# IO500 #1 DAOS Update

Intel Optane Persistent Memory Technical Solution Specialist: Fumiyasu Ishibashi

# 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.

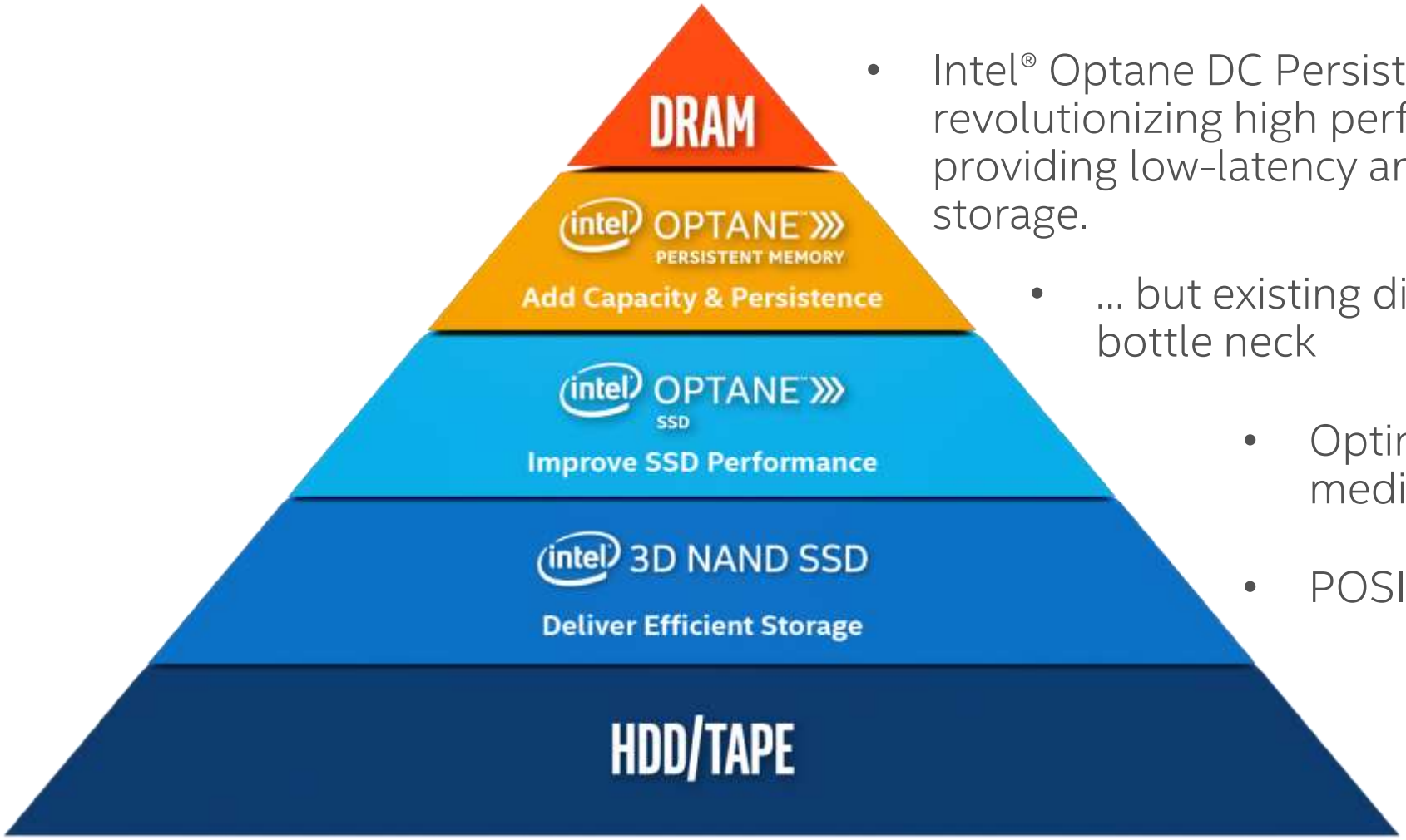
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality,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

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

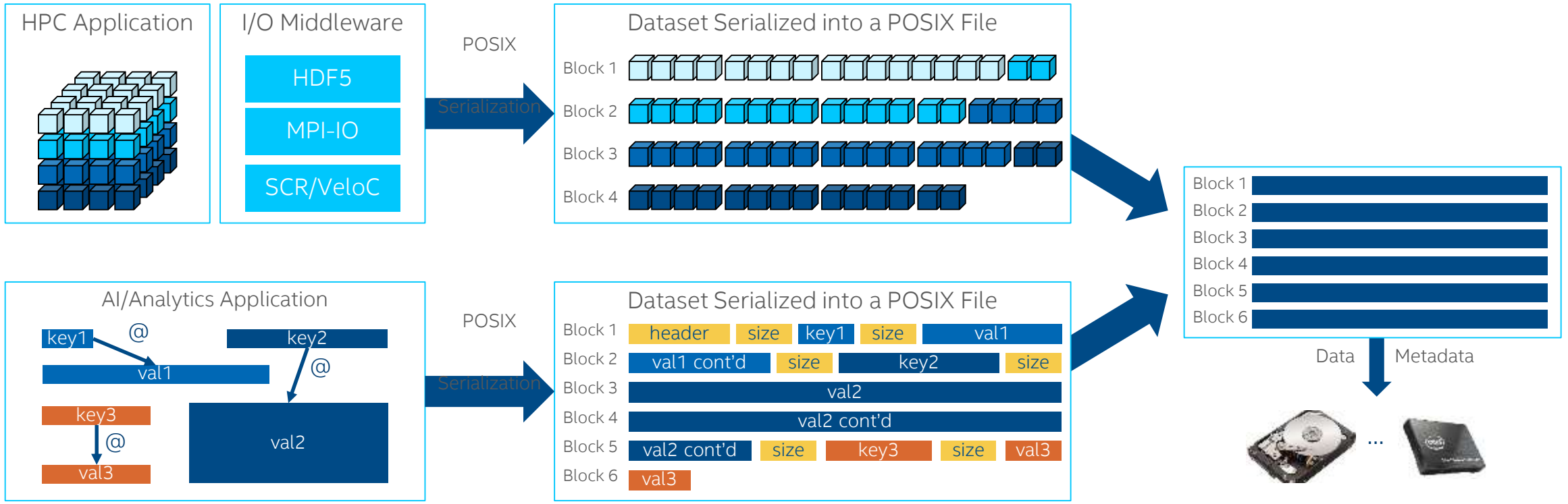
# Revolutionizing High Performance Storage




- Intel® Optane DC Persistent Memory is revolutionizing high performance storage by providing low-latency and fine-grained persistent storage.
  - … but existing distributed storage software is a bottle neck
    - Optimized for millisecond rotating media
    - POSIX constraints limit performance

Scale out storage needs to be built from the ground up for new NVM technology

# The Problem with POSIX and Blocks

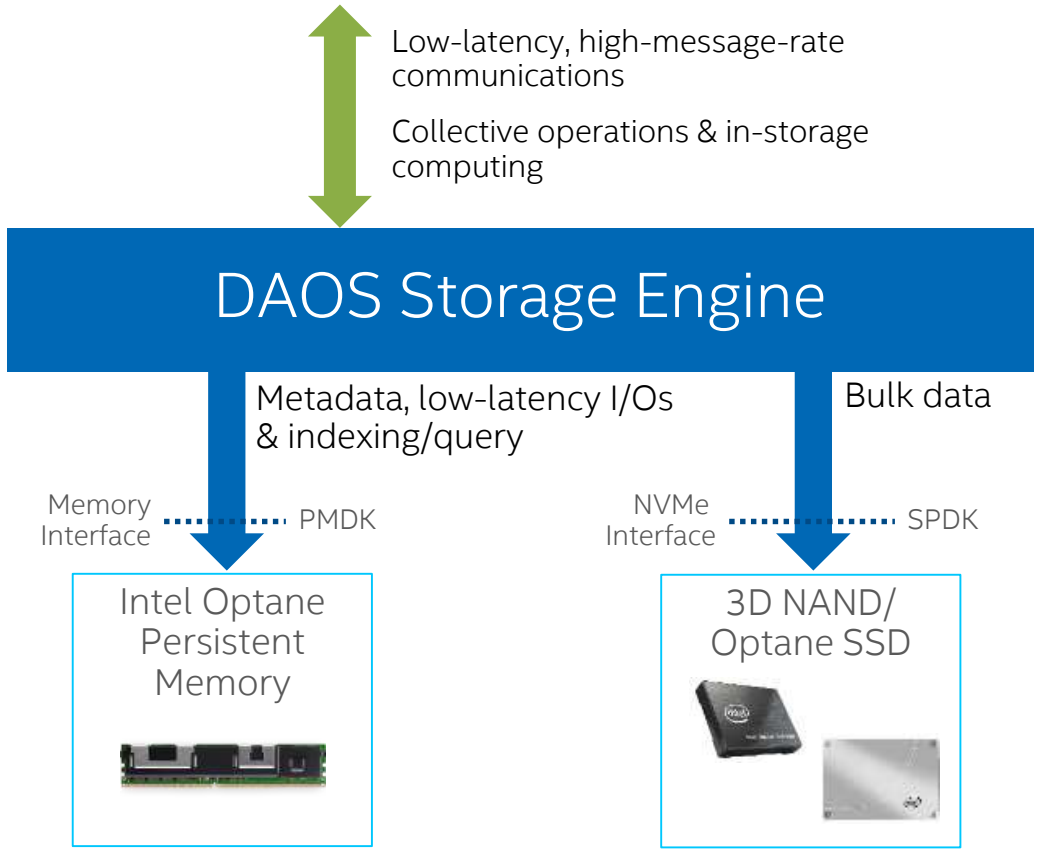


# DAOS history

- 2012 - 2015
  - Build DAOS storage model within Lustre stack
    - Fully in kernel space, block device
- 2015 – 2017
  - Re-architecture of DAOS for new storage technologies
    - Persistent memory, NVMe SSDs
  - Prototyping core algorithms
- 2017 – now
  - Development of DAOS

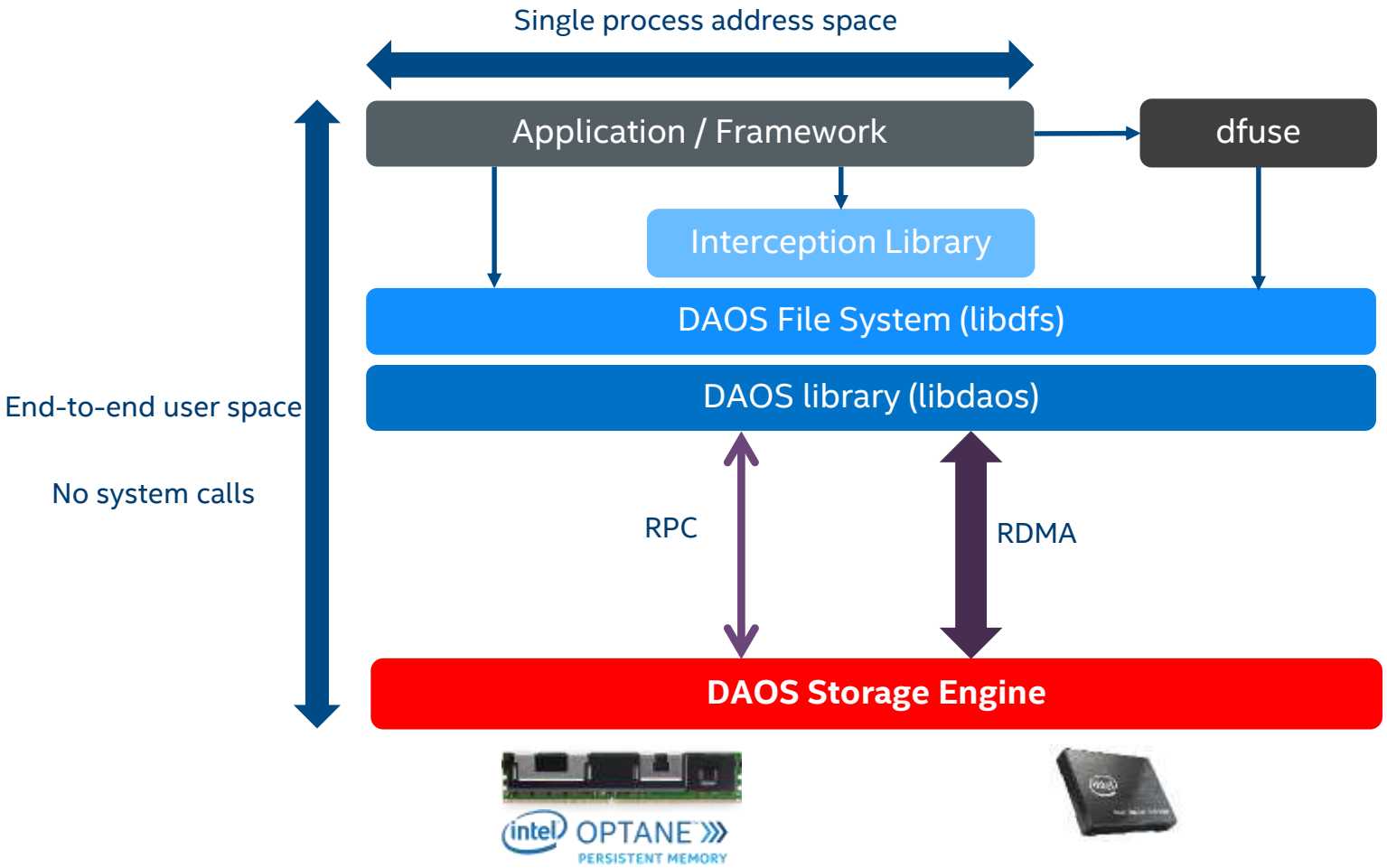
# What is DAOS?

- A new, innovative distributed parallel file system based on Intel Optane Persistent Memory and NVMe SSDs
- Coordinates parallel IO across many nodes presented to the user as a single filesystem
- 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

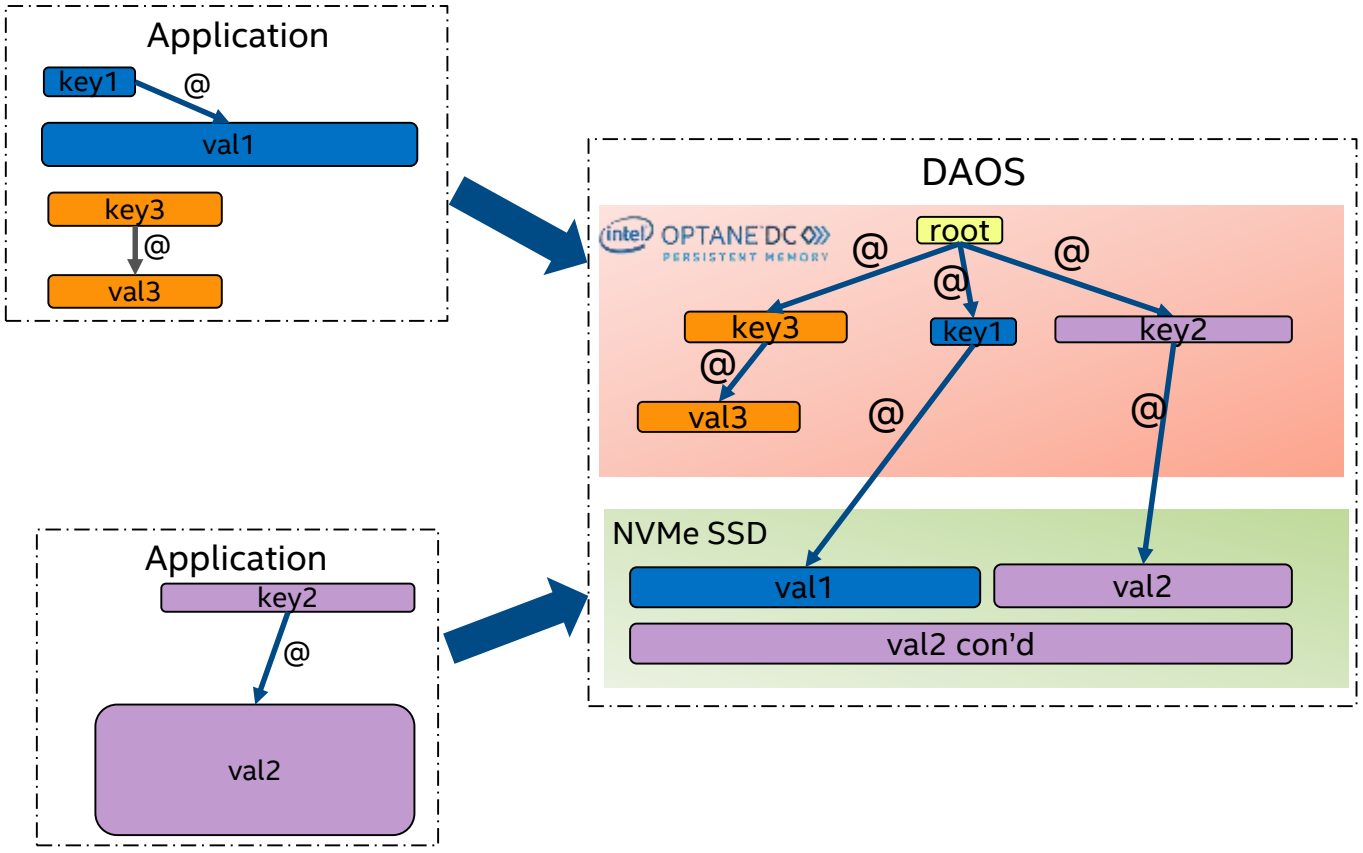



**More IOPs and bandwidth per dollar**

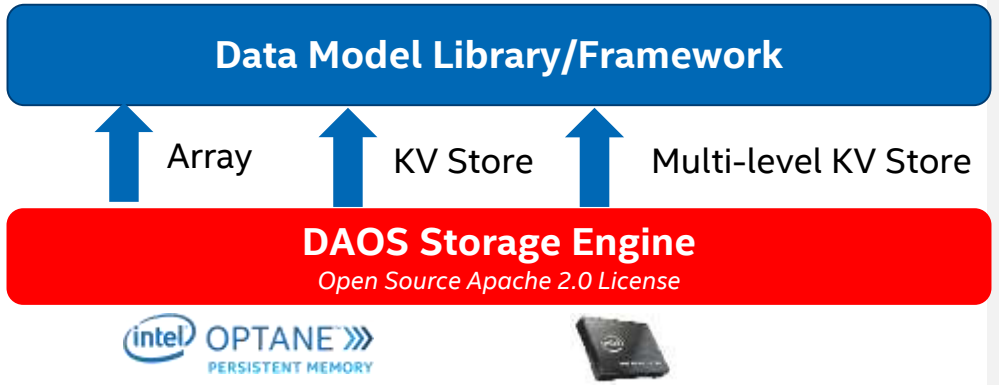
# POSIX I/O Support



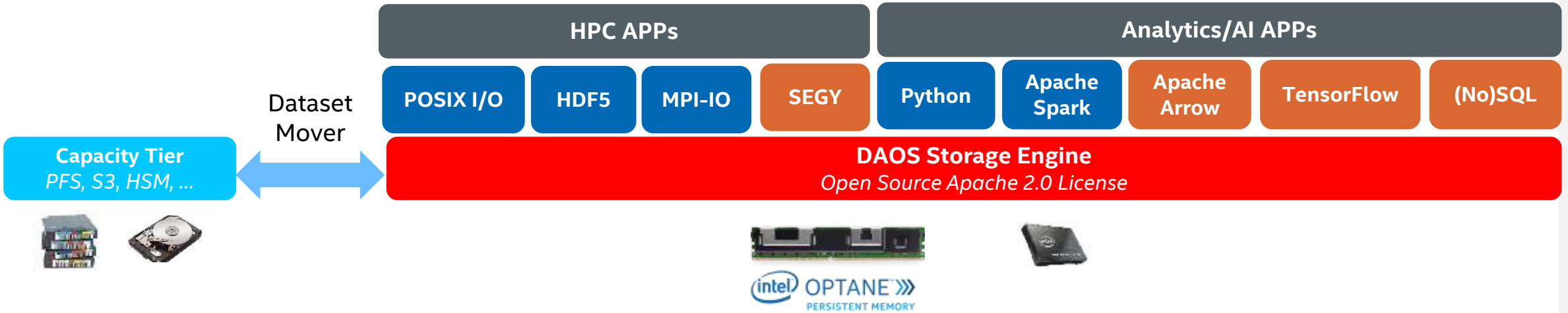
# DAOS API




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

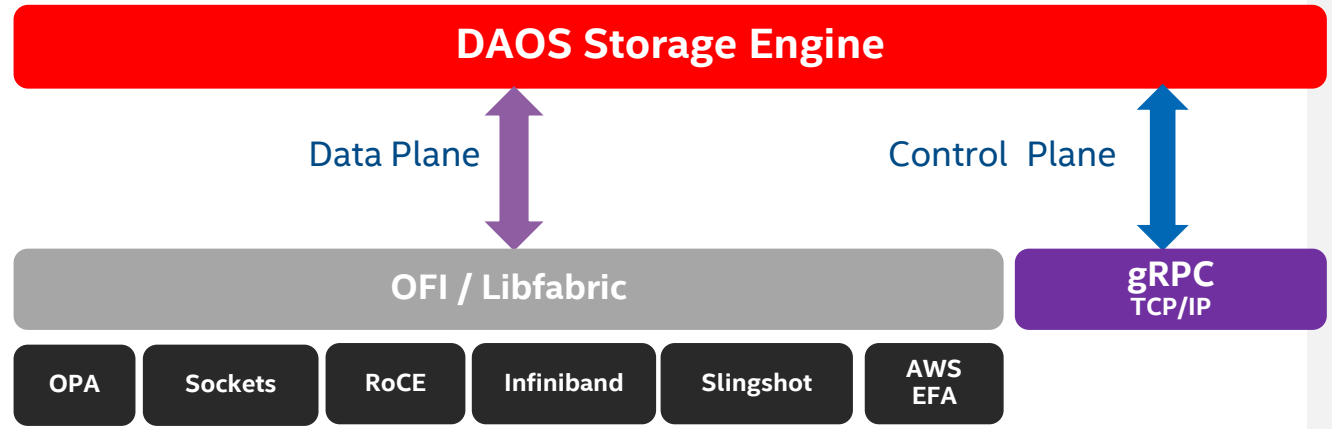


# Application Interface



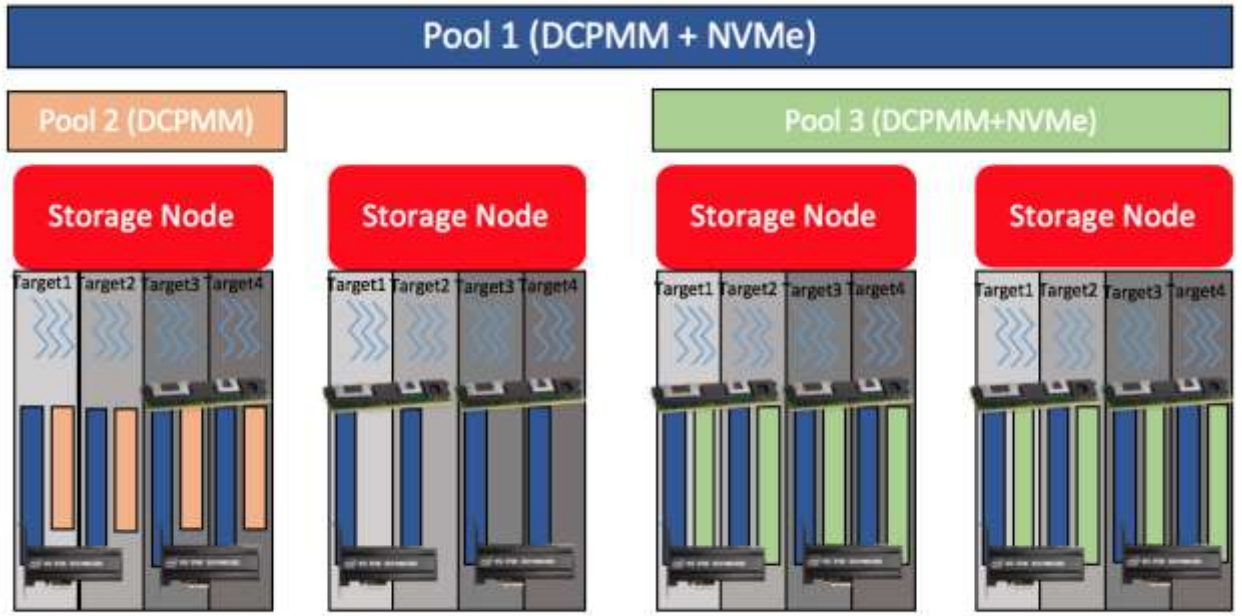
# Network Support

- Performance-critical I/O path over libfabric
  - Low-latency messaging
    - End-to-end in userspace
  - Native support for RDMA
    - True zero-copy I/O
  - Non-blocking
  - Scalable collective communications
- Out-of-band channel for administration
  - Manage hardware, service & pools
  - Telemetry & troubleshooting
  - Secured with TLS & certificate



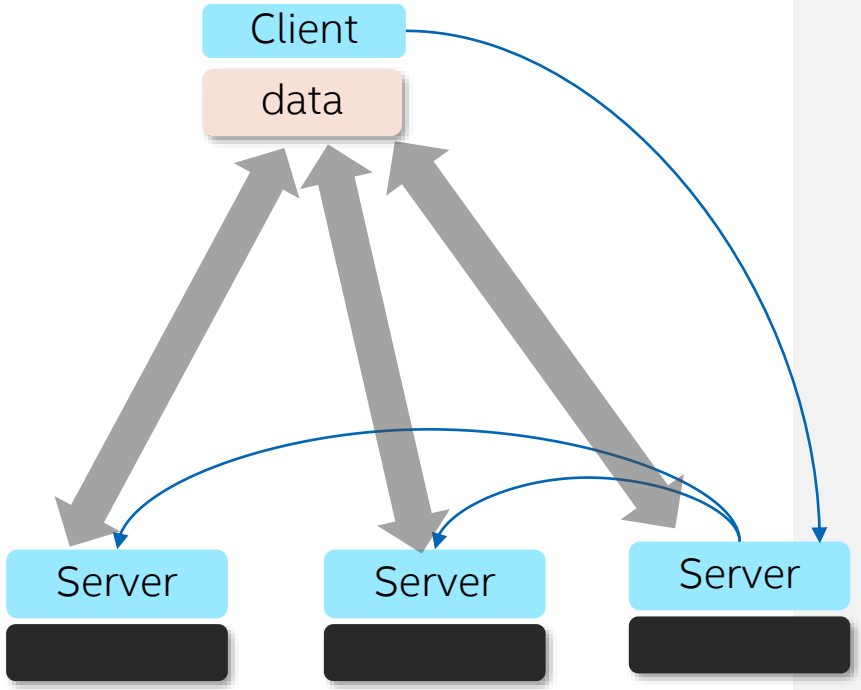
# Storage virtualization & Multi-tenancy

- Distributed storage reservation
  - AEP
  - NVMe SSD
- Predicatable capacity
  - Can be resized
  - Can be extended to span more servers
- Multi-tenancy
  - NFSv4-type ACLs
- Typically 1 pool = 1 project
  - Can have a single pool or 100's



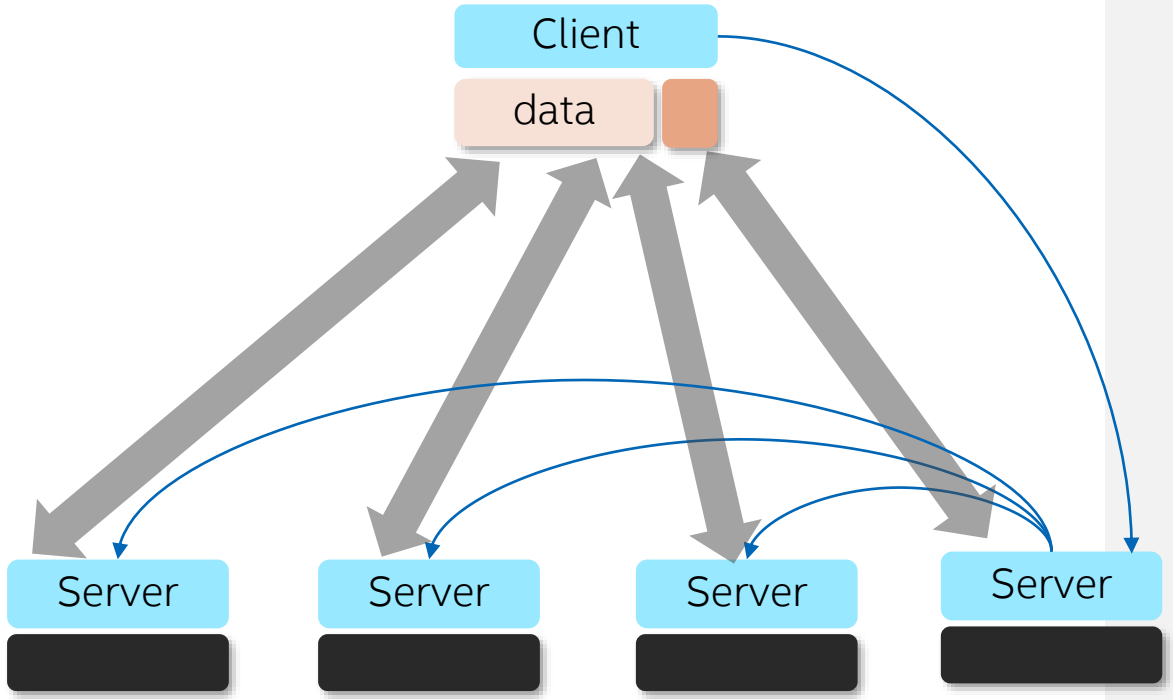
# Data Protection and self-healing - Replication

- Data replication in DAOS
  - Primary-slave replication
  - Distributed transaction for atomicity
- Degraded mode
  - Non-blocking protocol for server fail-out
- Online data recovery
  - Low impact on ongoing I/O
  - Declustered data reconstruction

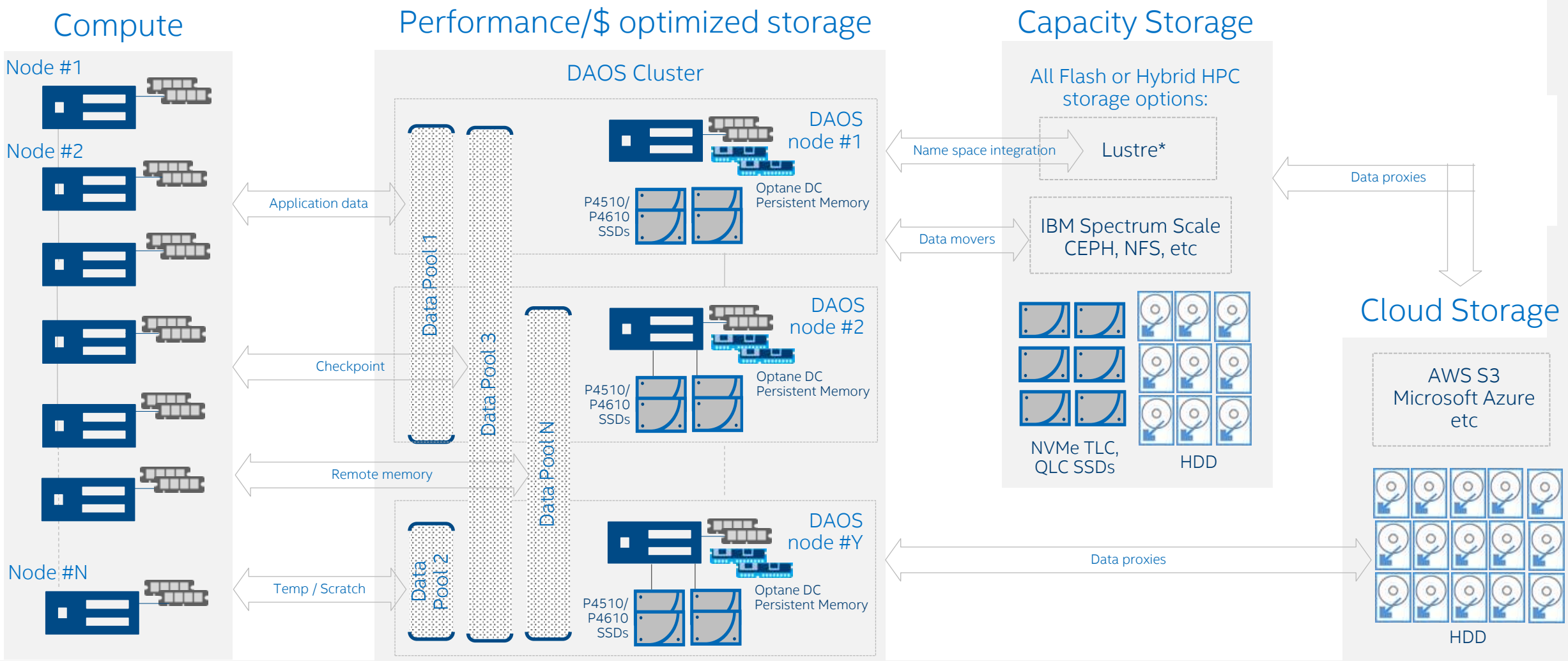


# Data Protection and self-healing – Erasure Code

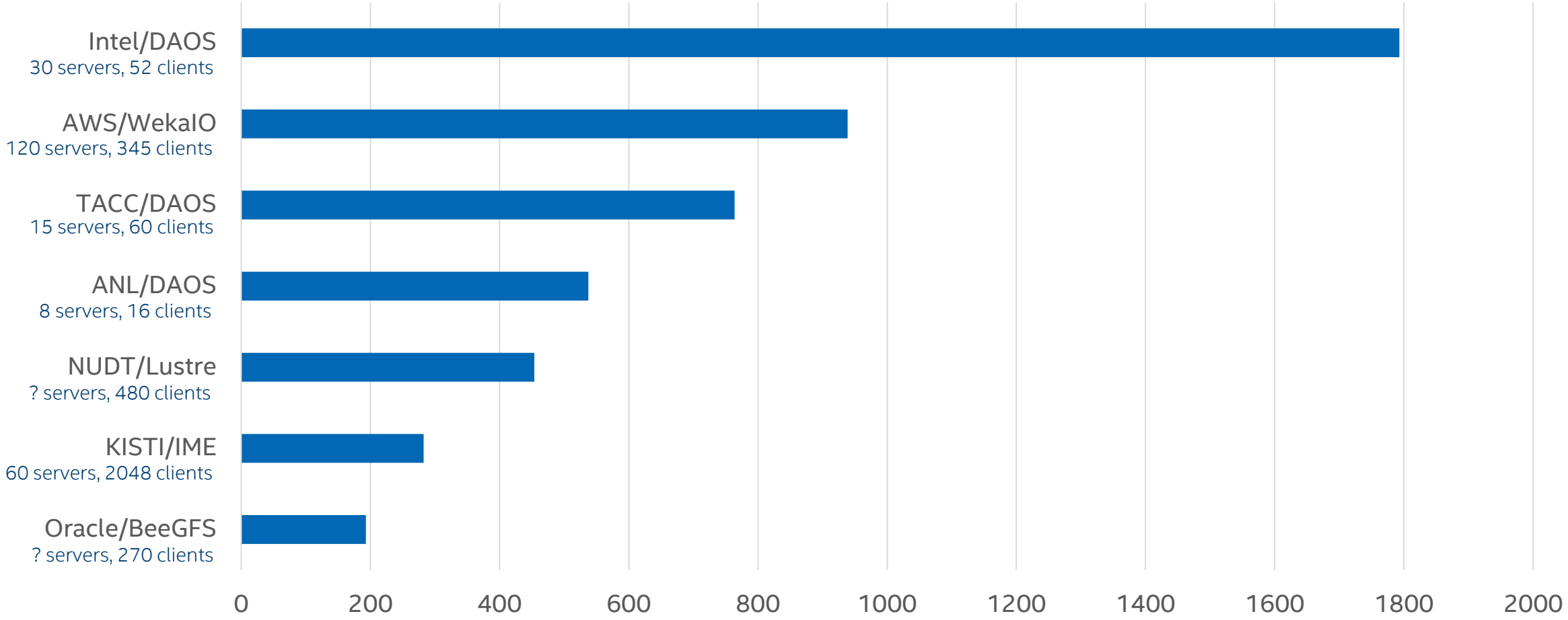
- Erasure code in DAOS
  - Computed by client on write
  - Distributed transaction for atomicity
  - Replication for partial write
    - Merge and encode by server
- Degraded mode
  - Non-blocking protocol for server fail-out
  - Client side inflight data reconstructing
- Online data recovery
  - Server side data exchange and reconstruction



# DAOS in the overall cluster architecture

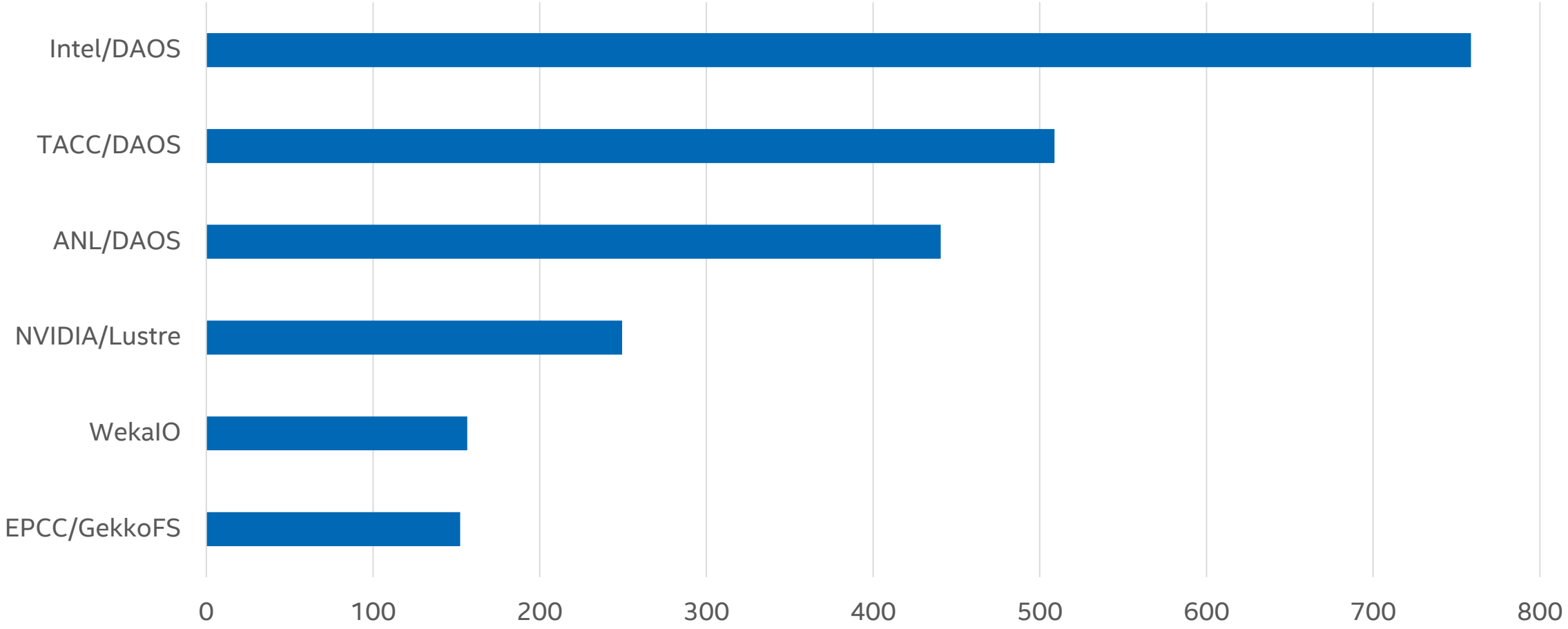


# IO-500 Full List Overall Score



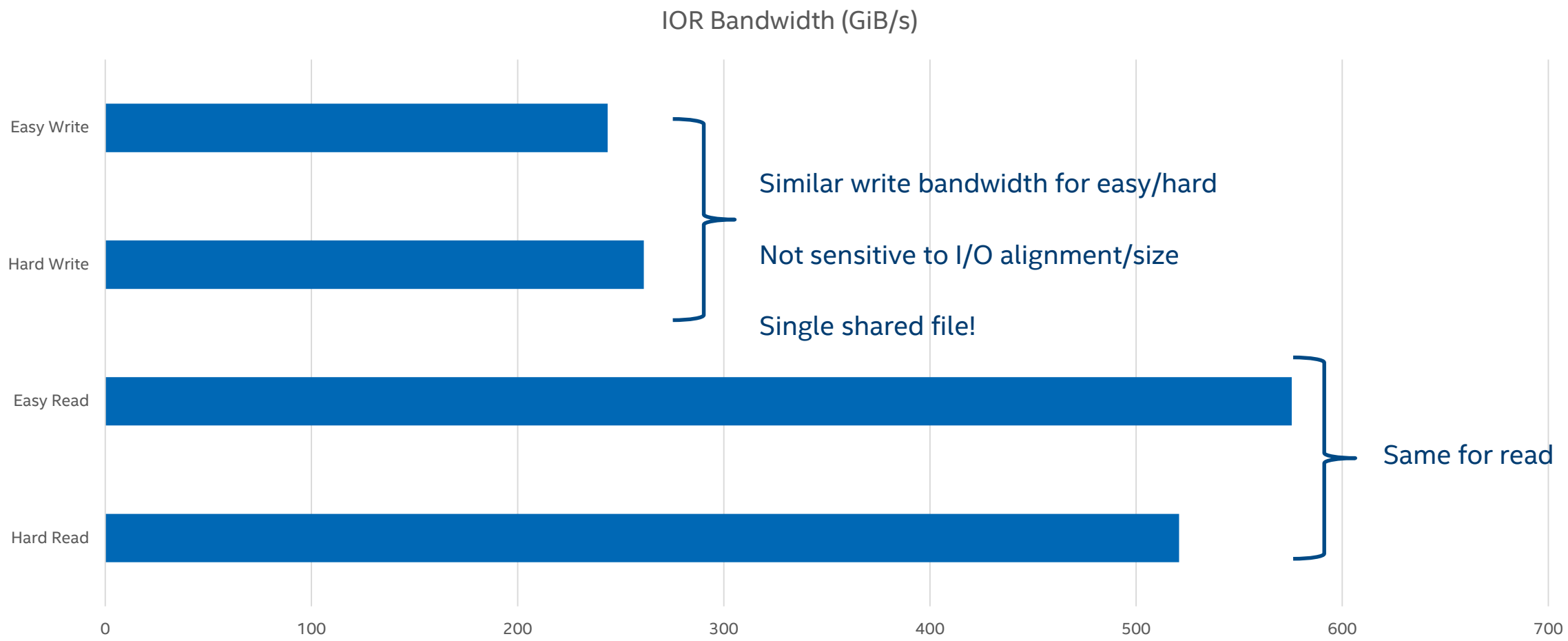

Data Source: https://www.vi4io.org/io500/start

# IO-500 10 Node Challenge Overall Score

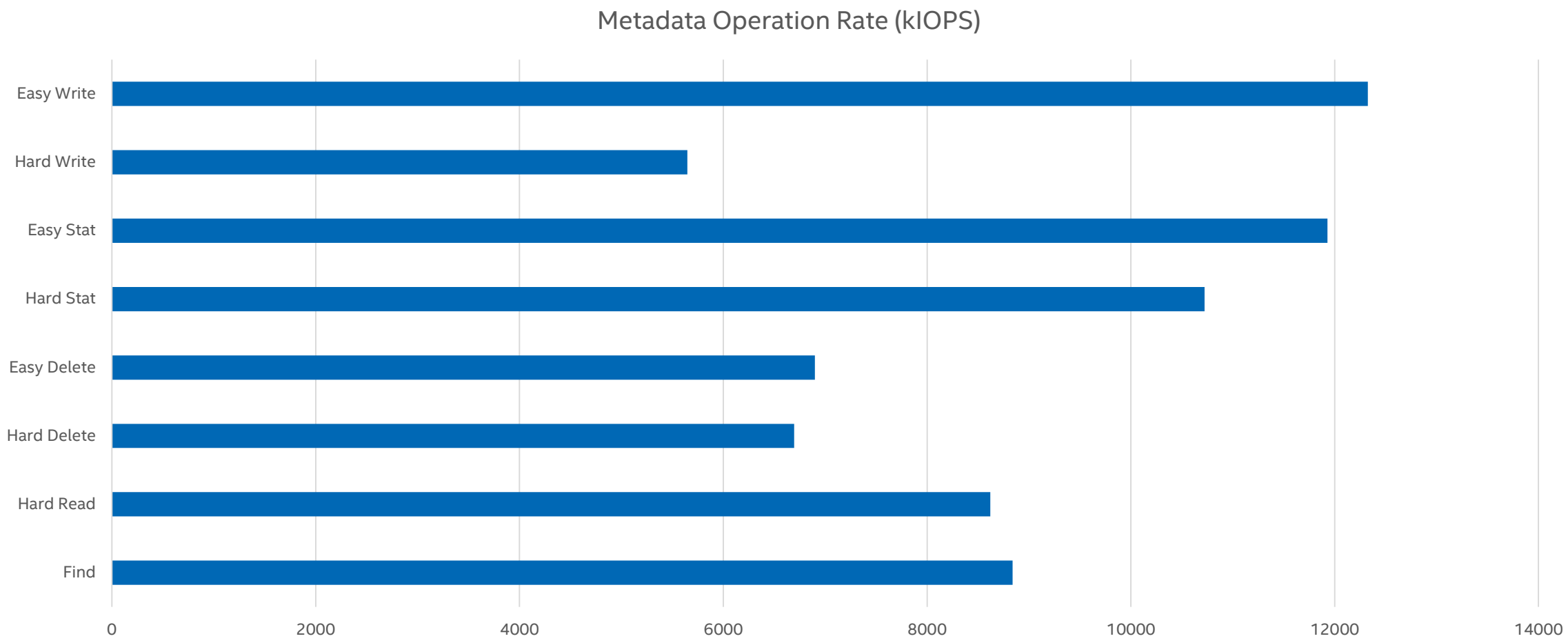



Data Source: https://www.vi4io.org/io500/start

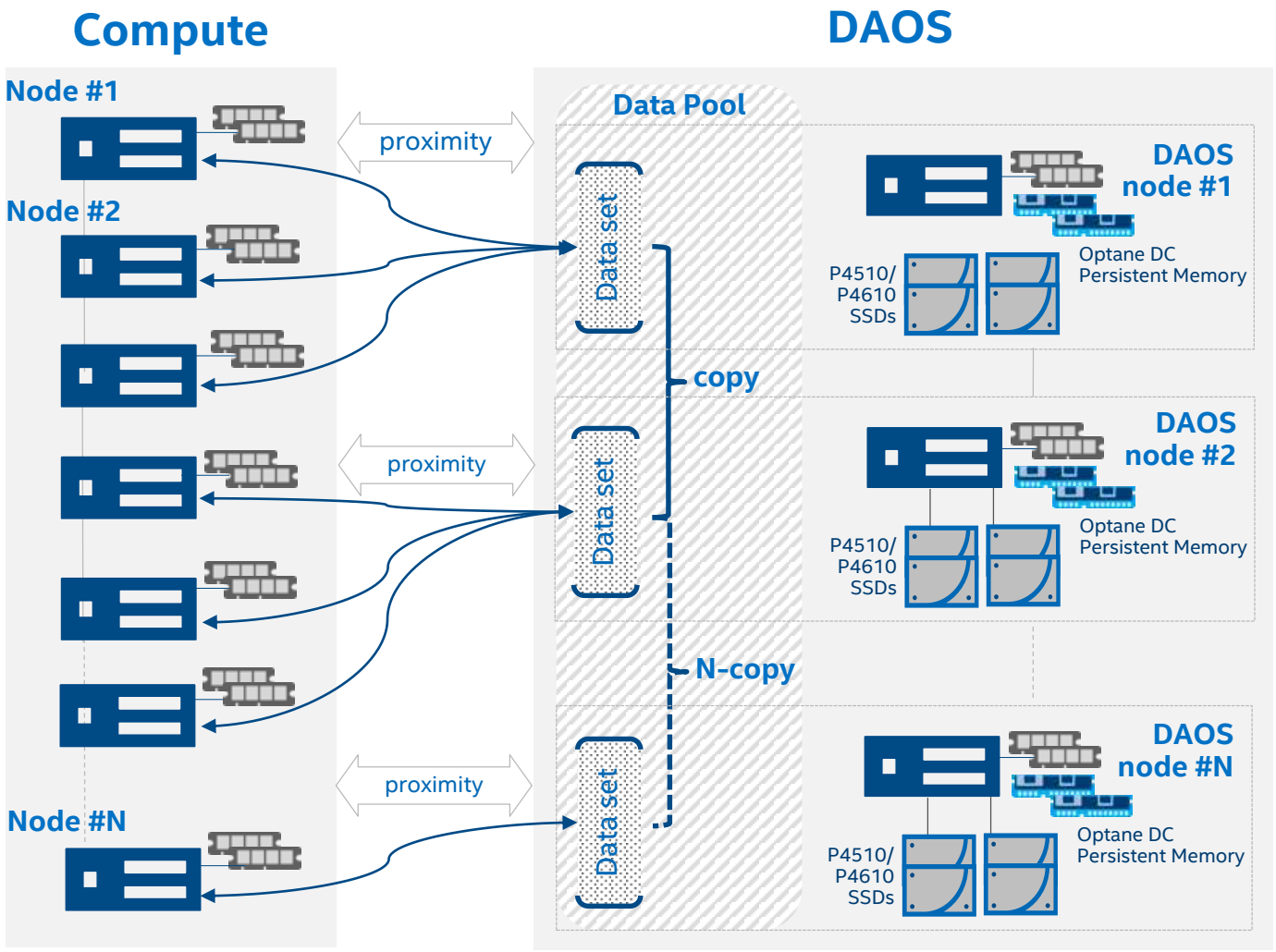
# IOR Bandwidth



# Metadata IOPS



# DAOS and AI Best play together




AI training workloads dominate more reads than writes

Parallel throughput becomes critical at training stage for a shared data set

Traditional parallel filesystems are not able to keep up with performance demand

DAOS has a native capability to set the replication factor per object

This can be used to bring higher bandwidth for the given object and increase parallelism and proximity to the data.

- In addition to replication, DAOS has optimized object placement planned in the roadmap for version 2.2.

DAOS further accelerates All Flash Arrays already been used with GPU accelerated compute nodes.

# 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 – Q2'2020


**Released on June 18**

***DAOS:***
- NVMe & DCPMM support
- Per-pool ACL
- UNS in DAOS via dfuse

***Application Interface:***
- MPI-IO Driver
- HDF5 Support
- Basic POSIX I/O support

***DAOS:***
- End-to-end data integrity
- Per-container ACL
- Improved control plane
- Lustre/UNS integration
- Replication & self-healing
- Erasure code (preview)
- Conditional updates

***Application Interface:***
- Replication & self-healing (preview)
- HDF5 DAOS Connector
- Spark

***DAOS:***
- Online server addition
- Advanced control plane
- Multi OFI provider support

***Application Interface:***
- POSIX data mover
- Async HDF5 operations over DAOS

***DAOS:***
- Erasure code
- Telemetry & per-job statistics
- Distributed transactions

***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.***

# Resources

ISC demonstration

- IOR + Spark workloads

- https://youtu.be/e69Rgz2FMbE

DAOS solution brief

- https://www.intel.com/content/www/us/en/high-performance-computing/

Source code on GitHub

- https://github.com/daos-stack/daos

Admin Guide

- https://daos-stack.github.io/

Community mailing list on Groups.io

- daos@daos.groups.io

Support

- https://jira.hpdd.intel.com


intel