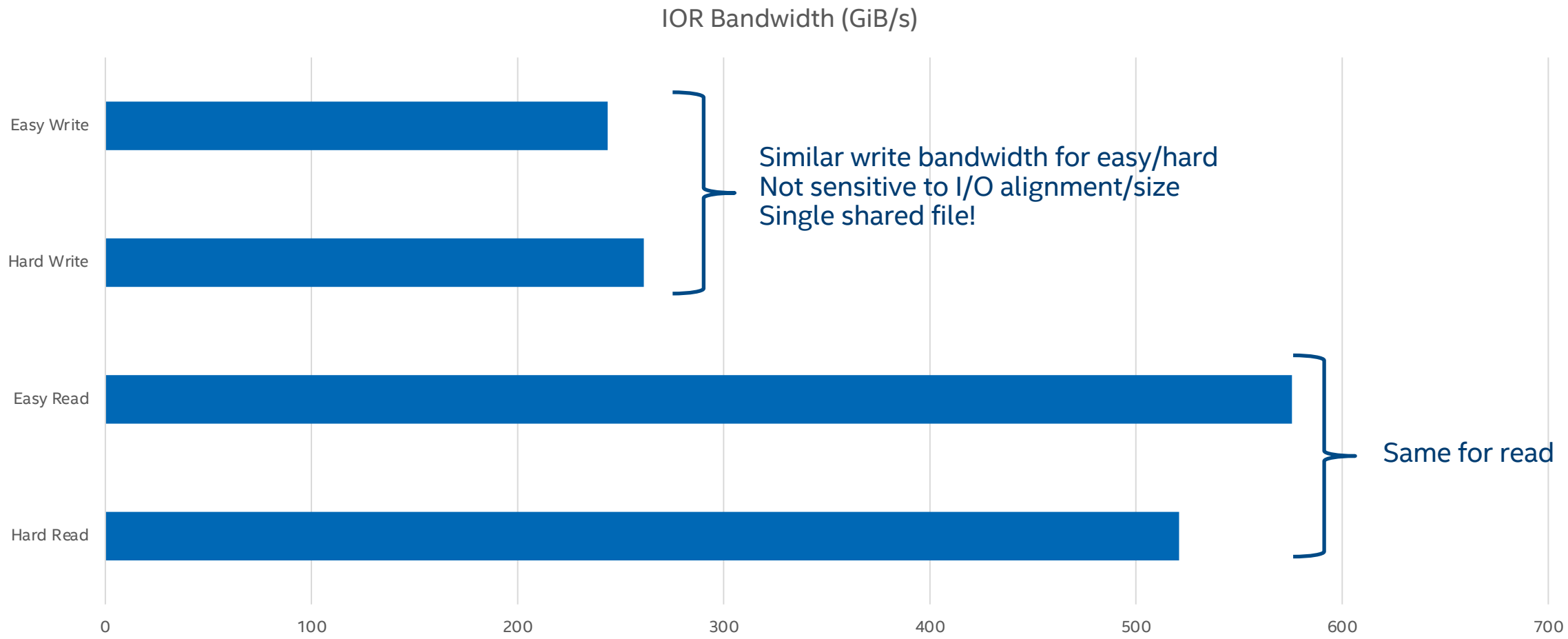
DAOS Data Model: Objects



DAOS Software Ecosystem

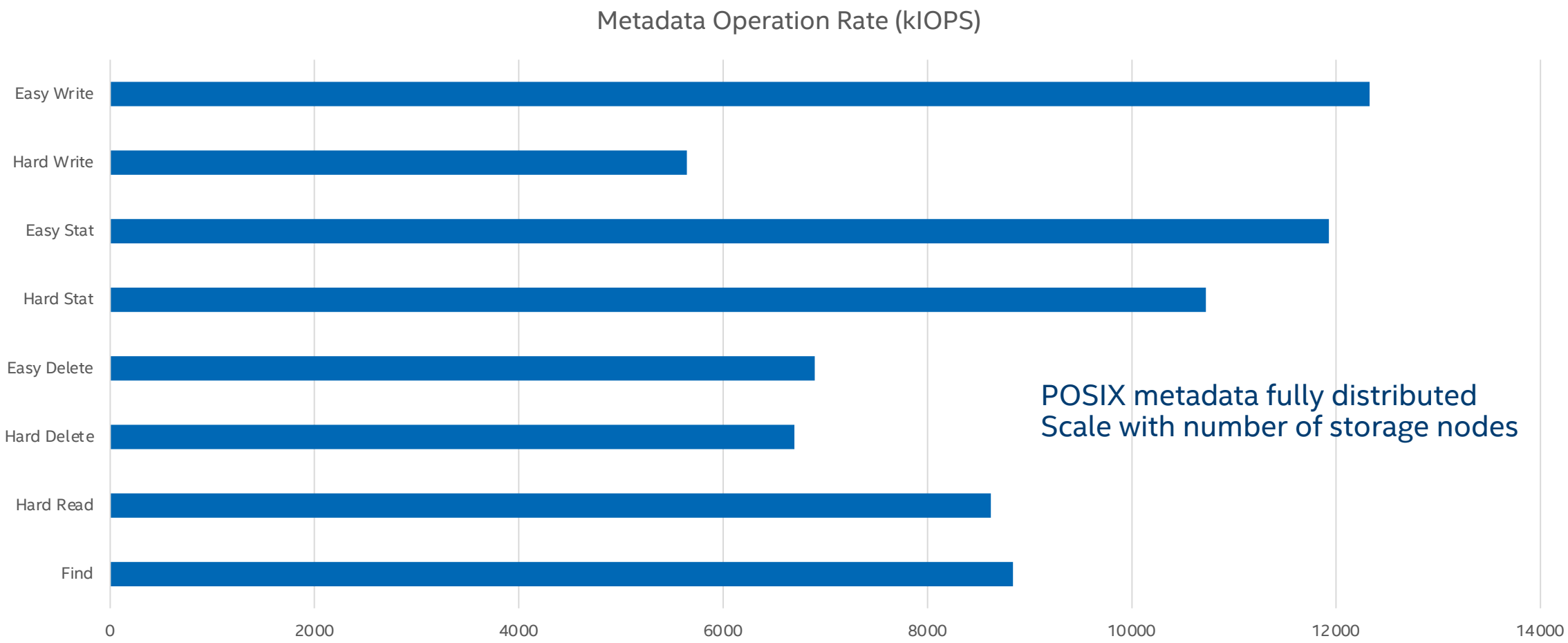


DAOS Bandwidth on IO500



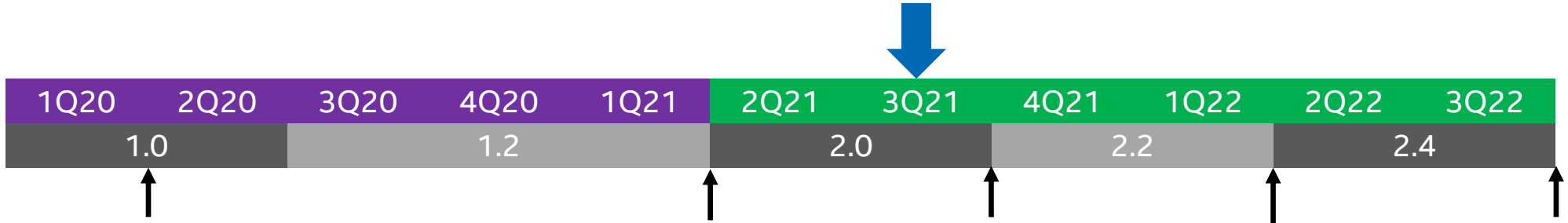
Source: https://io500.org/submissions/view/1

DAOS Metadata Performance on IO500



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

DAOS Community Roadmap



Released on June 18, 2020

DAOS:
- CentOS7 support
- 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

Released on April 30, 2021

DAOS:
- + Leap 15 support
- End-to-end data integrity
- Per-container ACL
- Replication & self-healing
- Quiescent server reintegration
- Erasure Code (Preview)

Application Interface:
- HDF5 DAOS VOL Connector
- POSIX I/O improvements
- POSIX data mover
- Apache Spark support

DAOS:
- + CentOS8 support
- Erasure code
- Telemetry
- Improved usability

DAOS:
- Checksum scrubbing
- Distributed transactions
- Lustre/UNS integration
- Online server addition/reintegration

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

DAOS:
- Catastrophic recovery tools
- Multi OFI provider support
- Progressive layout
- Placement optimizations

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

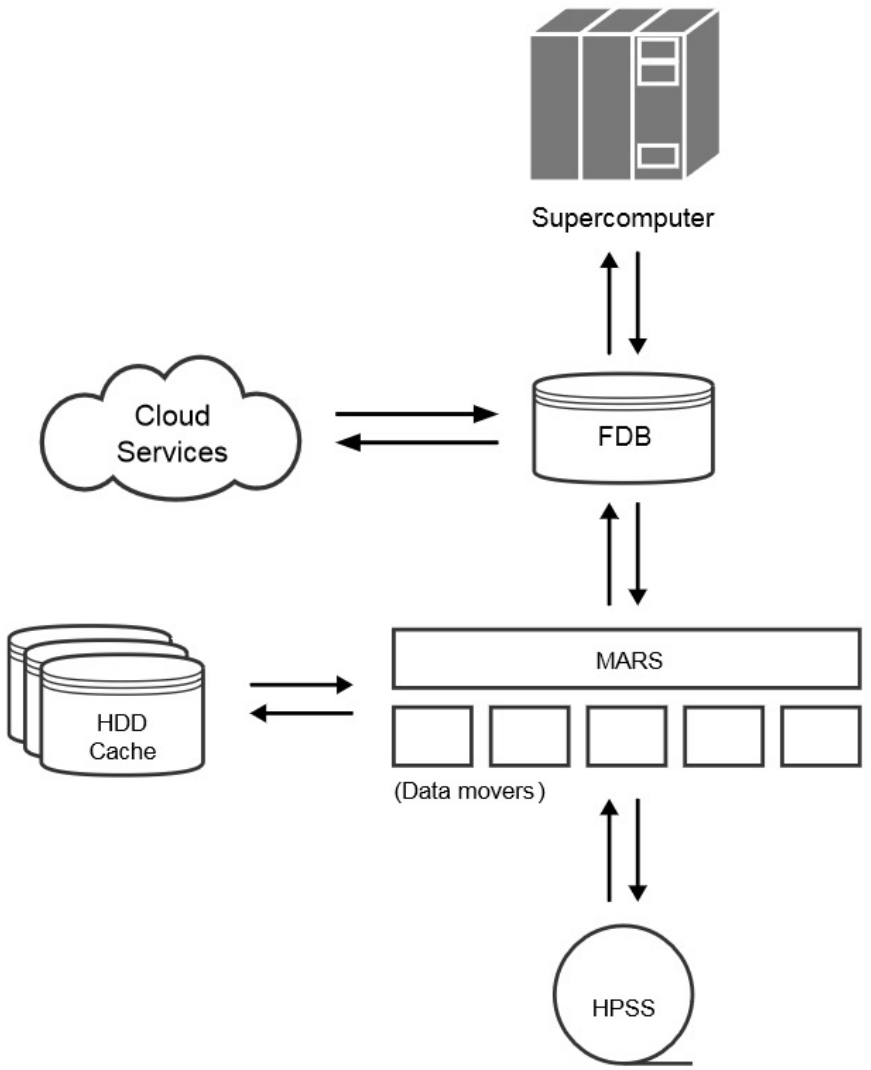
FDB and ECMWF's high-performance data infrastructure

- FDB5 is a software-defined domain-specific object store developed at ECMWF for storing, indexing and retrieving GRIB data

- Currently runs on Lustre at ECMWF

- FDB acts as a "hot" storage layer (RAM + HDD)
  - higher performance than MARS. Lower capacity

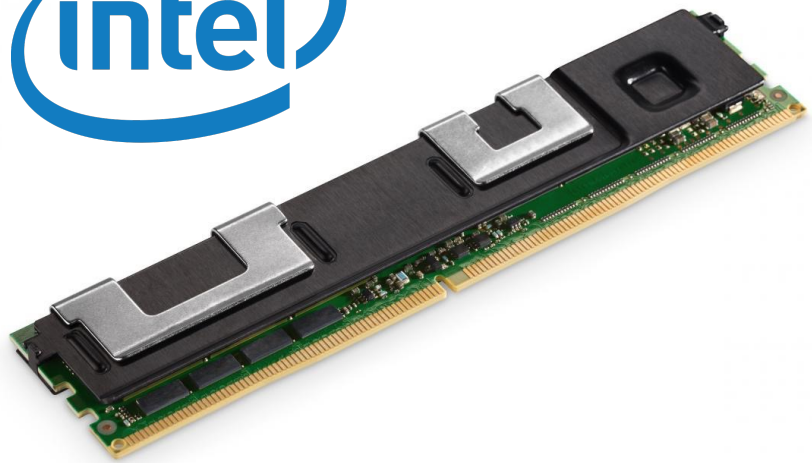
- Effort over the past 5 years in FDB5, particularly:
  - new design to support custom non-POSIX indexing and storage backends
  - e.g. object storages which potentially leverage SCM

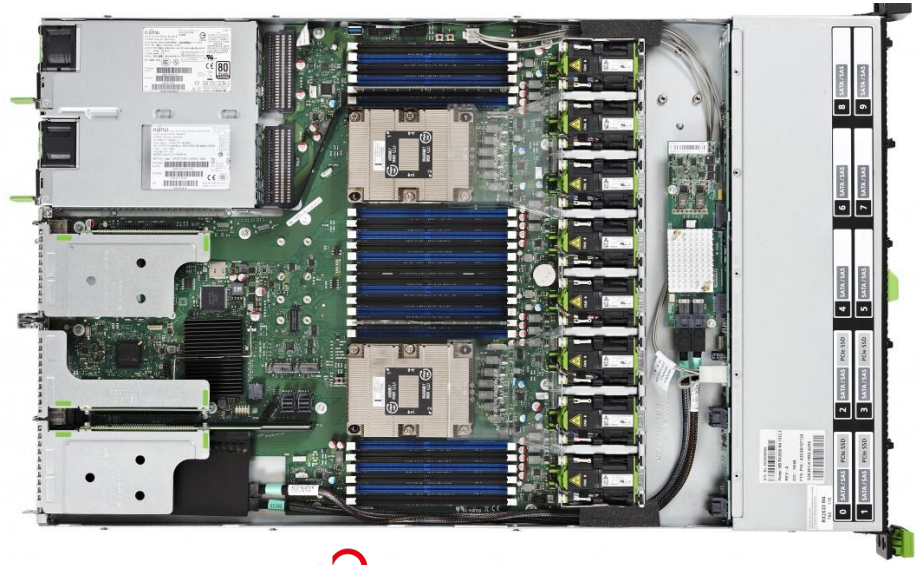


DAOS and NEXTGenIO

- We are assessing DAOS as a backend for FDB

- NEXTGenIO is the platform we test DAOS on
  - Dual-CPU Intel® Xeon® SP nodes (48 cores)
  - OmniPath network
  - 192GB DRAM
  - **3TiB / node of NVRAM DIMMs – Intel Optane DCPMM**
  - **34 compute nodes**
  - Hosted @ EPCC, Edinburgh





34 x 3 TiB Byte Addressable Storage

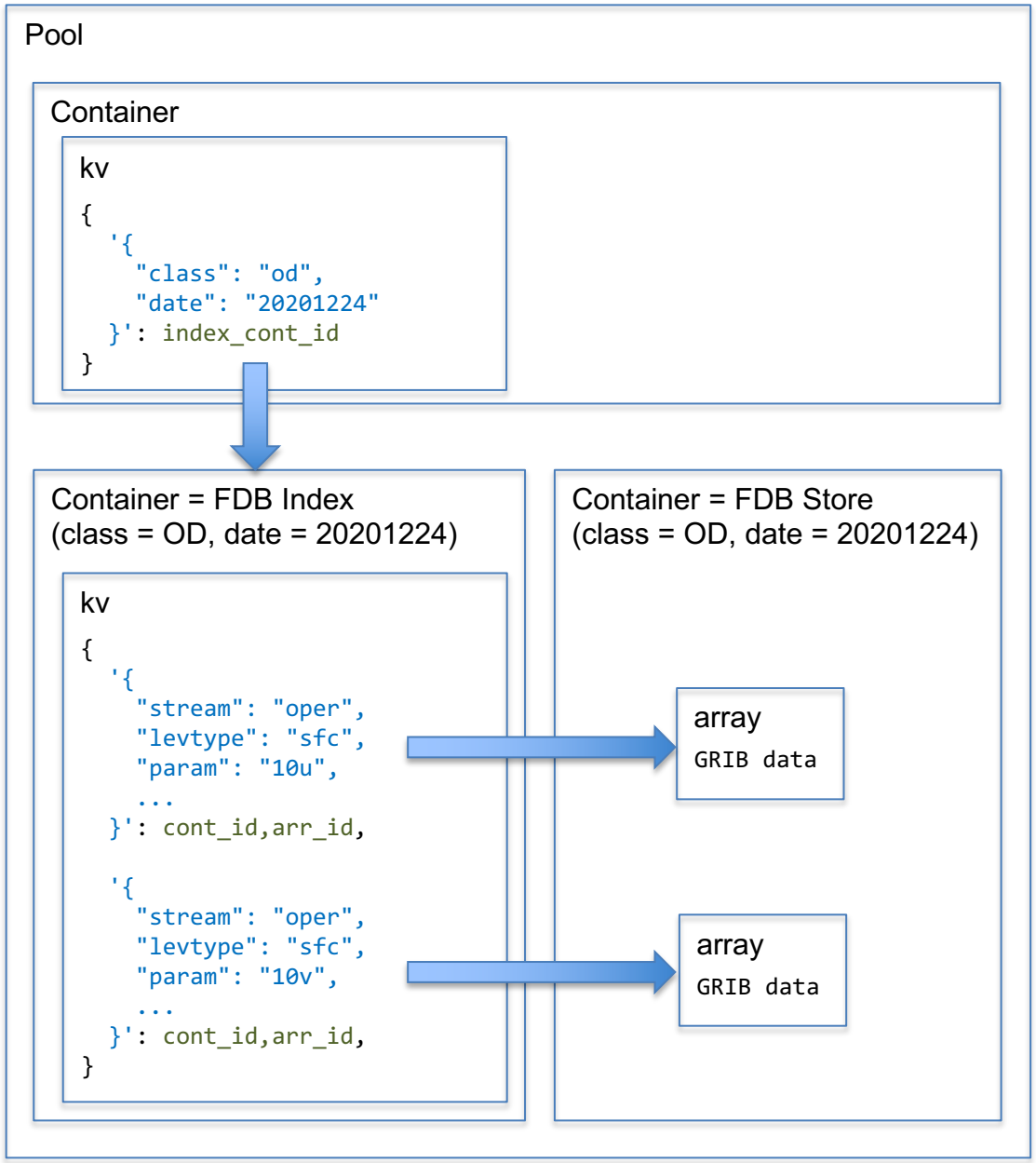
ECMWF

nextgenio

FUJITSU

DAOS as an FDB backend

- Developed C functions for field IO from/to DAOS
  - using DAOS high-level APIs (KV and array)
  - having FDB's architecture in mind for easy integration later
- In close collaboration with Intel DAOS and EPCC

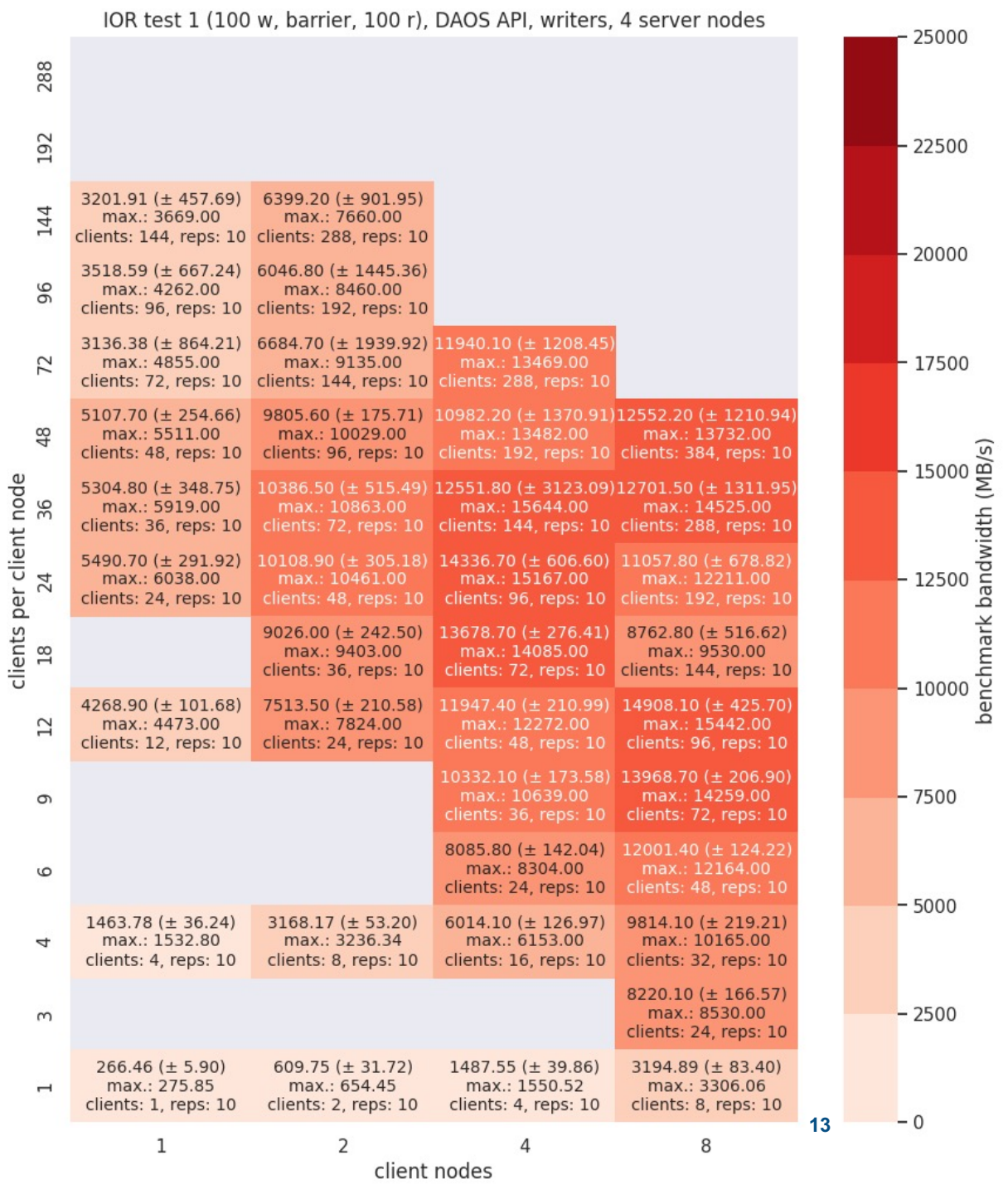


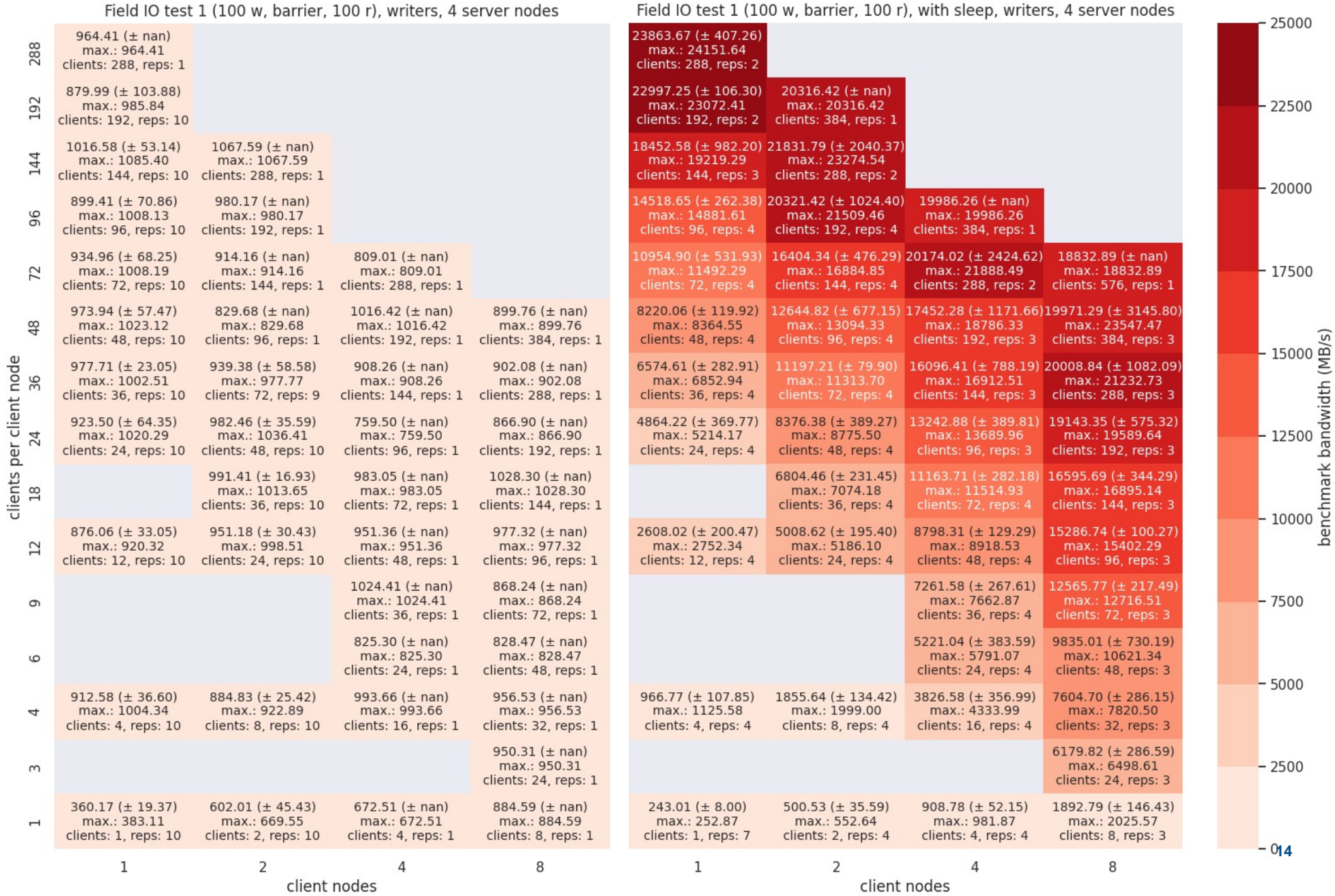
Assessment in progress

- Scalability and performance tests ongoing

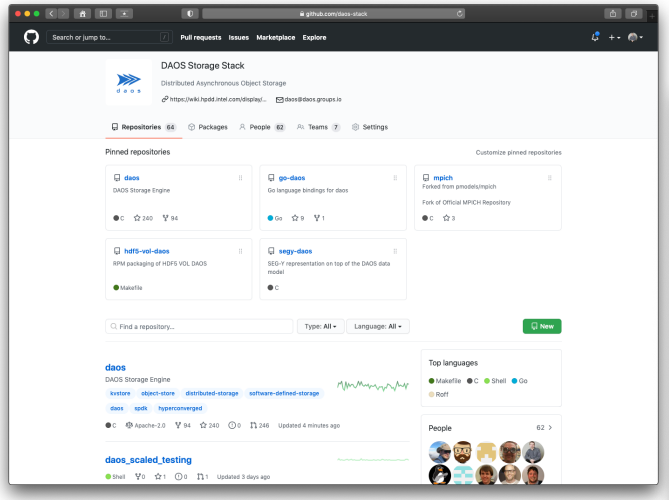
- Goals:
  - assessing if DAOS can be a good replacement for Lustre
  - assessing complexity of porting to DAOS
  - assessing robustness of DAOS

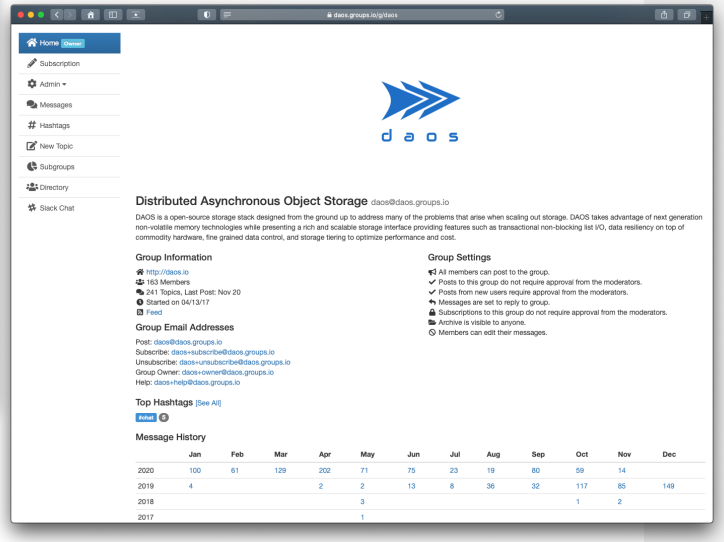
- Using IOR + our own benchmark with field IO functions

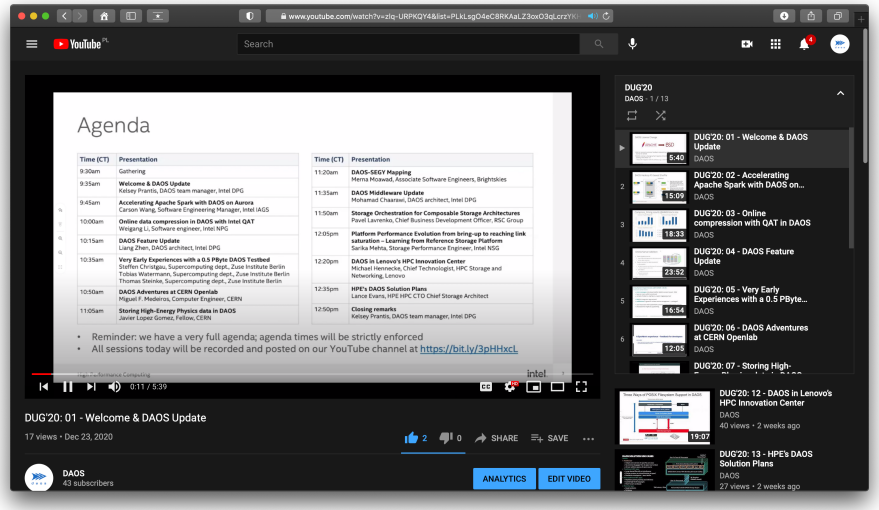



Field IO test 1 (100 w, barrier, 100 r), writers, 4 server nodes
964.41 (± nan) max.: 964.41 clients: 288, reps: 1
879.99 (± 103.88) max.: 985.84 clients: 192, reps: 10
1016.58 (± 53.14) max.: 1085.40 clients: 144, reps: 10
1067.59 (± nan) max.: 1067.59 clients: 288, reps: 1
899.41 (± 70.86) max.: 1008.13 clients: 96, reps: 10
980.17 (± nan) max.: 980.17 clients: 192, reps: 1
934.96 (± 68.25) max.: 1008.19 clients: 72, reps: 10
914.16 (± nan) max.: 914.16 clients: 144, reps: 1
809.01 (± nan) max.: 809.01 clients: 288, reps: 1
973.94 (± 57.47) max.: 1023.12 clients: 48, reps: 10
829.68 (± nan) max.: 829.68 clients: 96, reps: 1
1016.42 (± nan) max.: 1016.42 clients: 192, reps: 1
899.76 (± nan) max.: 899.76 clients: 384, reps: 1
977.71 (± 23.05) max.: 1002.51 clients: 36, reps: 10
939.38 (± 58.58) max.: 977.77 clients: 72, reps: 9
908.26 (± nan) max.: 908.26 clients: 144, reps: 1
902.08 (± nan) max.: 902.08 clients: 288, reps: 1
923.50 (± 64.35) max.: 1020.29 clients: 24, reps: 10
982.46 (± 35.59) max.: 1036.41 clients: 48, reps: 10
759.50 (± nan) max.: 759.50 clients: 96, reps: 1
866.90 (± nan) max.: 866.90 clients: 192, reps: 1
991.41 (± 16.93) max.: 1013.65 clients: 36, reps: 10
983.05 (± nan) max.: 983.05 clients: 72, reps: 1
1028.30 (± nan) max.: 1028.30 clients: 144, reps: 1
876.06 (± 33.05) max.: 920.32 clients: 12, reps: 10
951.18 (± 30.43) max.: 998.51 clients: 24, reps: 10
951.36 (± nan) max.: 951.36 clients: 48, reps: 1
977.32 (± nan) max.: 977.32 clients: 96, reps: 1
1024.41 (± nan) max.: 1024.41 clients: 36, reps: 1
868.24 (± nan) max.: 868.24 clients: 72, reps: 1
825.30 (± nan) max.: 825.30 clients: 24, reps: 1
828.47 (± nan) max.: 828.47 clients: 48, reps: 1
912.58 (± 36.60) max.: 1004.34 clients: 4, reps: 10
884.83 (± 25.42) max.: 922.89 clients: 8, reps: 10
993.66 (± nan) max.: 993.66 clients: 16, reps: 1
956.53 (± nan) max.: 956.53 clients: 32, reps: 1
950.31 (± nan) max.: 950.31 clients: 24, reps: 1
360.17 (± 19.37) max.: 383.11 clients: 1, reps: 10
602.01 (± 45.43) max.: 669.55 clients: 2, reps: 10
672.51 (± nan) max.: 672.51 clients: 4, reps: 1
884.59 (± nan) max.: 884.59 clients: 8, reps: 1
clients per client node
288
192
144
96
72
48
36
24
18
12
9
6
4
3
1
1
2
4
8
client nodes
Field IO test 1 (100 w, barrier, 100 r), with sleep, writers, 4 server nodes
23863.67 (± 407.26) max.: 24151.64 clients: 288, reps: 2
22997.25 (± 106.30) max.: 23072.41 clients: 192, reps: 2
20316.42 (± nan) max.: 20316.42 clients: 384, reps: 1
18452.58 (± 982.20) max.: 19219.29 clients: 144, reps: 3
21831.79 (± 2040.37) max.: 23274.54 clients: 288, reps: 2
14518.65 (± 262.38) max.: 14881.61 clients: 96, reps: 4
20321.42 (± 1024.40) max.: 21509.46 clients: 192, reps: 4
19986.26 (± nan) max.: 19986.26 clients: 384, reps: 1
10954.90 (± 531.93) max.: 11492.29 clients: 72, reps: 4
16404.34 (± 476.29) max.: 16884.85 clients: 144, reps: 4
20174.02 (± 2424.62) max.: 21888.49 clients: 288, reps: 2
18832.89 (± nan) max.: 18832.89 clients: 576, reps: 1
8220.06 (± 119.92) max.: 8364.55 clients: 48, reps: 4
12644.82 (± 677.15) max.: 13094.33 clients: 96, reps: 4
17452.28 (± 1171.66) max.: 18786.33 clients: 192, reps: 3
19971.29 (± 3145.80) max.: 23547.47 clients: 384, reps: 3
6574.61 (± 282.91) max.: 6852.94 clients: 36, reps: 4
11197.21 (± 79.90) max.: 11313.70 clients: 72, reps: 4
16096.41 (± 788.19) max.: 16912.51 clients: 144, reps: 3
20008.84 (± 1082.09) max.: 21232.73 clients: 288, reps: 3
4864.22 (± 369.77) max.: 5214.17 clients: 24, reps: 4
8376.38 (± 389.27) max.: 8775.50 clients: 48, reps: 4
13242.88 (± 389.81) max.: 13689.96 clients: 96, reps: 3
19143.35 (± 575.32) max.: 19589.64 clients: 192, reps: 3
6804.46 (± 231.45) max.: 7074.18 clients: 36, reps: 4
11163.71 (± 282.18) max.: 11514.93 clients: 72, reps: 4
16595.69 (± 344.29) max.: 16895.14 clients: 144, reps: 3
2608.02 (± 200.47) max.: 2752.34 clients: 12, reps: 4
5008.62 (± 195.40) max.: 5186.10 clients: 24, reps: 4
8798.31 (± 129.29) max.: 8918.53 clients: 48, reps: 4
15286.74 (± 100.27) max.: 15402.29 clients: 96, reps: 3
7261.58 (± 267.61) max.: 7662.87 clients: 36, reps: 4
12565.77 (± 217.49) max.: 12716.51 clients: 72, reps: 3
5221.04 (± 383.59) max.: 5791.07 clients: 24, reps: 4
9835.01 (± 730.19) max.: 10621.34 clients: 48, reps: 3
966.77 (± 107.85) max.: 1125.58 clients: 4, reps: 4
1855.64 (± 134.42) max.: 1999.00 clients: 8, reps: 4
3826.58 (± 356.99) max.: 4333.99 clients: 16, reps: 4
7604.70 (± 286.15) max.: 7820.50 clients: 32, reps: 3
6179.82 (± 286.59) max.: 6498.61 clients: 24, reps: 3
243.01 (± 8.00) max.: 252.87 clients: 1, reps: 7
500.53 (± 35.59) max.: 552.64 clients: 2, reps: 4
908.78 (± 52.15) max.: 981.87 clients: 4, reps: 4
1892.79 (± 146.43) max.: 2025.57 clients: 8, reps: 3
1
2
4
8
client nodes
benchmark bandwidth (MB/s)
25000
22500
20000
17500
15000
12500
10000
7500
5000
2500
0

Resources







- Github
  - https://github.com/daos-stack/daos
- DAOS online documentation
  - http://daos.io
- Community mailing list & slack
  - https://daos.groups.io
- DAOS ISC'21 Demo
  - https://www.youtube.com/watch?v=1ryUB5VsclcI
- DAOS User Group
  - https://www.youtube.com/playlist?list=PLkLsgO4eC8RKAaLZ3oxO3qLcrzYKHXNDm


intel