DAOS Foundation

ISC 2024 Meeting

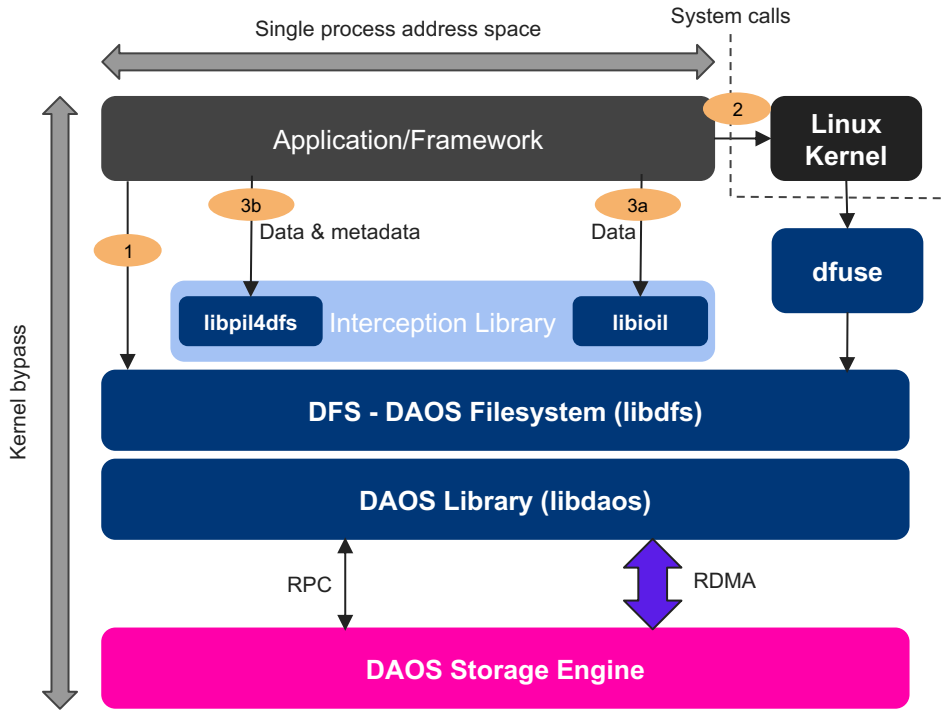
THE LINUX FOUNDATION

https://foundation.daos.io

daos

Roadmap

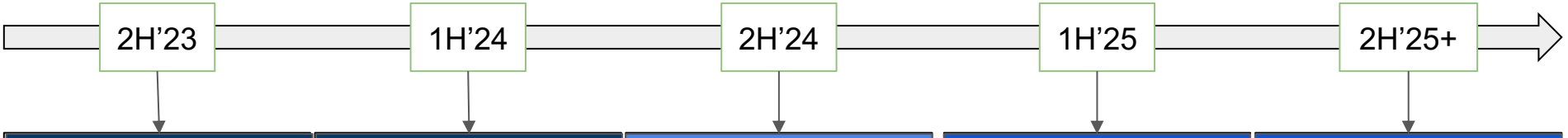
POSIX Support & Interception



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



DAOS 2.4
(Intel Release)

OS Packages:
- Leap 15.4/15.5
- RHEL/Rocky/Alma 8.6/8.8

Networking:
- HPE Slingshot support
- UCX Support
- Multiple engines per socket

Features:
- online drain/reintegration/add
- per-server fault domain
- EC parity rotation
- Dfuse readdir caching
- Streaming functions in libioil

Tech preview:
- pmem-less support phase 1
- pil4dfs

UX Improvements:
- LED management via Intel VMD

DAOS 2.6
(Intel Release)

OS Packages:
- Leap 15.5
- RHEL/Rocky/Alma 8.8/9.2

Networking:
- Change provider w/o reformat
- MD duplicate RPC detection

Features:
- pmem-less support phase 1
- libpil4dfs
- Intel VMD hotplug
- Delayed rebuild

Tech preview:
- Catastrophic recovery

UX Improvements:
- Improved version interoperability

DAOS 2.8
(DAOS Foundation Release)

OS Packages:
- Leap 15.6
- RHEL/Rocky/Alma 8.10/9.4
- Ubuntu 22.04

Networking:
- VNI support
- iWARP support

Features:
- Optimized object placement
- Mount POSIX snapshots RO
- Client telemetry

Tech preview:
- pmem-less support phase 2

UX Improvements:
- Reintegration of all pools
- Additional RAS events

DAOS 3.0
(DAOS Foundation Release)

OS Packages:
- Leap 15.6 (x86_64)
- RHEL/Rocky/Alma 8.x
- Ubuntu 22.04
- RHEL/Rocky/Alma 9.x (ARM64/x64_64)

Networking:
- Multihomed network support

Features:
- pmem-less support phase 2
- WORM containers
- DSA optimizations
- Lustre integration
- Container parking/serialization
- Catastrophic recovery
- Fault domains beyond servers
- Engine scalability improvements

Tech preview:
- Rolling upgrade
- Incremental rebuild/reintegration
- Omni-Path Express support

DAOS 3.x
(DAOS Foundation Release)

OS Packages:
- Leap 15.7
- RHEL/Rocky/Alma 8.x
- Ubuntu 22.04
- RHEL/Rocky/Alma 9.x (ARM64/x64_64)

Features:
- Pool resizing
- Inline compression (QAT)
- Inline encryption (QAT)
- Inline deduplication
- Middleware catastrophic recovery
- Progressive layout
- Pipeline API
- SQL support with predicate pushdown
- Distributed transactions
- Pool/container freeze
- CXL SSD support / QLC
- Tiered container phase 1

THE LINUX FOUNDATION

Color coding schema:
- Committed (or released) release/features
- In-planning release/features
- Future release/features

daos

2.8 Plan

- 2.8 milestones
  - Test builds
  - Feature freeze
  - Code freeze
  - Release candidates
  - Release: shooting for Q4’24
- Generate more release candidates
  - Give opportunity for more community testing
- Train model

Networking

- Slingshot VNI support
  - Isolate jobs running on the fabric with virtual
  - All jobs should still access the multi-tenant storage system
  - Block of VNIs allocated to DAOS at initialization time
- iWARP support
  - RDMA over tcp/ip
  - Supported via libfabric verbs provider
  - To be tested/validated

Client Telemetry

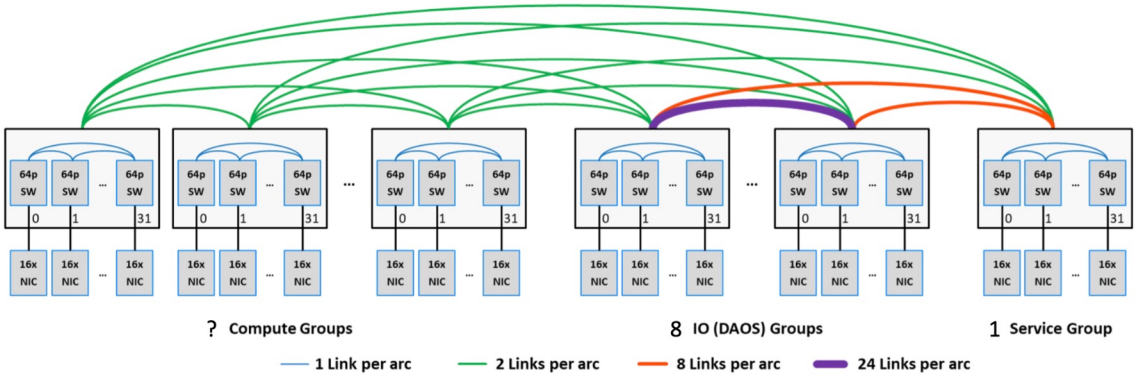
- Infrastructure already landed for 2.6
- Leverage initial contribution from HPE
- Restructured by Intel and Google
- Client metrics accessible via multiple ways:
  - Prometheus endpoint exported by agent
  - Exported to csv file when job terminates
- Exploring alternatives:
  - Darshan
  - Dump to a “metrics” container?

POSIX Snapshot

- Snapshot already supported for any container type via daos utility
  - Identified by epoch number
- Add ability to access a snapshot of a POSIX container
- libdfs
  - Mount libdfs and “switch” mountpoint to a snapshot
  - Must have no active dfs objects to perform the switch
- dfuse
  - All snapshots accessible via .snapshot directory under the root
  - Accessible read-only
  - Can copy files/subtree from .snapshots/ back to main namespace
- Can delete snapshots via daos utility
  - Will automatically disappear from .snapshots directory

Optimized Object Placement

- Placement aware of the network topology
  - E.g. DragonFly fabric
- Performance domains complementary to fault domains
- Different strategies
  - Spread redundancy groups as widely as possible across different performance domains
  - Spread redundancy groups across as fewer performance domains as possible



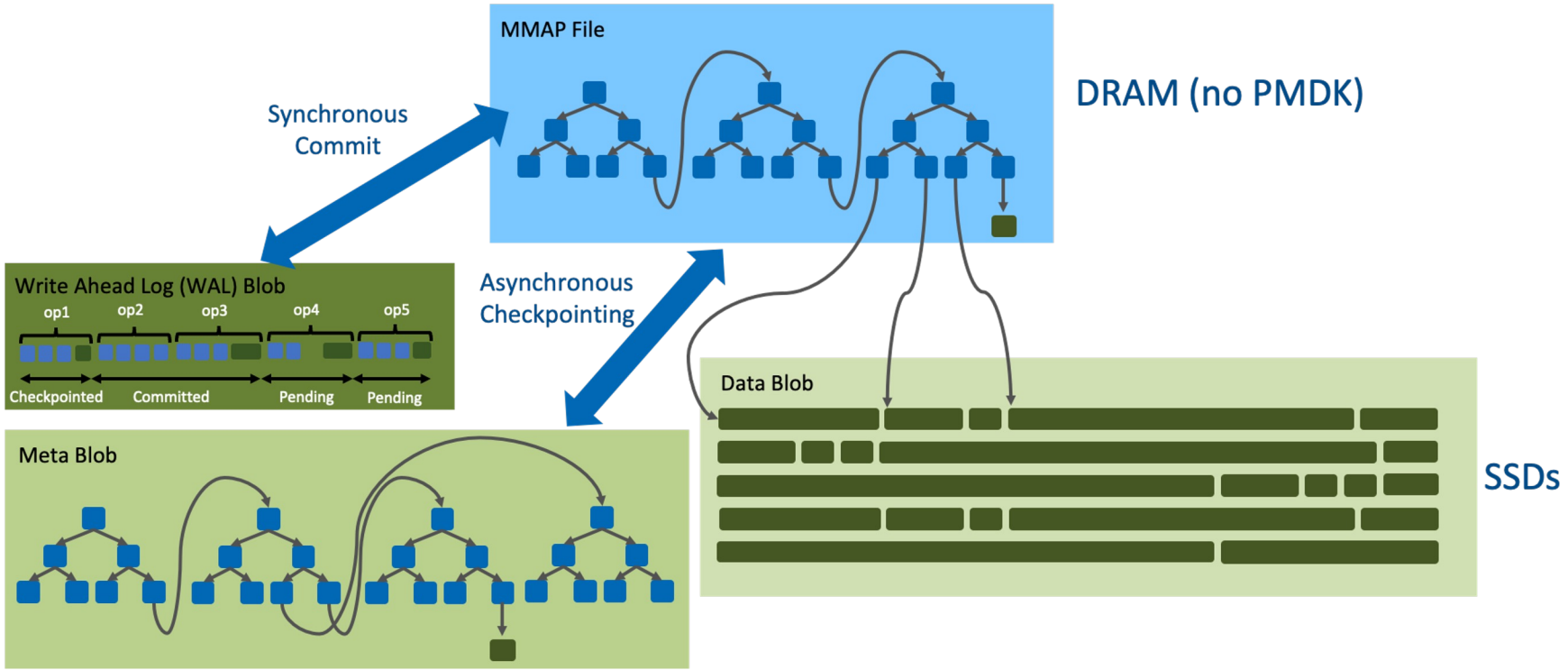
Aurora Network Topology
(from Kevin Harms' presentation at DUG'22)

daos

UX Improvements

- Reintegration of all pools
  - Upon exclusion of an engine from a pool, it must be reintegrated manually
  - No automatic reintegration (bad idea, right?)
  - Must be reintegrate from each pool individually
    - dmg pool reintegrate pool1 –rank=20; dmg pool reintegrate pool2 –rank=20; …
  - Provide a way to reintegrate an engine from all the pools it was excluded in a single command line
- Additional RAS events
  - Generate notifications when some operations are completed on the cluster
  - E.g. a new pool or container is created

Pmem-less Support Phase 2 (tech preview)



Pmem-less Support Phase 2 (tech preview)

- Phase 1 GA in 2.6
  - Meta blob size = mmap size
  - Work to reduce VOS memory footprint
  - ~600M 4K files per TB of RAM (using S1)
- Phase 2
  - Meta blob size > mmap size
  - Memory bucket allocator with dynamic eviction
    - 64MB bucket flushed to metadata blob
    - Ability to flush to data blob (3.0 scope)
  - Object flattening and eviction (3.0 scope)
    - Small objects flattened in metadata blob
    - Fetch entirely on cache miss