

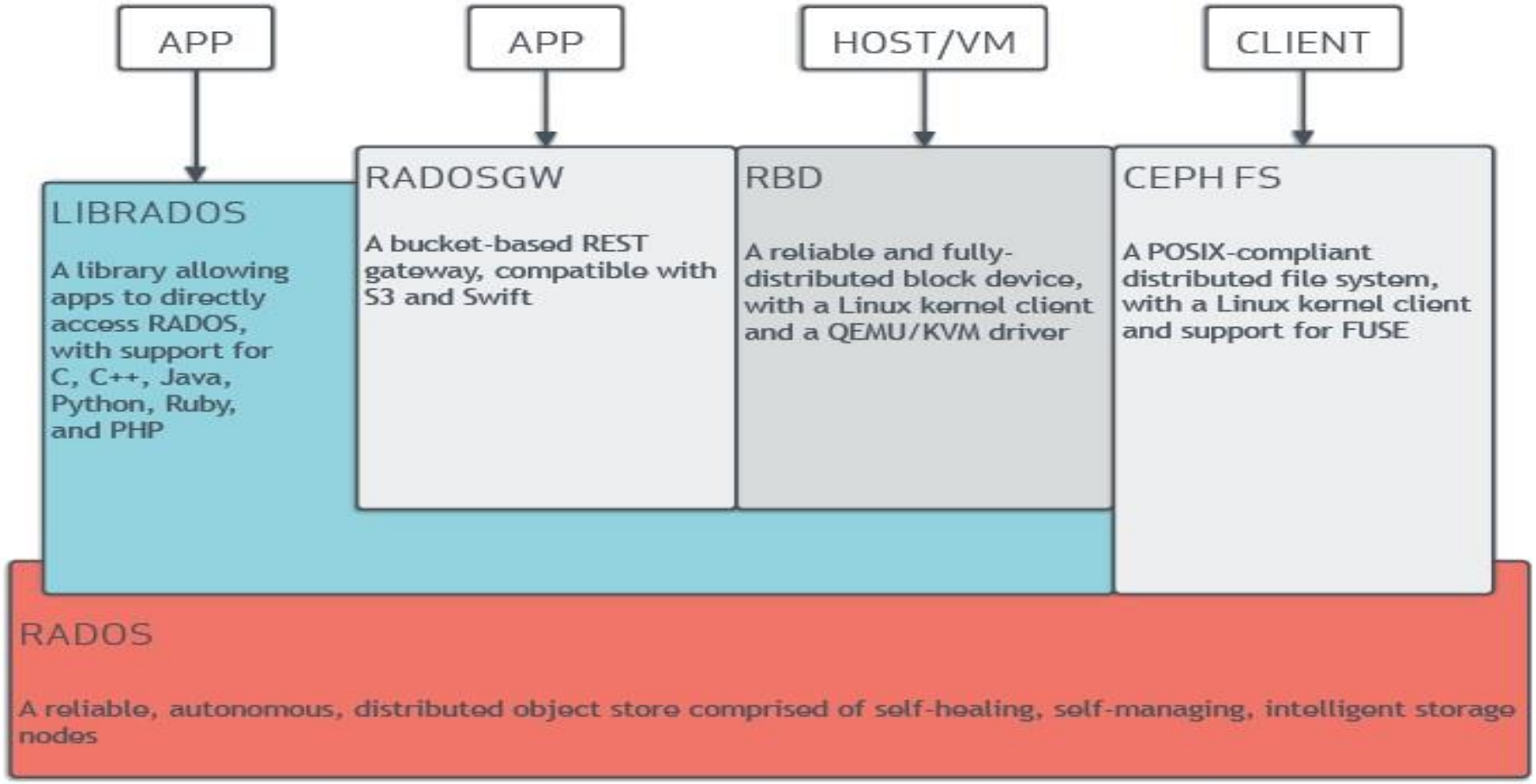
Introduction to client-side caching in Ceph

Mahati Chamarthy, Software Engineer, Intel

Agenda

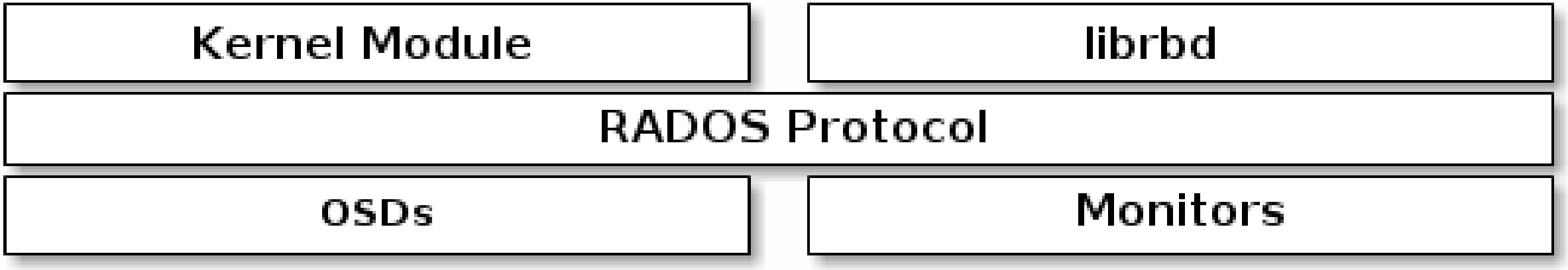
- Ceph, RBD
- Caching Status
- Writeback cache
  - SSD based writeback cache
  - Replicated Write Log (RWL) based on persistent memory – Intel Optane DC persistent memory module
- Performance

Ceph Architecture



Rados Block Device (RBD)

- RBD images – block devices striped over objects and stored in RADOS object store
- Stripes images across the entire cluster
- Common use-case: disks for VMs



Rados Block Device (RBD) Features

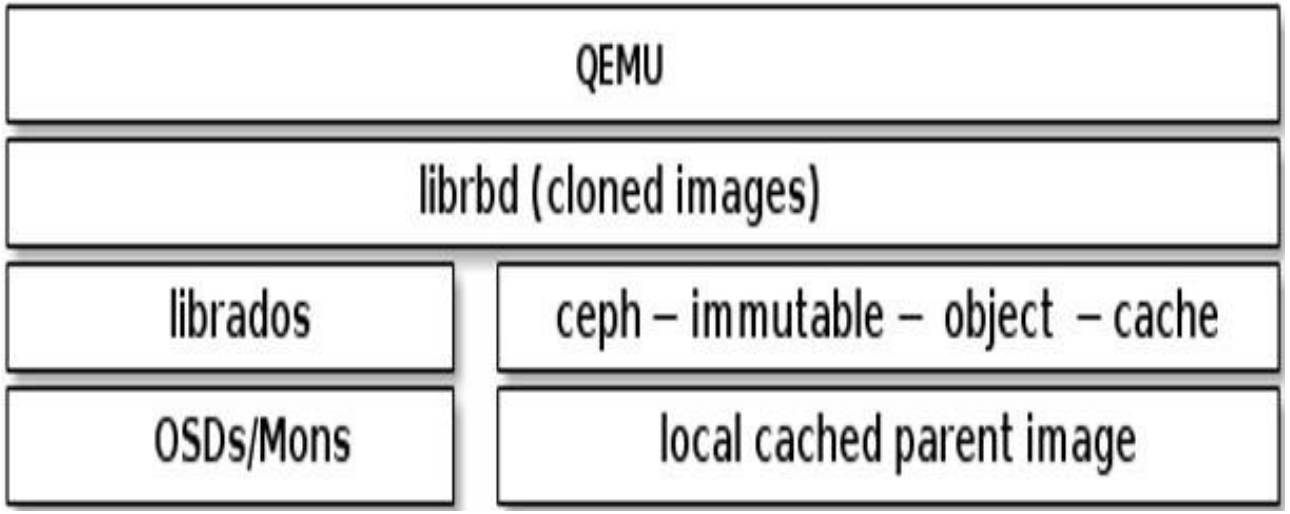
- Snapshots
- Copy-on-write
- Exclusive-lock
- Mirroring
  - Per-pool, per-image
- Journaling

Caching Status

- In-memory caching
  - Writearound
  - Writethrough
  - Writeback
- Smaller Capacity
  - non persistent

Caching Status

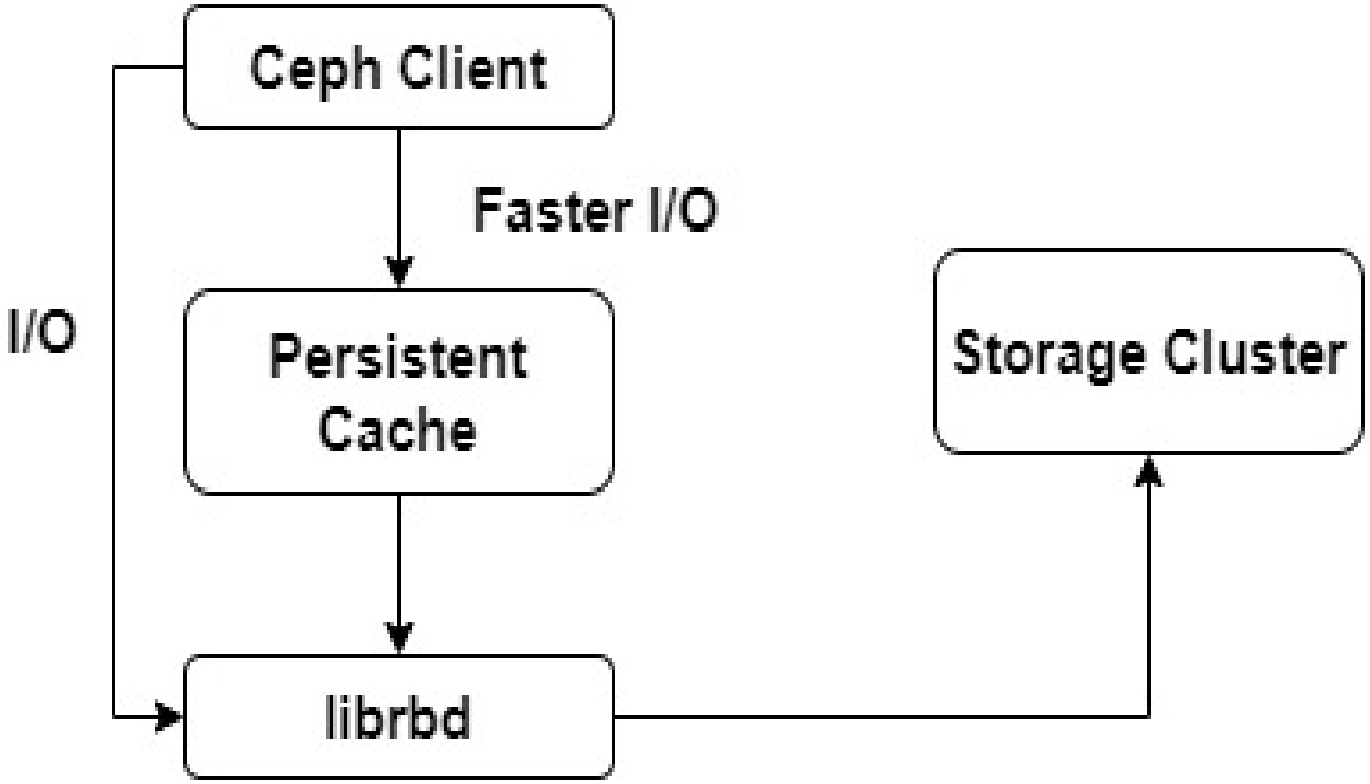
- Read-only parent image-cache
  - Use-case for cloned images
- Requires explicit enabling
- Ceph-immutable-object-cache daemon



WRITEBACK CACHE

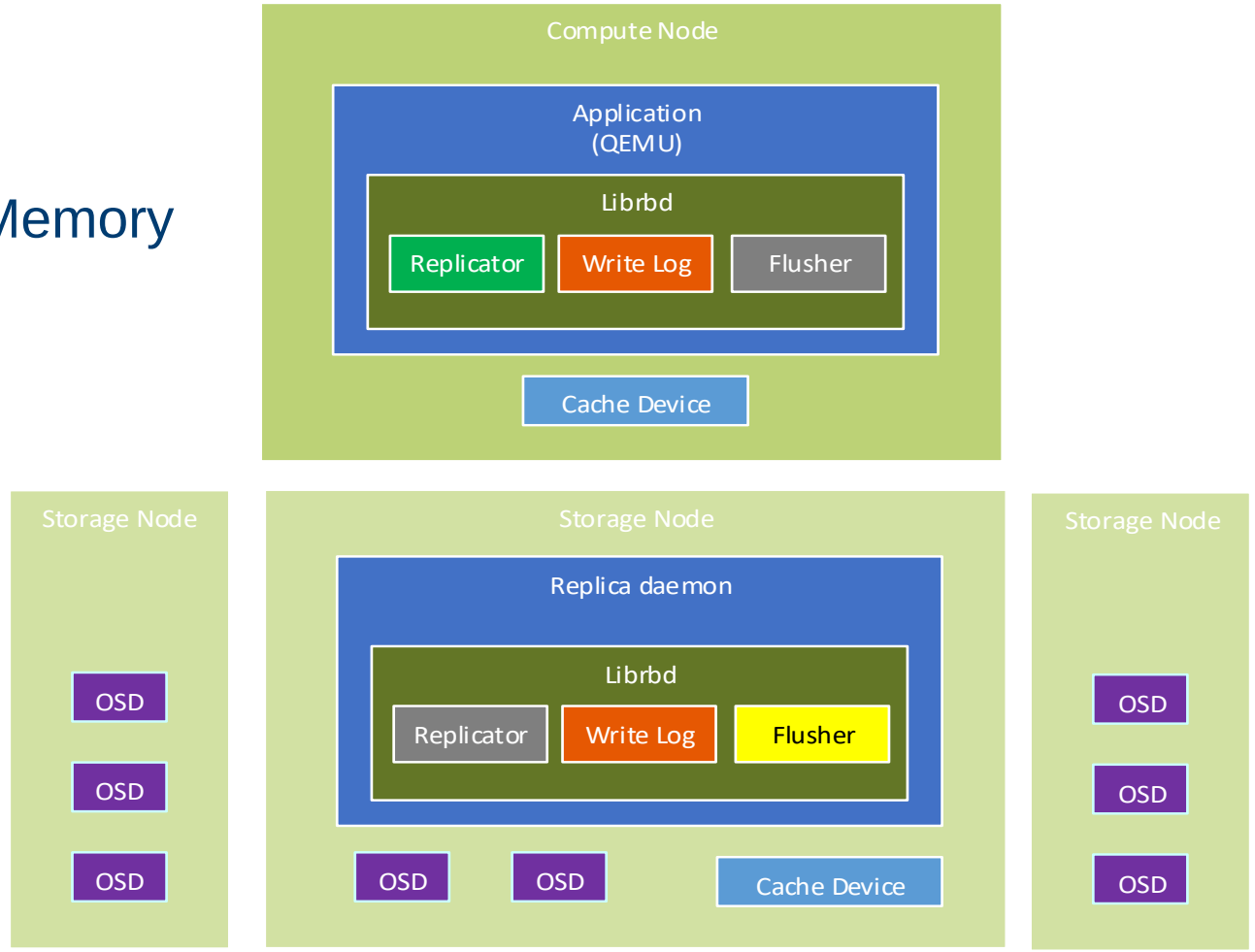
Writeback cache

- RWL – Replicated Write Log
- SSD writeback cache



RWL - Architecture

- Ordered, crash-consistent WB
  - Intel® Optane DC Persistent Memory module
- Replication via RDMA

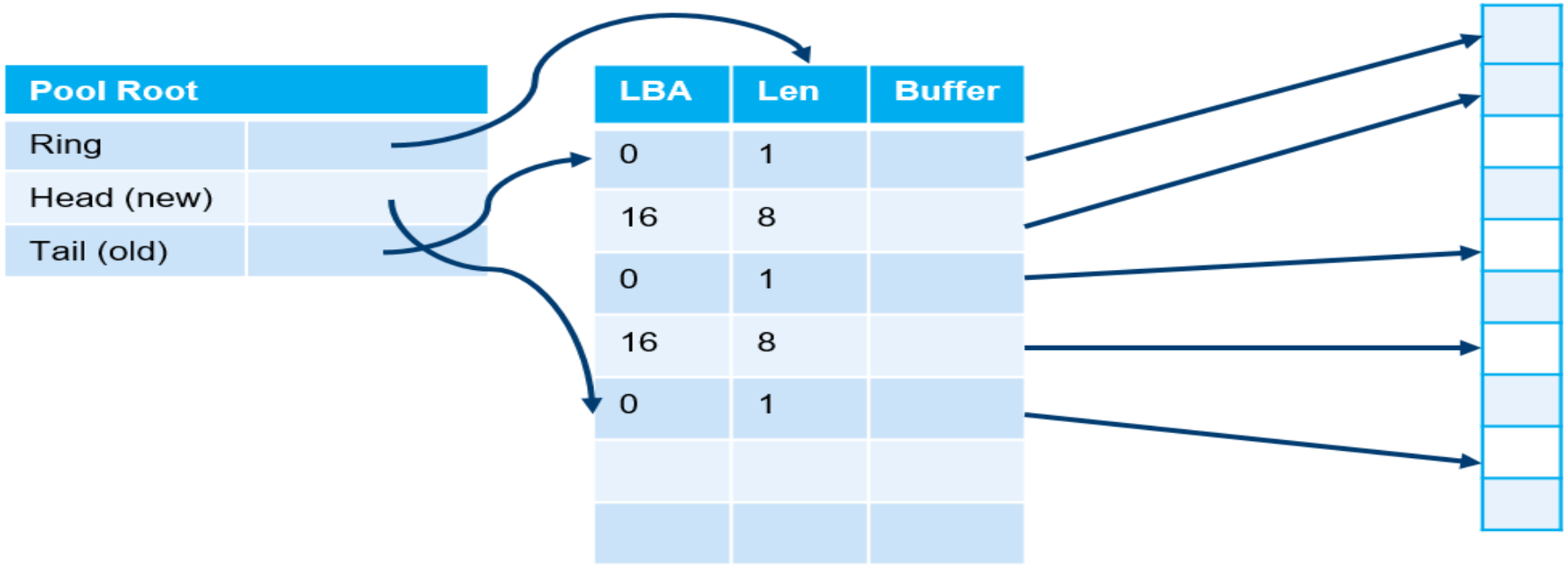


Writeback cache – persistent memory

- The Persistent Memory Development Kit (PMDK) - growing collection of libraries and tools.
- Libraries part of pmdk
  - **libpmemobj**
- https://pmem.io/pmdk/

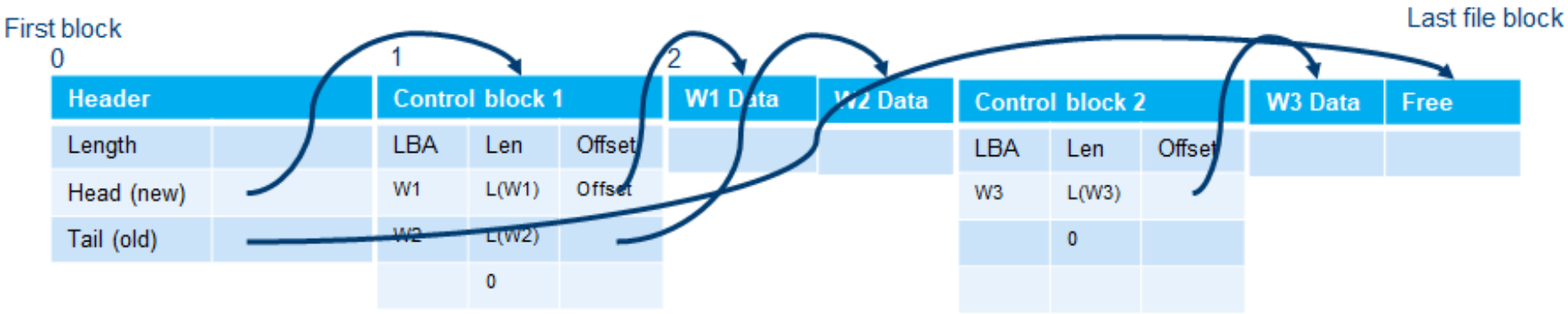
Data layout – persistent memory

- Byte-addressable
- Pool root, journal and data buffer
- libpmemobj used to write to the data buffer



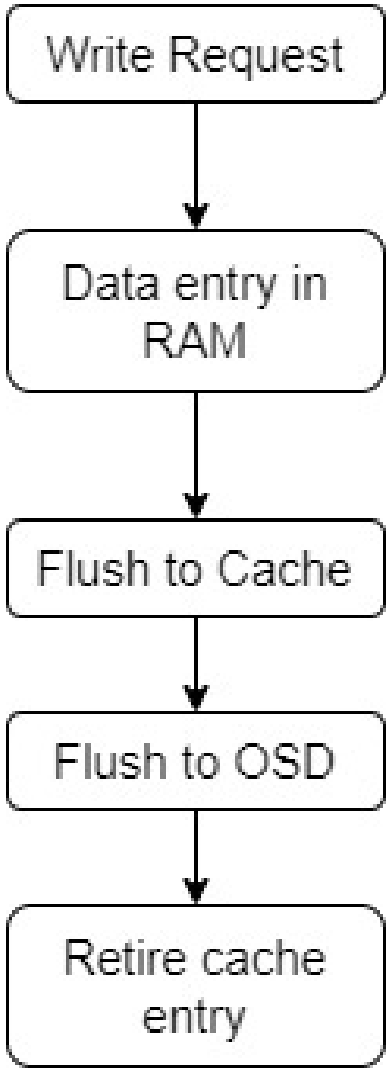
Data layout – SSDs

- block-addressable
- Superblock, control block, data buffer
- filesystem access, no page cache



Request flow

- User request to librbd
- Reserve data buffer - create an entry in RAM
- Copy and persist or replicate data
- Schedule log entry
- Persist log entries
- Flush to cache
- Flush to OSDs
- Retire cache data

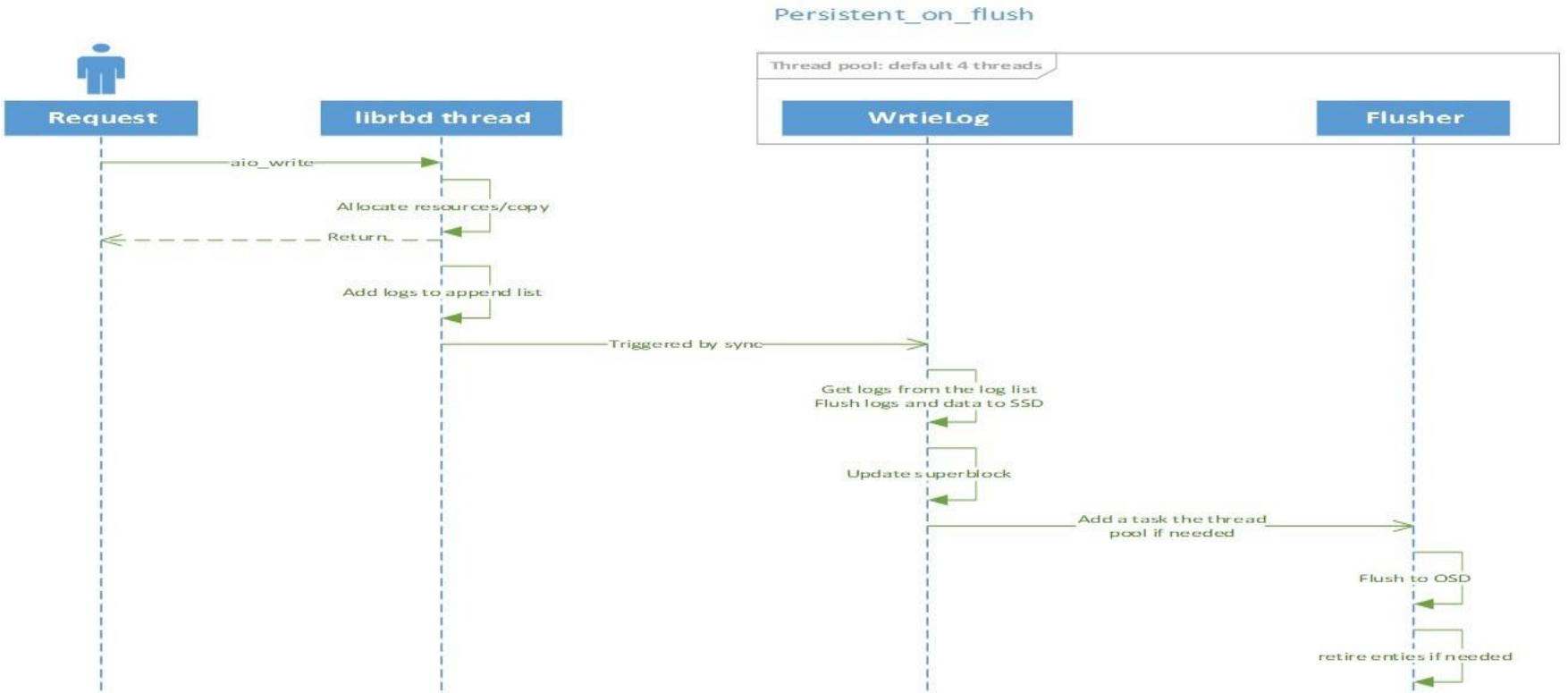


Flush Modes

- Persist-on-write
  - barrier (sync point) for every write request
- Persist-on-flush
  - Barrier (sync point) on every user flush request

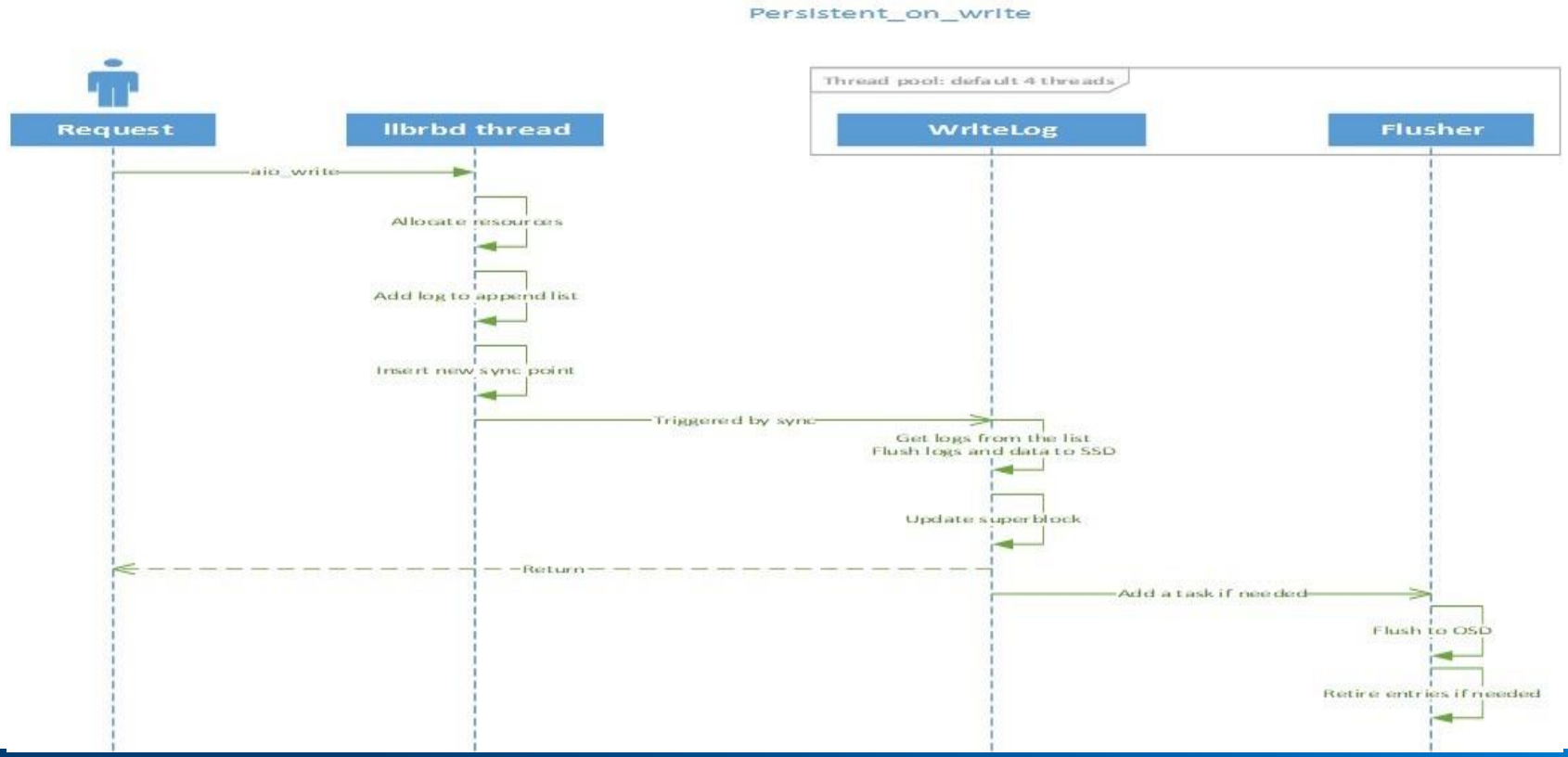
Persist-on-Flush

- Entry created in RAM, along with buffers holding user data
- Request is queued and is completed to user
- Flushed to cache on user flush request



Persist-on-write

- Entry created in RAM, along with buffers holding user data
- Create barrier for every write request and flush to cache
- Complete request to user



Failure scenarios for RWL

- Writing node failure
  - Flushed on node reboot or by the replica node
- Replica node or replica NVDIMM failure
  - Writing node will proceed
  - New replica is chosen
- Writing node NVDIMM failure
  - Takes advantage of persist-on-flush mode

PERFORMANCE

Test config – SSD

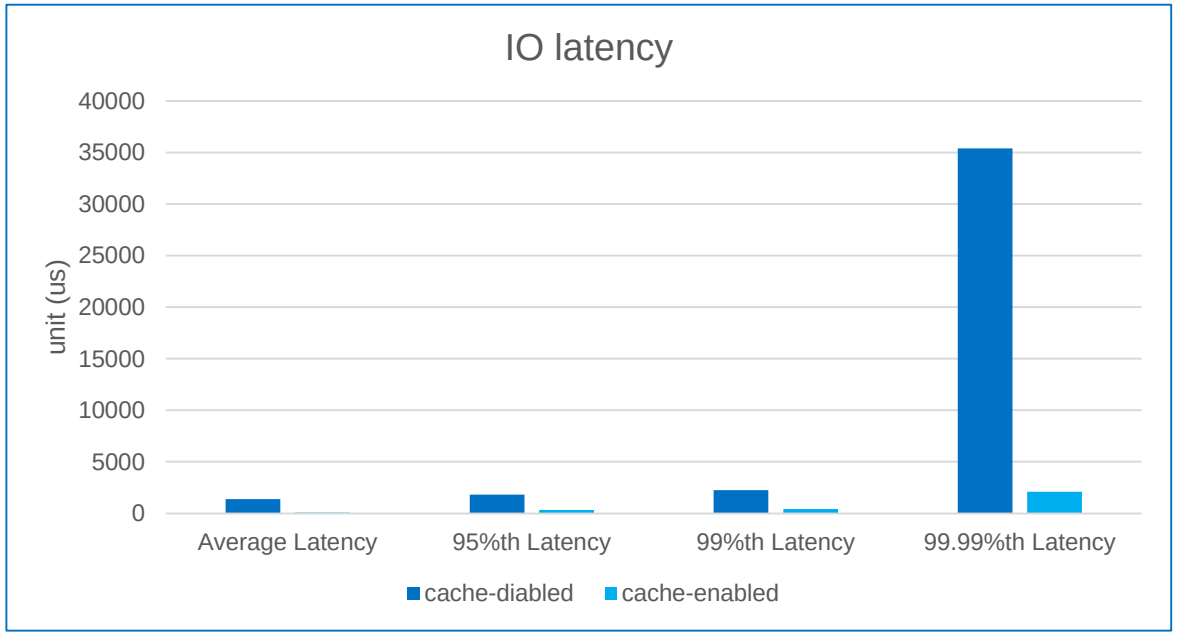
- Ceph Client
  - Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz
  - Intel® Optane™ SSD DC P4800X 375G
- Ceph cluster – one storage node
  - Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz
  - One Intel® SSD DC P3700 Series 400G
- IO
  - 10G rbd image size, Cache_size = 1G
  - FIO + librbd, 4K random writes
  - RBD IOPS Limit = 4000; sync request every 32 IOS
  - ramptime = 10 min, runtime=10 min

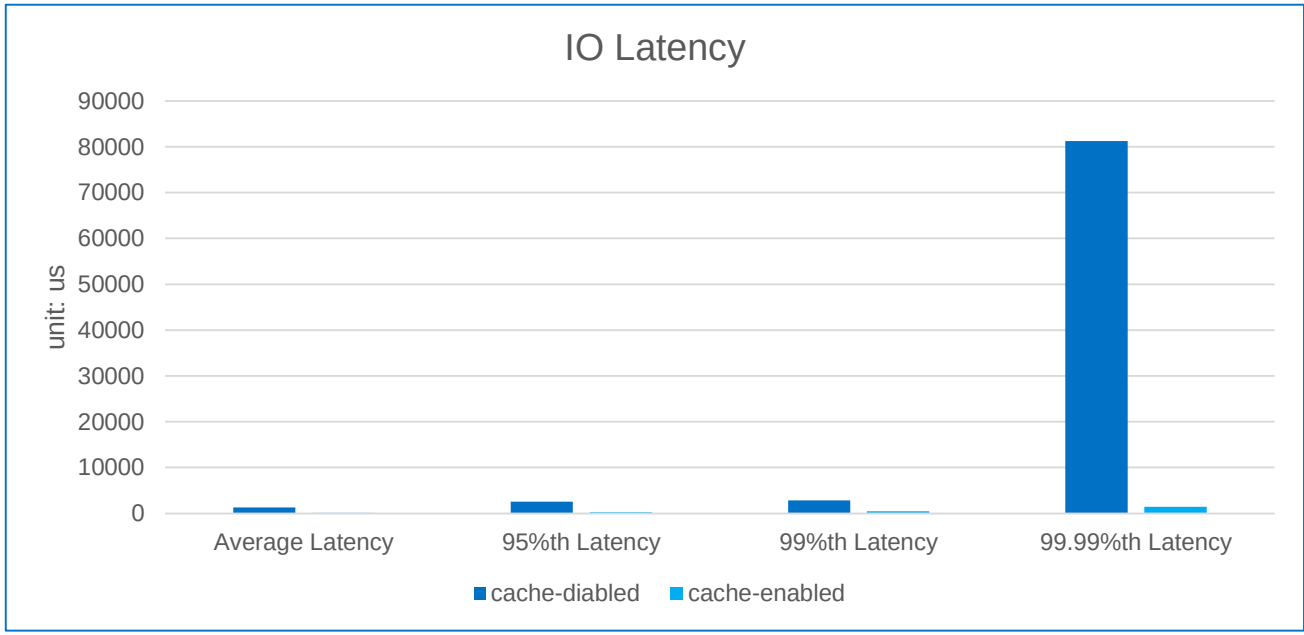
Fio – single rbd image
94% better tail write latency

	Average Latency	95%th Latency	99%th Latency	**99.99%th Latency**
Cache-disabled	1382	1811	2245	**35390**
Cache-enabled	89	326	424	**2089**

Fio – multiple rbd images
98% better tail write latency

cache	Average Latency	95%th Latency	99%th Latency	**99.99%th Latency**
cache-diabled	1320	2540	2868	**81265**
cache-enabled	83	285	429	**1450**

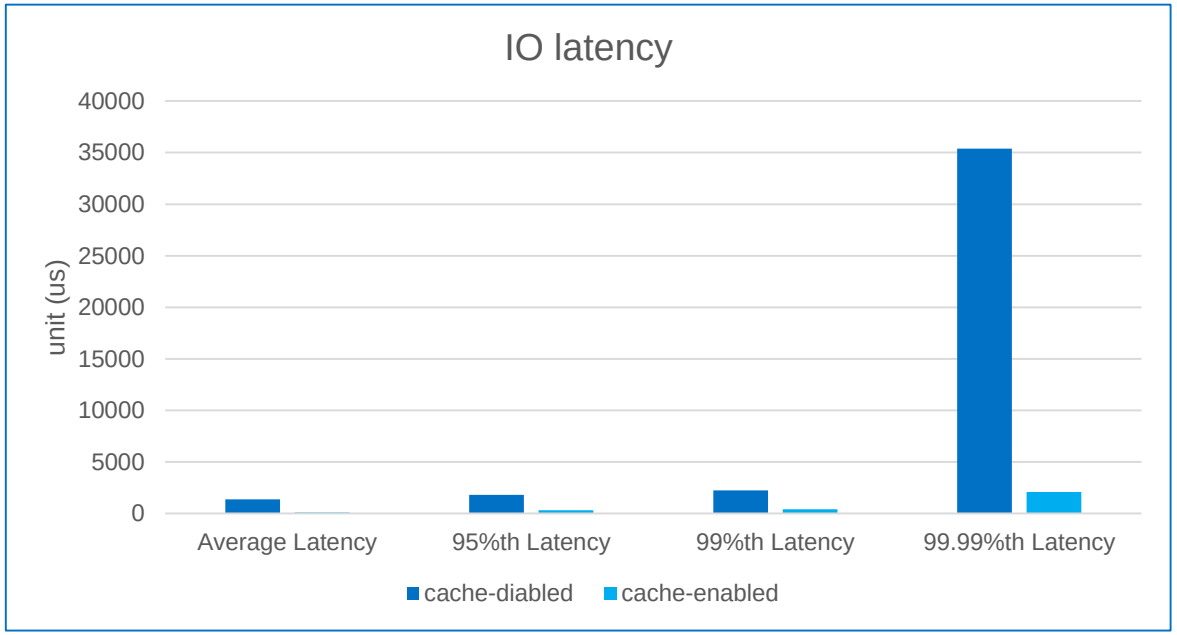




Test config - RWL

- Ceph Client
  - Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz, 187GB Memory
  - 1x AEP 246G
- Ceph cluster – 3 storage nodes
  - Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz, 376G Memory
  - 4x Intel NVME SSD 1.5T
- IO
  - 100G rbd image size, Cache_size = 1G
  - FIO + librbd, 4K random writes
  - ramptime = 0, runtime=10 min

	Average Latency	90%th Latency	99%th Latency	**99.99%th Latency**
Cache-disabled	1382	1811	2245	**35390**
Cache-enabled	182	321826	529	**2409**



Upstream Status

- RWL (persistent memory) – merge under way
  https://github.com/ceph/ceph/pull/31963/
- SSD cache – development under way with 2020 merge target

THANK YOU!

Questions?