

DAOS Overview

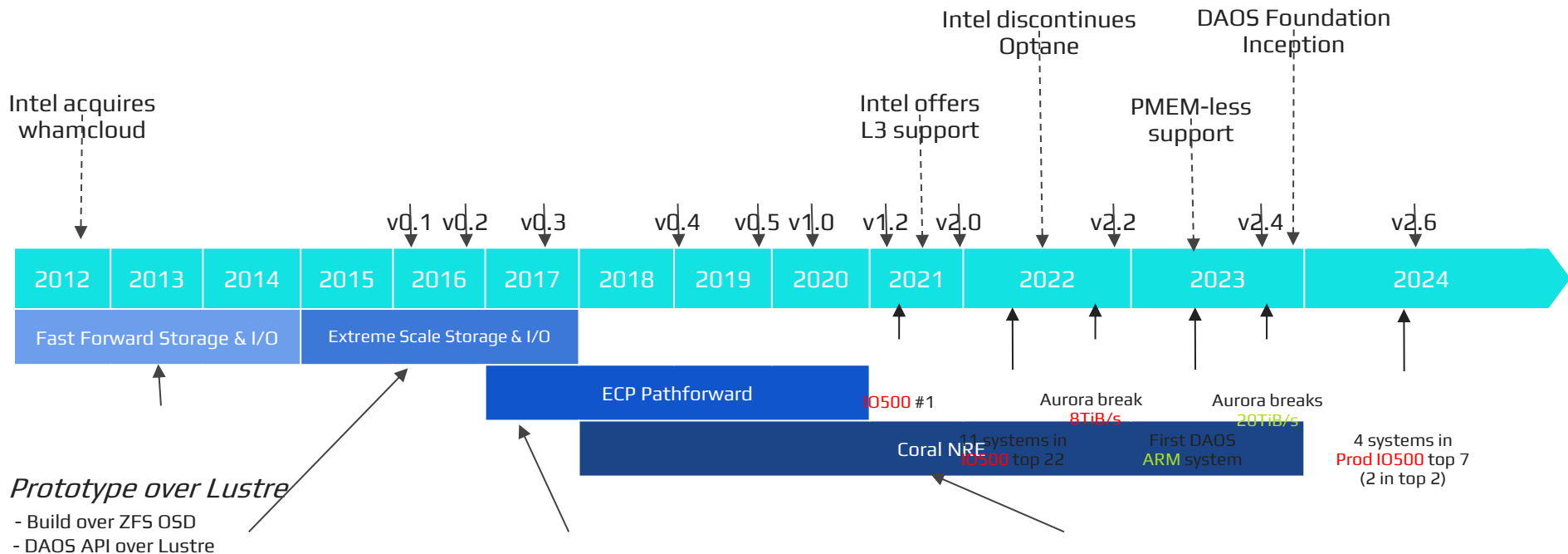
ISC25 Tutorial



<https://foundation.daos.io>



DAOS History



Standalone prototype

- OS-bypass
- Persistent memory via PMDK
- Replication & self healing

DAOS embedded on FPGA

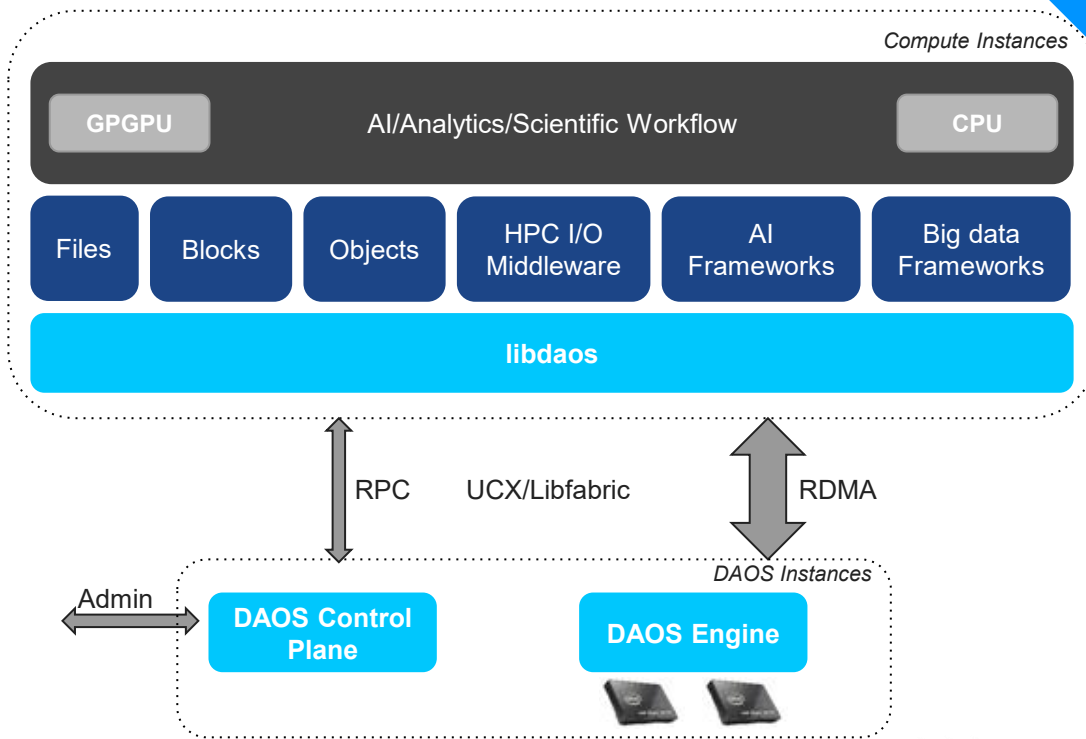
- Disaggregated I/O
- Monitoring
- NVMe SSD support via SPDK

DAOS Productization for Aurora

- Hardening
- 10+ new features
- Support for extra AI/Big data frameworks

DAOS: Nextgen Open Storage Platform

- Platform for innovation
- Files, blocks, objects and more
- Full end-to-end userspace
- Flexible built-in data protection
 - EC/replication with self-healing
- Flexible network layer
- Efficient single server
 - O(100)GB/s and O(1M) IOPS per server
- Highly scalable
 - TB/s and billions IOPS of aggregated performance
 - O(1M) client processes
- Time to first byte in O(10) μ s



DAOS Design Fundamentals

- No read-modify-write on I/O path (use versioning)
- No locking/DLM (use MVCC)
- No client tracking or client recovery
- No centralized (meta)data server
- No global object table
- Non-blocking I/O processing (futures & promises)
- Serializable distributed transactions
- Built-in multi-tenancy
- User snapshot

Scalability &
Performance

High IOPS

Unique
Capabilities

Aurora DAOS System



- 1024x DAOS Storage nodes
 - 2x Xeon 5320 CPUs (ICX)
 - 512GB DRAM
 - 8TB Optane Persistent Memory 200
 - 244TB NVMe SSDs
 - 2x HPE Slingshot NICs
- Supported data protection schemes
 - No data protection
 - All EC flavors: 2+1, 2+2, 4+1, 4+2, 8+1, 8+2, 16+1 and 16+2
 - N-way replication
- Usable DAOS capacity
 - between 220PB and 249PB depending on redundancy level chosen

Aurora System Specifications

Compute Node

2 Intel Xeon scalable "Sapphire Rapids" processors;
6 Xe arch-based GPUs; Unified Memory
Architecture; 8 fabric endpoints; RAMBO

CPU-GPU Interconnect

CPU-GPU: PCIe; GPU-GPU: Xe Link

Peak Performance

≈ 2 Exaflop DP

Platform

HPE Cray EX supercomputer

System Size (# Nodes)

> 9,000

Software Stack

HPE Cray EX supercomputer software stack + Intel
enhancements + data and learning

System Interconnect

Slingshot 11; Dragonfly topology with adaptive
routing

High-Performance Storage

≈ 230 PB, ≈ 25 TB/s (DAOS)

Aggregate System Memory

> 10 PB

GPU Architecture

Xe arch-based "Ponte Vecchio" GPU; Tile-based
chipslets, HBM stack, Foveros 3D integration, 7nm

Network Switch

25.6 Tb/s per switch, from 64–200 Gbs ports (25
GB/s per direction)

Programming Models






Intel oneAPI, MPI, OpenMP, C/C++, Fortran,
SYCL/DPC++

Node Performance (TF)

> 130



DAOS Performance - ISC'24 Production List

# ↑	INFORMATION								IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW	MD	REPRO.
									(GiB/s)	(KIOP/s)	
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	
3	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	
4	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	
5	ISC24	Zuse Institute Berlin	Lise	Megware	DAOS	10	960	324.54	65.01	1,620.13	

IOR & FIND

EASY WRITE	20,693.63 GiB/s
EASY READ	12,122.87 GiB/s
HARD WRITE	4,216.34 GiB/s
HARD READ	9,706.55 GiB/s
FIND	229,672.10 KiOP/s

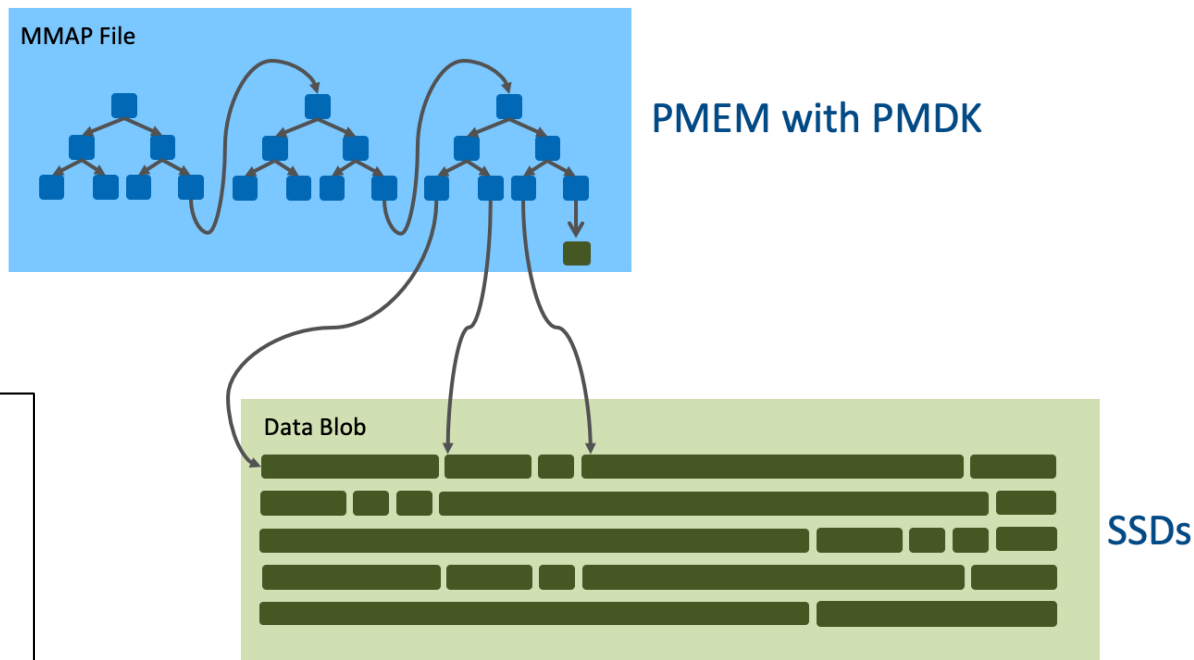
METADATA

EASY WRITE	60,985.13 KiOP/s
EASY STAT	225,295.35 KiOP/s
EASY DELETE	57,648.44 KiOP/s
HARD WRITE	33,827.19 KiOP/s
HARD READ	141,467.16 KiOP/s
HARD STAT	230,086.03 KiOP/s
HARD DELETE	62,196.78 KiOP/s

Aurora IO500 Run

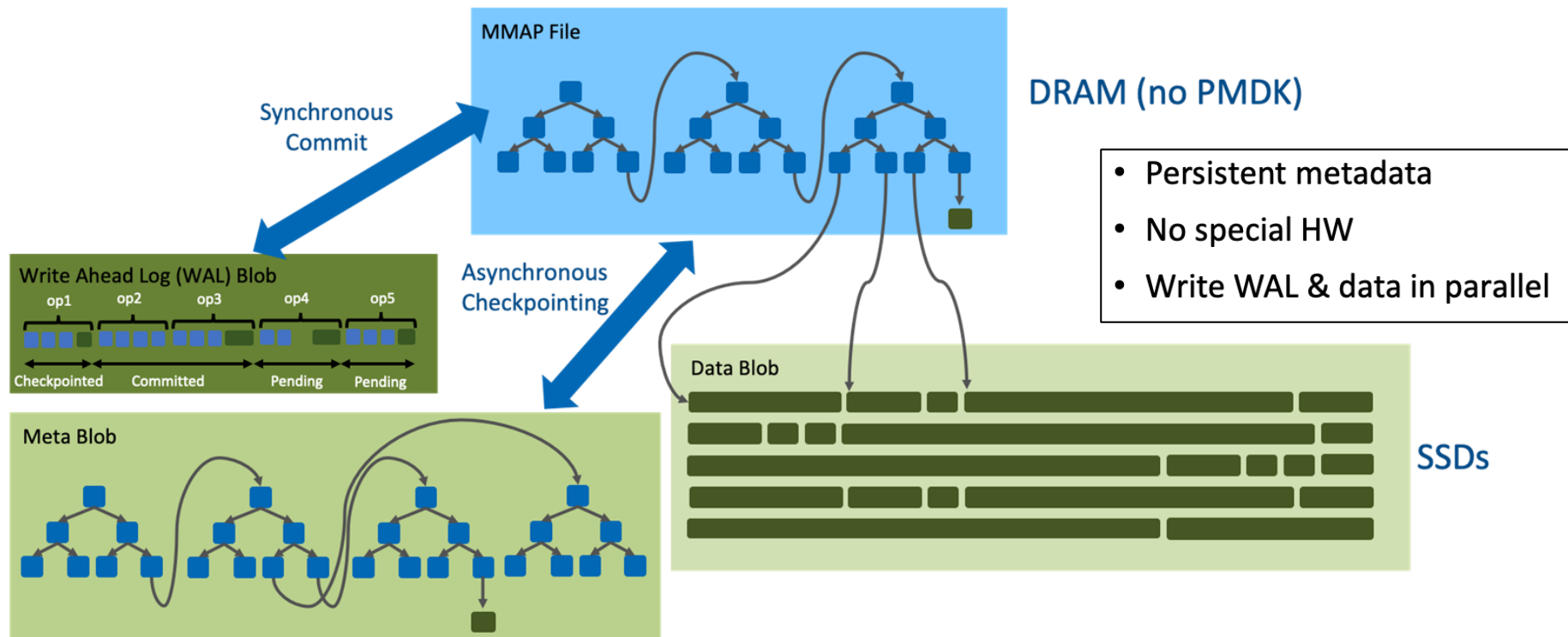
Features	Limits
Number of client nodes	512
Number of client endpoints	4k
Number of client processes	53k
Number of DAOS servers	642
Number of DAOS engines	1284
Largest Pool	160PiB
Largest file	8.5PiB
Total number of files	177 Billions
Number of files in a single directory	33 Billions

DAOS Architecture Evolution: Pmem Mode

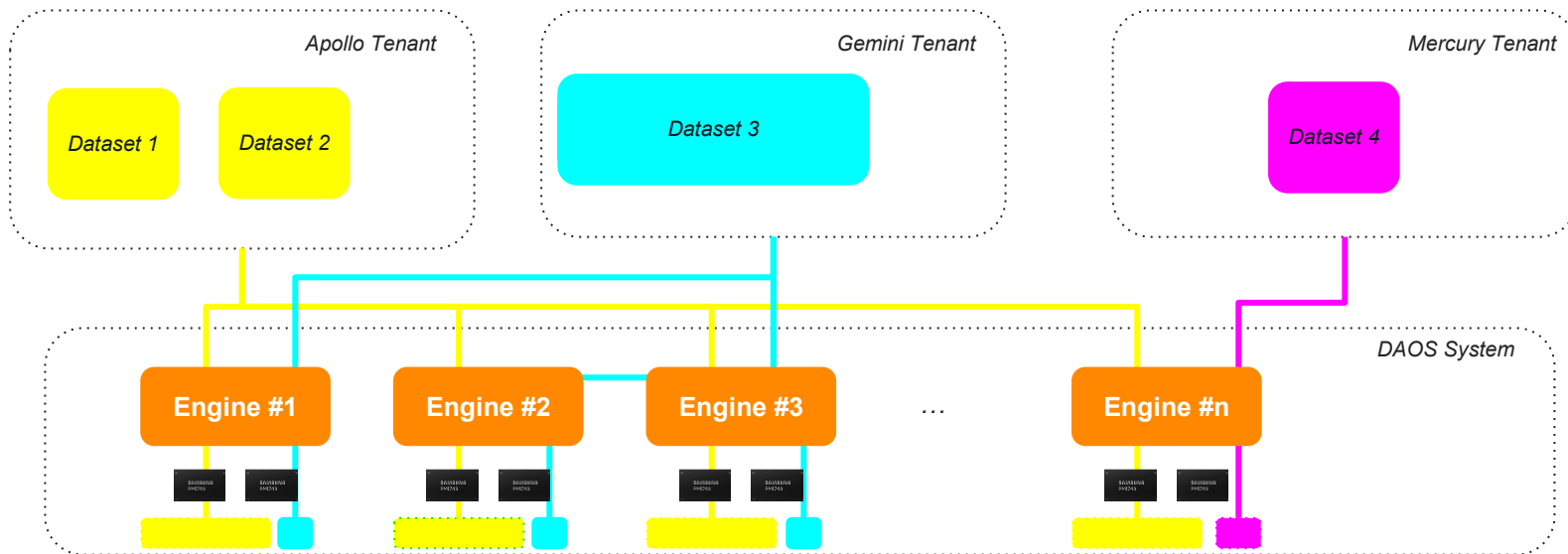





- Persistent metadata
- Require Intel Optane PMEM (or NVDIMM-N)
- App Direct mode
- Mode used on Aurora

DAOS Architecture Evolution: Pmem-less Mode



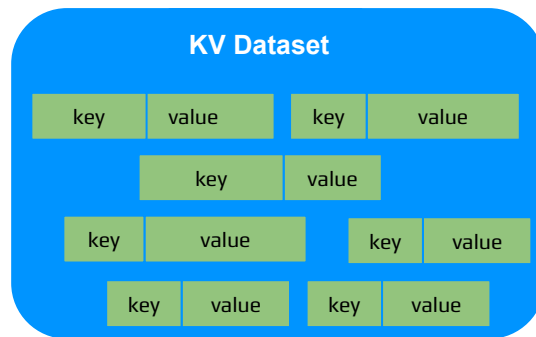
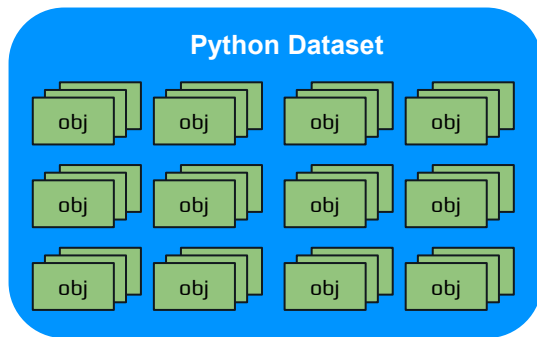
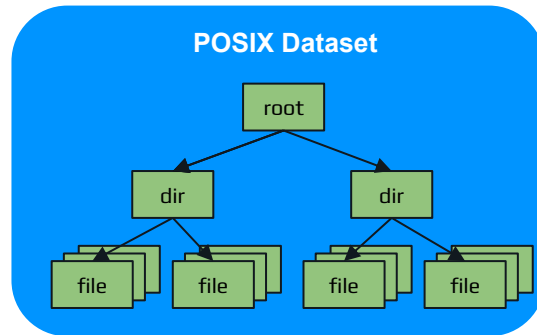
Storage Pooling - Multi-tenancy



Pool 1		Apollo Tenant	100PB	20TB/s	200M IOPS
Pool 2		Gemini Tenant	10PB	2TB/s	20M IOPS
Pool 3		Mercury Tenant	30TB	80GB/s	2M IOPS

Dataset Management

- New data model to unwind 30+y of file-based management
- Introduce notion of dataset
- Basic unit of storage
- Datasets have a type
- POSIX datasets can include trillions of files/directories
- Advanced dataset query capabilities
- Unit of snapshots
- ACLs/IAM



Object Interface

Middleware/Framework View

DAOS Layout View

Mapping

Object

128-bit
object Identifier

Array

Multi-dimensional
Array

Key-value
Store

Multi-level
Key-value
Store

- No object create/destroy
- No size, permission/ACLs or attributes
- Sharded and erasure-coded/replicated
- Algorithmic object placement
- Very short Time To First Byte (TTFB)



python

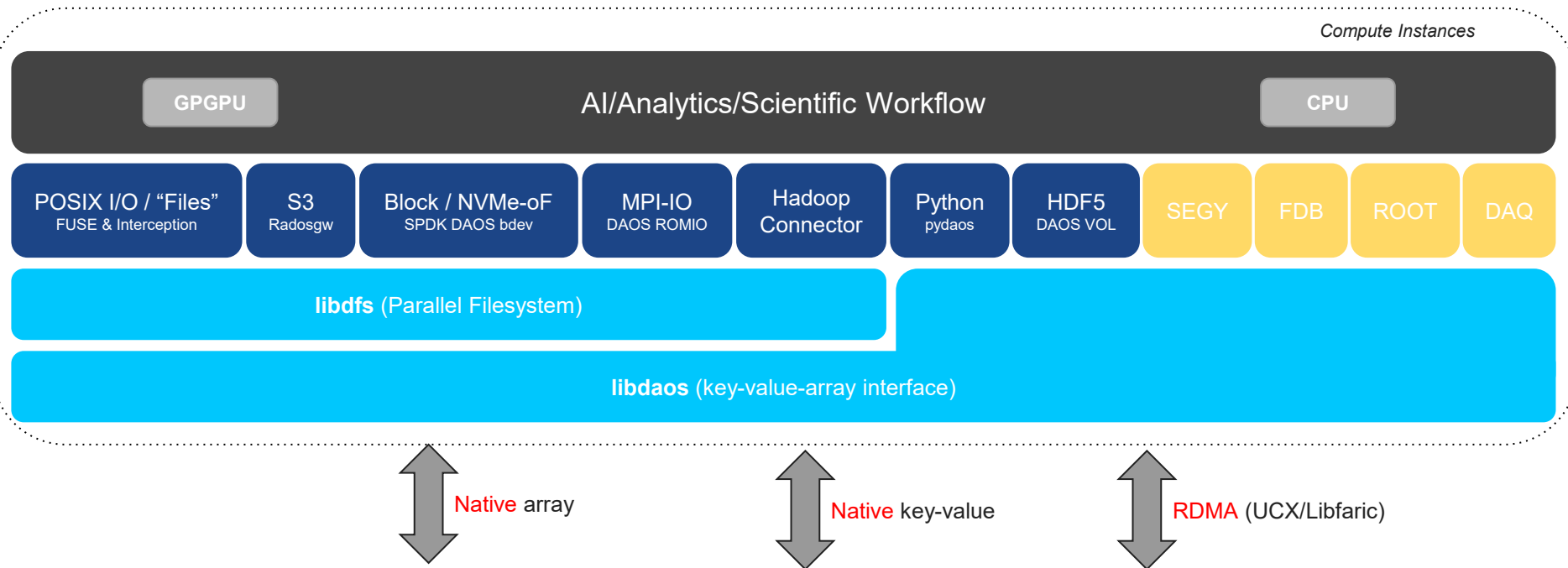
PyTorch



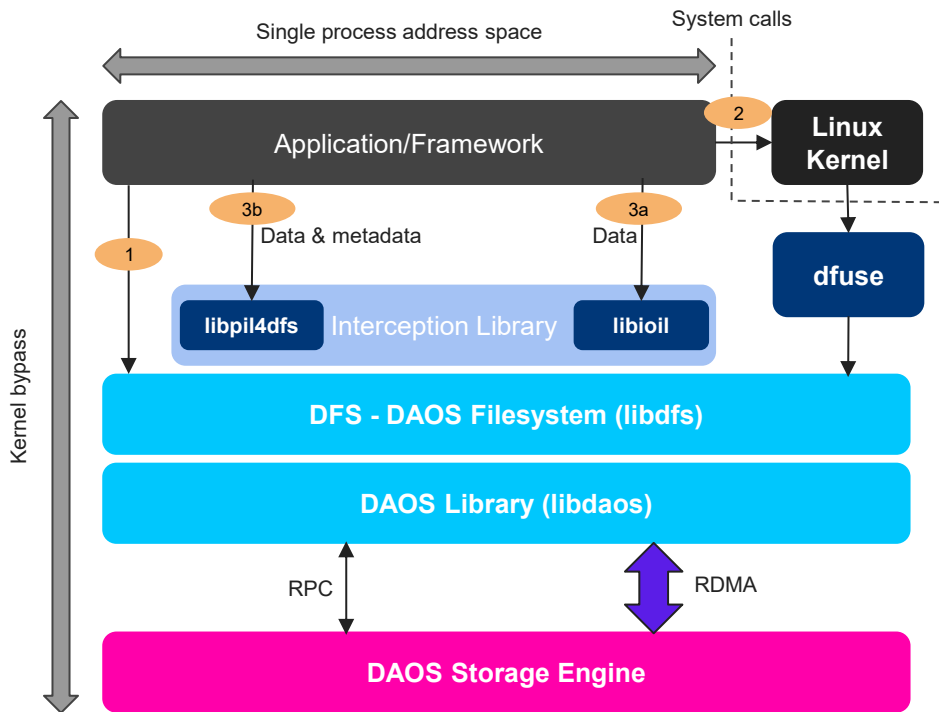
TensorFlow



Software Ecosystem



POSIX Support & Interception



1. Userspace DFS library with API like POSIX
 - **Require** application changes
 - Low latency & high concurrency
 - No caching
 2. DFUSE daemon to support POSIX API
 - **No** application changes
 - VFS mount point & high latency
 - Caching by Linux kernel
 3. DFUSE + Interception library
 - **No** application changes
 - 2 flavors using LD_PRELOAD
- 3a libioil
- (f)read/write interception
 - Metadata via dfuse
- 3b libpil4dfs
- Data & metadata interception
 - Aim at delivering same performance as #1 w/o any application change
 - Mmap & binary execution via fuse

Resources

- Foundation website: <https://daos.io/>
- Github: <https://github.com/daos-stack/daos>
- Online doc: <https://docs.daos.io>
- Mailing list & slack: <https://daos.groups.io>
- YouTube channel: <http://video.daos.io>
- 9th DAOS User Group (DUG'25) at SC'25 in St Louis

