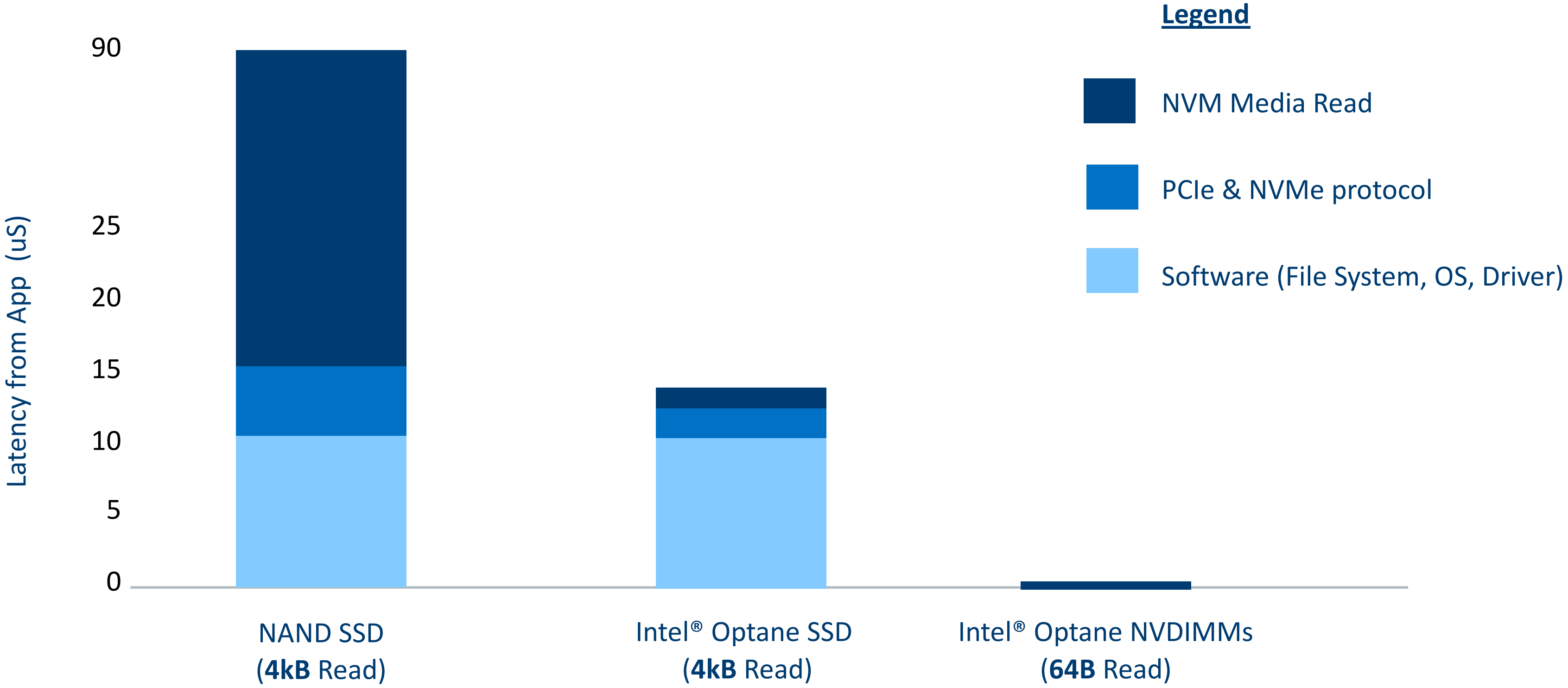
DAOS with PMDK

Di Wang **E**xtreme **S**torage **A**rchitecture & **D**evelopment (ESAD), Intel

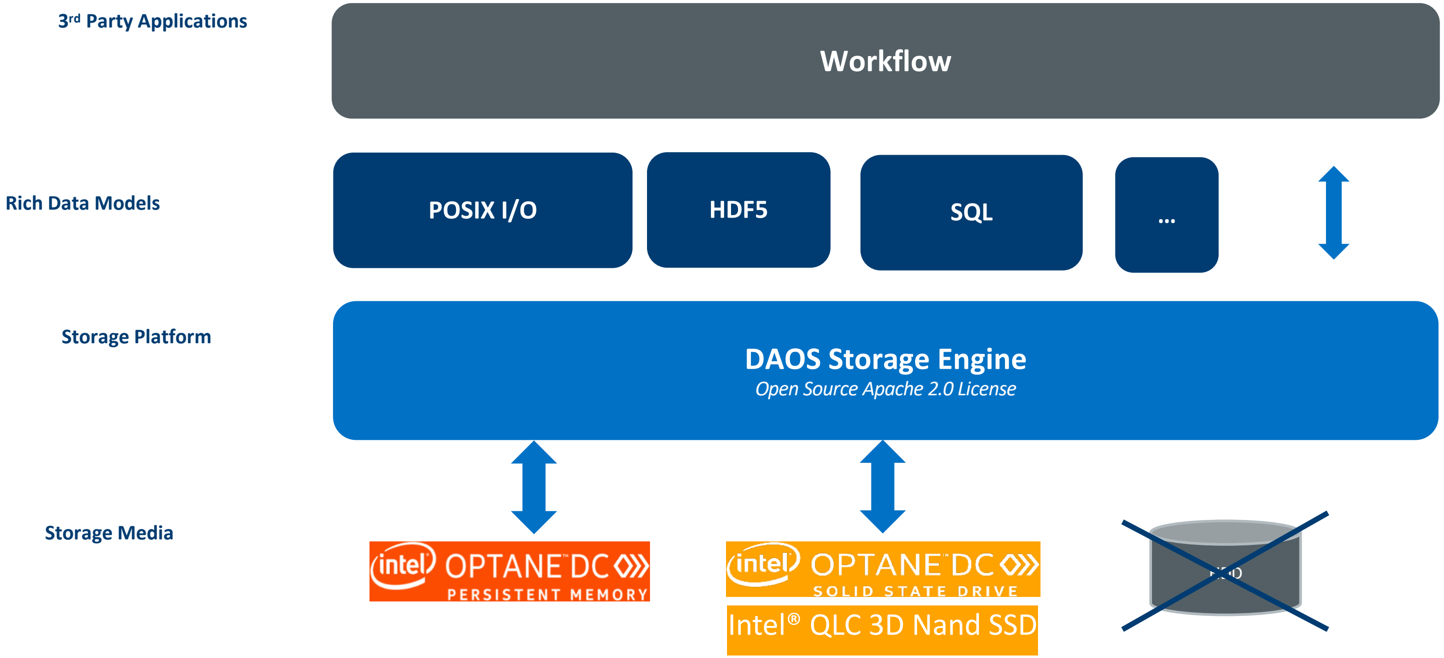
Agenda

- DAOS (Distributed Asynchronous Object Storage) Overview
- DAOS Architecture & features
- DAOS Storage Model
- DAOS with PMDK & SPDK
- Current Performance & Resource

Storage revolution



DAOS overview



3rd Party Applications

Workflow

Rich Data Models

POSIX I/O

HDF5

SQL

...

Storage Platform

DAOS Storage Engine

Open Source Apache 2.0 License

Storage Media

Lightweight I/O

Mercury userspace function shipping

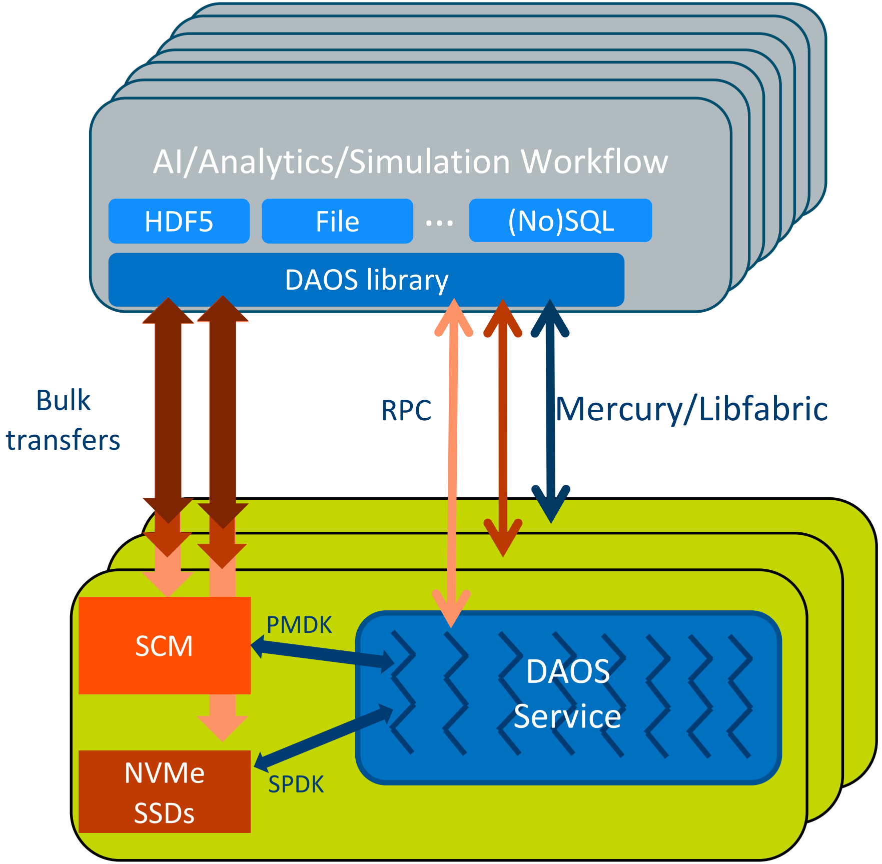
- **MPI** equivalent communications **latency**
- Built over libfabric

Applications link directly with DAOS lib

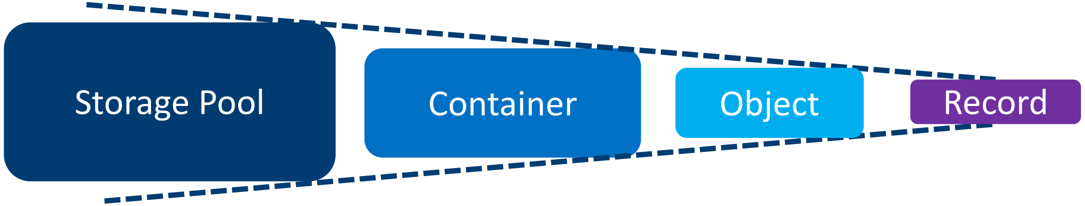
- Direct call, no context switch
- **Small** memory **footprint**
- No locking, caching or data copy

Userspace DAOS server

- Mmap non-volatile memory via PMDK
- NVMe access through SPDK/Blobstore



Storage Model



DAOS provides a **rich** storage API

- New 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
- **Ad-hoc** concurrency control mechanism

Pool

- **Reservation** of distributed storage
- Predictable/extendable **performance/capacity**

Container

- Aggregate **related** datasets into manageable entity
- Unit of **snapshot**/transaction

Object

- **Key-array store** with own distribution/resilience schema
- **Multi-level** key for fine-grain control over **colocation** of related data

Record

- Arbitrary binary **blob** from single byte to several Mbytes

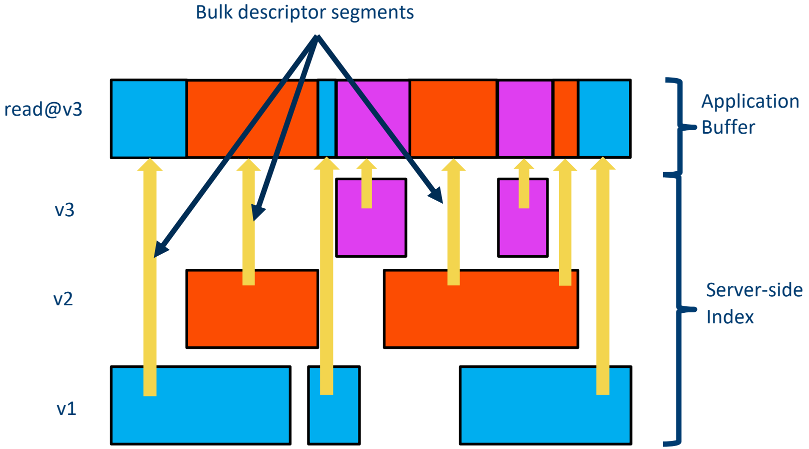
Fine-grained I/O

Mix of storage technologies

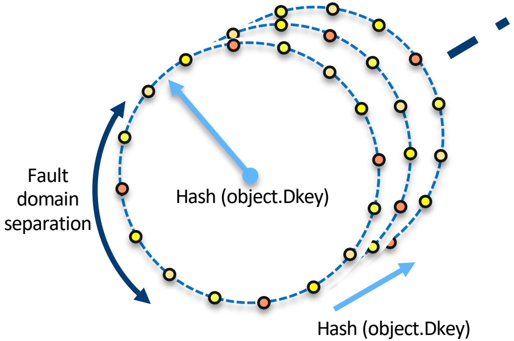
- Storage Class Memory
  - DAOS **metadata** & application **metadata**
  - **Byte-granular** application **data**
- NVMe SSD (*NAND)
  - Cheaper storage for **bulk** data (e.g. checkpoints)
  - Multi-KB

I/Os are **logged** & inserted into **persistent index**

- **Non-destructive** write & **consistent** read
- **No alignment** constraints
- **No read-modify-write**



DATA Management



Data Distribution

- Algorithmic placement

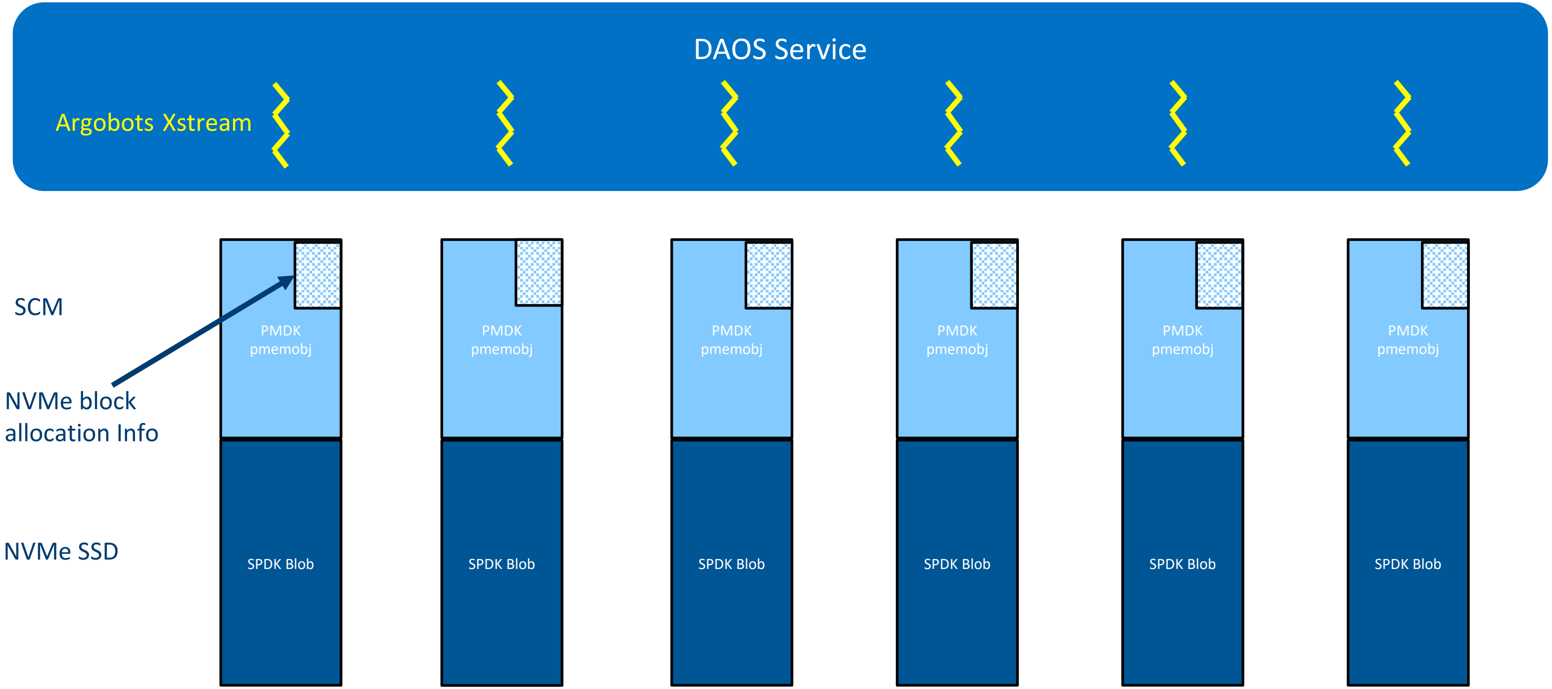
Data Protection

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

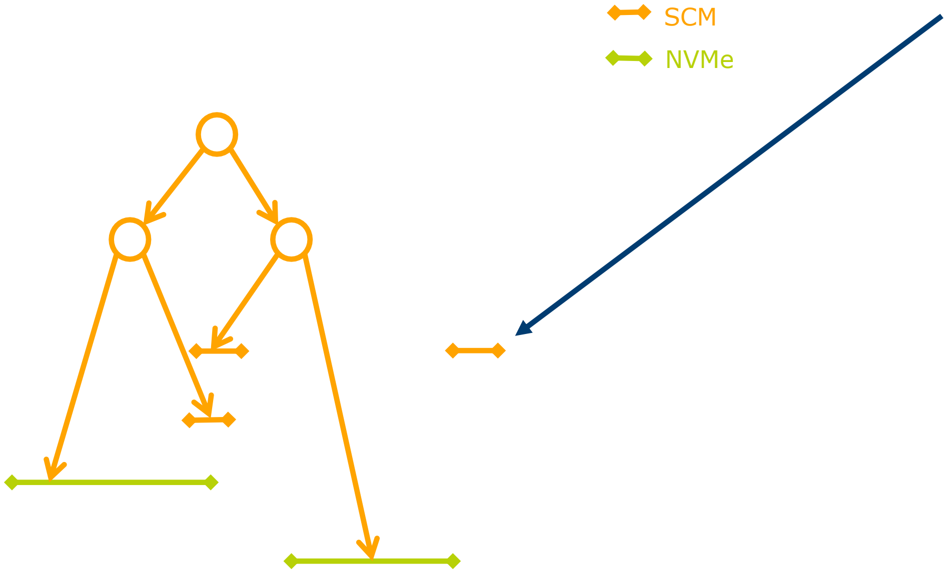
Data Security & Reduction

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

Pool Storage on DAOS Server



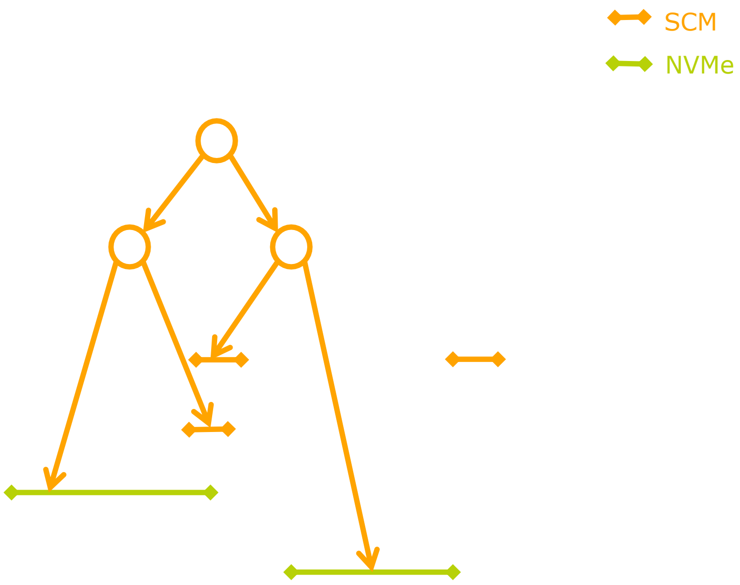
DAOS I/O over PMDK/SPDK



DAOS Xstream

- Reserve new buffer
  - Either reserve by pmemobj_reserve
  - Or reserve in NVME SSD

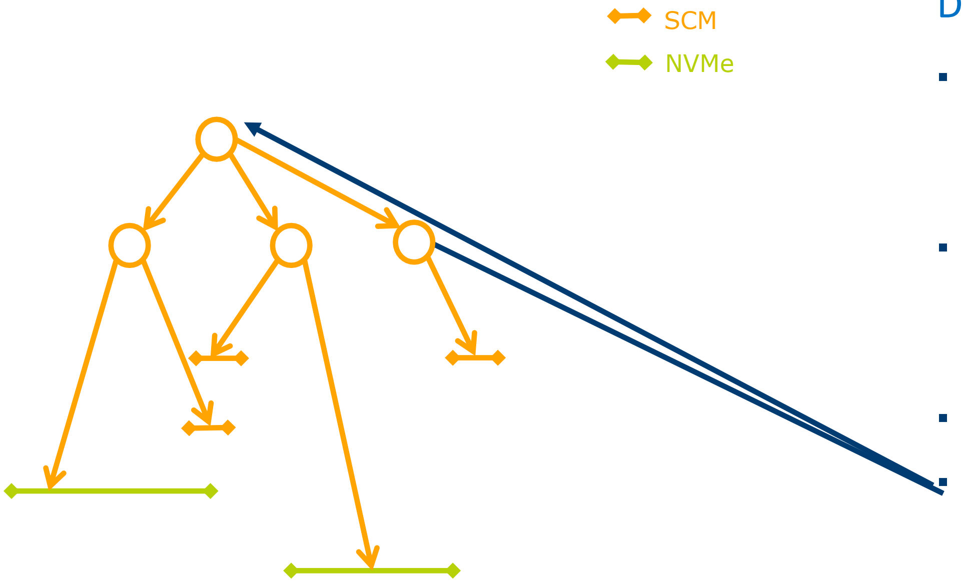
DAOS I/O over PMDK/SPDK



DAOS Xstream

- Reserve new buffer
  - Either reserve by pmemobj_reserve
  - Or reserve in NVME SSD
- Start RDMA transfer to newly allocated buffer
  - Either transfer to PMEM
  - Or transfer to DMA buffer then to NVME SSD
- Start pmemobj transaction

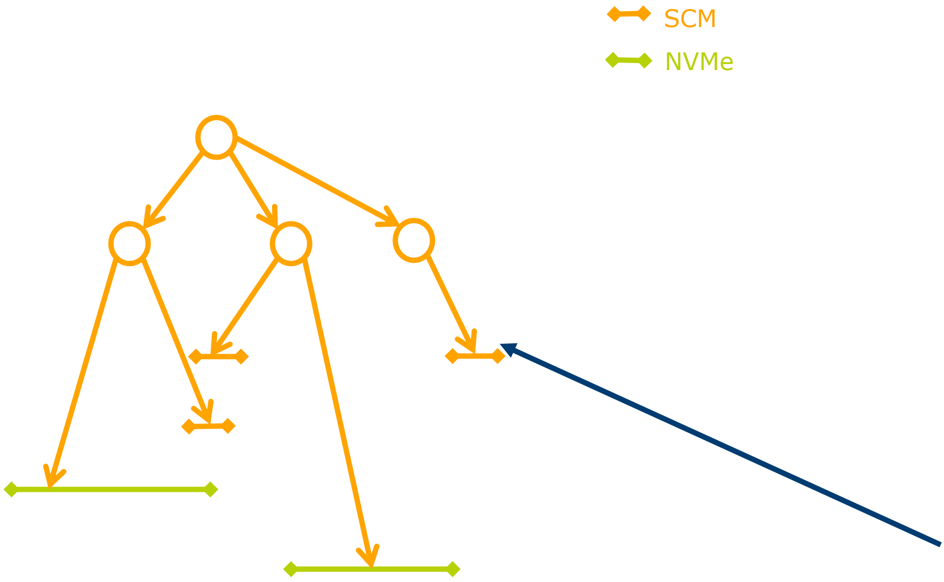
DAOS I/O over PMDK/SPDK



DAOS Xstream

- Reserve new buffer
  - Either reserve by pmemobj_reserve
  - Or reserve in NVME SSD
- Start RDMA transfer to newly allocated buffer
  - Either transfer to PMEM
  - Or transfer to DMA buffer then to NVME SSD
- Start pmemobj transaction
- Modify index to insert new extent

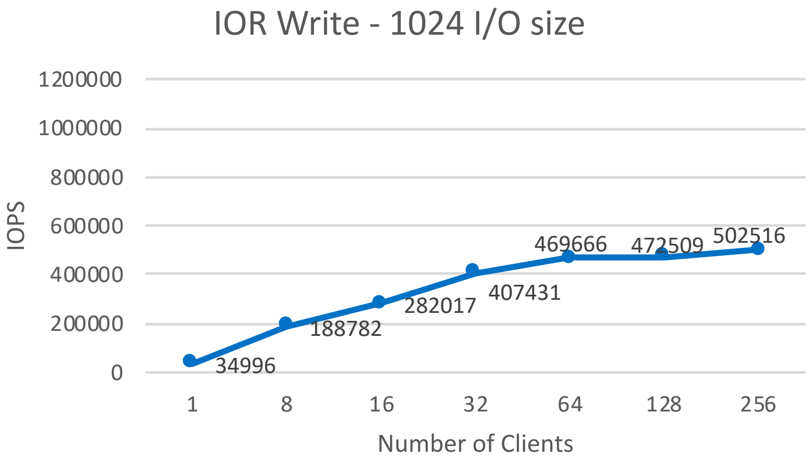
DAOS I/O over PMDK/SPDK

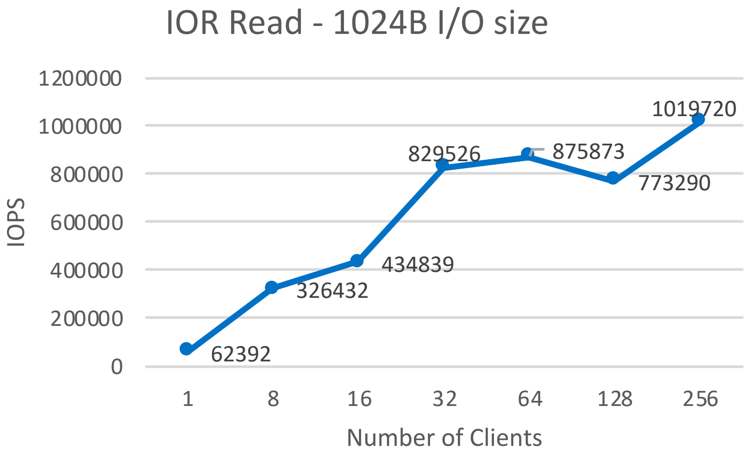


DAOS Xstream

- Reserve new buffer
  - Either reserve by pmemobj_reserve
  - Or reserve in NVME SSD
- Start RDMA transfer to newly allocated buffer
  - Either transfer to PMEM
  - Or transfer to DMA buffer then to NVME SSD
- Start pmemobj transaction
- Modify index to insert new extent
- Publish the reserve the space.
  - Either pmemobj_tx_publish() for SCM.
  - Or publish the space for NVMe SSD.
- Commit pmemobj transaction and reply to client

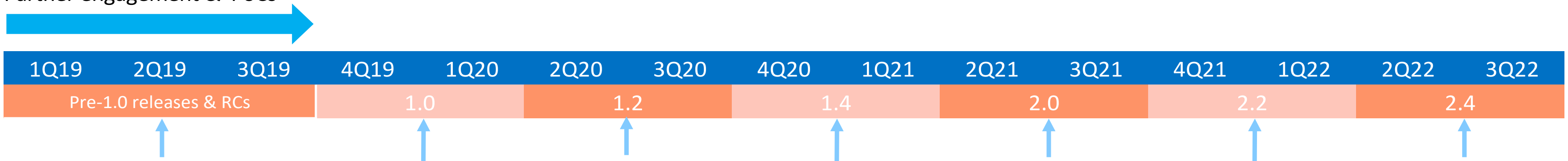
DAOS Performance


IOR Write - 1024 I/O size


IOR Read - 1024B I/O size

- IOR runs on remote clients sending the I/O requests to the single DAOS server over the fabric
  - Intel Omni-Path Host Adapter 100HFA016LS
- Using the DAOS MPI-IO driver with the full DAOS stack (client, network, server)
- Cascade Lake CPUs, 6 Dimms 512G AEP NMA1XBD512GQSE

DAOS Community Roadmap



Partner engagement & PoCs

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	

DAOS:
- Replication with self-healing
- Persistent Memory support
- NVMe SSD support
- Self monitoring & bootstrap
- Initial control plane
- python/golang API bindings

I/O Middleware:
- MPI-IO driver
- HDF5 DAOS Connector (proto)
- POSIX I/O (proto)

DAOS:
- Per-pool ACL
- Lustre integration

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

DAOS:
- End-to-end data integrity
- Per-container ACL
- SmartNICs & accelerators
- 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

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.

Resource

Source code on GitHub

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

Community mailing list on Groups.io

daos@daos.groups.io or https://daos.groups.io/g/daos

Wiki

http://daos.io or https://wiki.hpdd.intel.com

Bug tracker

https://jira.hpdd.intel.com