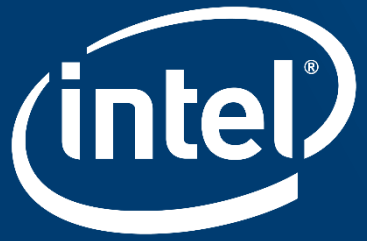


DISRUPTING HIGH PERFORMANCE STORAGE WITH INTEL DC PERSISTENT MEMORY & DAOS

John Carrier, Johann Lombardi

Extreme Storage Architecture & Development
Intel Corporation

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 .

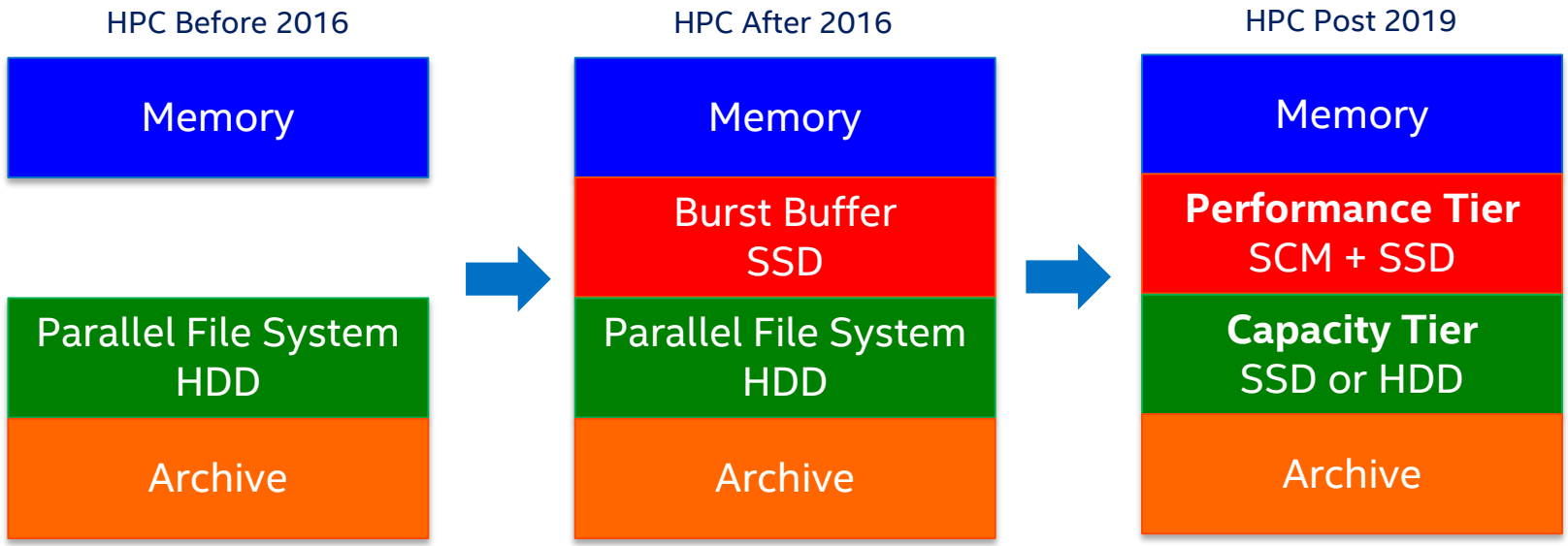
Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology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

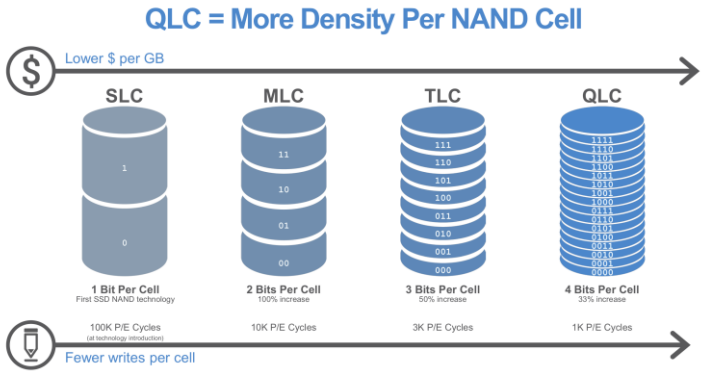
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality,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

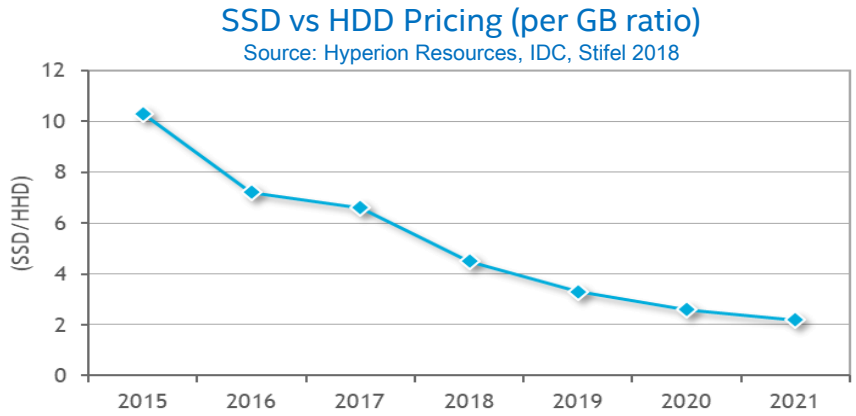
Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

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.

HIGH PERFORMANCE STORAGE EVOLUTION

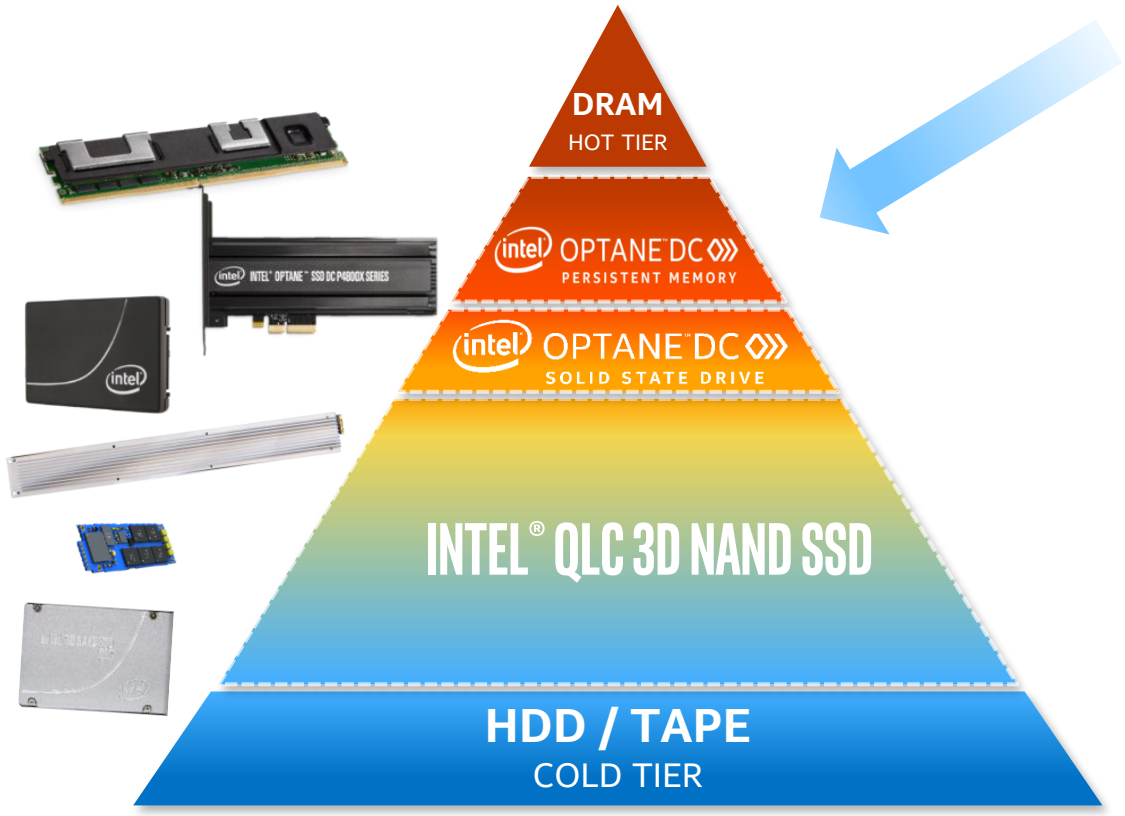






Source: Inspired from Gary Grider's presentation at SSIO'18 workshop

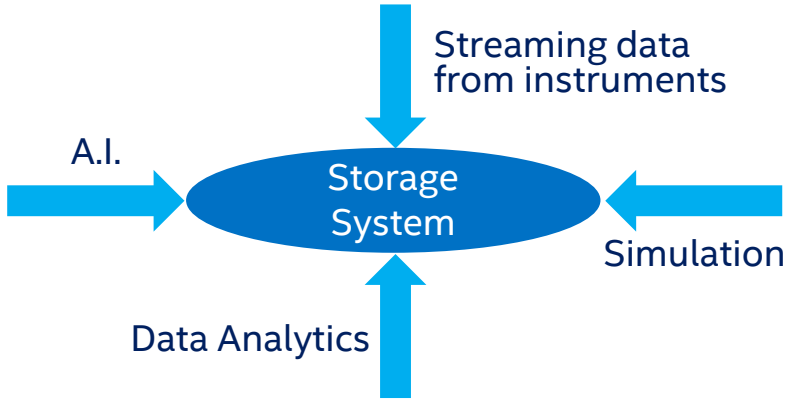
EVOLVING STORAGE TECHNOLOGIES



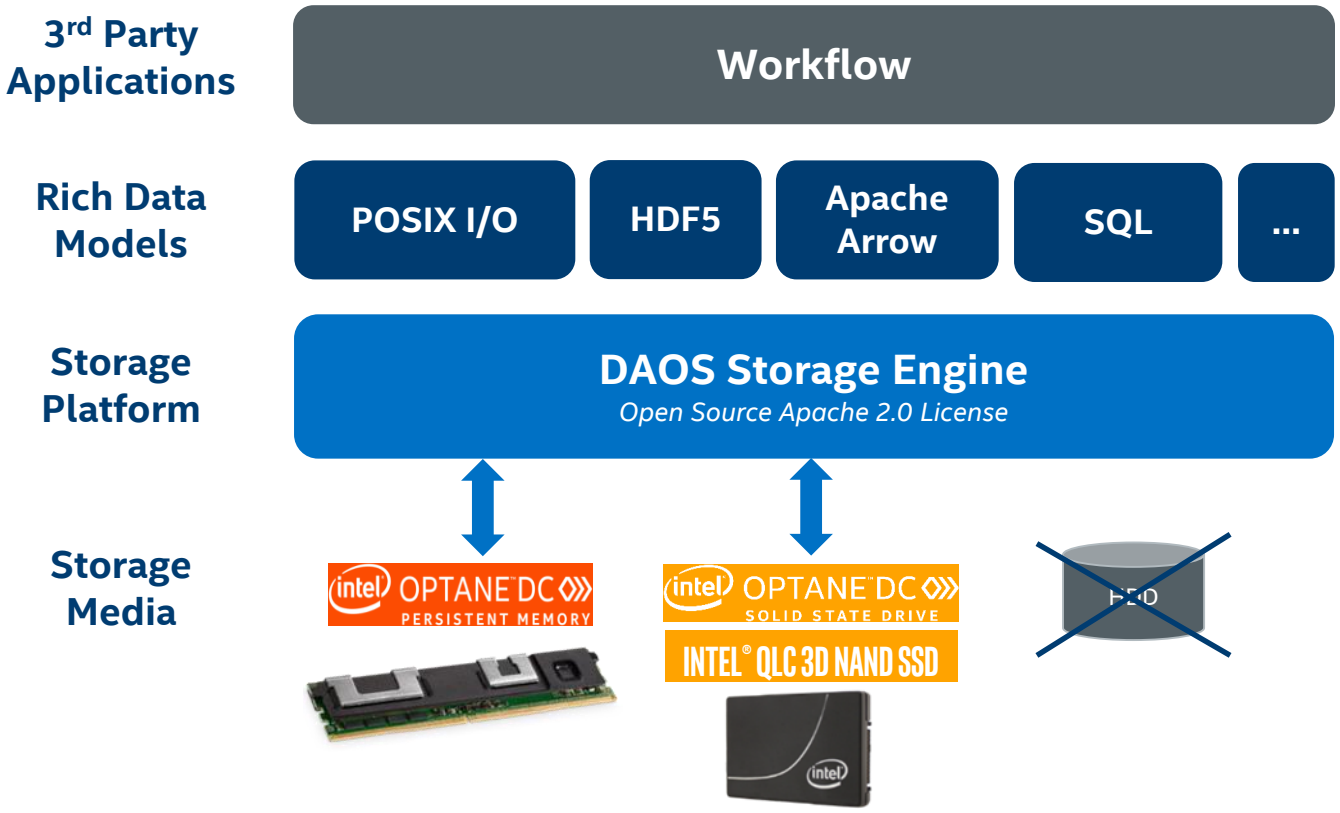
Storage Class Memory (SCM)

- Persistent, like storage
- Byte-addressable, like memory
- Lower latency, higher bandwidth, greater endurance than Flash
- Creates a new storage tier between DRAM and NAND SSDs

Challenge : exploit SCM for evolving storage workflows



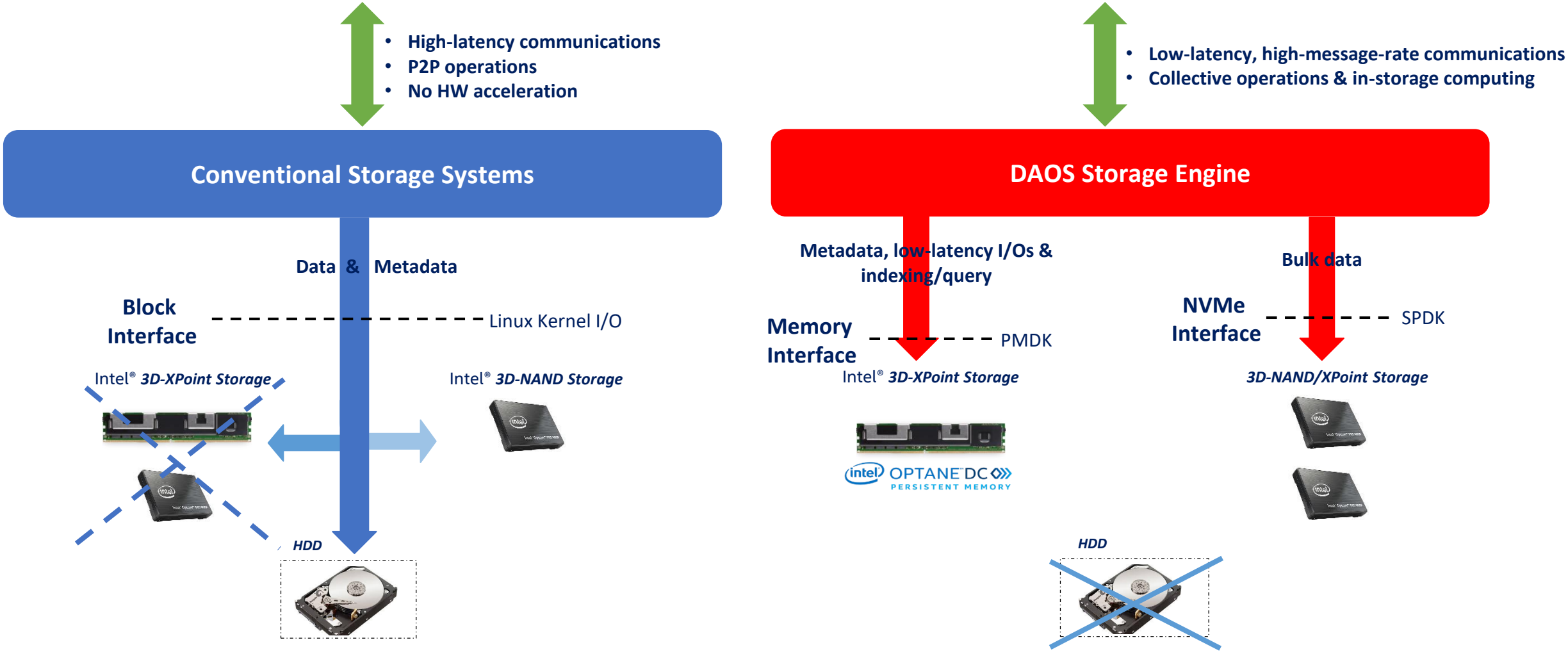
DAOS: SCALE-OUT SOFTWARE-DEFINED STORAGE



DAOS = Distributed Asynchronous Object Storage

- Intel DCPMM for metadata and small block data
- Intel NVMe SSDs for large block data
- Built natively over userspace PMEM/NVMe software stacks
- Ultra-fine grained I/O
- High throughput/IOPS @arbitrary alignment/size
- Software-managed redundancy
- Rich storage semantics
- Rely on COTS server hardware

DAOS ARCHITECTURE



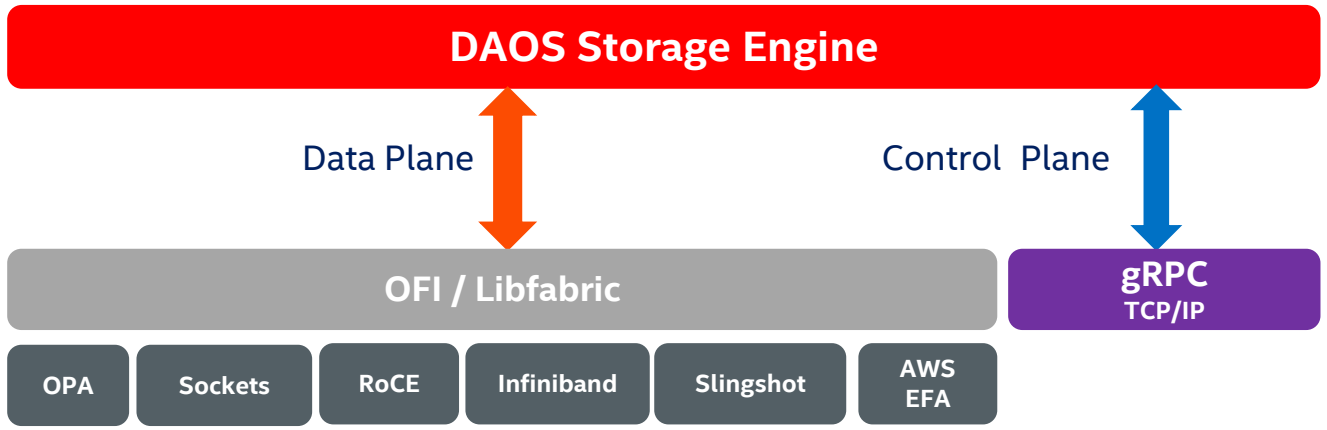
DAOS NETWORK SUPPORT

Performance-critical I/O path over libfabric

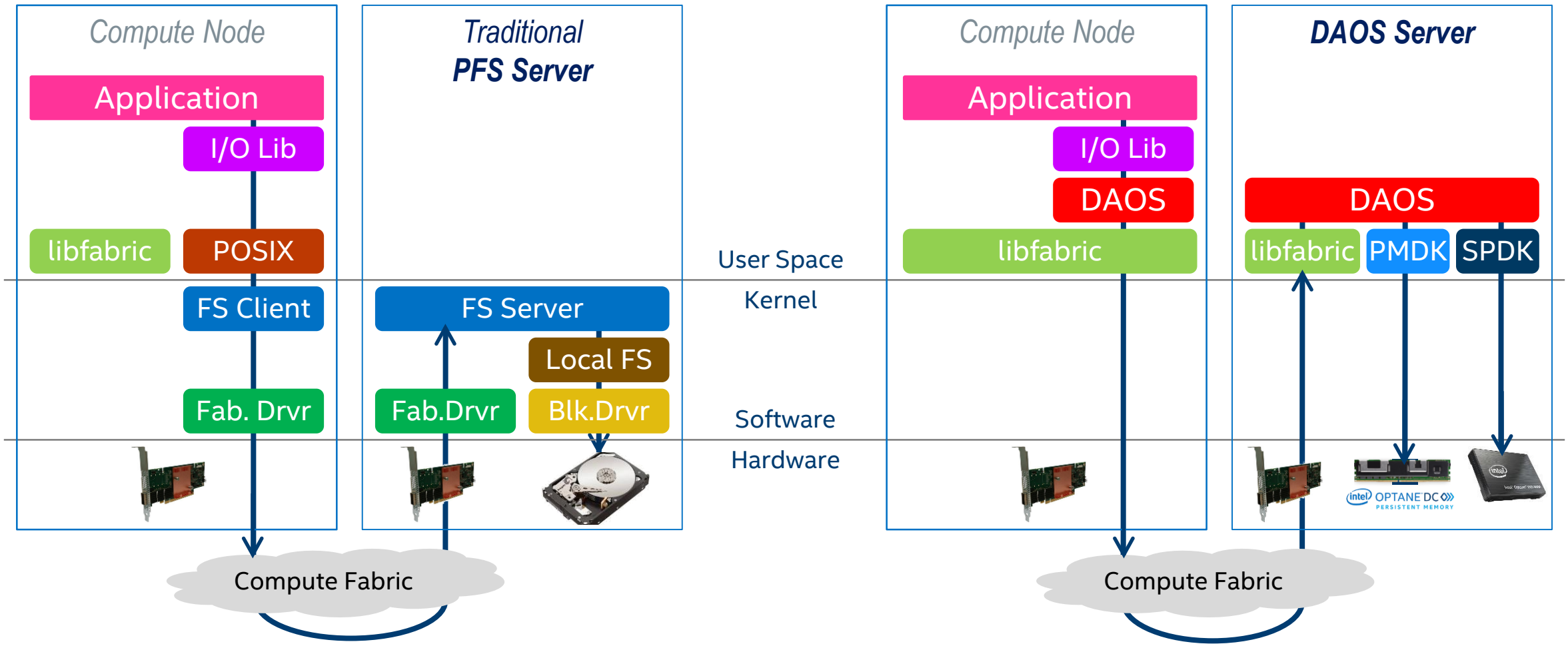
- Low-latency messaging
  - End-to-end in userspace
- Native support for RDMA
  - True zero-copy I/O
- Non-blocking
- Scalable collective communications

Out-of-band channel for administration

- Manage hardware, service & pools
- Telemetry & troubleshooting
- Secured with TLS & certificate



STORAGE SOFTWARE STACKS



DAOS & DATA ANALYSIS

Fast data retrieval

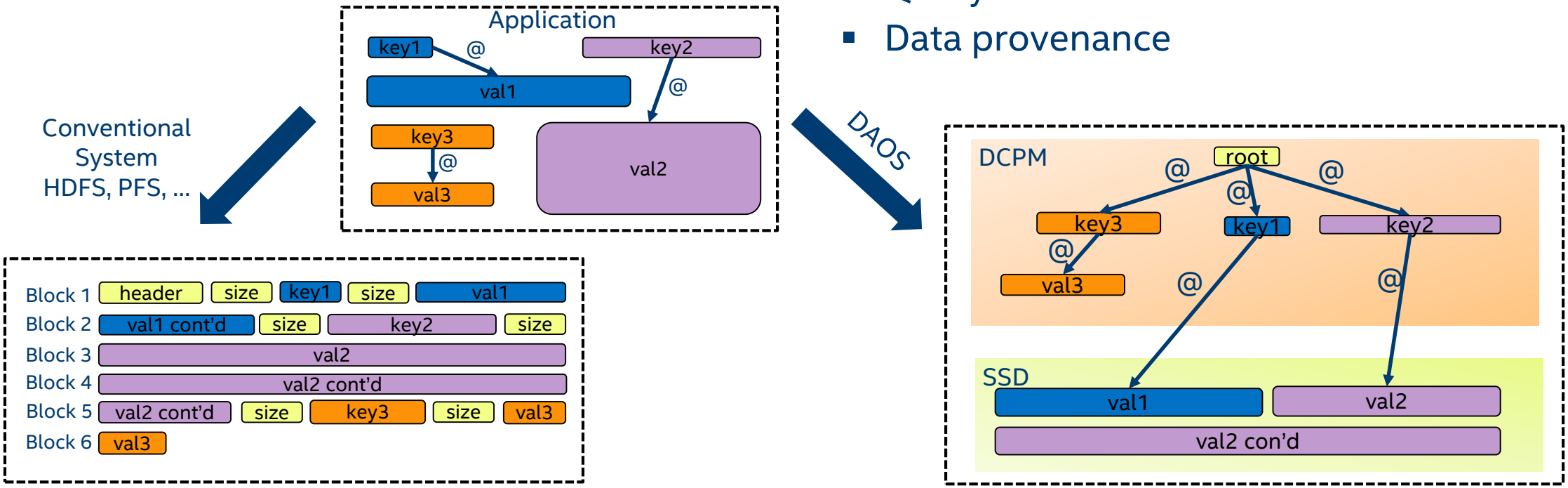
- Avoid file serialization and offset management
- Keys can be of any size/type
- Keys can be ordered with range query support

Scalable insert

- Allow concurrent access/update
- Distributed transactions keep KV store always consistent

Data indexing

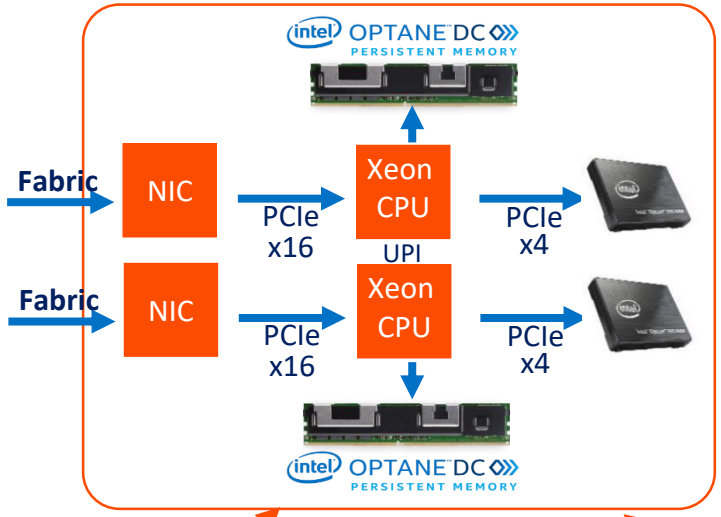
- Query & custom index
- Data provenance



DAOS DEPLOYMENT OPTIONS

Pooled Storage

- Storage pooled in separate racks
- High density storage servers
- Uniform storage access



Disaggregated Storage

- Storage integrated in compute rack
- Highly distributed storage
- Non-uniform storage access

DAOS Nodes
Intel® Xeon servers
with DCPMM & NVMe SSDs

Fabric

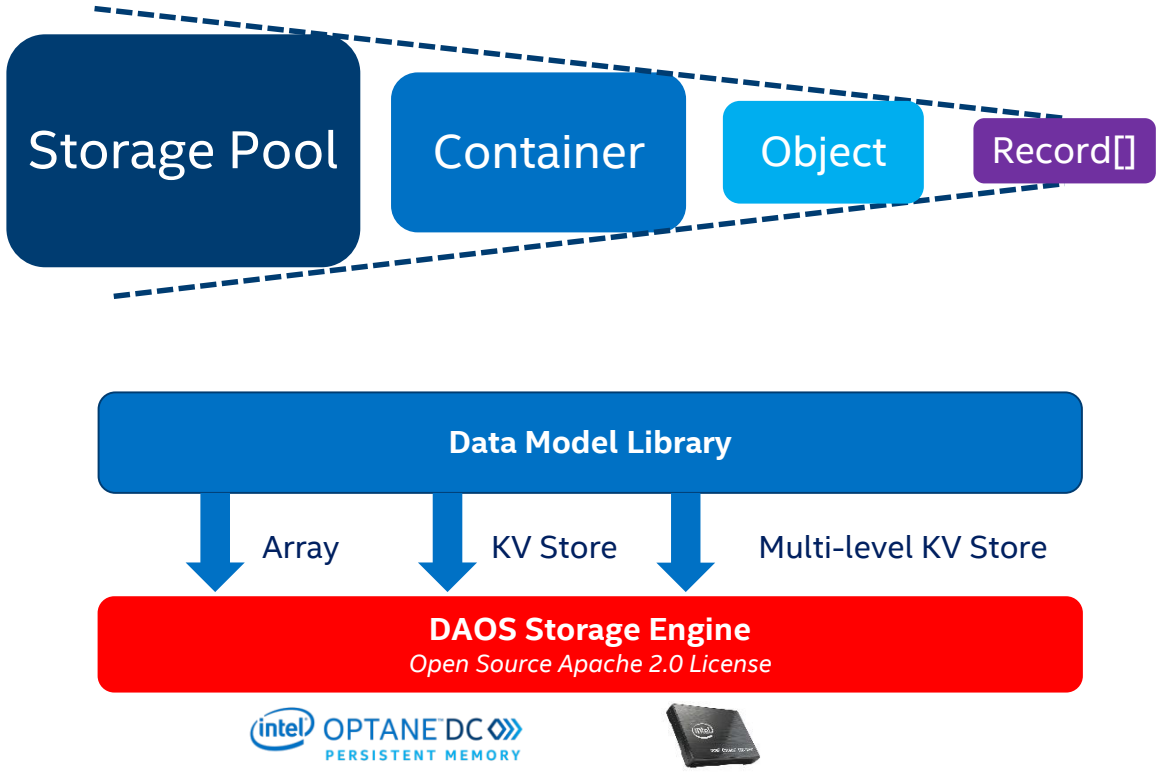
Gateway Nodes (GNs)
Intel® Xeon servers with no local storage

Capacity Tier
PFS, Cloud Object Store, ...

DAOS DATA MODEL

Non-POSIX rich storage API as the new foundation

- Scalable storage model suitable for both **structured & unstructured** data
  - key-value stores, multi-dimensional arrays, columnar databases, …
  - Accelerate data analytic/AI frameworks
- **Non-blocking** data & metadata operations
- **Extendable** through microservice architecture



STORAGE VIRTUALIZATION & MULTI-TENANCY

Distributed storage reservation

- Intel® Optane ™ DC Persistent Memory (DCPMM)
- NVMe SSD

Predicatable capacity

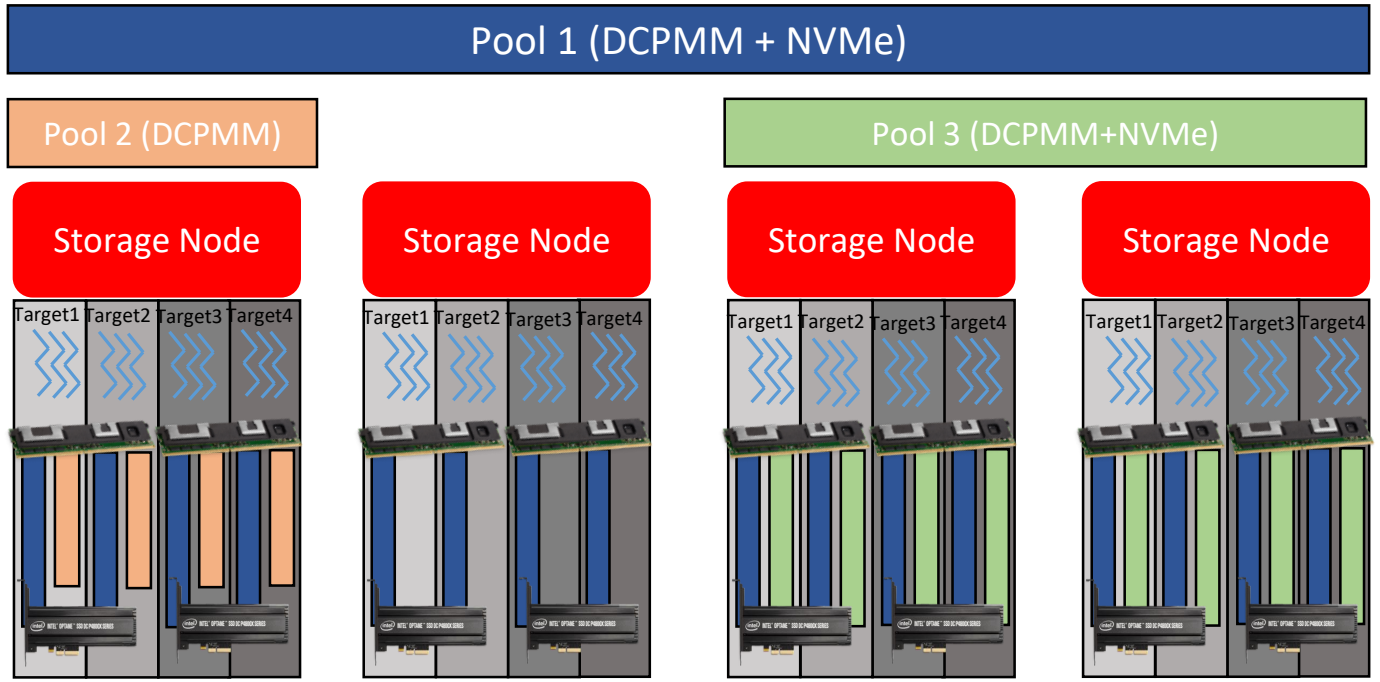
- Can be resized
- Can be extended to span more servers

Multi-tenancy

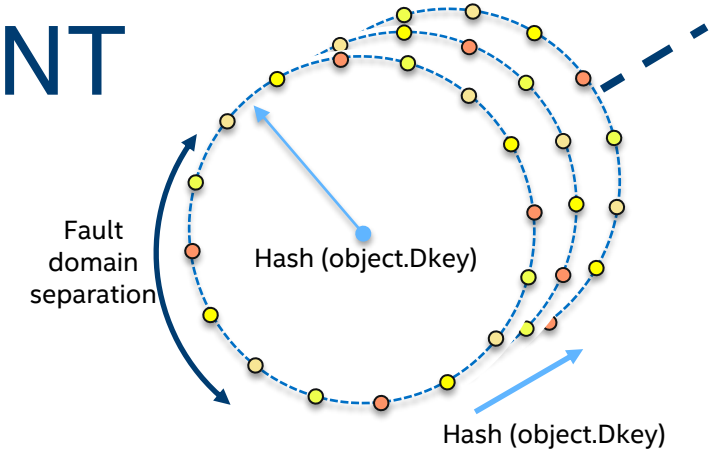
- NFSv4-type ACLs

Typically 1 pool = 1 project

- Can have a single pool or 100's
- Can be ephemeral (per-job) or persistent



DAOS DATA MANAGEMENT



End-to-end Data Integrity

- Supported algorithms (ISA-L)
  - CRC32, CRC64 & SHA
- Selectable on a per-container basis

Distributed transactions

- Serializable
  - Support database semantics
- Can involve any objects in a container

Data Sharding

- Algorithmic placement
- Progressive layout with GIGA+
- Selectable on a per-object basis

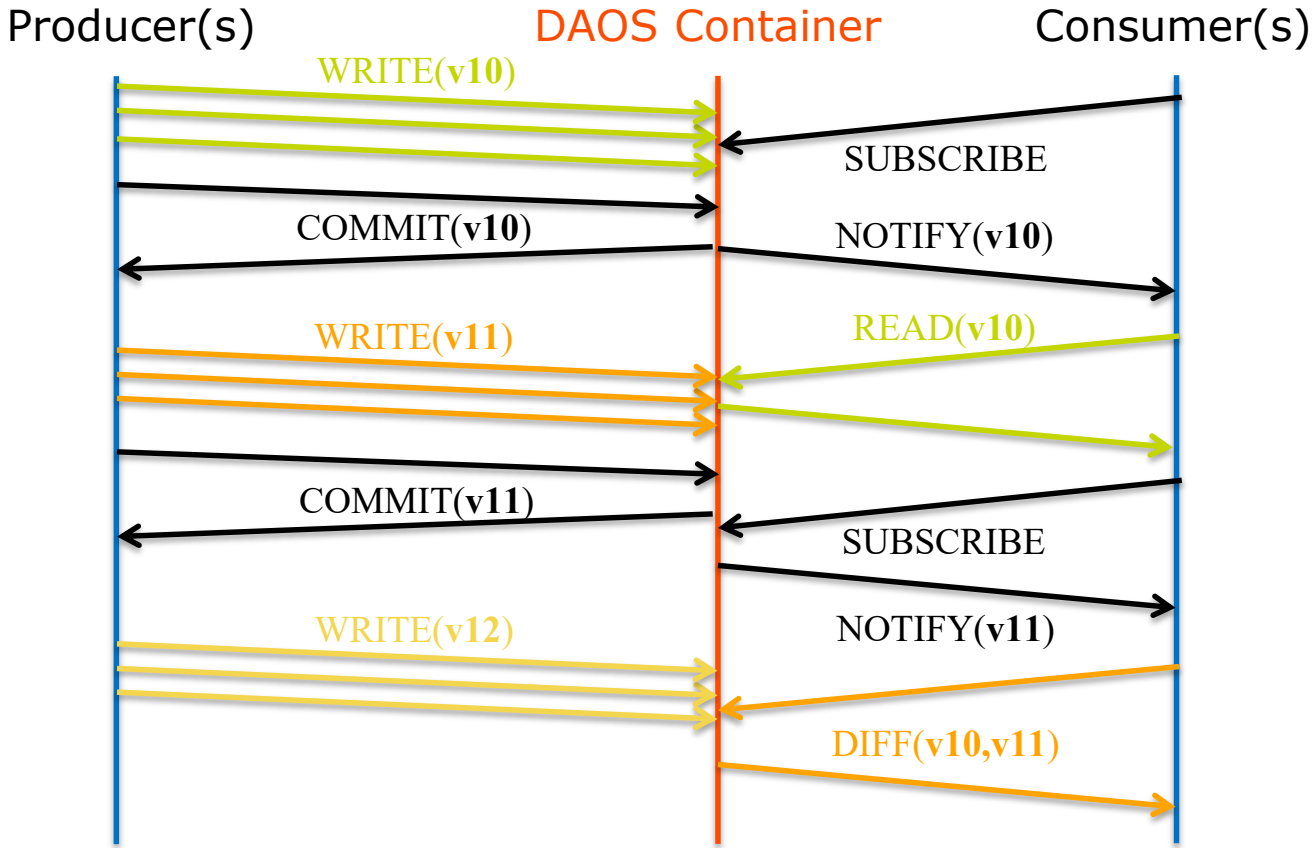
Data Protection

- Declustered replication & erasure code
- Fault-domain aware placement
- Self-healing
- Selectable on a per-object basis

DATA-DRIVEN WORKFLOW

DAOS simplifies developing complex workflows

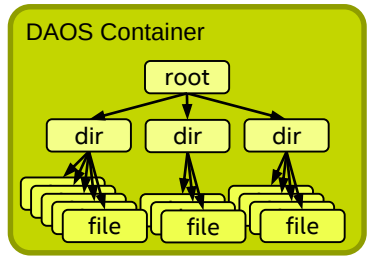
- **Native** producer/consumer **pipeline** support
- **Concurrency** & **dataflow** control
- Container **versioning**
  - Allow incremental update when maintaining external data structures
  - Many use cases like data indexing, visualization, ML, incremental backup, …



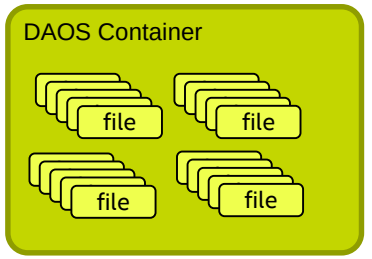
DAOS DATASET MANAGEMENT

Aggregate related datasets into manageable and coherent entities

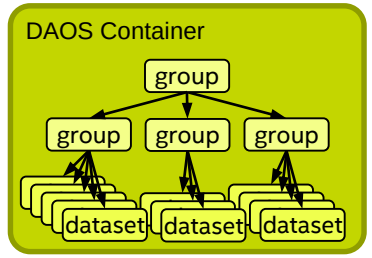
- Distributed consistency & automated recovery
- Full Versioning
- Simplified data management
  - Snapshot
  - Cross-tier Migration
  - Indexing
  - Needs for clones?



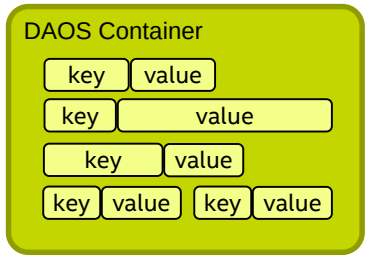
Encapsulated POSIX Namespace



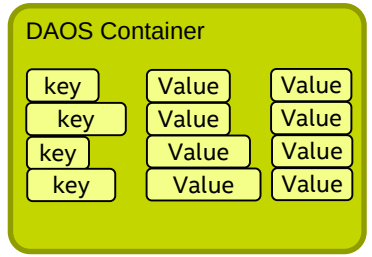
File-per-process



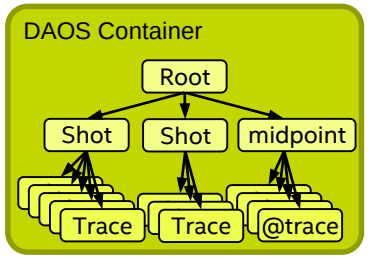
HDF5 « File »



Key-value store

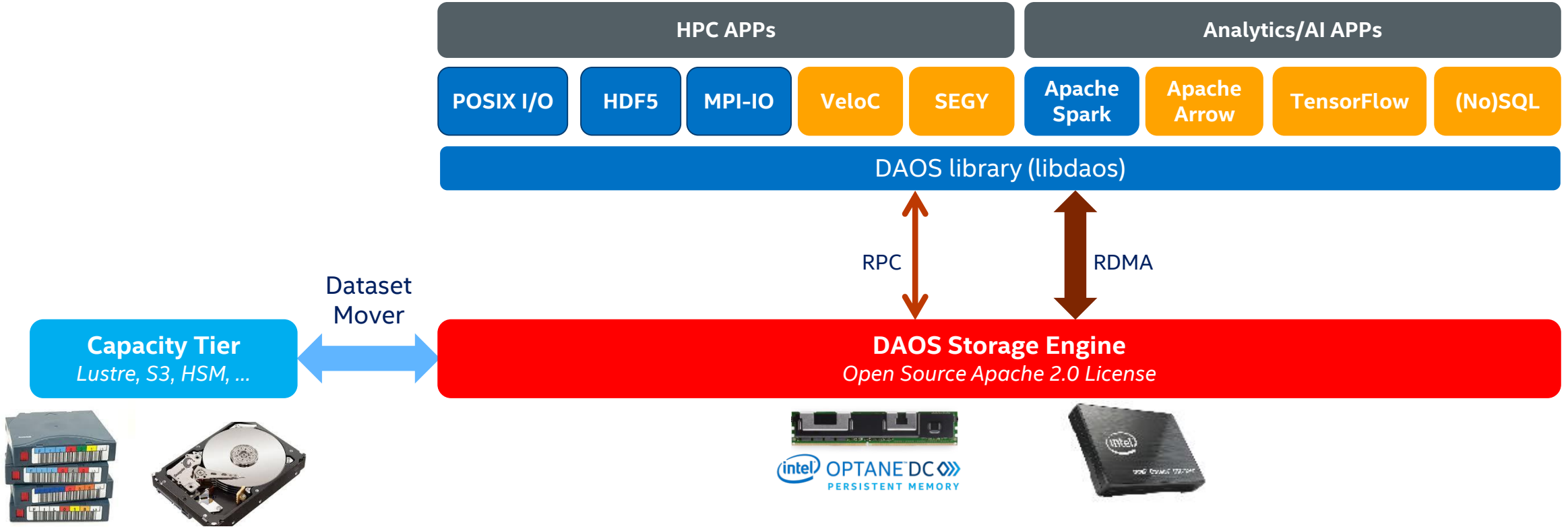


Columnar Database

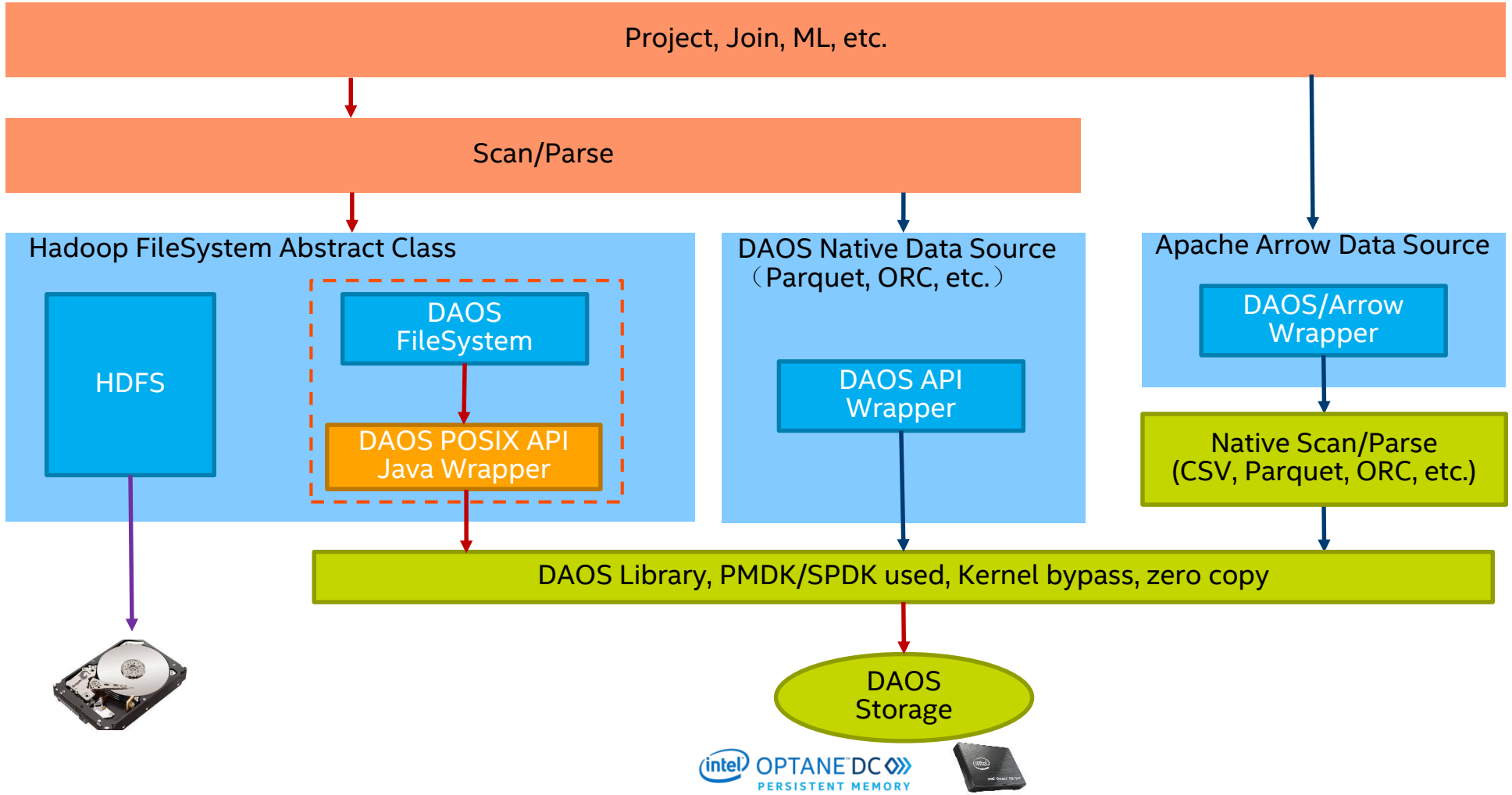


SEGY

DAOS APPLICATION INTERFACES



DAOS & BIG DATA / AI



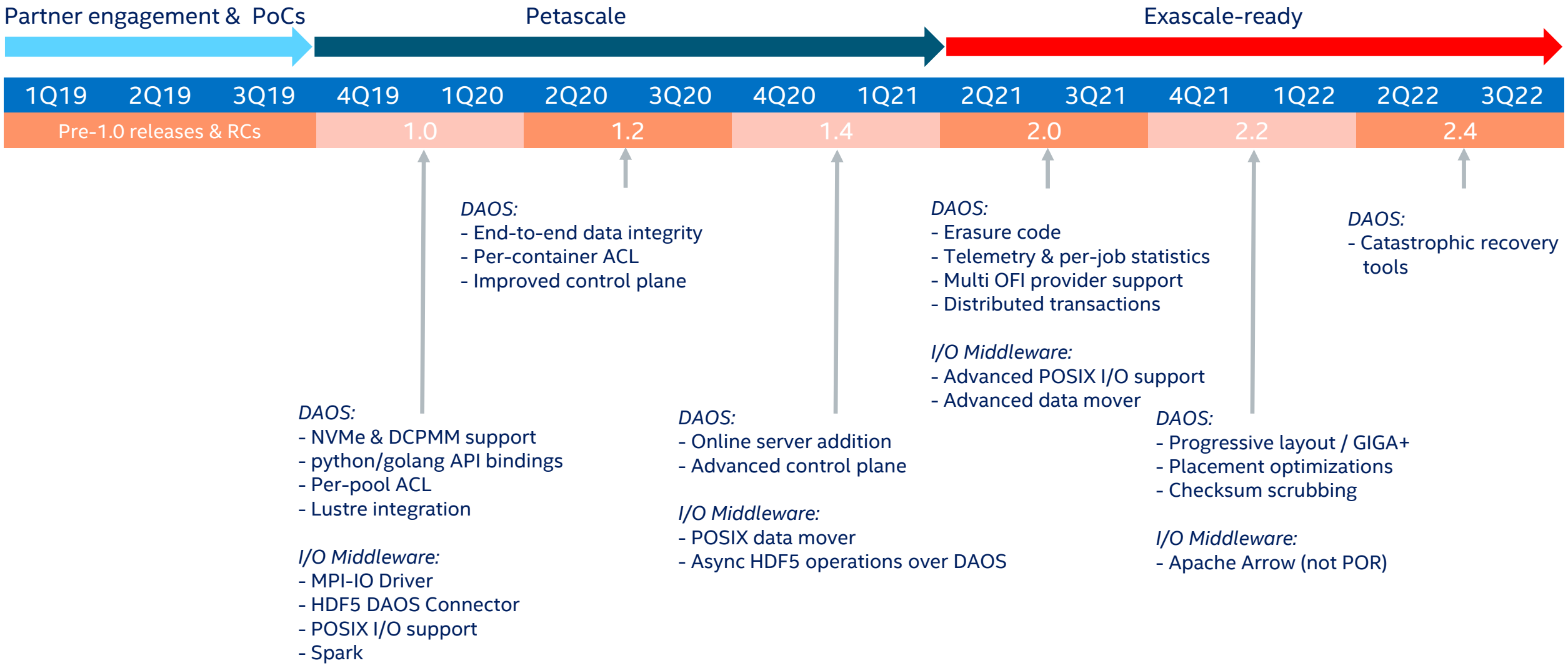
DAOS: PRIMARY STORAGE FOR AURORA

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



DAOS RESOURCES

Source code on GitHub

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

Documentation

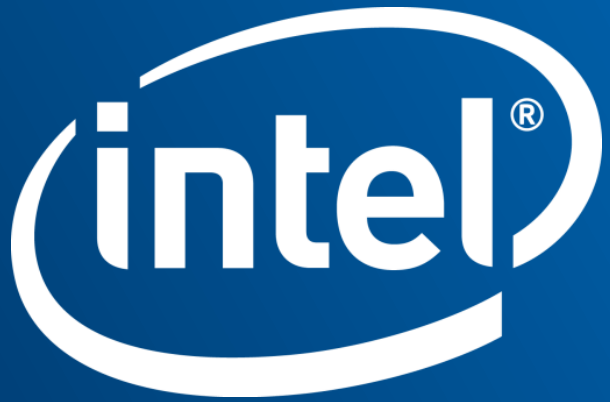
- http://daos.io

Community mailing list on Groups.io

- daos@daos.groups.io

Bug tracker & support

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


intel

POSIX I/O SUPPORT

DAOS File System (libdfs)

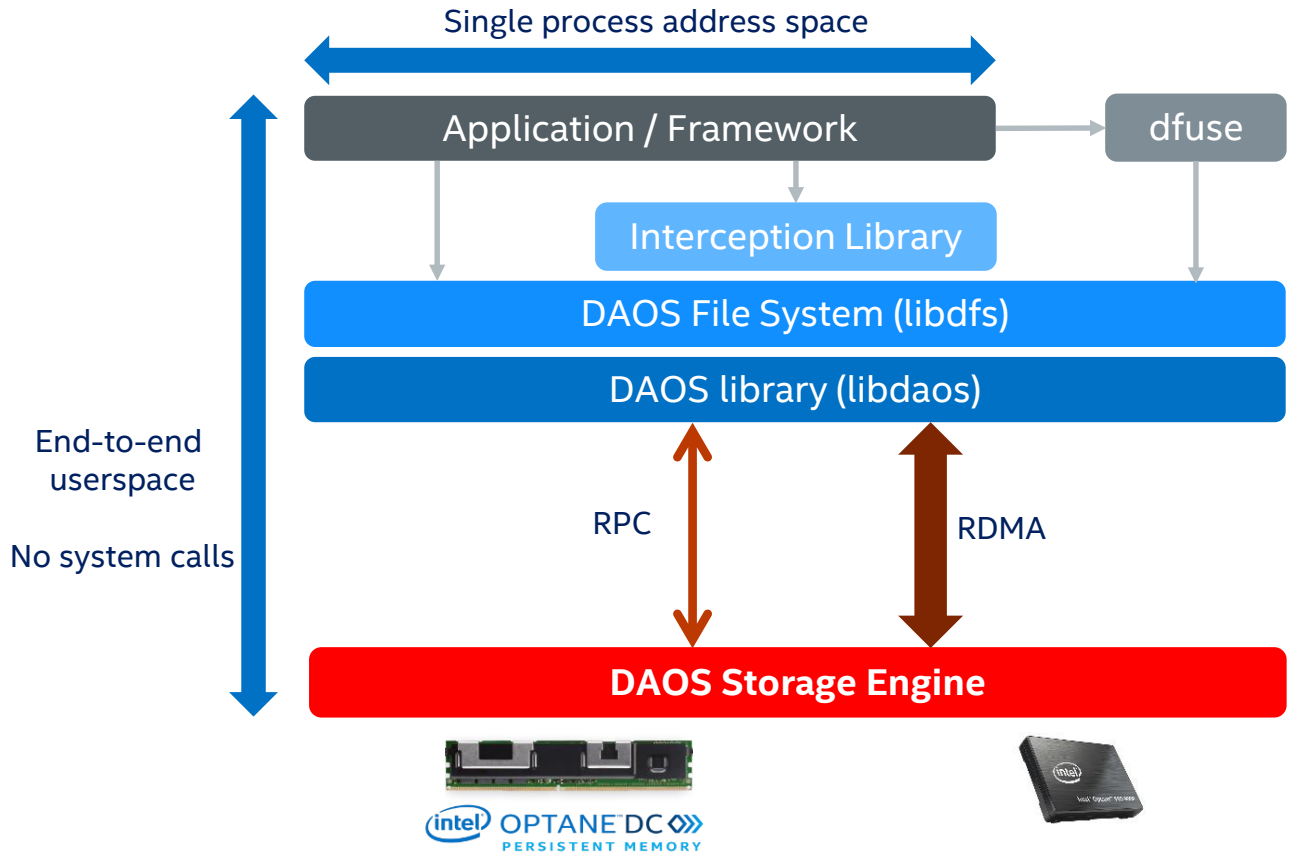
- Encapsulated POSIX namespace
- Application/framework can link directly with libdfs
  - ior/mdtest backend provided
  - MPI-IO driver leveraging collective open
  - TensorFlow, …

FUSE Daemon (dfuse)

- Transparent access to DAOS
- Involve system calls

I/O interception library

- OS bypass for read/write operations



PERFORMANCE

Demonstrated at ISC

- Deliver HW performance
  - Saturate SSD bandwidth with large blocks
  - Latency/IOPS of persistent memory for metadata & small I/Os
  - Only need a few clients to reach max performance
- ISC demo available on-line
  - https://www.youtube.com/watch?v=EMGBcvnftwQ
  - https://www.youtube.com/watch?v=e69Rgz2FMbE