DAOS

Scalable Software-Defined Storage for HPC/Big Data/AI Convergence

Jeff Olivier

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 computer system 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.

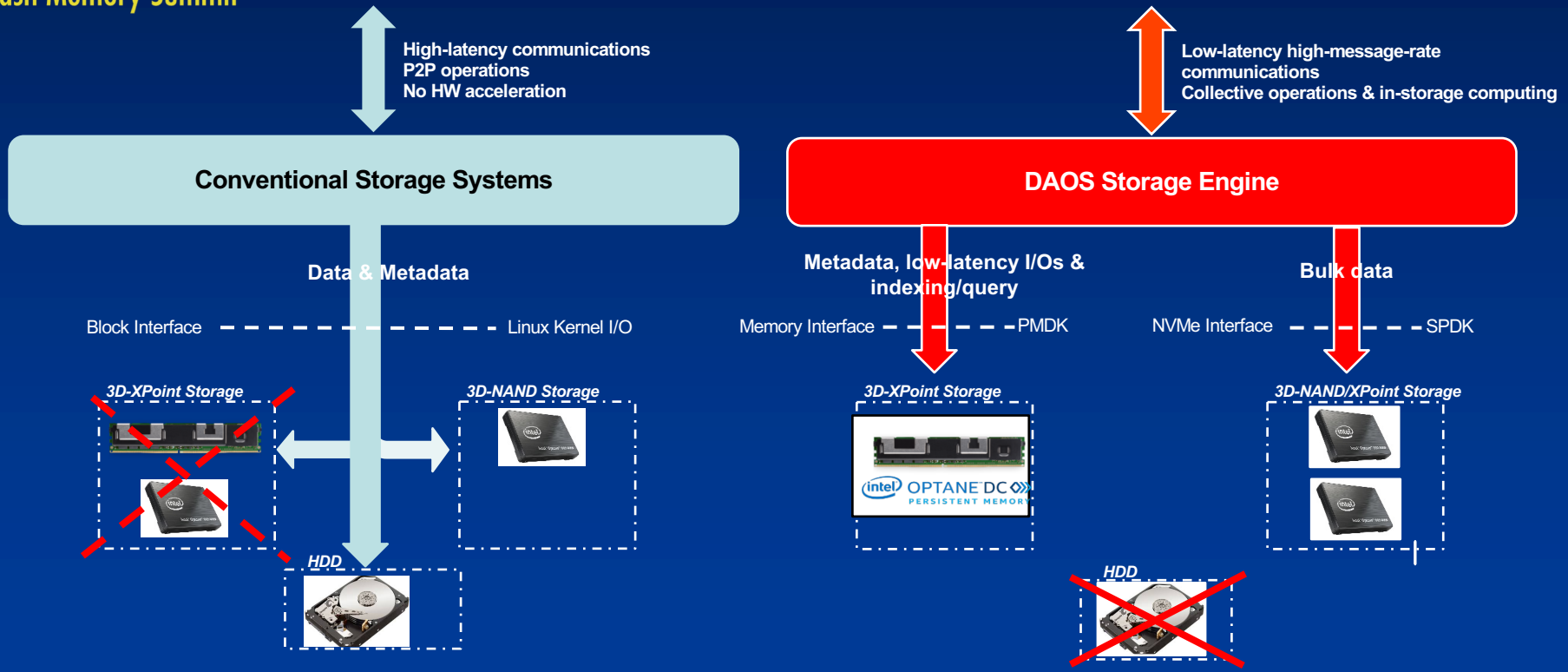
Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

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.

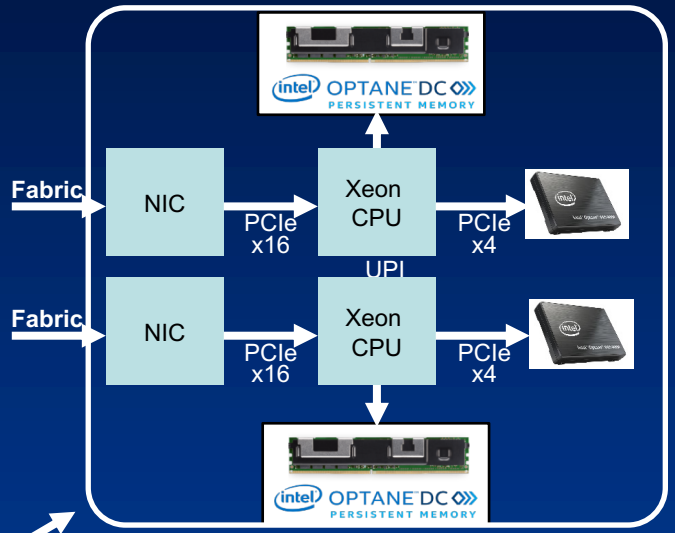
Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

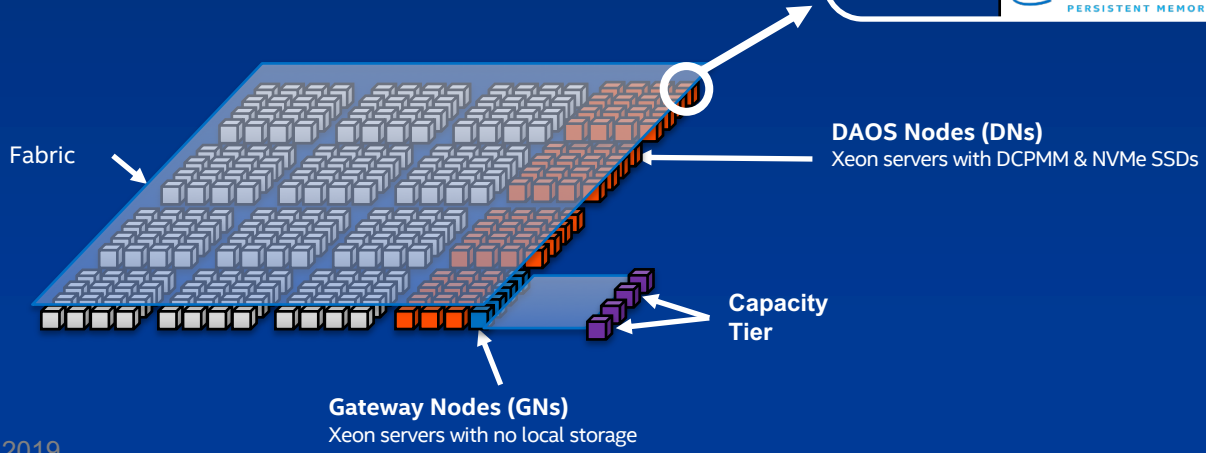
*Other names and brands may be claimed as property of others.

DAOS Architecture



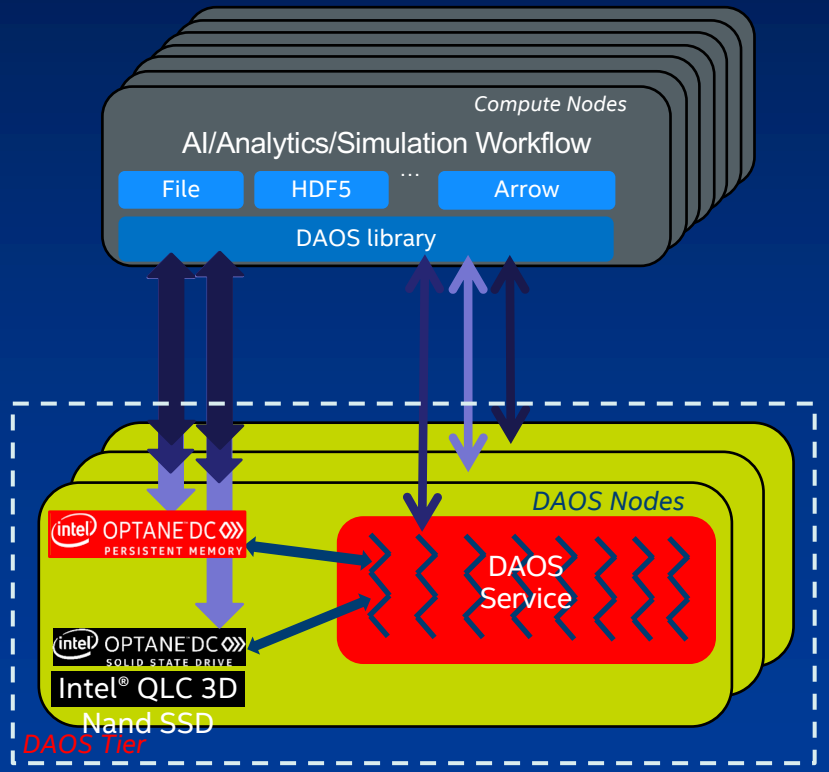
Scale-out Storage





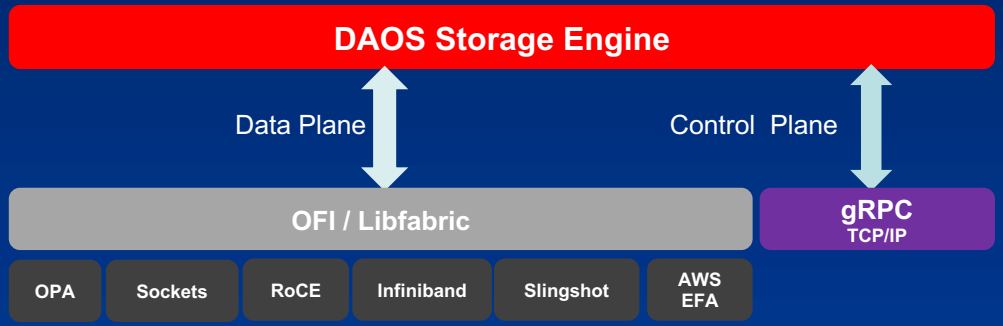
DAOS Tier Anatomy

- DAOS Tier
  - Globally accessible from any compute nodes
  - Large capacity (100's PB)
- DAOS Nodes
  - COTS Xeon servers running the DAOS service
  - RNIC attached for communications
    - Support multiple RNICs per server to sustain backend storage IOPS/bandwidth
  - Mix of storage technologies attached
    - Optane DC Persistent Memory (DCPM)
    - NVMe SSD (*NAND, Optane SSDs)



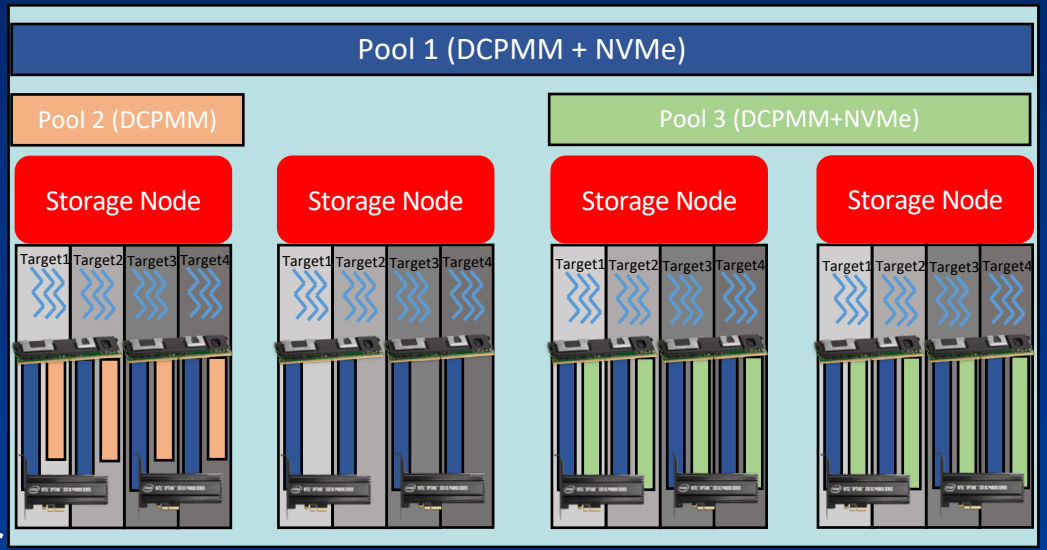
Network Support

- 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



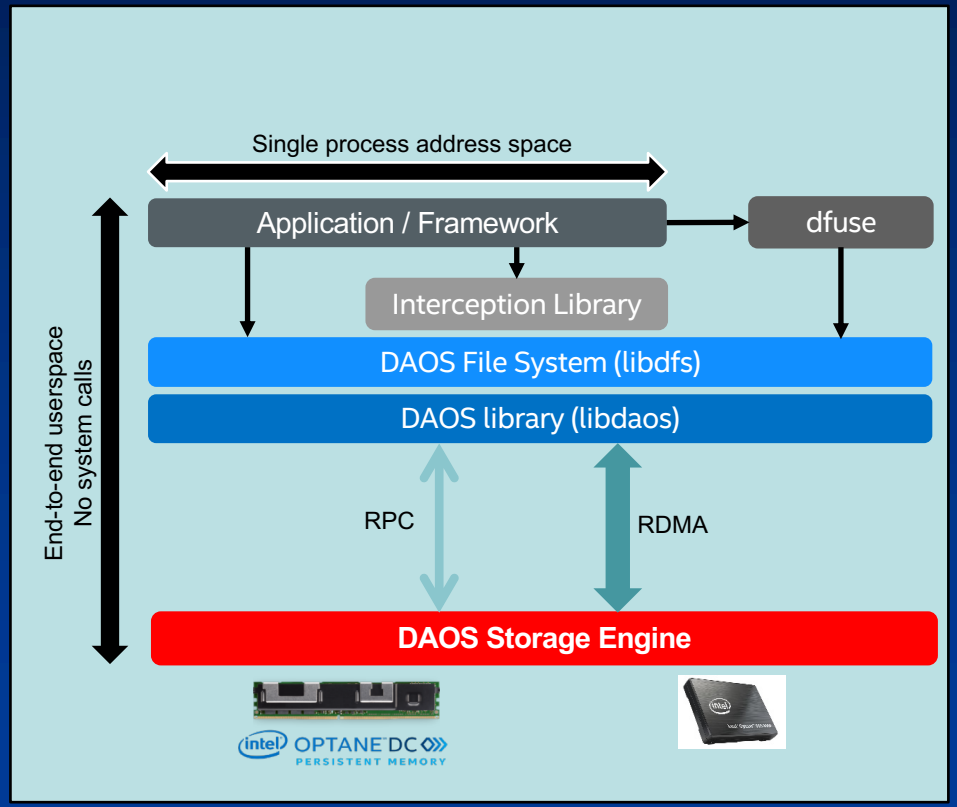
Storage virtualization & Multi-tenancy

- Distributed storage reservation
  - Persistent memory / DCPMM
  - 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

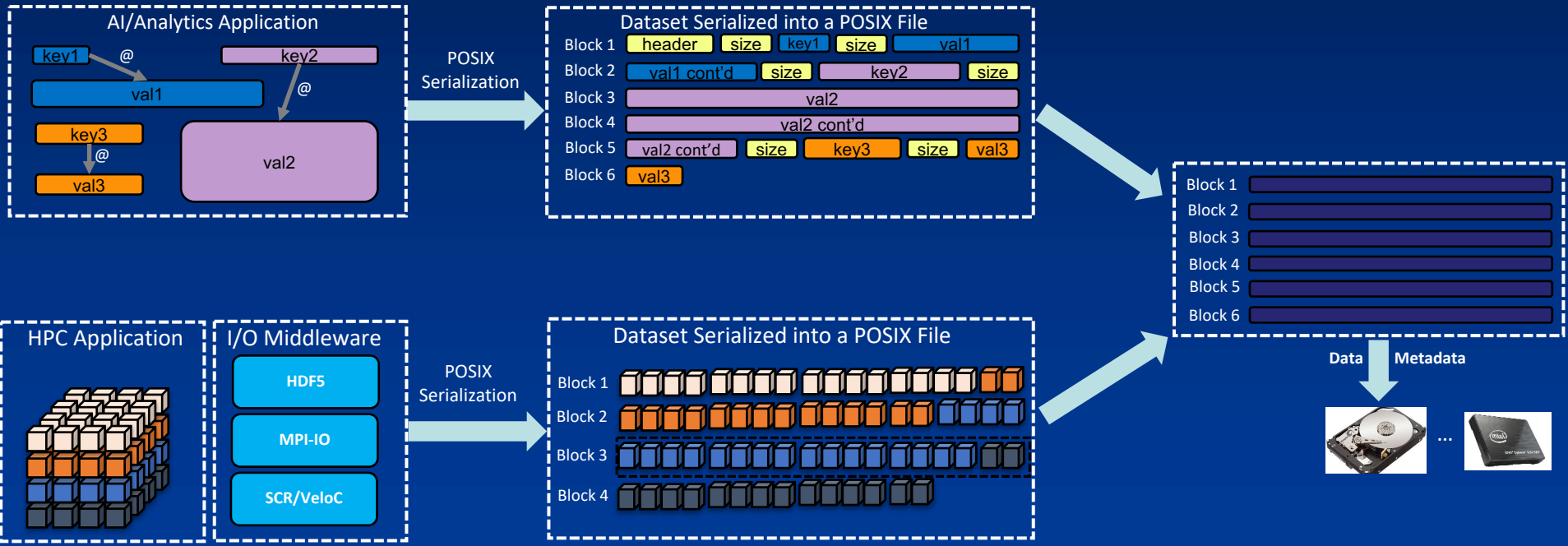


POSIX I/O Support

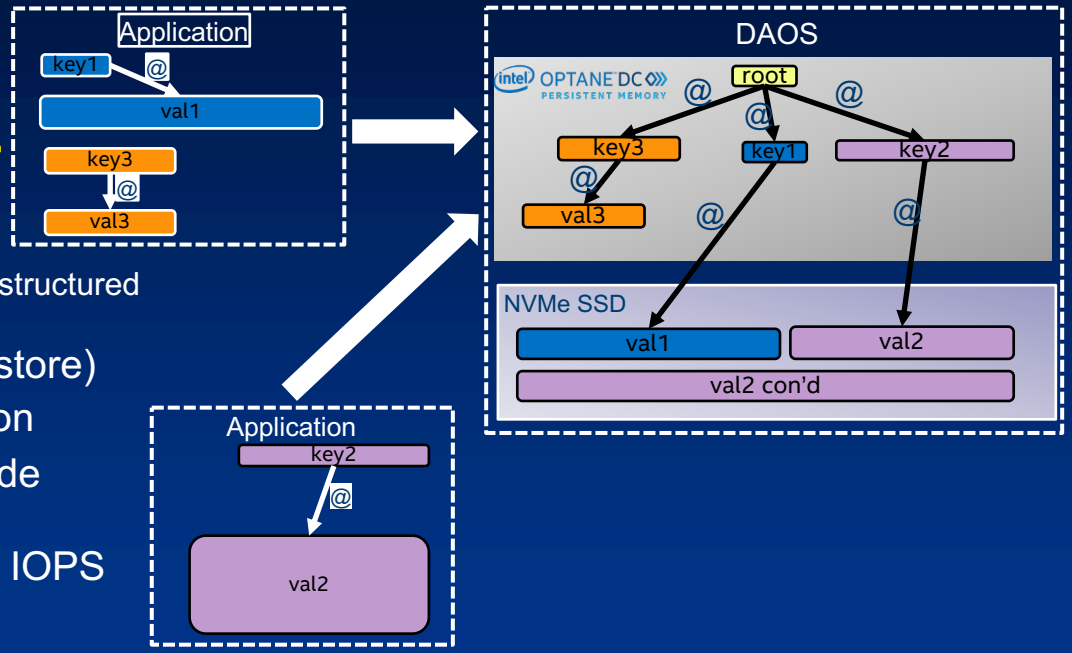
- DAOS File System (libdfs)
  - Encapsulated POSIX namespace
  - Application/framework can link directly with libdfs
    - ior/mdtest backend provided
    - MPI-IO driver leveraging collective open
    - TensorFlow, …
- FUSE Daemon (dfuse)
  - Transparent access to DAOS
  - Involve system calls
- I/O interception library
  - OS bypass for read/write operations



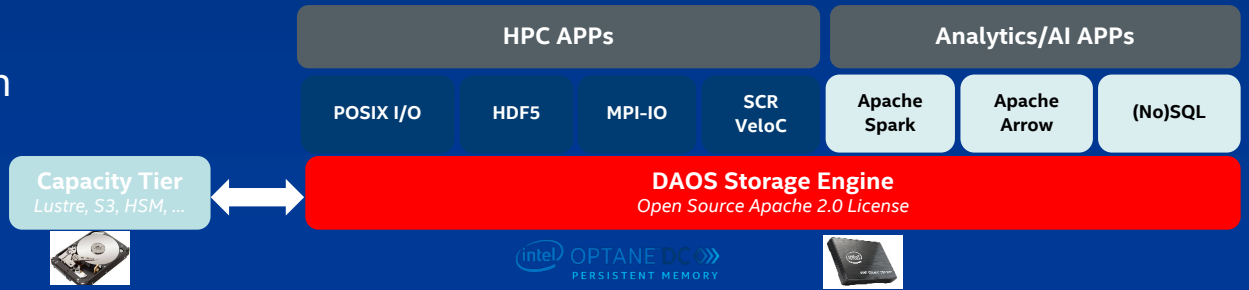
POSIX I/O Limitations



Beyond POSIX …



- Native support for structured, semi-structured & unstructured data models
  - Built on top of DCPMM (direct load/store)
  - Unconstrained by POSIX serialization
  - Data access time orders of magnitude faster (µs)
  - Scalable concurrent updates & high IOPS
  - Enable in-storage computing
- Multi-tier support & Lustre integration
  - Dataset mover
  - Smooth migration path



Flash Memory Summit

DAOS: Primary storage on Aurora

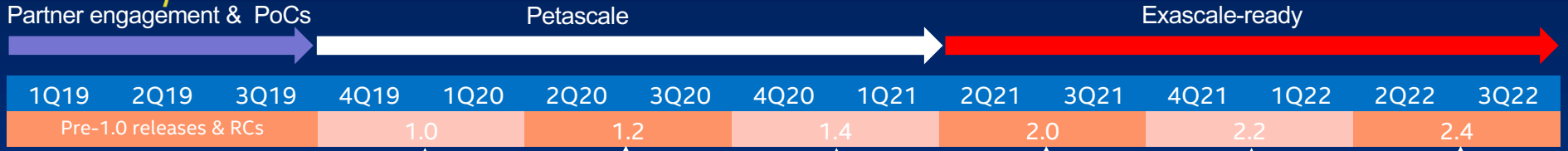
"We have already begun testing the open source DAOS system software on our internal test systems in collaboration with Intel and the results are as expected."
Kevin Harms, ALCF Performance Engineer Team Lead

- Aurora DAOS configuration
  - Capacity: 230PB
  - Bandwidth: >25TB/s

"The Argonne Leadership Computing Facility will be the first major production deployment of the DAOS storage system as part of Aurora, the first US exascale system coming in 2021. The DAOS storage system is designed to provide the levels of metadata operation rates and bandwidth required for I/O extensive workloads on an exascale-level machine."
Susan Coghlan, ALCF-X Project Director/Exascale Computing Systems Deputy Director

DAOS Community Roadmap – Q2 2019

Flash Memory Summit



DAOS:
- NVMe & DCPMM support
- python/golang API bindings
- Per-pool ACL
- Lustre integration

I/O Middleware:
- MPI-IO Driver
- HDF5 DAOS Connector
- POSIX I/O support
- Spark

DAOS:
- End-to-end data integrity
- Per-container ACL
- Improved control plane

DAOS:
- Online server addition
- Advanced control plane

I/O Middleware:
- POSIX data mover
- Async HDF5 operations over DAOS

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

I/O Middleware:
- Advanced POSIX I/O support
- Advanced data mover

DAOS:
- Progressive layout / GIGA+
- Placement optimizations
- Checksum scrubbing

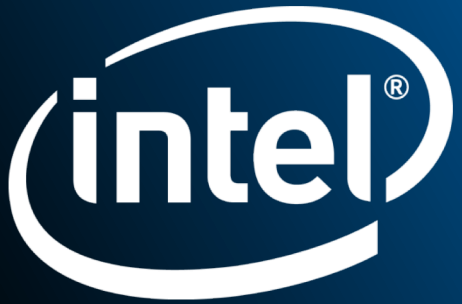
I/O Middleware:
- Apache Arrow (not POR)

DAOS:
- Catastrophic recovery tools

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

Resources

- ISC Demonstration animation
  - https://youtu.be/5RJbHwtHos0
- 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
  - http://daos.io/doc
- Community mailing list on Groups.io
  - daos@daos.groups.io
- Support
  - https://jira.hpdd.intel.com


intel®