E&G ECO ENGINEERING

DISRUPTING HIGH PERFORMANCE STORAGE WITH INTEL DC PERSISTENT MEMORY & DAOS

Mohamad Chaarawi

September 26, 2019

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.

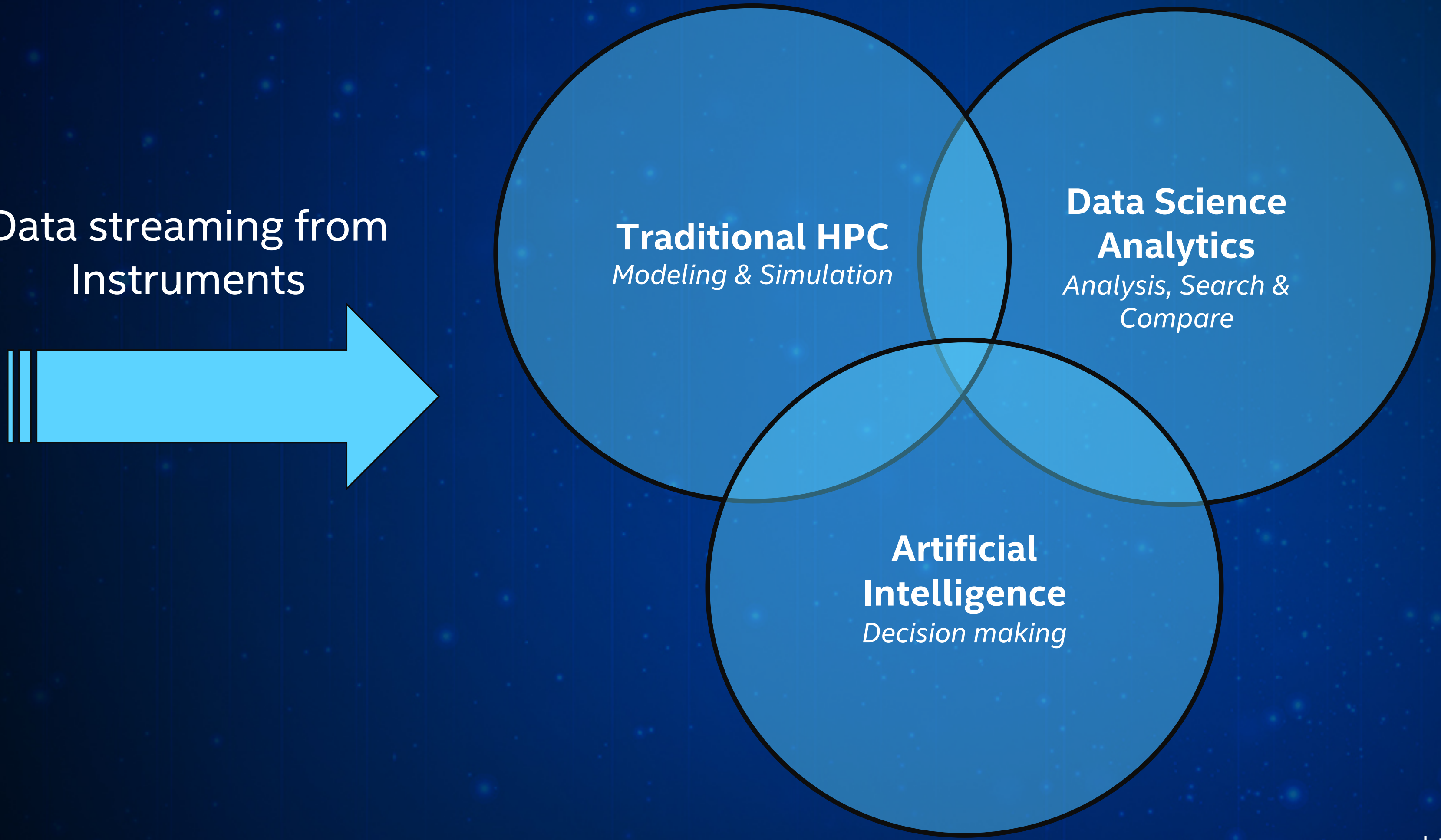
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality,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

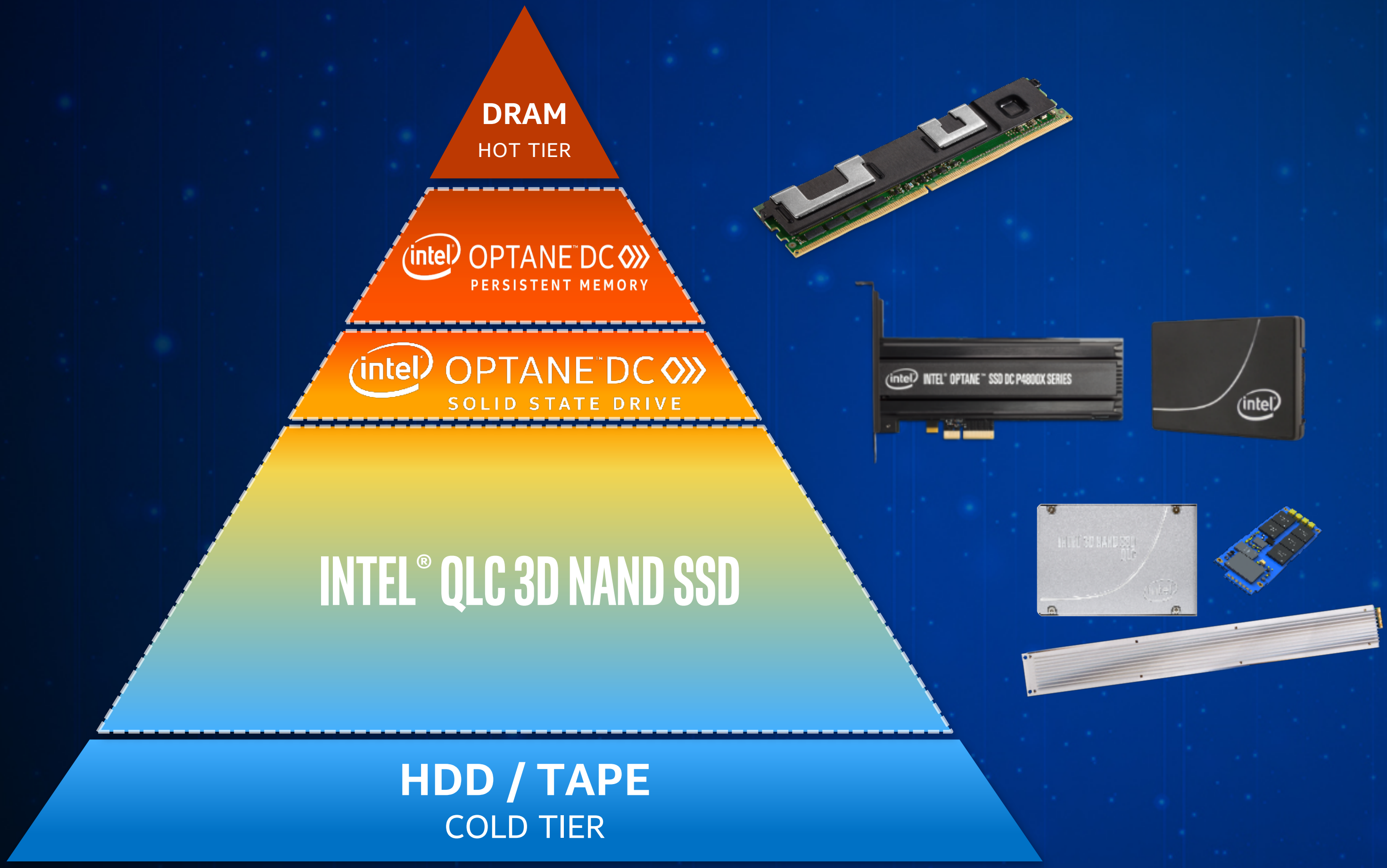
Intel, the Intel logo, Intel Optane and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as property of others.

EVOLVING STORAGE NEEDS


Data streaming from Instruments
Traditional HPC
Modeling & Simulation
Data Science Analytics
Analysis, Search & Compare
Artificial Intelligence
Decision making

https://www.orau.gov/ssioworkshop2018/agenda.htm

EVOLVING STORAGE TECHNOLOGIES

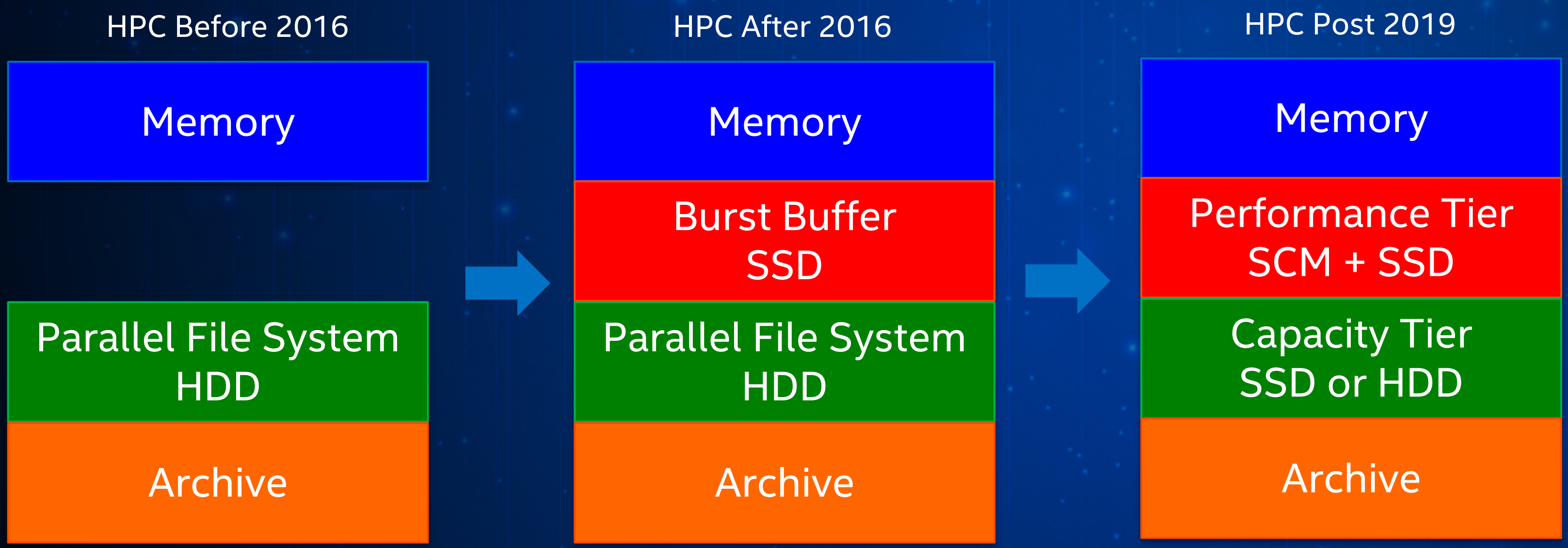


Storage Class Memory:

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

Challenge: exploit SCM for evolving storage workloads.

HIGH PERFORMANCE STORAGE EVOLUTION



HPC Before 2016

Memory

Parallel File System HDD

Archive

HPC After 2016

Memory

Burst Buffer SSD

Parallel File System HDD

Archive

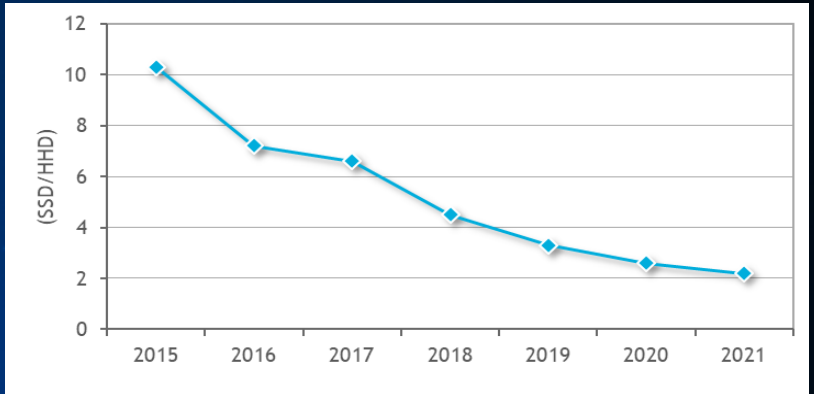
HPC Post 2019

Memory

Performance Tier SCM + SSD

Capacity Tier SSD or HDD

Archive

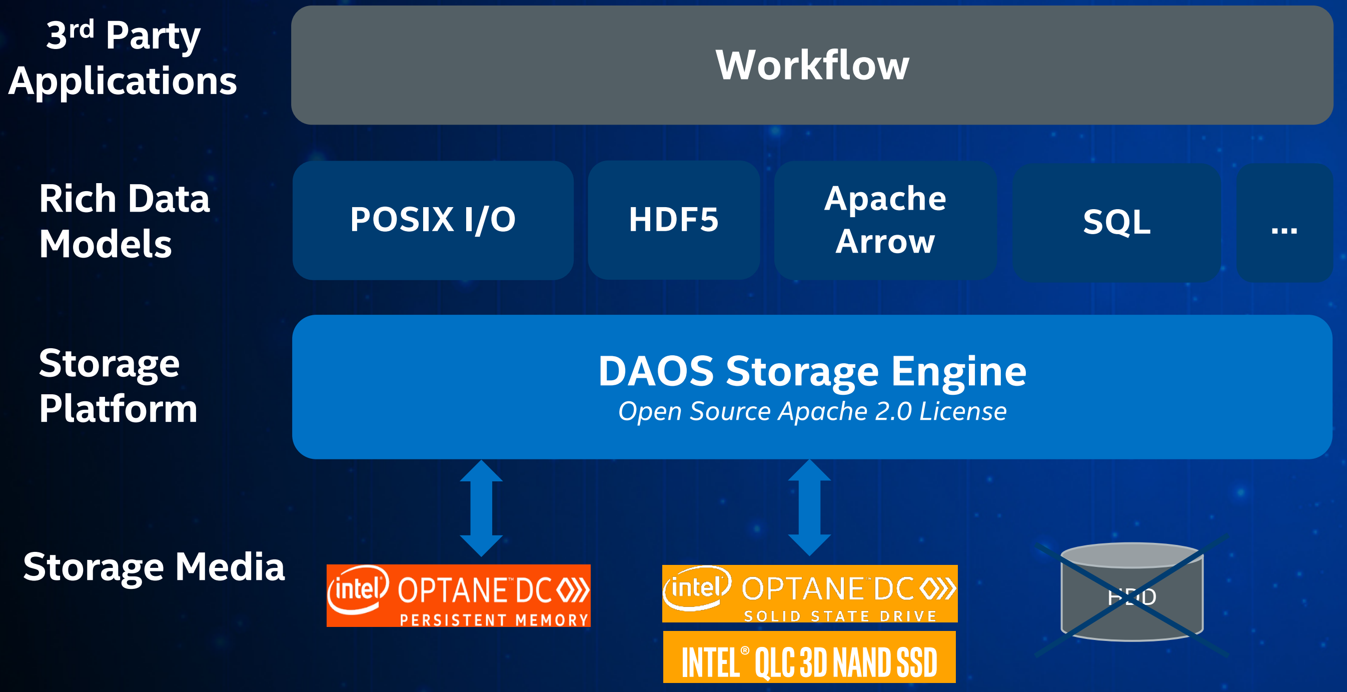


SSD vs HDD Pricing (per GB ratio)

Source: Hyperion Resources, IDC, Stifel 2018

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

DISTRIBUTED ASYNCHRONOUS OBJECT STORAGE

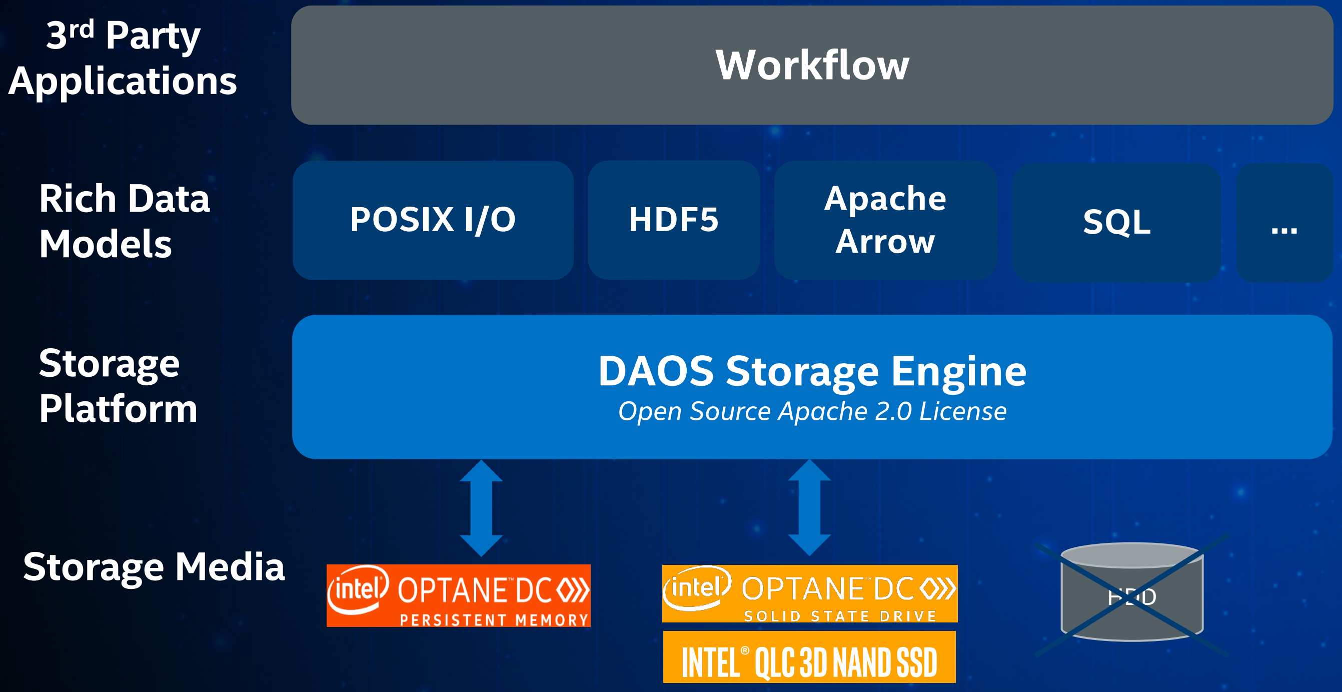


Benefits

- Built natively over **new userspace** PMEM/NVMe software stack
- **Rich** storage semantics
- High **throughput/IOPS @arbitrary** alignment/size
- **Fine-grained**, **low-latency** & True **zero-copy** I/Os
- **Scalable** communications
- **Software**-managed **redundancy**
- Rely on **COTS** hardware

open source initiative Approved License

DISTRIBUTED ASYNCHRONOUS OBJECT STORAGE



Benefits

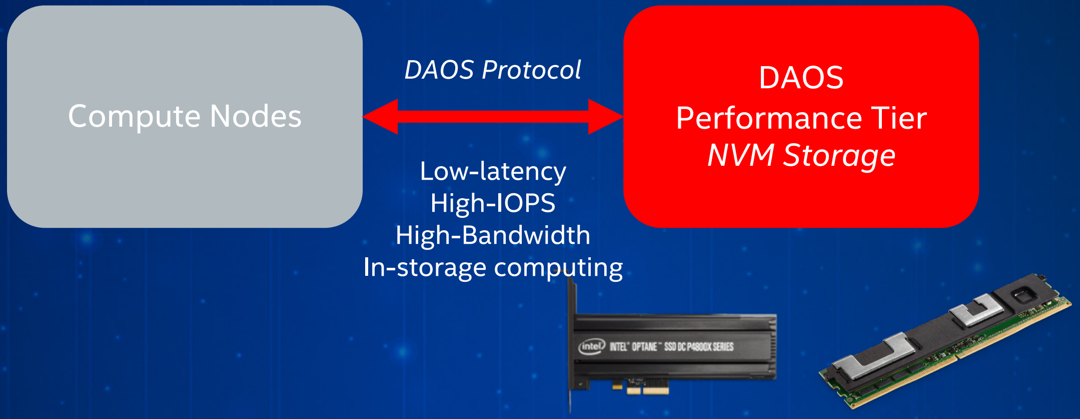
- Built natively over e PMEM/NVMe so
- **Rich** storage
- High **thro** **@arbitrary** alignm
- **Fin** **-latency** & T I/Os
- mmunications
- e-managed **redundancy**
- on **COTS** hardware

Open source

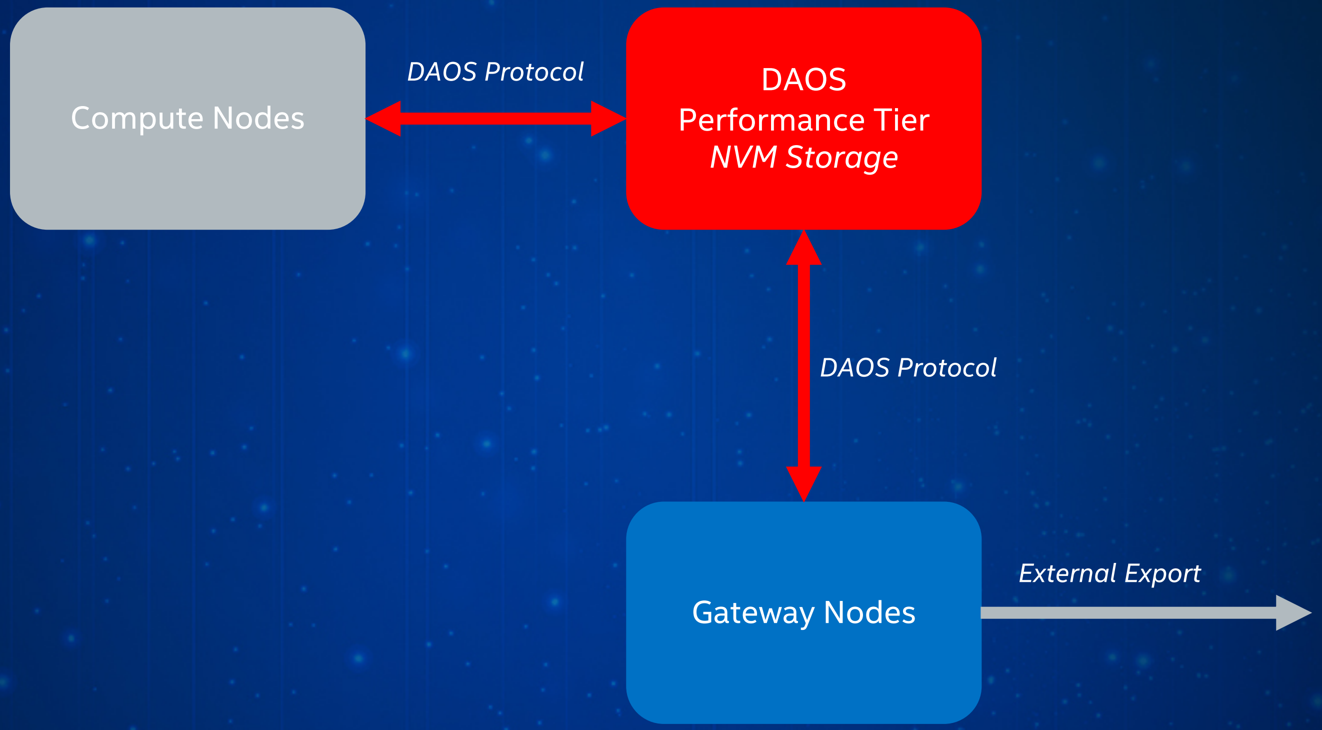
APACHE 2.0 License

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

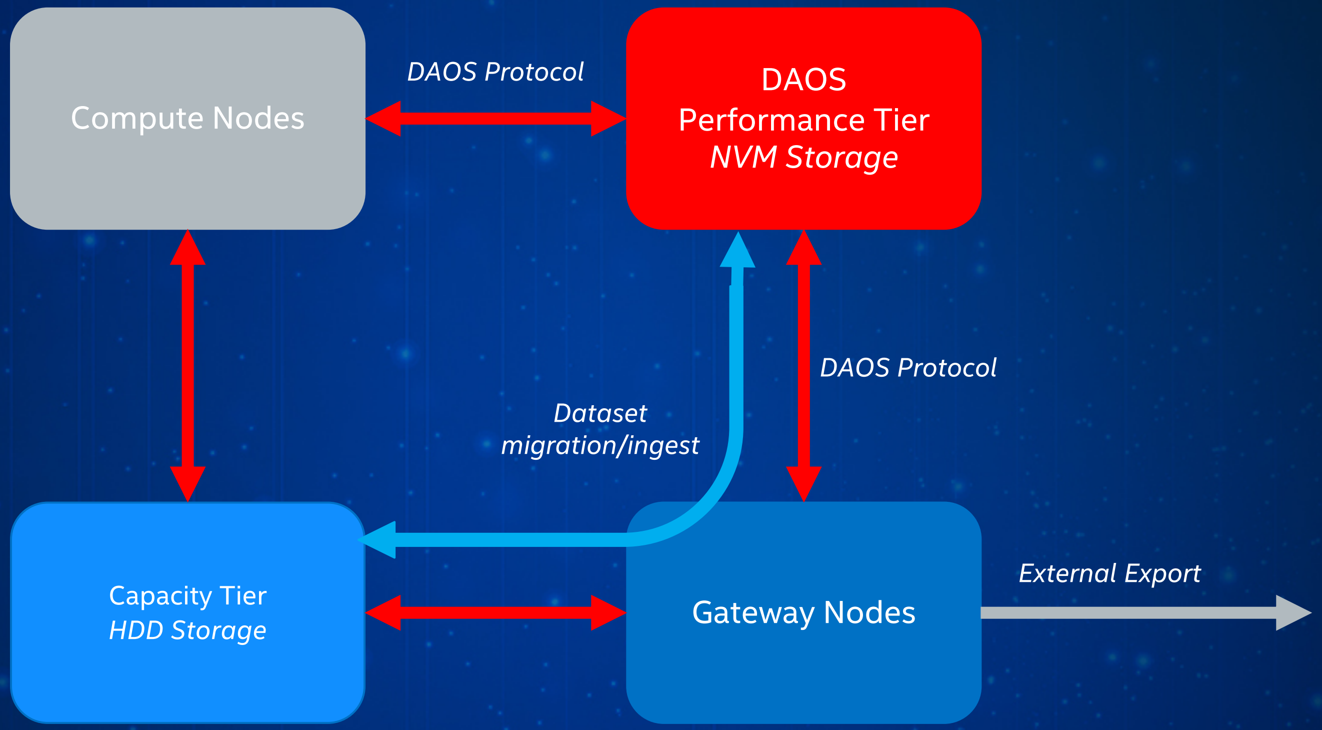
STORAGE ARCHITECTURE



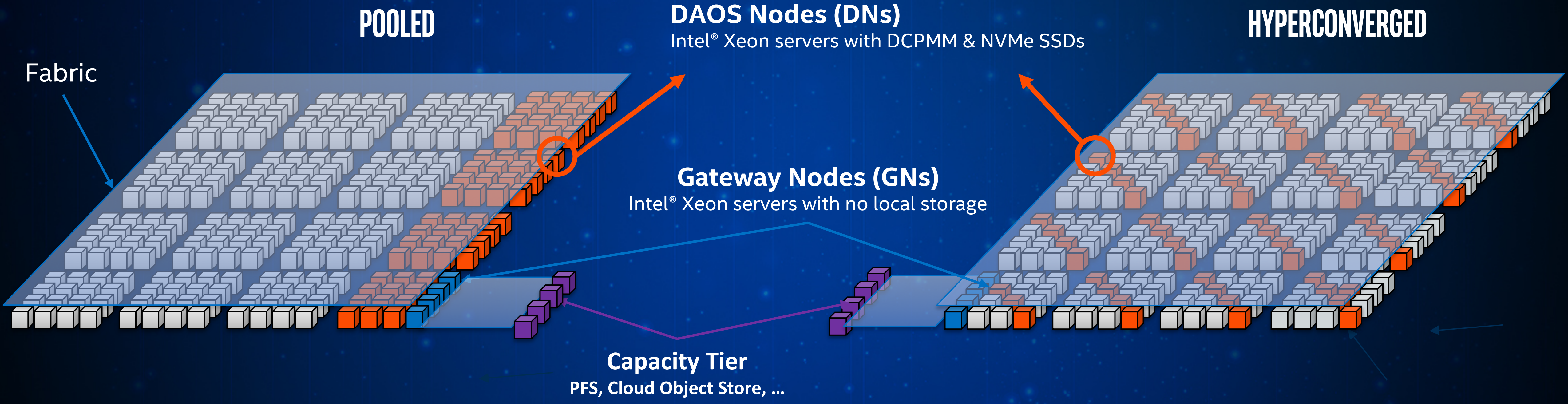
STORAGE ARCHITECTURE



STORAGE ARCHITECTURE



DAOS DEPLOYMENTS



DAOS Nodes (DNs)
Intel® Xeon servers with DCPMM & NVMe SSDs

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

Capacity Tier
PFS, Cloud Object Store, …

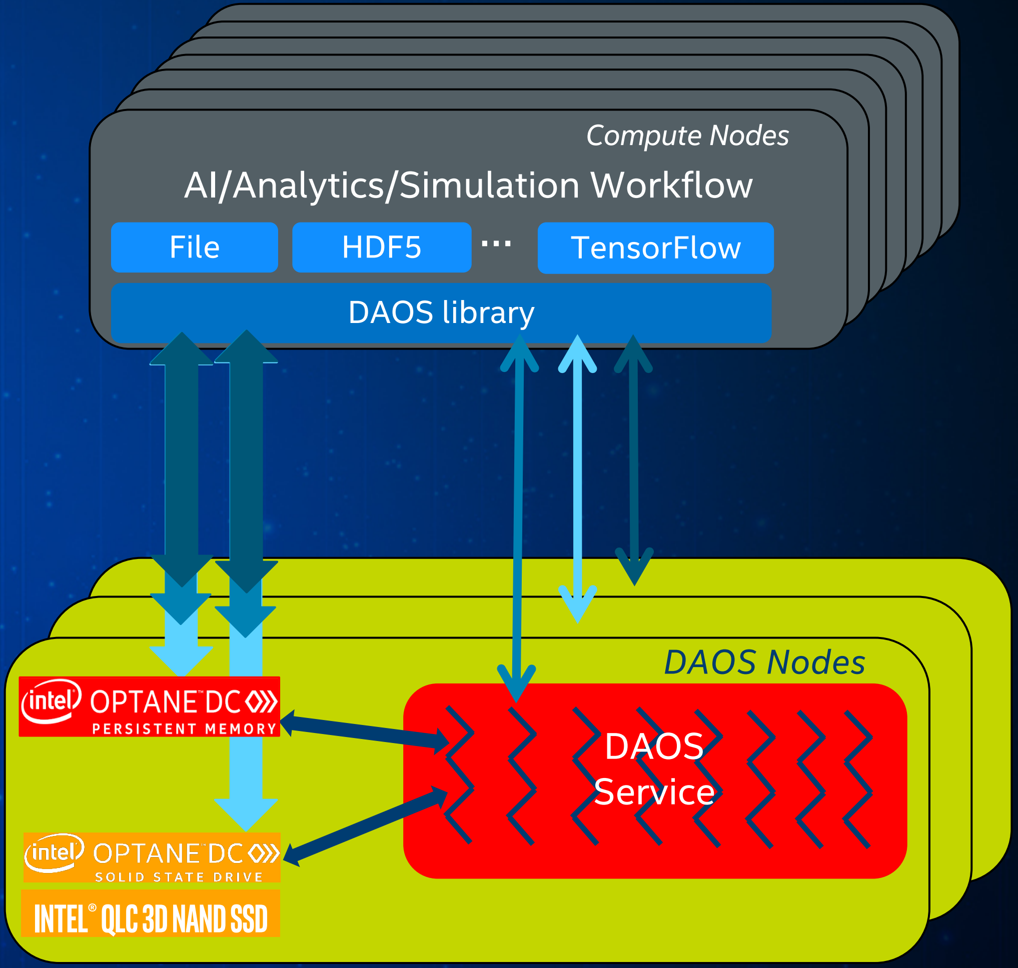
DAOS TIER ANATOMY

DAOS Tier

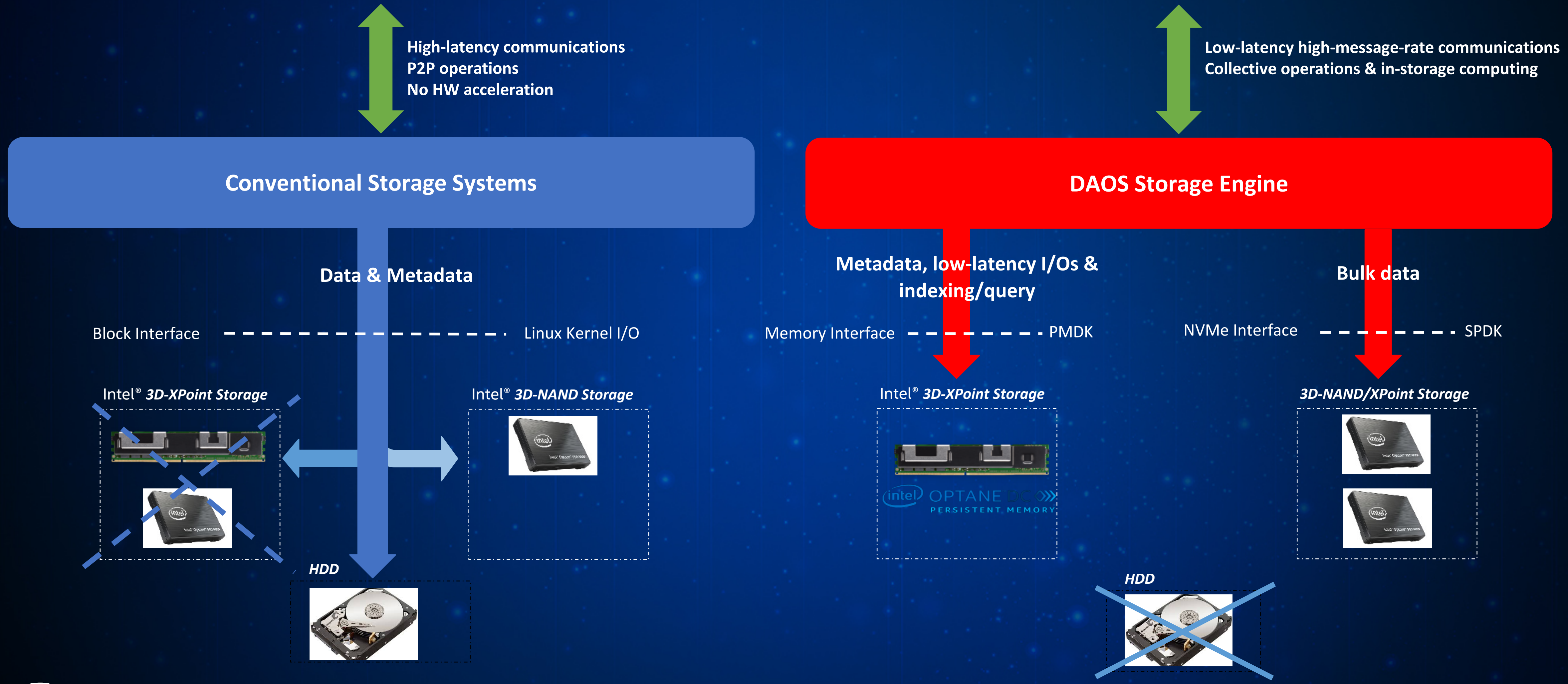
- Globally accessible from any compute nodes
- Large capacity (100's PB)

DAOS Nodes

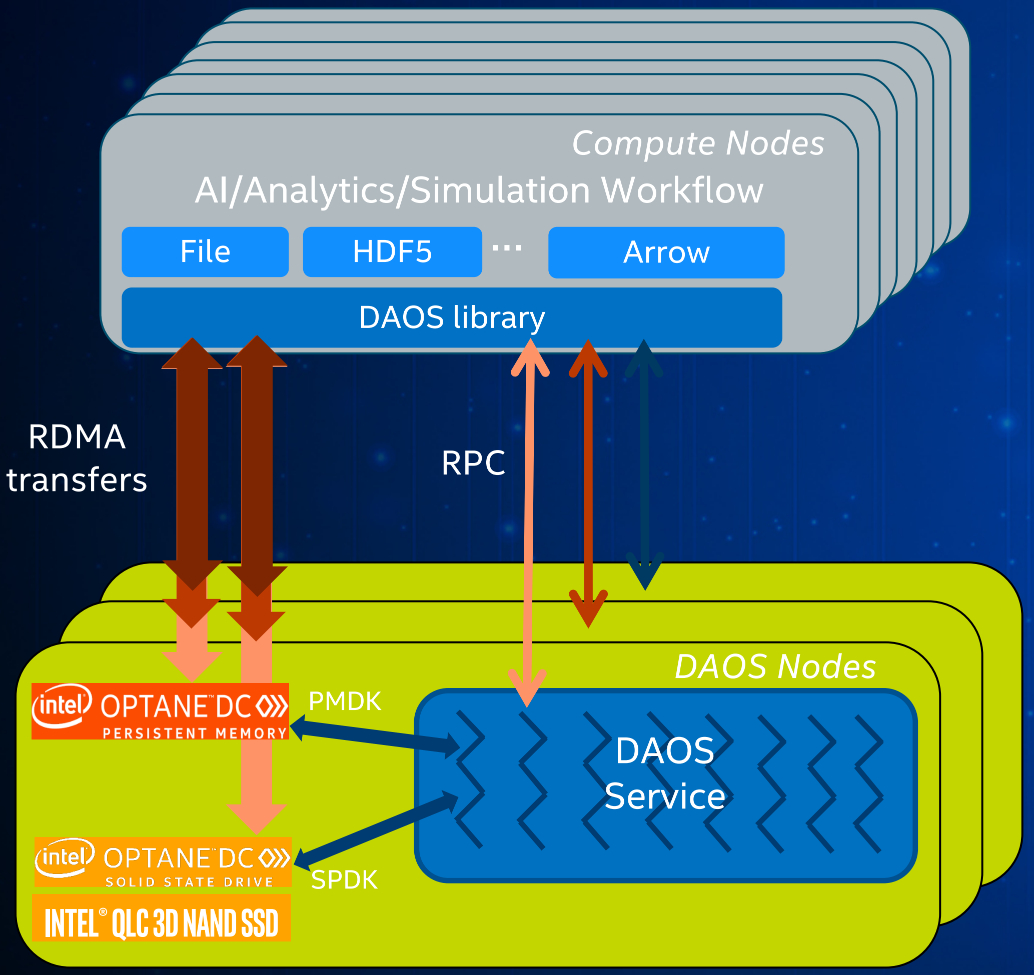
- COTS Intel® Xeon servers running the DAOS service
- RNIC attached for communications
  - Support multiple RNICs per server to sustain backend storage IOPS/bandwidth
- Mix of storage technologies attached
  - Intel® Optane™ DC Persistent Memory (DCPMM)
  - NVMe SSD (*NAND, Intel® Optane™ SSDs)

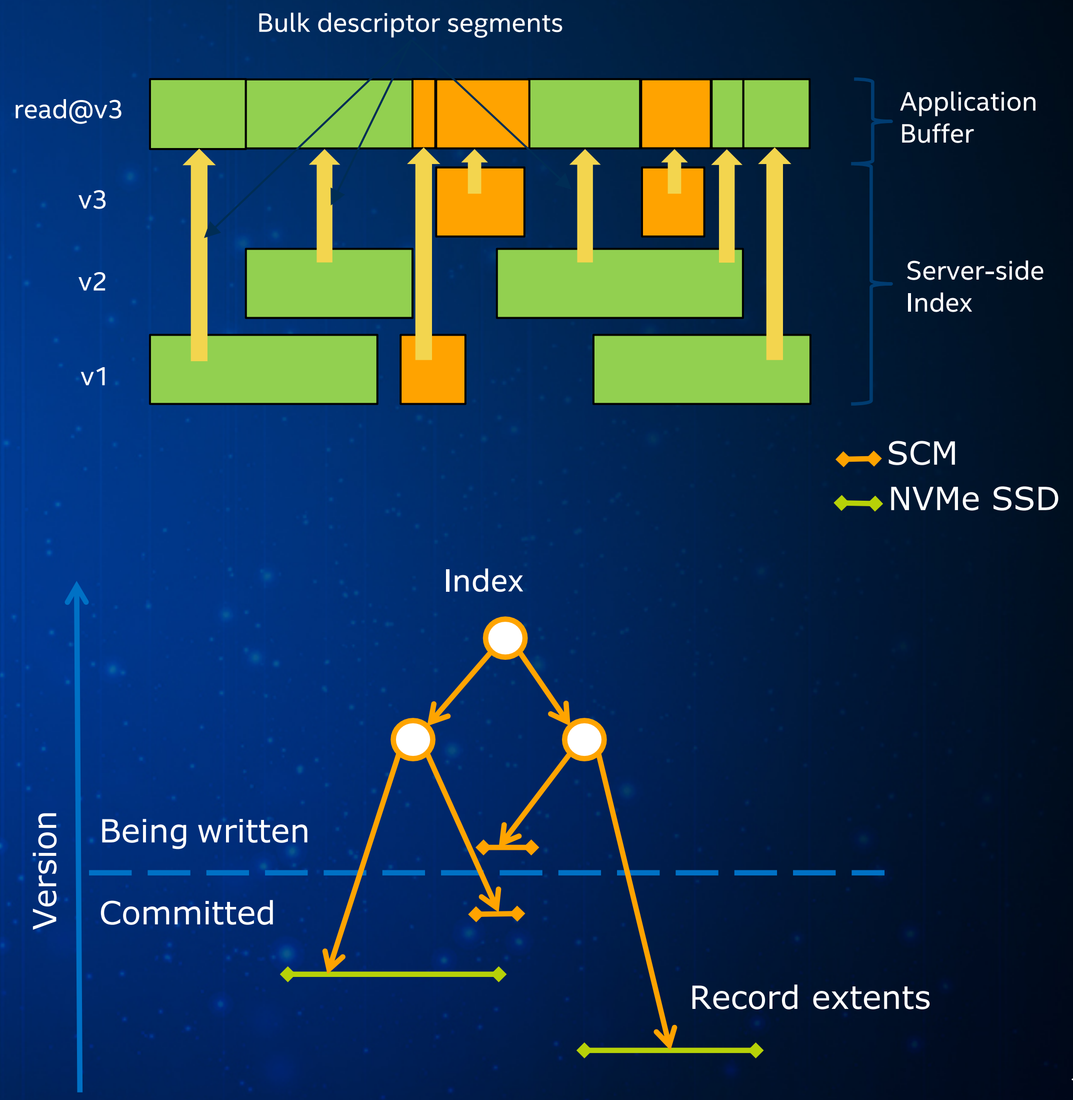

Compute Nodes
AI/Analytics/Simulation Workflow
File
HDF5
…
TensorFlow
DAOS library
DAOS Nodes
intel OPTANE DC PERSISTENT MEMORY
DAOS Service
intel OPTANE DC SOLID STATE DRIVE
INTEL® QLC 3D NAND SSD

DAOS ARCHITECTURE



LIGHTWEIGHT I/O STACK & FINE-GRAINED I/O





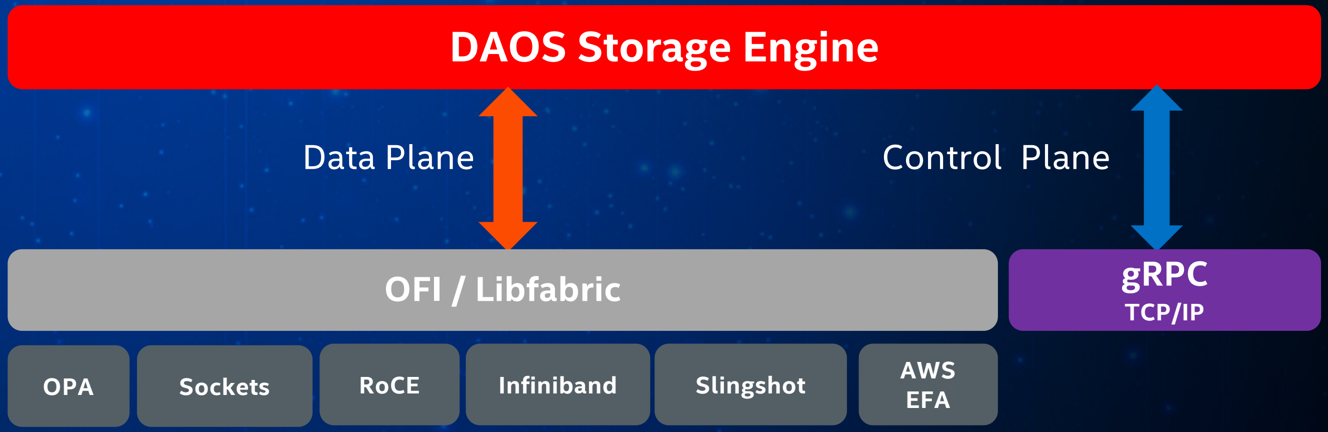
NETWORK SUPPORT

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

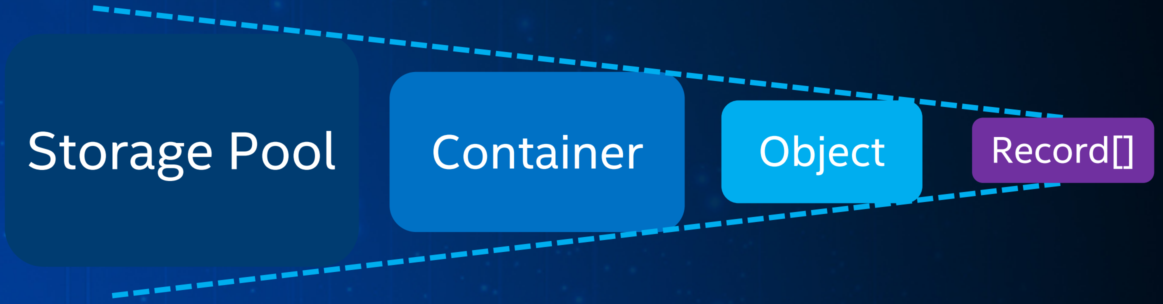
- Manage hardware, service & pools
- Telemetry & troubleshooting
- Secured with TLS & certificate



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

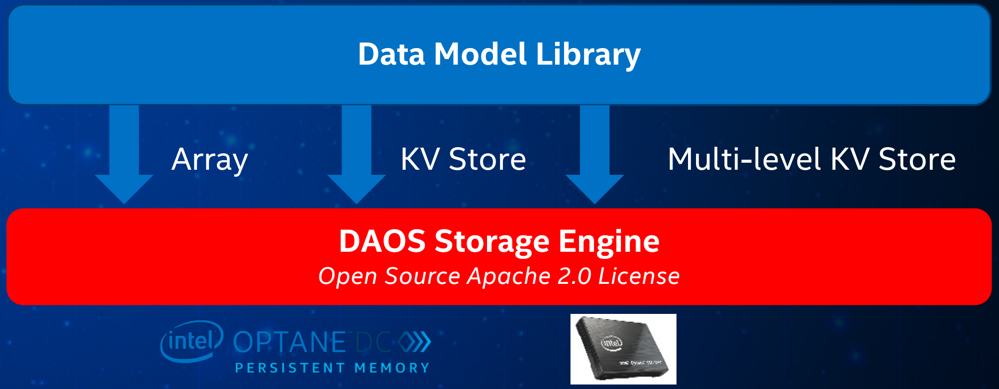


Storage Pool

Container

Object

Record[]



Data Model Library

Array

KV Store

Multi-level KV Store

DAOS Storage Engine

Open Source Apache 2.0 License

STORAGE VIRTUALIZATION & MULTI-TENANCY

Distributed storage reservation

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

Predictable capacity

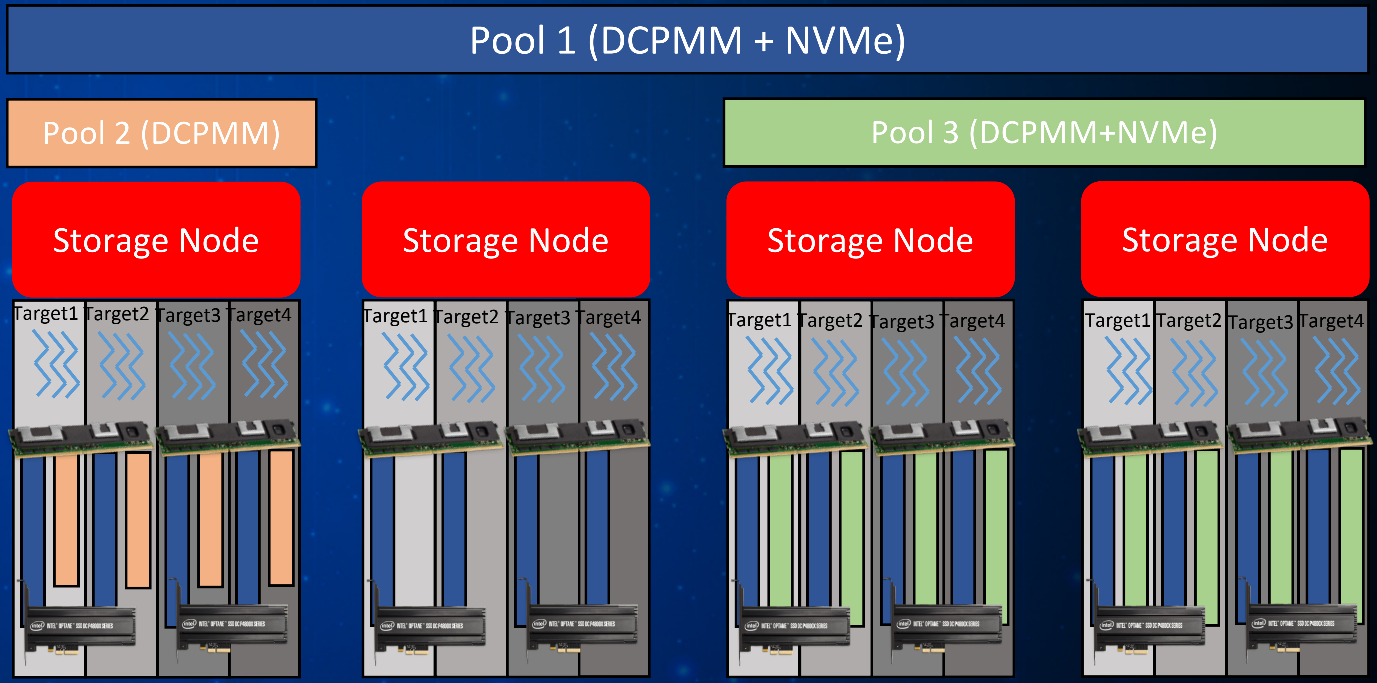
- Can be resized
- Can be extended to span more servers

Multi-tenancy

- NFSv4-type ACLs

Typically 1 pool = 1 project

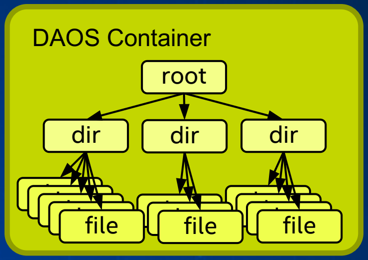
- Can have a single pool or 100's
- Can be ephemeral (per-job) or persistent



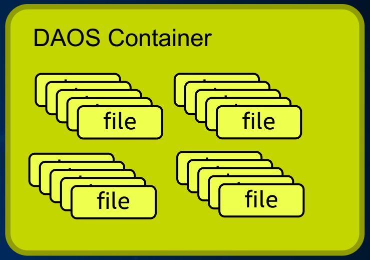
DATASET MANAGEMENT

Aggregate related datasets into manageable and coherent entities

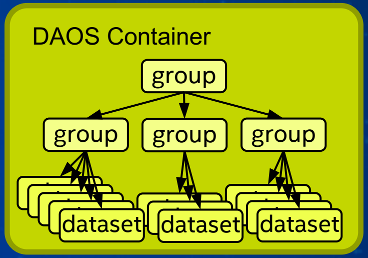
- Distributed consistency & automated recovery
- Full Versioning
- Simplified data management
  - Snapshot
  - Cross-tier Migration
  - Indexing



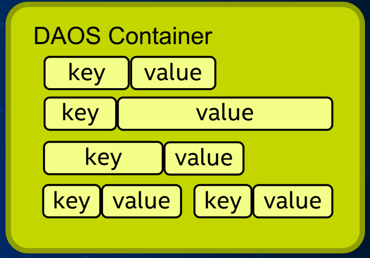
Encapsulated POSIX Namespace



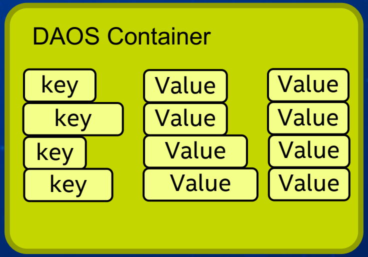
File-per-process



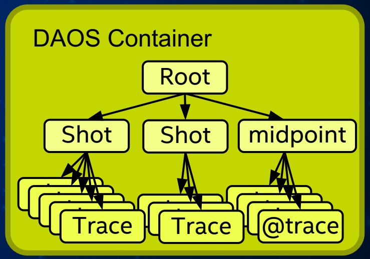
HDF5 « File »



Key-value store

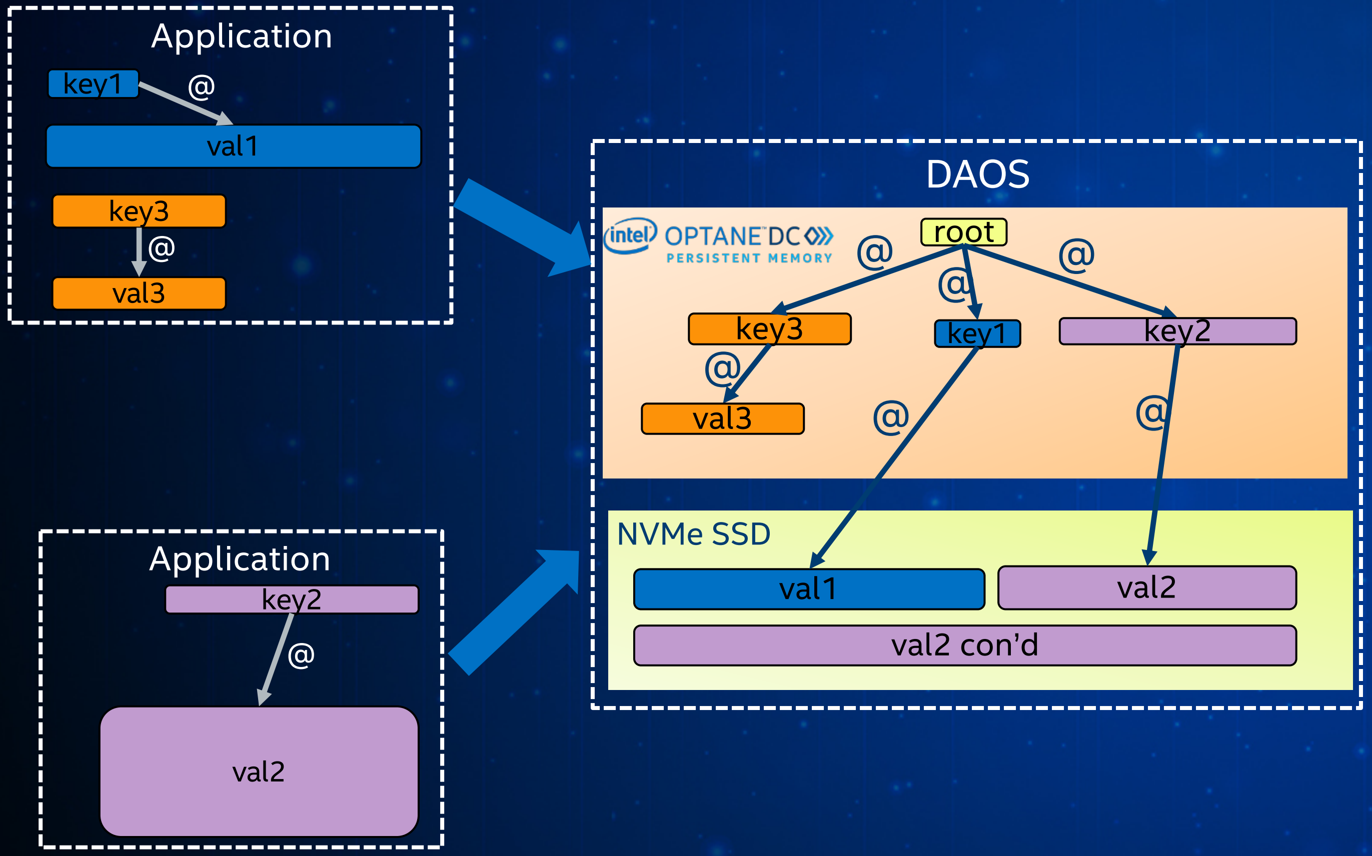


Columnar Database



SEGY

ADVANCED STORAGE API



Fast data retrieval

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

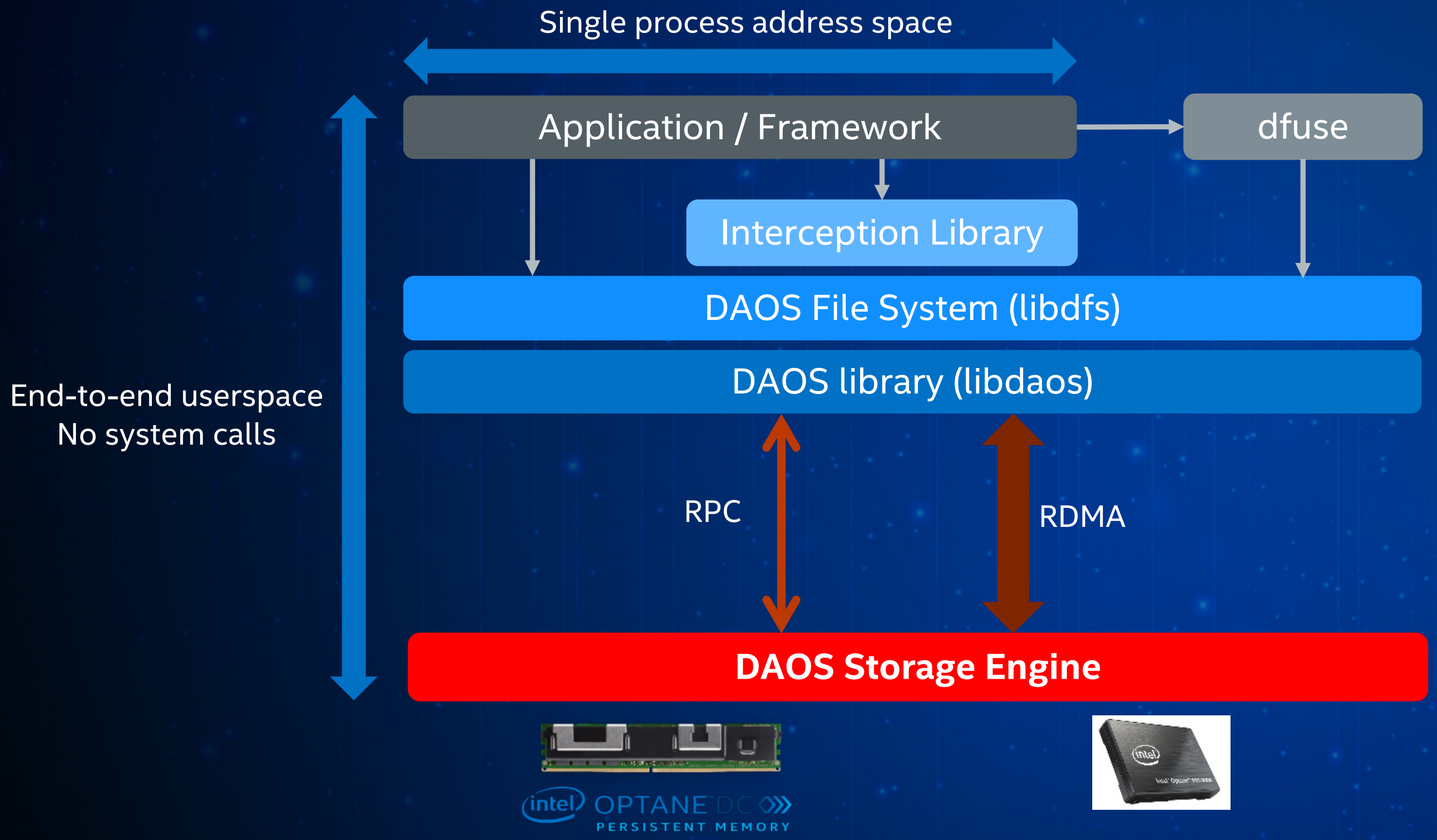
Scalable insert

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

Data indexing

- Query & custom index
- Data provenance

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, …

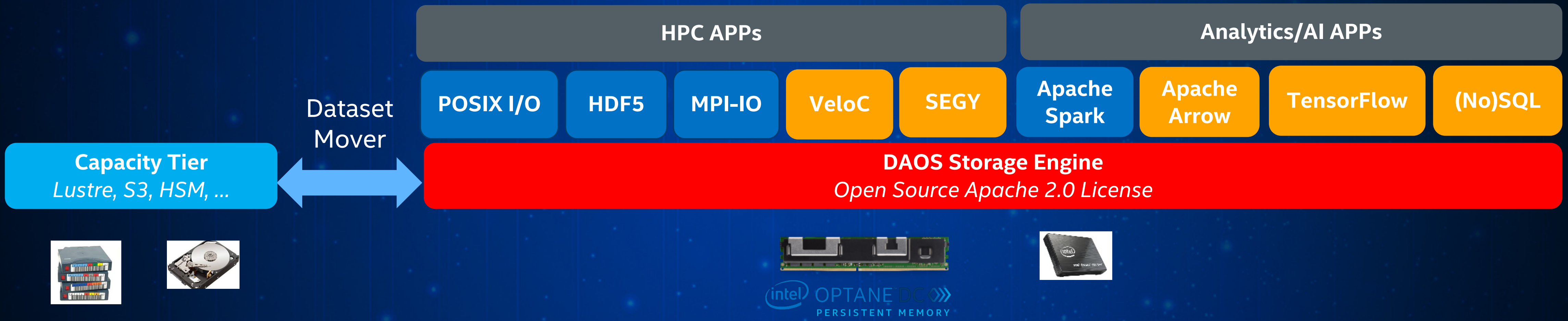
FUSE Daemon (dfuse)

- Transparent access to DAOS
- Involves system calls

I/O interception library

- OS bypass for read/write operations

APPLICATION INTERFACE



HPC APPs

POSIX I/O | HDF5 | MPI-IO | VeloC | SEGY

Analytics/AI APPs

Apache Spark | Apache Arrow | TensorFlow | (No)SQL

Capacity Tier
Lustre, S3, HSM, …

Dataset Mover

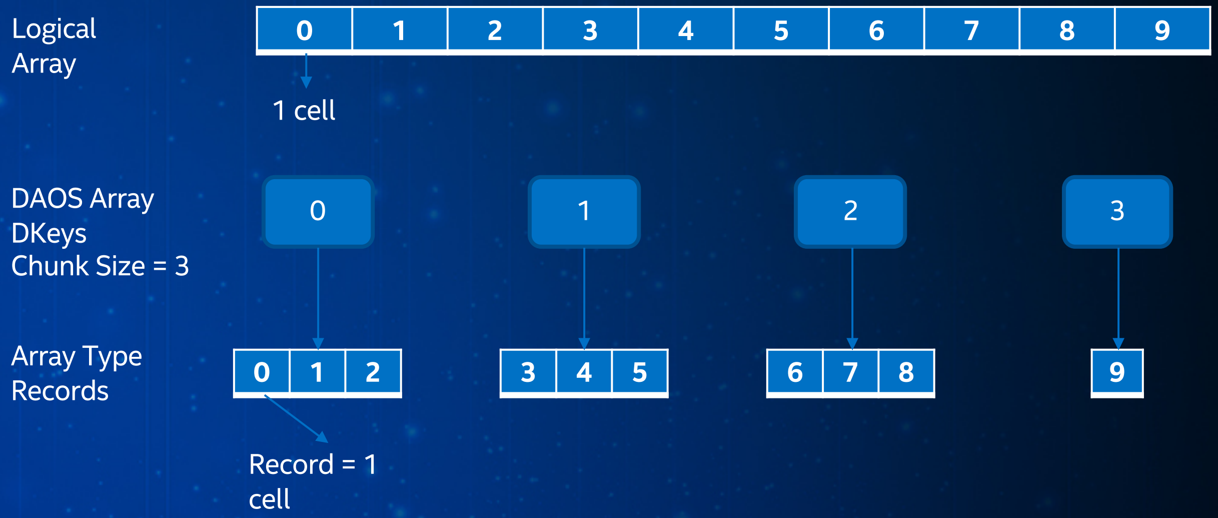
DAOS Storage Engine
Open Source Apache 2.0 License

MPI-IO DRIVER FOR DAOS

The DAOS MPI-IO driver is implemented within the I/O library in MPICH (ROMIO).

- Added as an ADIO driver
- Portable to Open-MPI, Intel MPI, etc.
- https://github.com/daos-stack/mpich
- daos_adio branch
- PR to mpich master in review

1 MPI File = 1 DAOS Array Object



Logical Array

0 1 2 3 4 5 6 7 8 9

1 cell

DAOS Array DKeys Chunk Size = 3

0 1 2 3

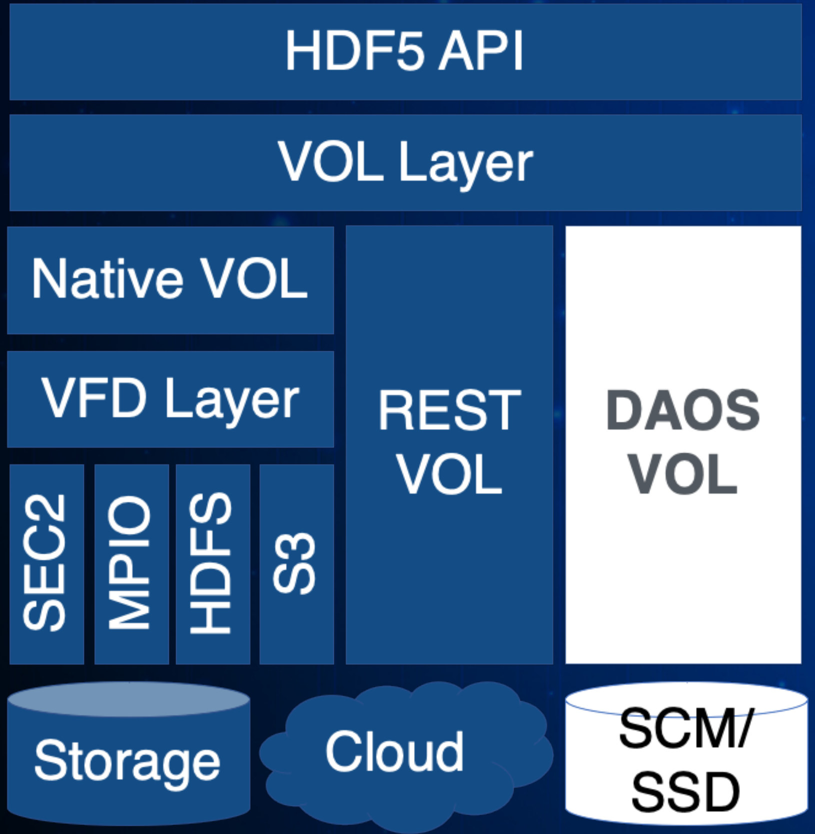
Array Type Records

0 1 2 | 3 4 5 | 6 7 8 | 9

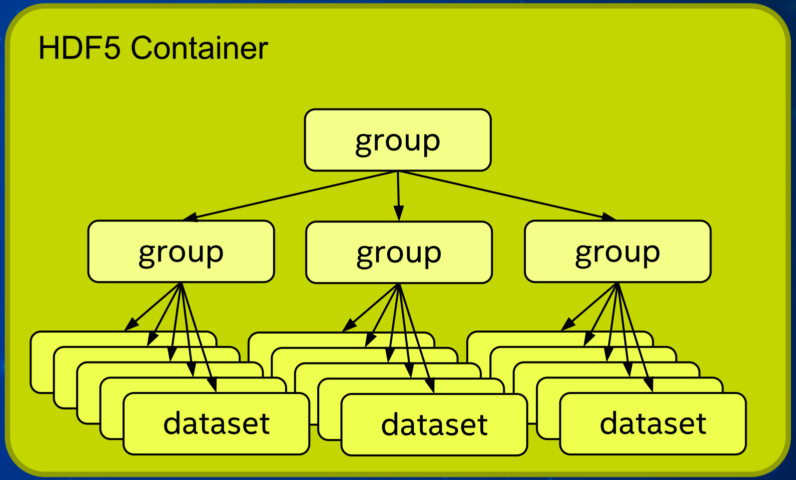
Record = 1 cell

Application works seamlessly otherwise by just specifying the use of the driver by appending “`daos:`” to the path.

HDF5



- Developing an HDF5 VOL Connector
  - Prototyped in ESSIO
- All applications or middleware I/O libraries (e.g. NetCDF4, PIO, etc.) that use HDF5 would be able to run over the DAOS stack with minimal changes.

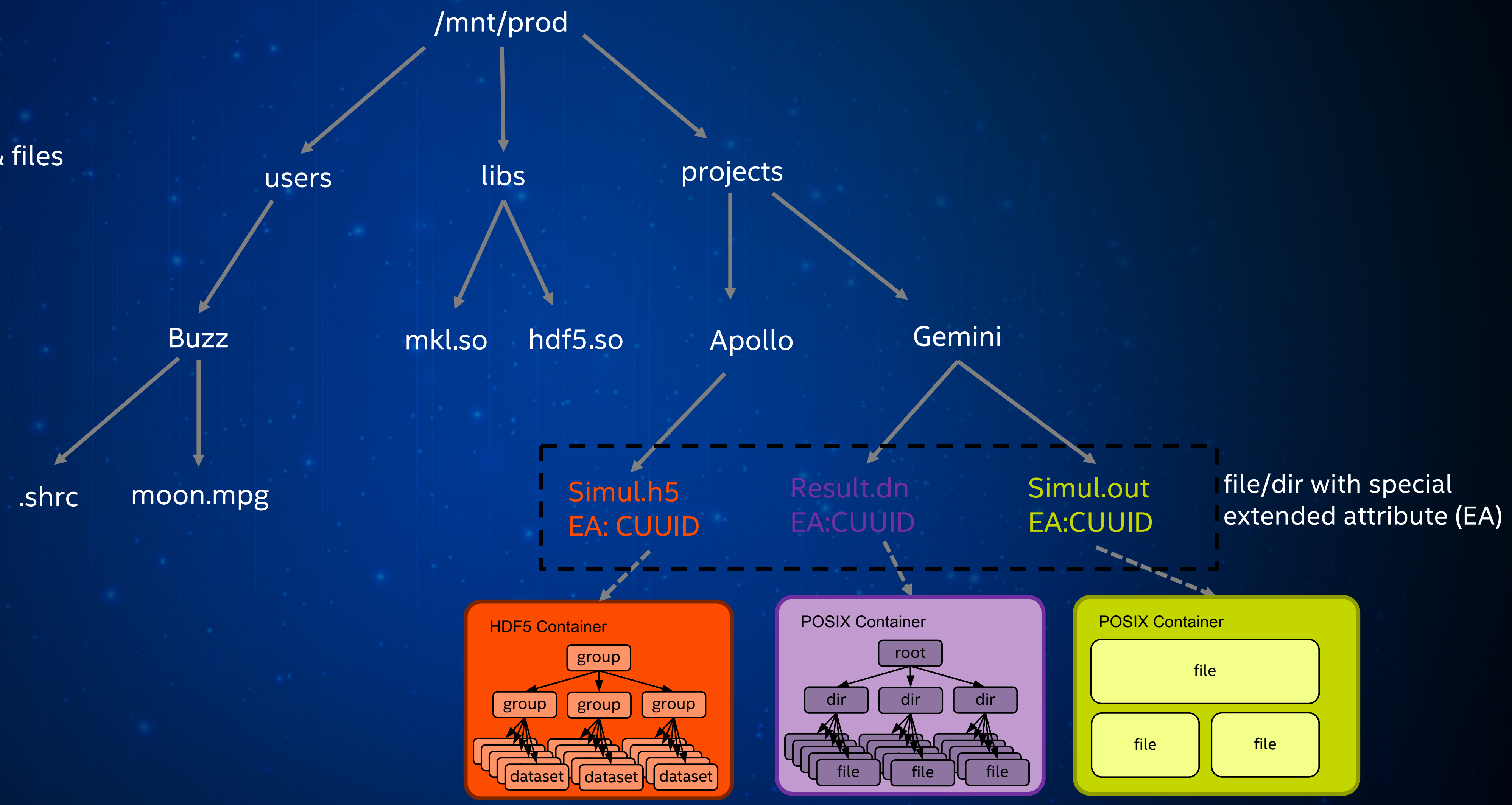


Adding new extensions to HDF5 that are not available to date without the DAOS VOL connector

- Asynchronous I/O for both metadata and raw data operations
- Container Snapshots
- Query & Indexing API

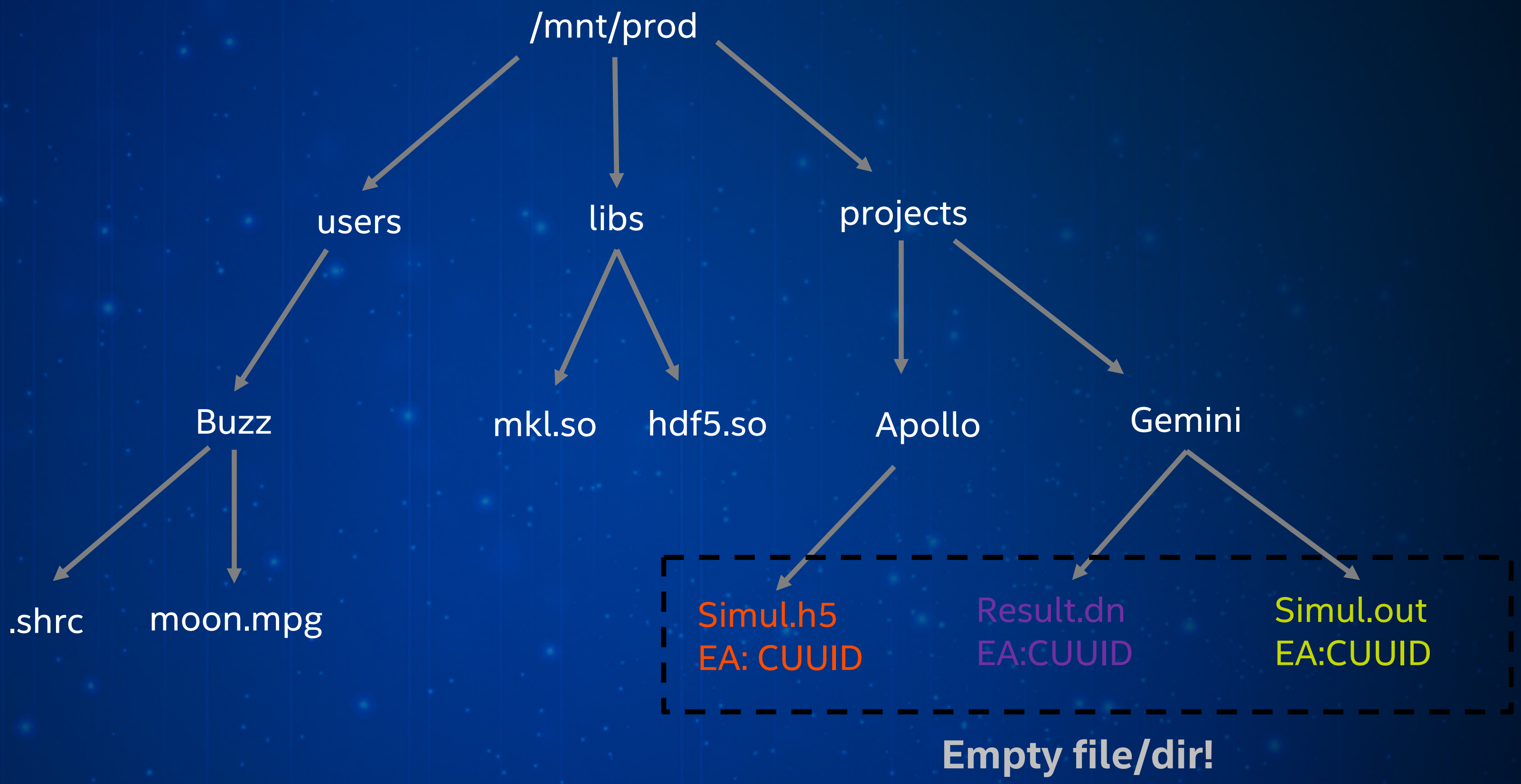
UNIFIED NAMESPACE CONCEPT

Regular Lustre directories & files
HDF5 Container
DAOS POSIX Container
DAOS POSIX Container

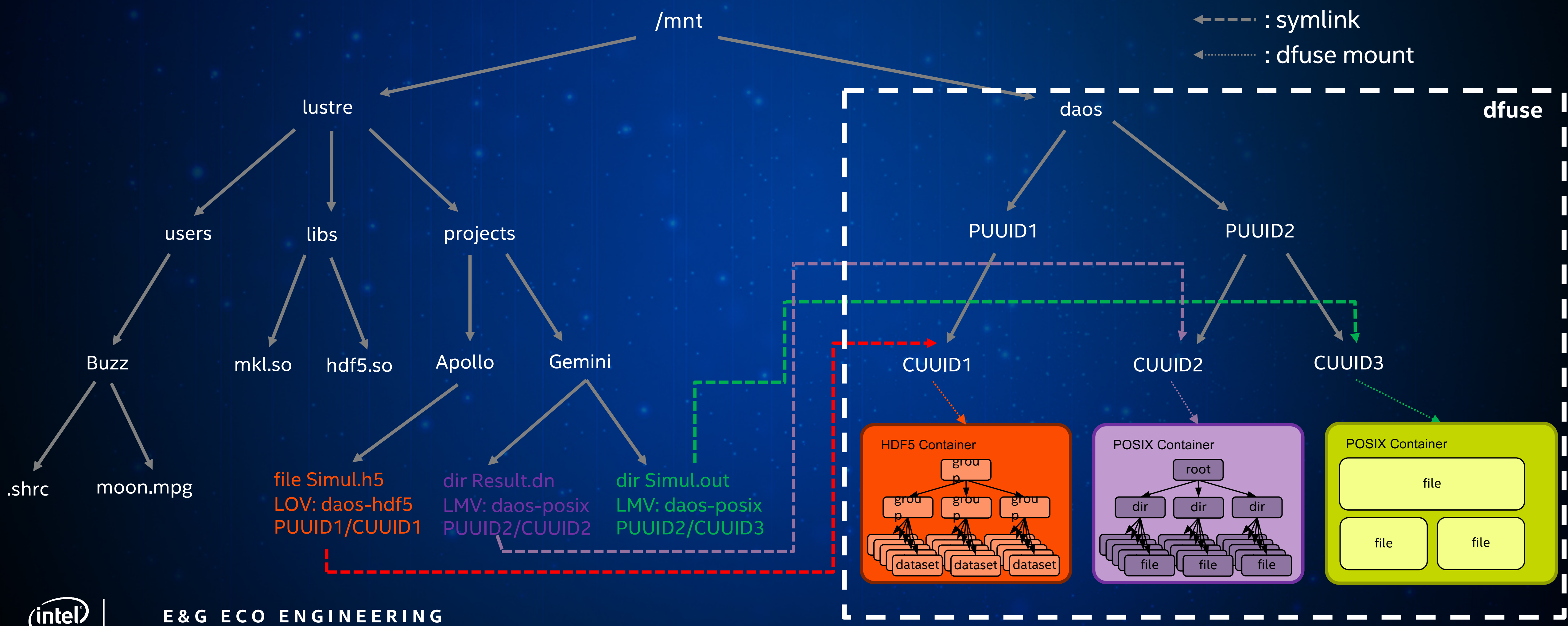


UNIFIED NAMESPACE CONCEPT

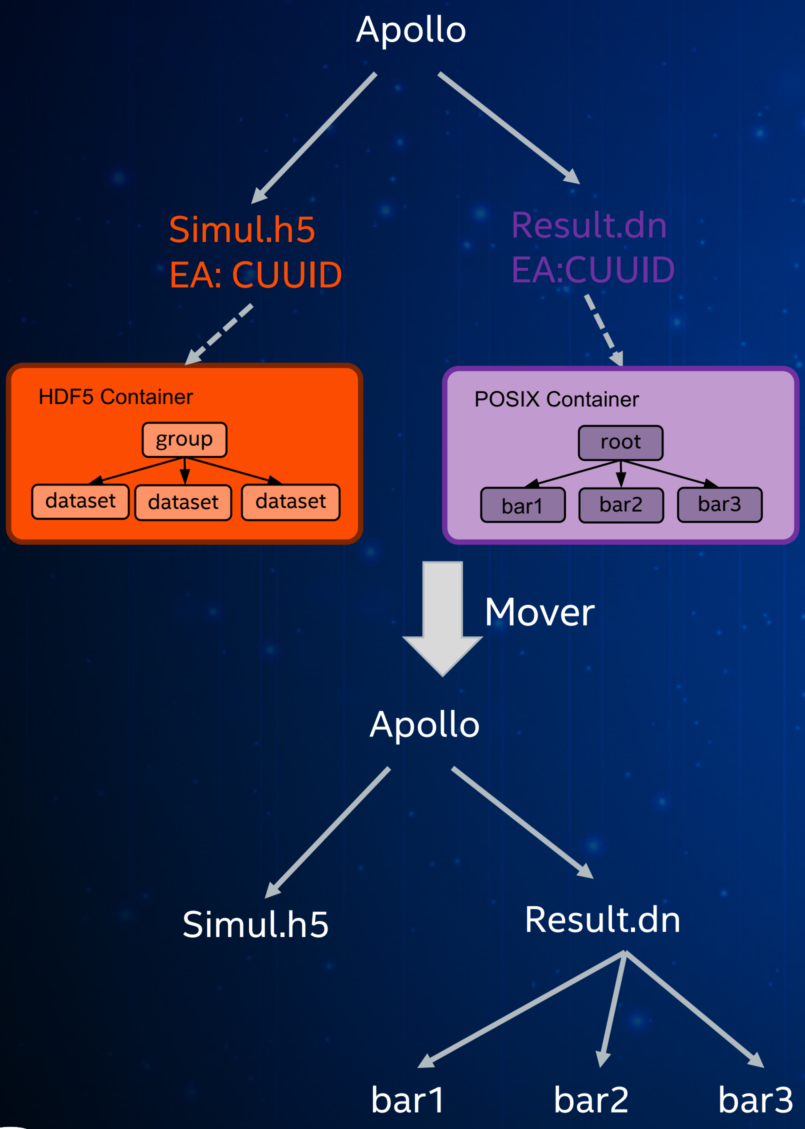
Regular Lustre directories & files
HDF5 Container
DAOS POSIX Container
DAOS POSIX Container

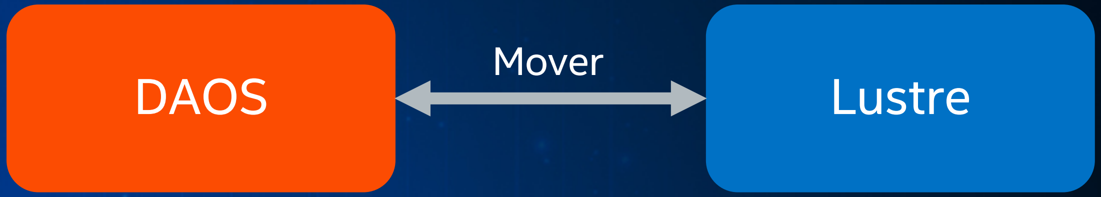


TRANSPARENT ACCESS OF DAOS STORAGE FROM LUSTRE



DATA MOVER





- Different use cases
  - POSIX container migration
  - Other middleware specific data migration (e.g. HDF5)
  - Cross-Pool Container Migration
- Develop an MPI application
  - Parallel movement of datasets between tiers.
- Provide a library and DAOS tool that allows integration with other data movement frameworks (e.g. Globus, DMF, etc.).

DAOS: PRIMARY STORAGE FOR AURORA

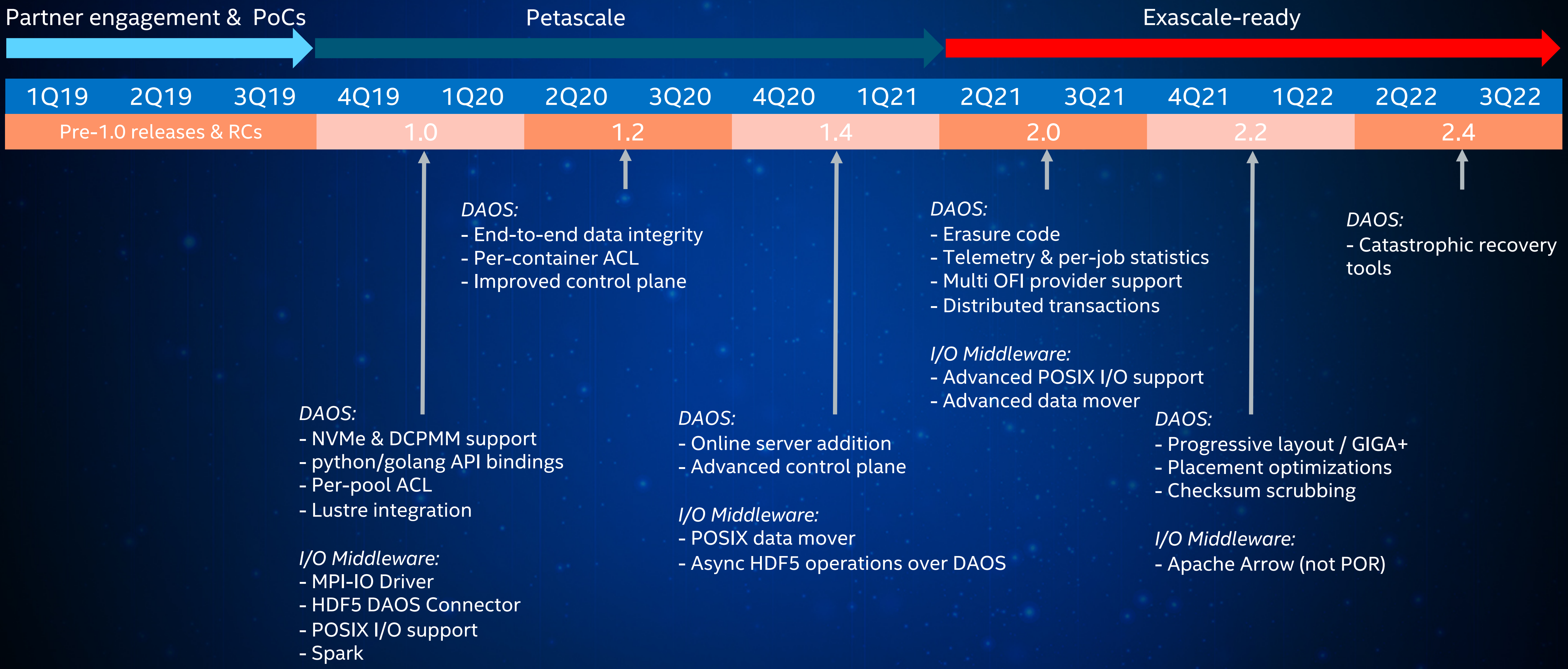
Aurora DAOS configuration

- 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



Partner engagement & PoCs

Petascale

Exascale-ready

1Q19	2Q19	3Q19	4Q19	1Q20	2Q20	3Q20	4Q20	1Q21	2Q21	3Q21	4Q21	1Q22	2Q22	3Q22
Pre-1.0 releases & RCs			1.0		1.2		1.4		2.0		2.2		2.4	

1.0

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

1.2

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

1.4

DAOS:
- Online server addition
- Advanced control plane

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

2.0

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

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

2.2

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

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

2.4

DAOS:
- Catastrophic recovery tools

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

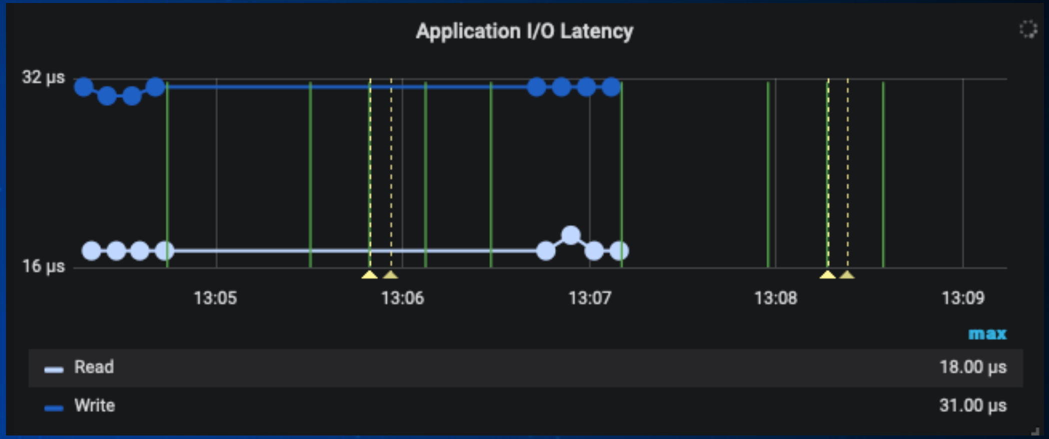
PERFORMANCE

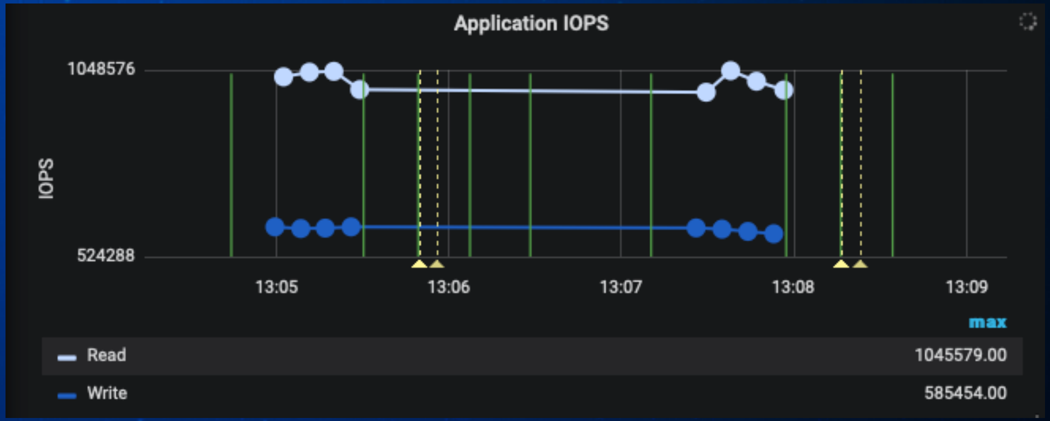
Demonstrated at ISC (½U server)

- https://www.youtube.com/watch?v=EMGBcvnftwQ
- https://www.youtube.com/watch?v=e69Rgz2FMbE

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
  - One task enough to reach 8GB+/s





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®

INTEL® OPTANE™ TECHNOLOGY

Intel® Optane™ SSD DC P4800X/P4801X

PCIe* 3.0 x4, NVMe*

100GB
200GB
375GB
750GB
1.5TB

Intel® Optane™ DC D4800X

PCIe* 3.0 2x2, NVMe*

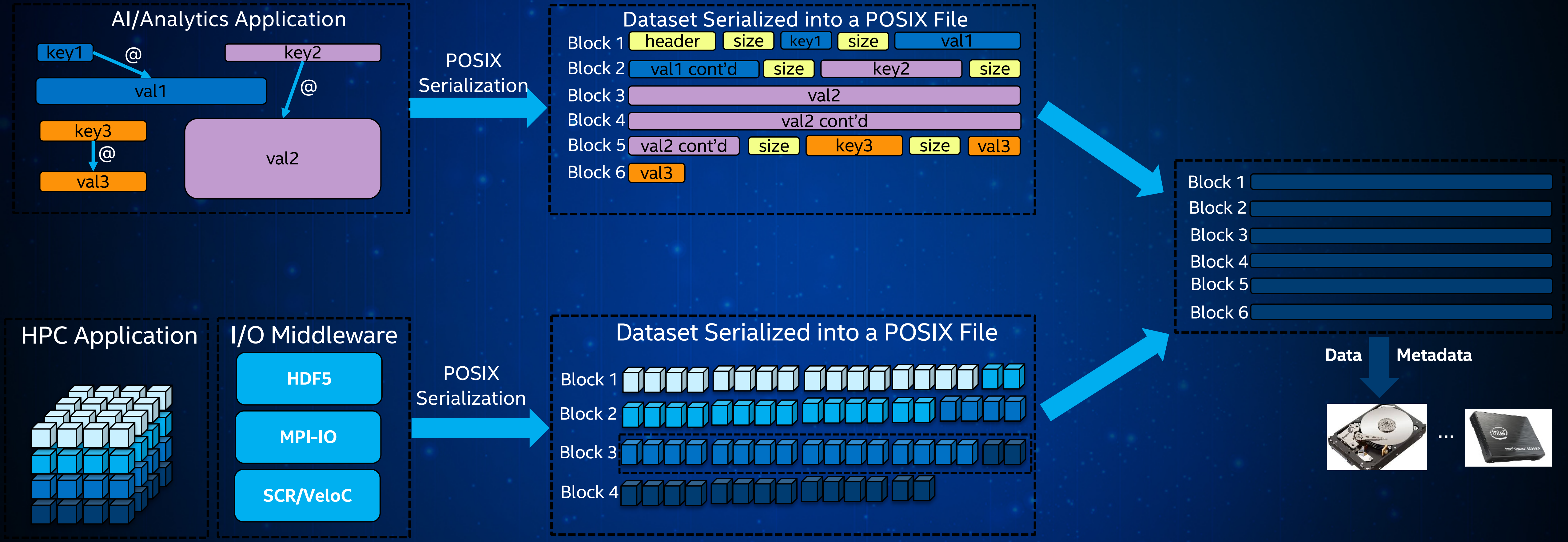
375GB
750GB
1.5TB

Intel® Optane™ DC Persistent Memory

128GB
256GB
512GB

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

POSIX LIMITATIONS



AI/Analytics Application

key1 @ val1

key2 @ val2

key3 @ val3

POSIX Serialization

Dataset Serialized into a POSIX File

Block 1 header size key1 size val1

Block 2 val1 cont'd size key2 size

Block 3 val2

Block 4 val2 cont'd

Block 5 val2 cont'd size key3 size val3

Block 6 val3

HPC Application

I/O Middleware

HDF5

MPI-IO

SCR/VeloC

POSIX Serialization

Dataset Serialized into a POSIX File

Block 1

Block 2

Block 3

Block 4

Block 1

Block 2

Block 3

Block 4

Block 5

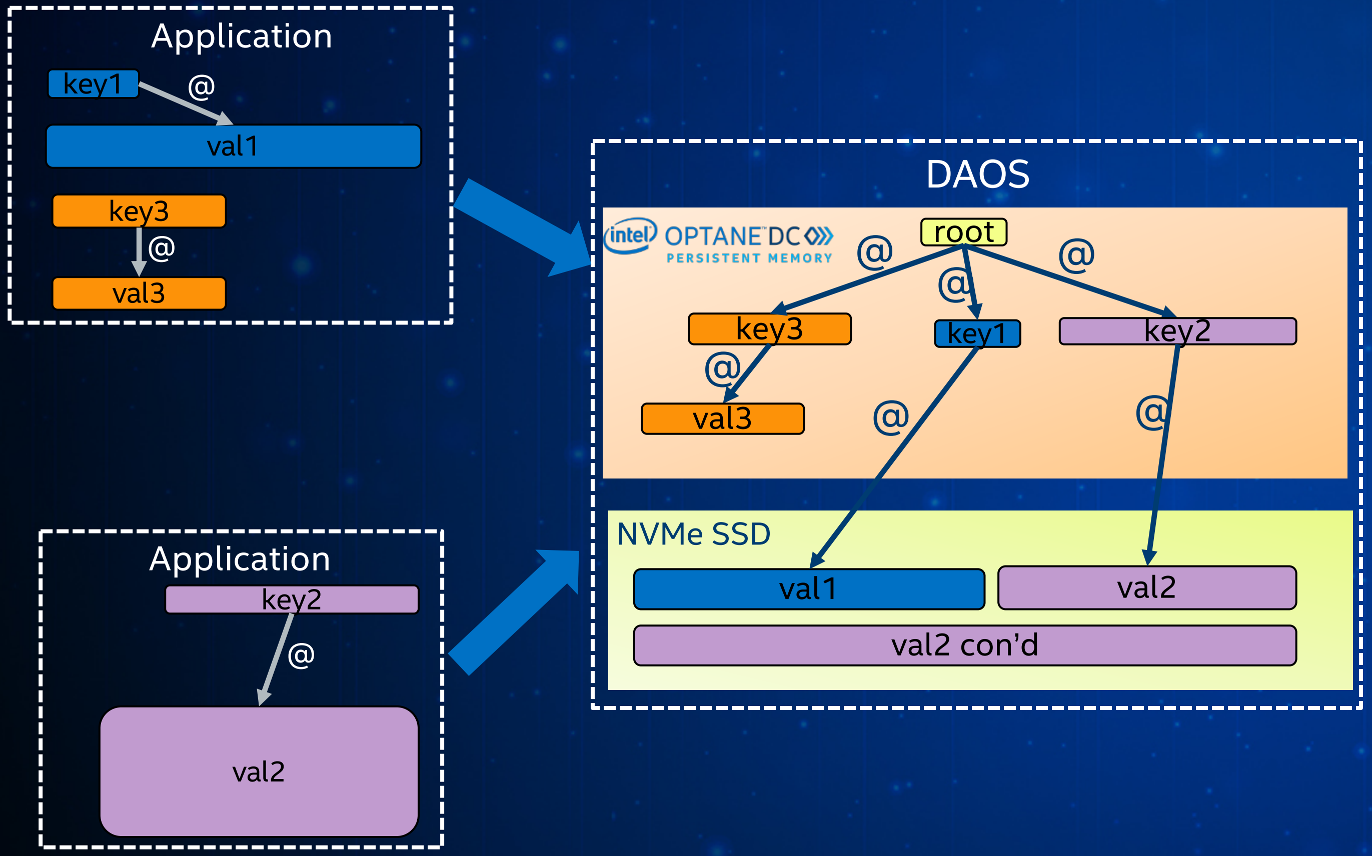
Block 6

Data

Metadata

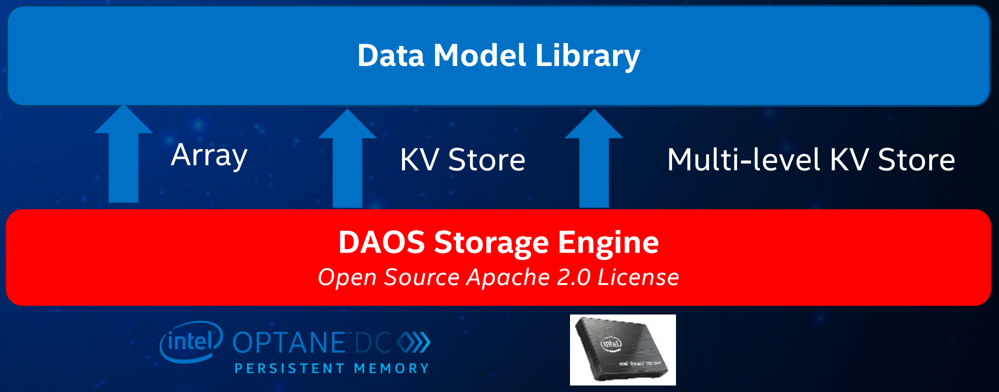
...

ADVANCED STORAGE API

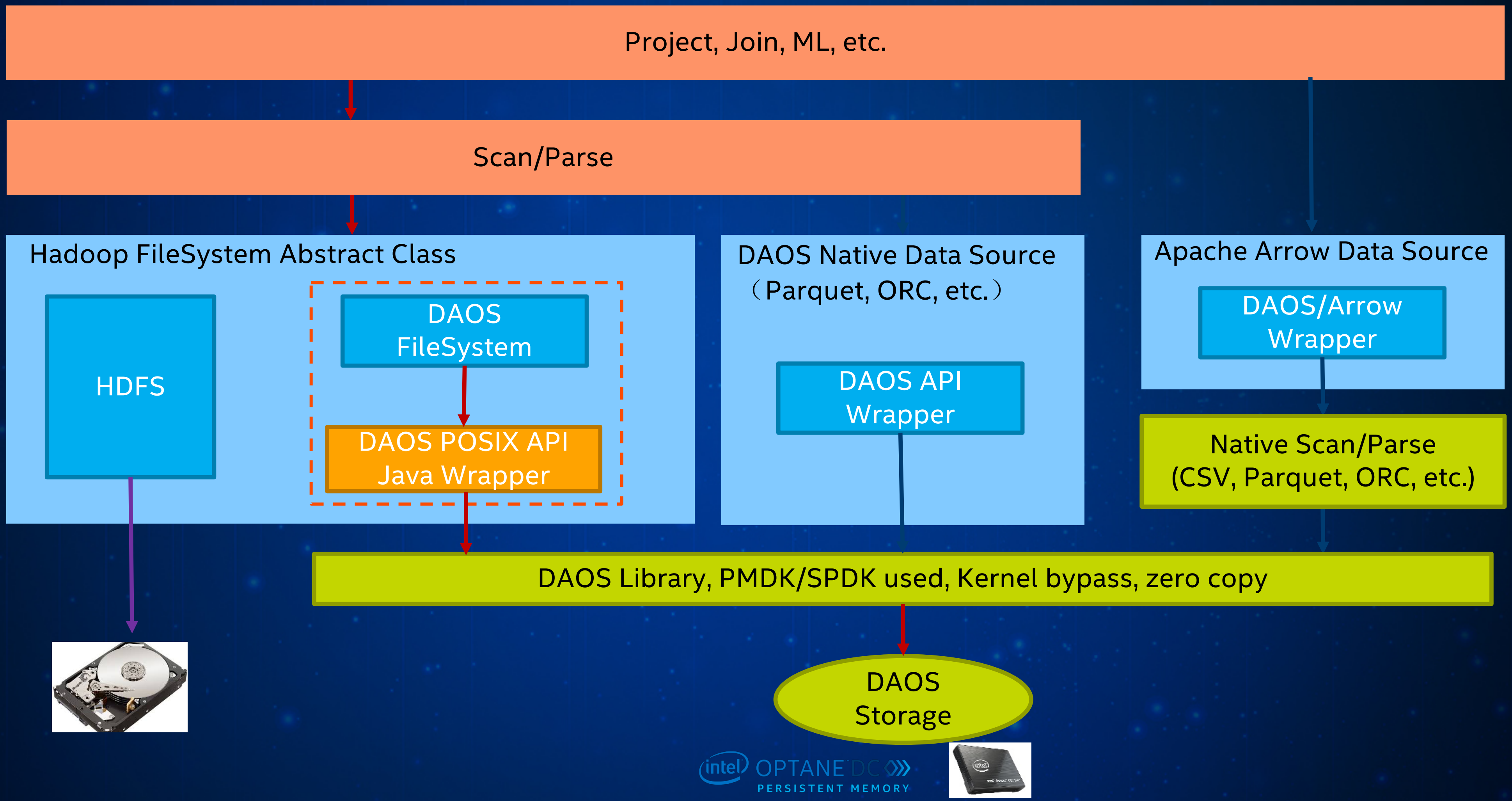


Native support for structured, semi-structured & unstructured data models

- 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 & BIG DATA / AI



Project, Join, ML, etc.

Scan/Parse

Hadoop FileSystem Abstract Class

HDFS

DAOS FileSystem

DAOS POSIX API Java Wrapper

DAOS Native Data Source （Parquet, ORC, etc.）

DAOS API Wrapper

Apache Arrow Data Source

DAOS/Arrow Wrapper

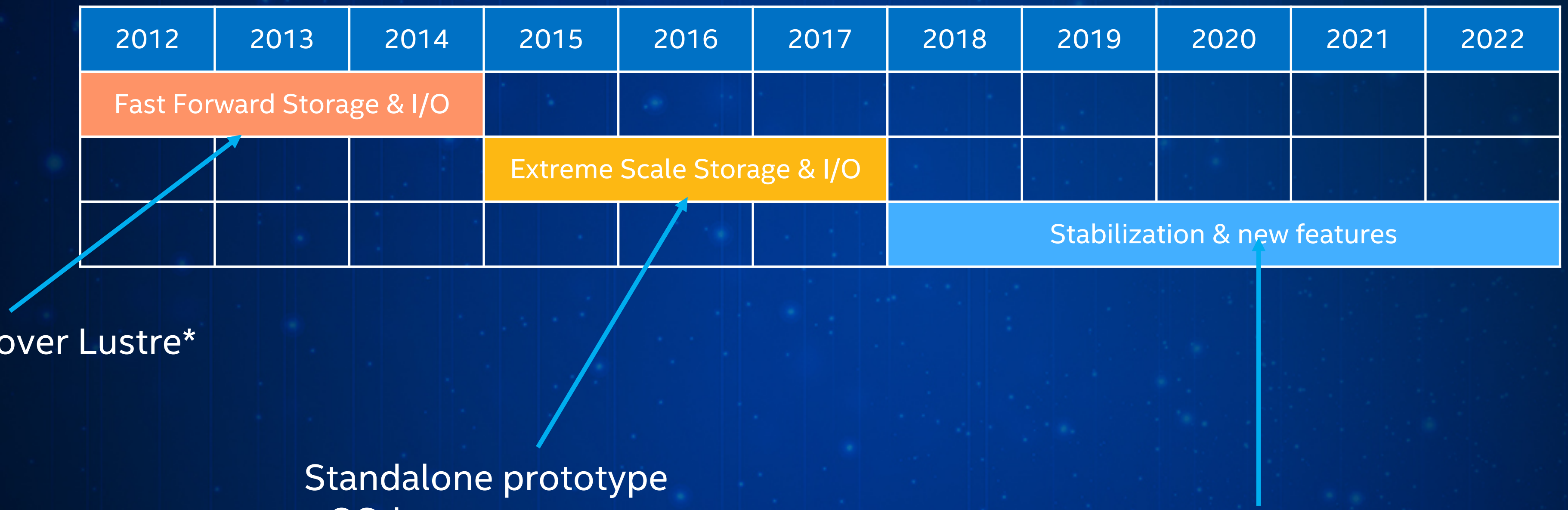
Native Scan/Parse (CSV, Parquet, ORC, etc.)

DAOS Library, PMDK/SPDK used, Kernel bypass, zero copy

DAOS Storage

Intel Proprietary

DAOS PROJECT HISTORY



2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Fast Forward Storage & I/O										
			Extreme Scale Storage & I/O							
						Stabilization & new features				

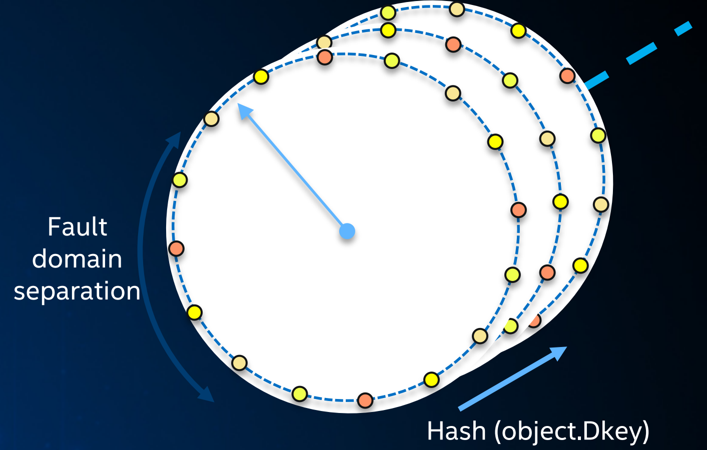
Prototype over Lustre*

Standalone prototype
- OS-bypass
- Persistent memory
- Replication & self healing

DAOS Productization

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

DATA MANAGEMENT



Data Distribution

- Algorithmic placement
- Progressive layout with GIGA+

Data Protection

- Declustered replication & erasure code
- Fault-domain aware placement
- Self-healing
- End-to-end data integrity

Data Versioning

- Non-destructive write & consistent read
- Native snapshot support

Data Security & Reduction (not POR)

- Online real-time data encryption & compression
- Hardware acceleration

CONTROL PLANE

Storage provisioning

- Detect SCM & NVMe storage
  - CPU/storage affinity
- Configure/format/mount SCM
  - Interleaved mode
- Configure NVMe SSDs
  - Firmware update
- Integrated storage burn-in capability

Fabric configuration

- Comm layer configuration
- Interface/CPU affinity

DAOS configuration

- zero-conf/auto-conf with device filters/manual-conf
- YAML configuration for admins

DAOS service management

- Manage/monitor/troubleshoot
- Integration with systemd & other frameworks

Telemetry

- Storage/service/fabric activity
- Per-job statistics

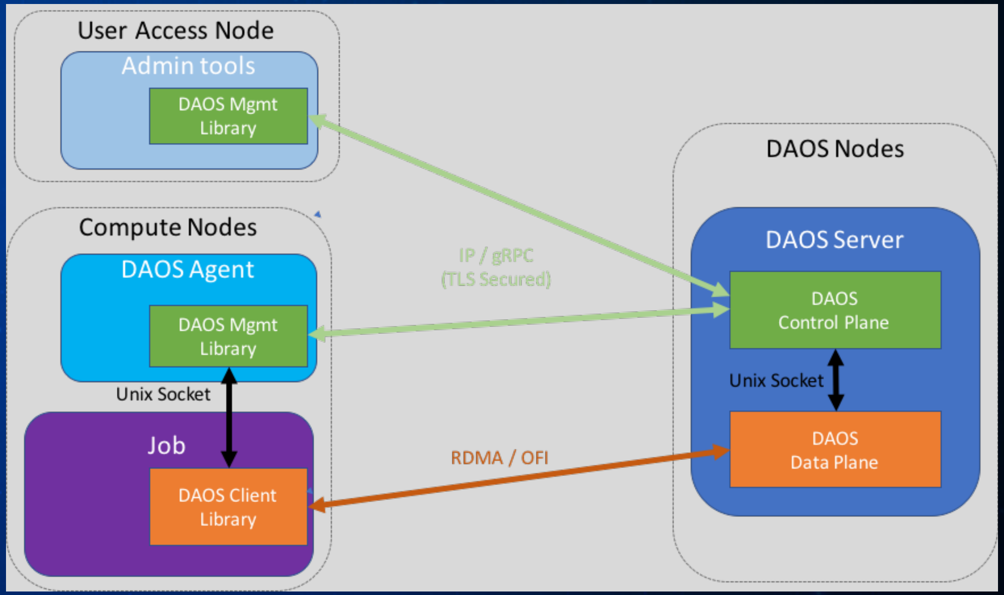
Storage API & tools

- CLI tools built over the control plane API

SECURITY

Flexible security framework

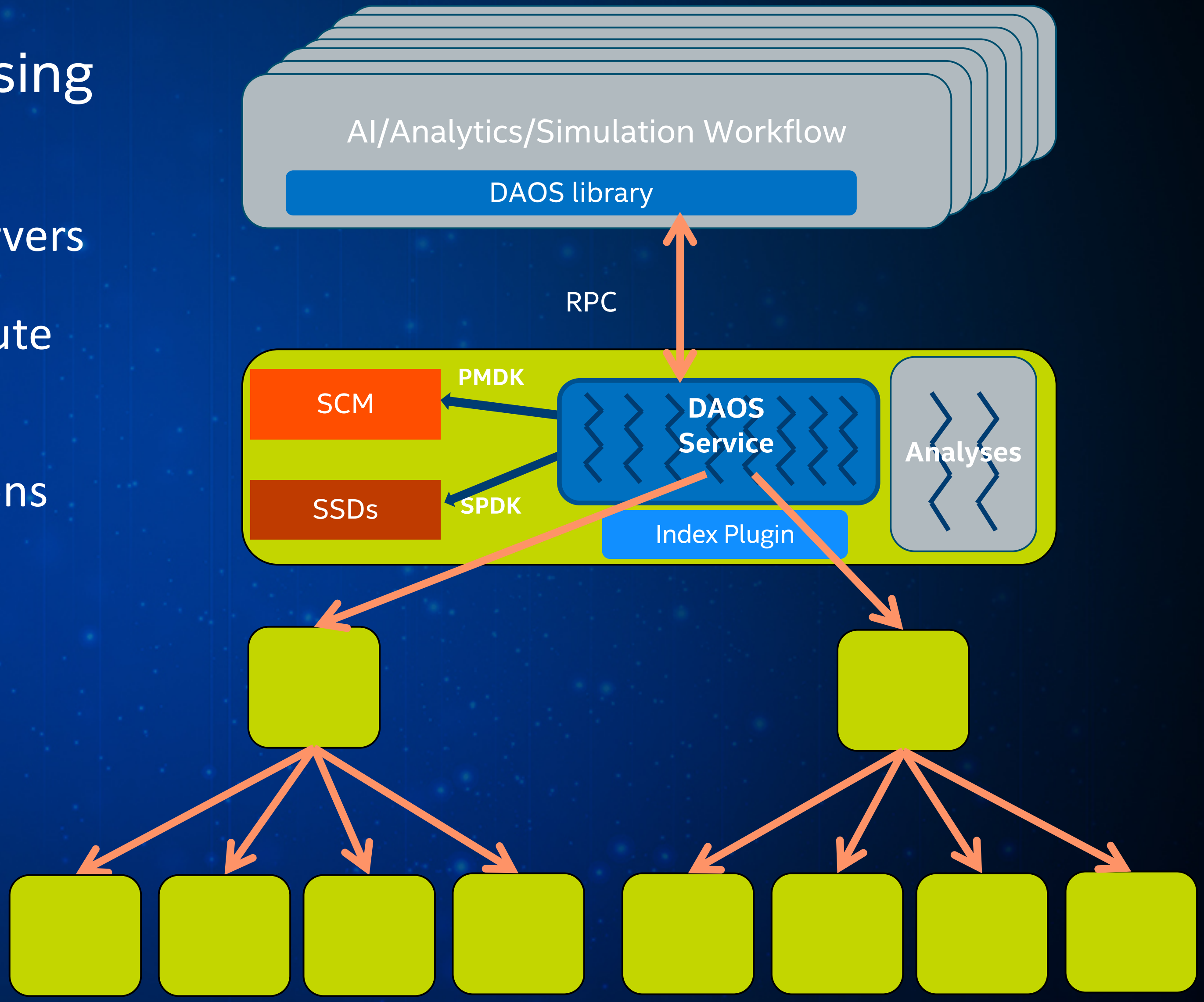
- Support different authentication methods
  - Local agent on compute node authenticating process through AUTH_SYS
  - Third party authentication service (e.g. munge)
- TLS-secured channel using certificates
- Very minimal impact expected on I/O path



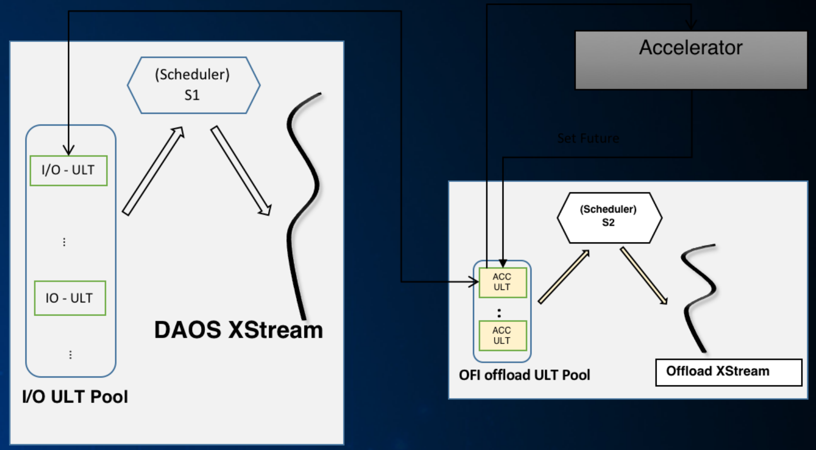
IN-STORAGE COMPUTING

Function shipping for in-situ processing

- Execute pre-defined / user-defined data processing function directly on storage servers
- Prevent loading entire dataset onto compute nodes
- Execute filtering/MapReduce-like operations where data is located
  - Collective with reply aggregation
- Send results back to caller
- Not POR

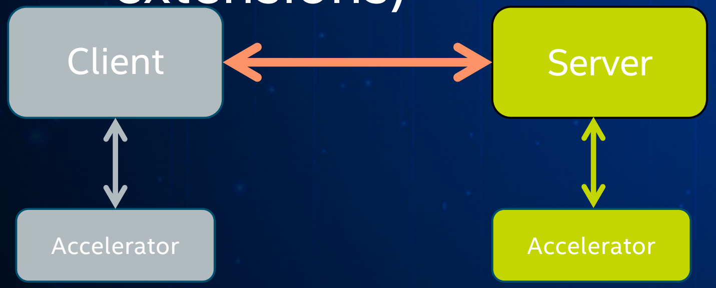


STORAGE ACCELERATION FRAMEWORK



Offload API for client and server

- ISA-L (software) on IA
- Accelerators (hardware)
  - Intel QuickAssist
  - GPGPU
  - FPGA/SmartNICs (libfabric extensions)



Use cases

- Erasure code
- Checksums
- Compression
- Encryption
- Data indexing/query
- Data transformation
  - Dropping floating point precision
  - AoS to SoA and vice-versa