SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

Mortimer: A High-Performance Scale out Storage for Persistent Memory and NVMe SSDs

Anjaneya “Reddy” Chagam, Cloud Architect
Intel Corporation

Notices & Disclaimers

Intel technologies may require enabled hardware, software or service activation. Your costs and results may vary.

No product or component 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'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.

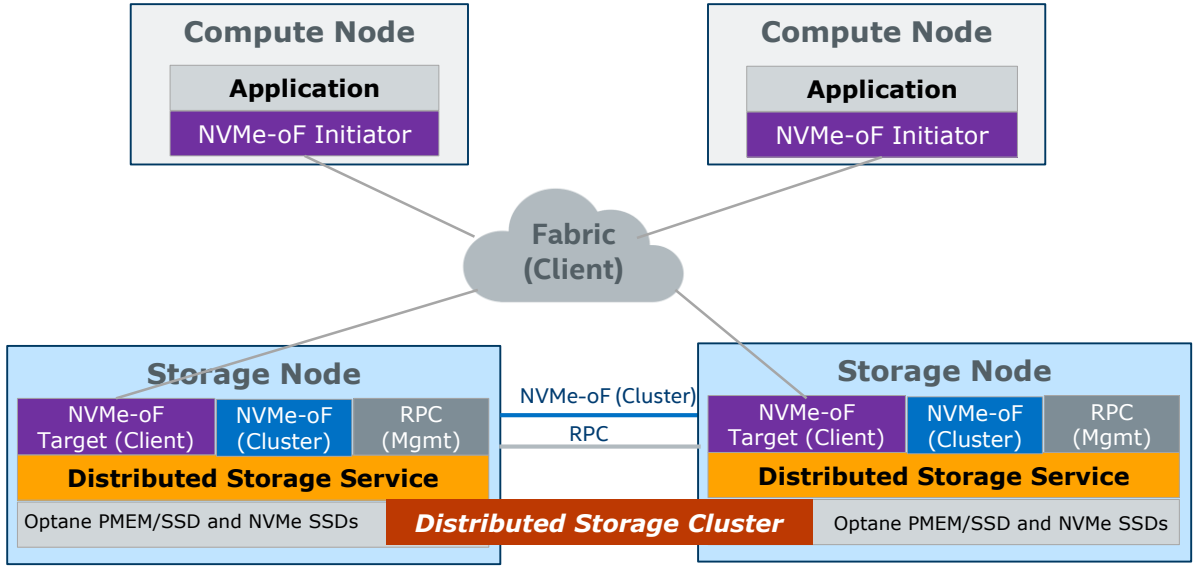
.

Agenda

- Mortimer Overview
- Motivation
- Mortimer Architecture
- Mortimer Metadata Design
- Demo
- Summary and Next Steps

Mortimer - Overview

Distributed all flash NVMe-oF storage software



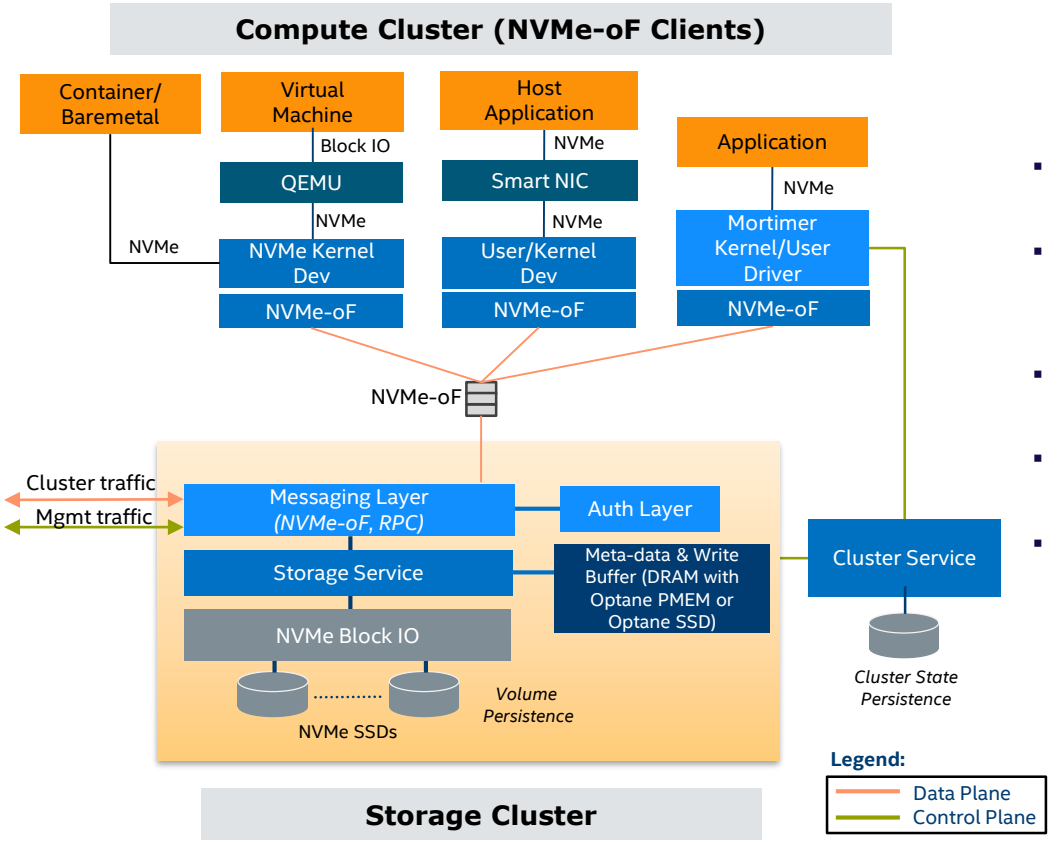
- NVMe-oF for all data plane operations (client & cluster)
- Optane PMEM (app direct) or Optane SSD for meta-data and fast write buffering
- Lockless and poll mode for all data plane operations

- Thin volume provisioning and replication
- Rebalancing and recovery for storage node failures
- Authentication and authorization
- Raw block access (i.e., no file-system overhead)

Mortimer – Motivation

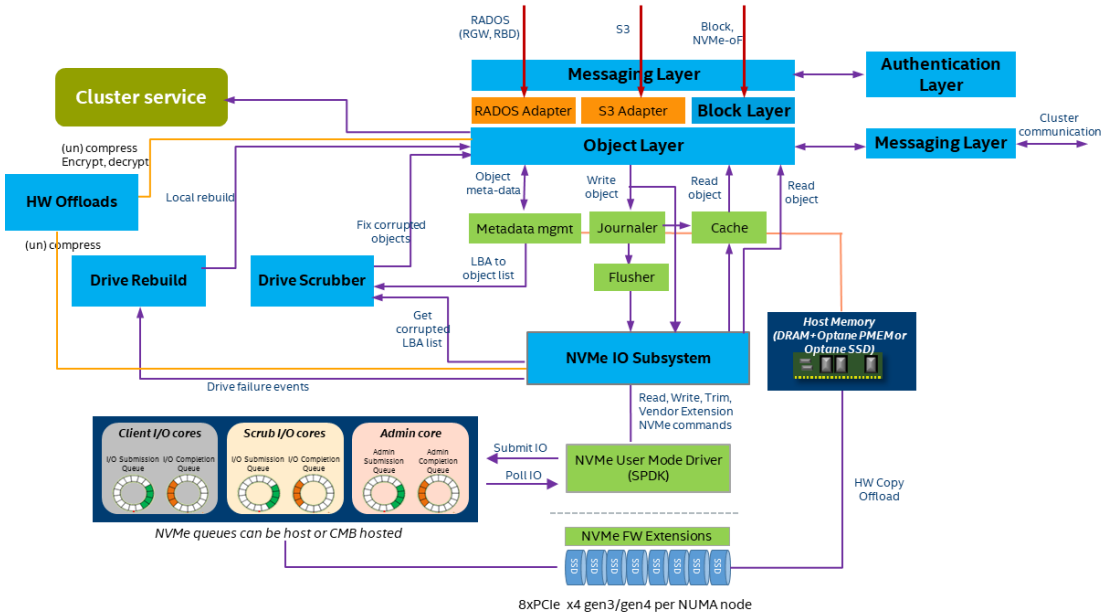
- **Emerging workloads** such as **AI, 5G, Edge** driving need for **all flash** storage
- **NVMe-oF protocol** is gaining traction for **network storage**
- **Persistent Memory**, **CXL** and **Platform innovations** can be exploited to deliver low latency distributed storage
- A **light-weight data path** coupled with **distributed NVMe-oF** semantics paves way for **computational offloads**
- Pluggable **NVMeoF distributed stateful storage** for scale out services (e.g., distributed databases) enables rapid innovation
- Mortimer design philosophy – **exploit NVMe-oF with distributed storage semantics** and **persistent memory (app direct)** to deliver low latency open source software

Mortimer – Architecture (Client)



- NVMe kernel driver can only connect to single storage target per volume
- SmartNIC client can support user mode driver with rich client-side services (e.g. encryption)
- Custom Kernel driver is needed for volume striping across storage nodes
- Cluster service uses Raft consensus protocol
- Storage service can take advantage of DRAM meta-data for battery backed platforms

Mortimer – Architecture (Storage)

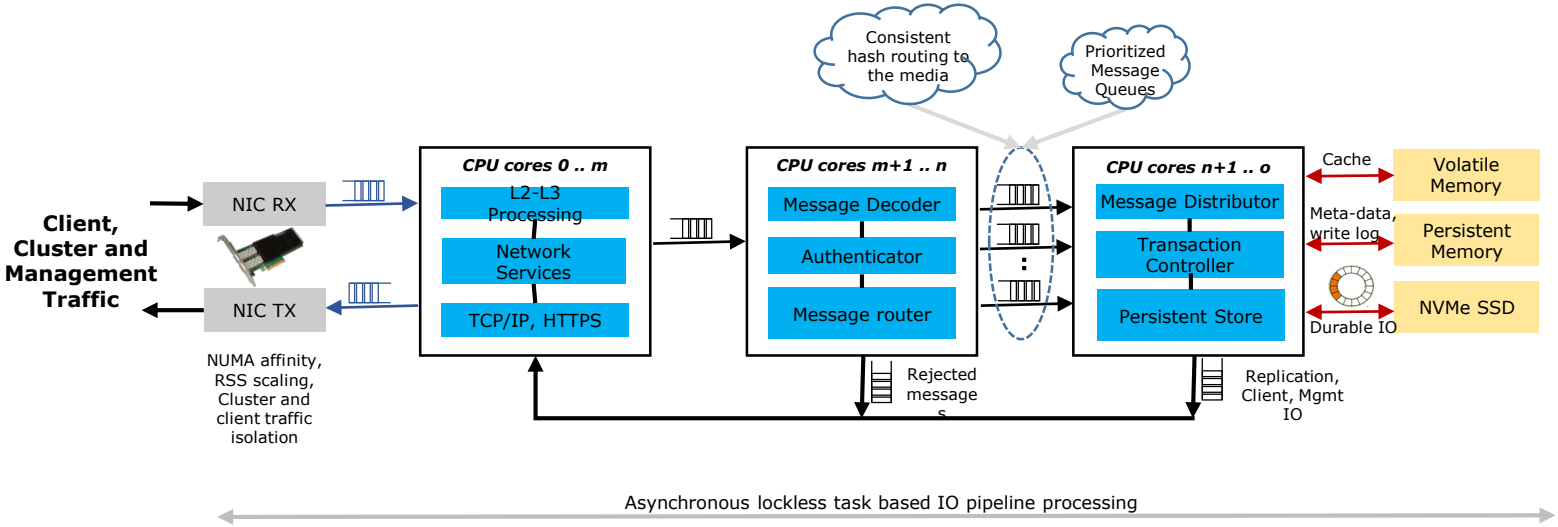


Component	Description
Messaging Layer	Provides interface to clients using NVMe-oF. Can be extended to other protocols such as RADOS, s3 in future Responsible for cluster communication
Object Layer	Responsible for coordinating read, write etc. operations, data distribution to other nodes, ensuring consistency guarantee and recovery
Authentication Layer	Provides access controls to tenant, internal services using TLS and external auth providers
Cluster Service	Responsible for maintaining distributed cluster state using consensus protocols (i.e., Raft)
NVMe IO Subsystem	Responsible for managing IO to multiple drives using poll mode drivers
Meta-data mgmt module	Responsible for meta-data management in persistent memory
Drive Scrubber module	Responsible for background data integrity checking, corrupted data recovery
Drive Rebuild module	Bulk copy, restore of rebuilds due to drive failures

- **Volume is sharded into objects and distributed among drives (avoids hot spots)**
- **Provides offload capabilities (e.g., drive assisted scrubbing, encryption, memory copy)**

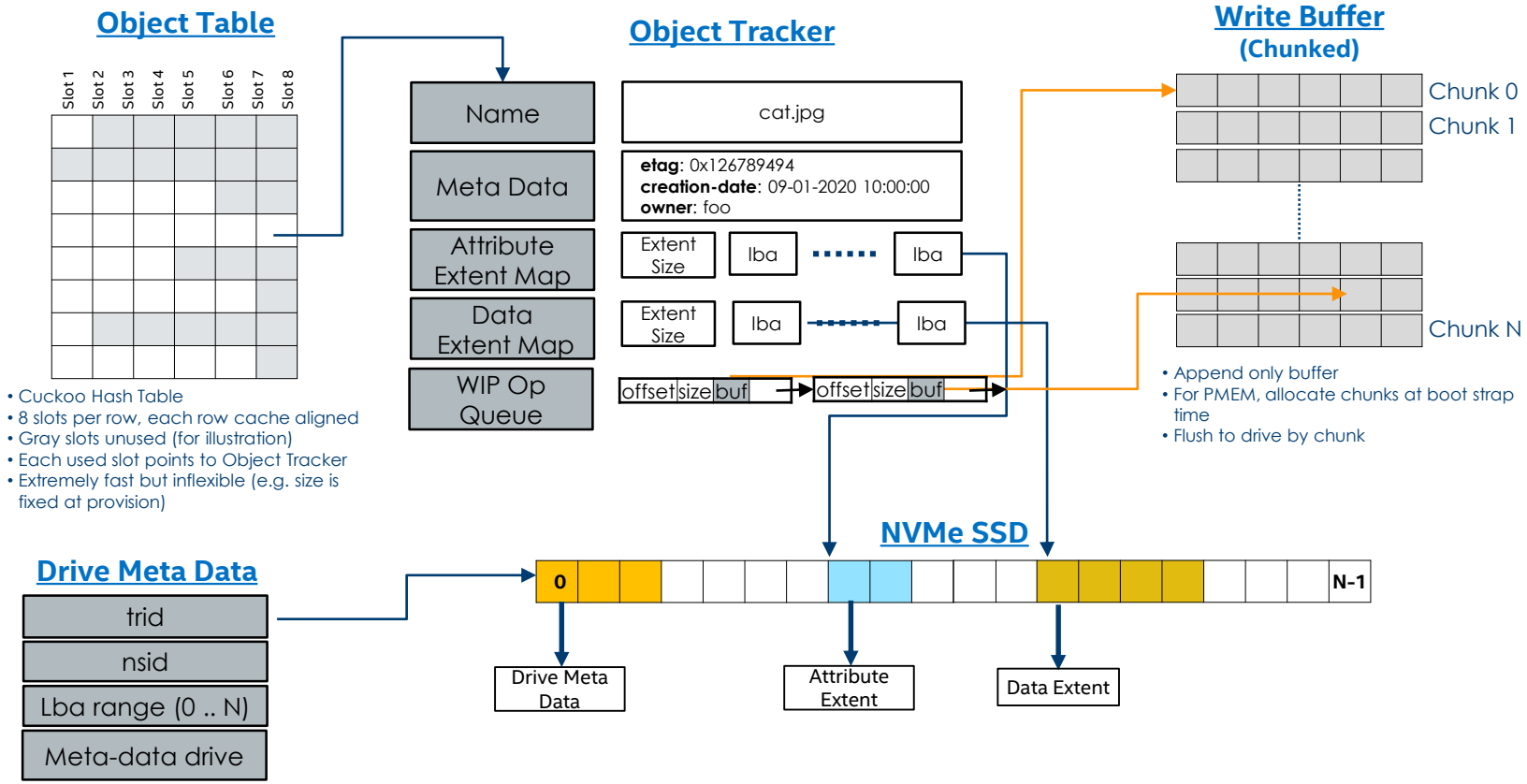
- **One storage service per NUMA domain**
- **Optane SSD meta-data mgmt. uses DRAM cache and journaling for durability**
- **Only small writes, pending updates get buffered**

Message Processing Pipeline

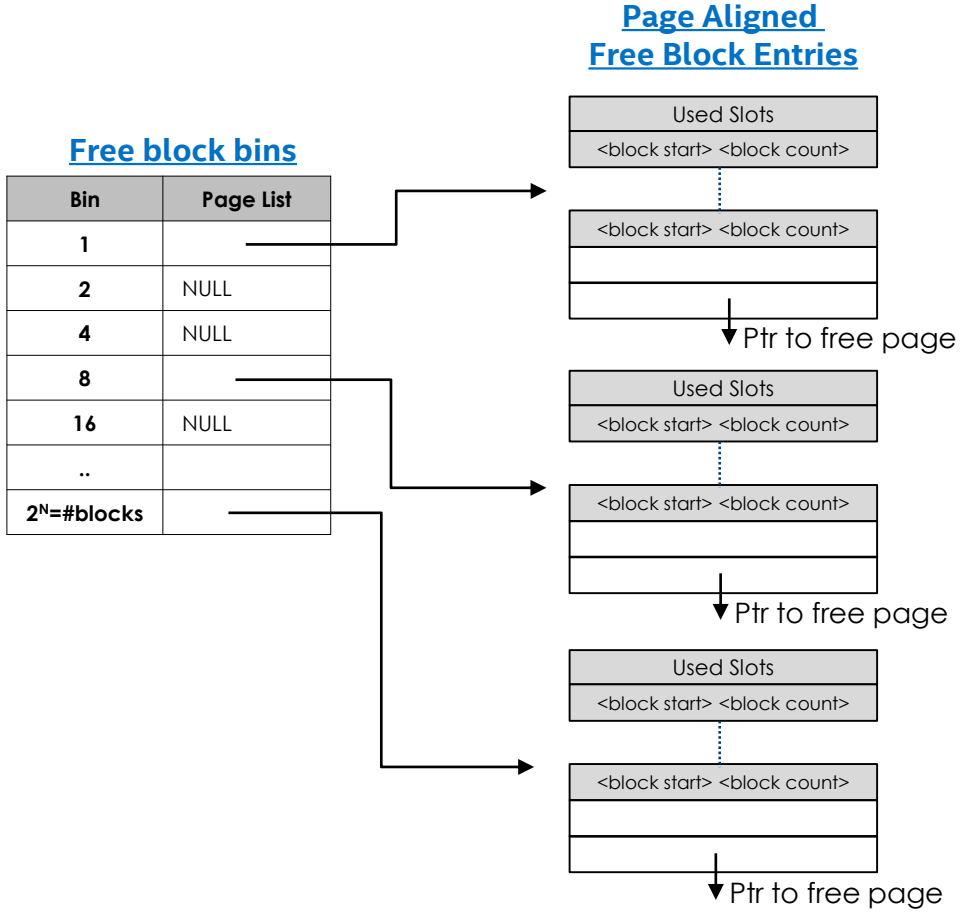


- **Stateful tasks (e.g., IO to specific drive) get executed on dedicated cores. No other tasks will run on these cores.**
- **Stateless tasks (e.g., management traffic, authentication) can get executed on pool of cores. Employs light-weight scheduling.**
- **It is possible to split drive into multiple logical partitions and run on dedicated cores**

PMEM Metadata: Core Data Structures



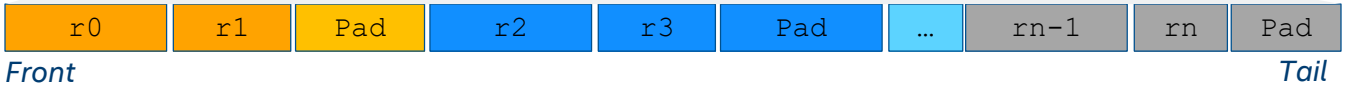
PMEM Metadata: Drive Free List



- Use power of 2 binning to index into free block page pool quickly
- Each page has at least one free entry
- Entries may gradually move down to lower bins if partially allocated
- Once all slots are used, page flag bit flipped to indicate it is not part of the pool
- Page pool linked list is dynamically constructed at boot strap with DRAM pointers for perf reasons

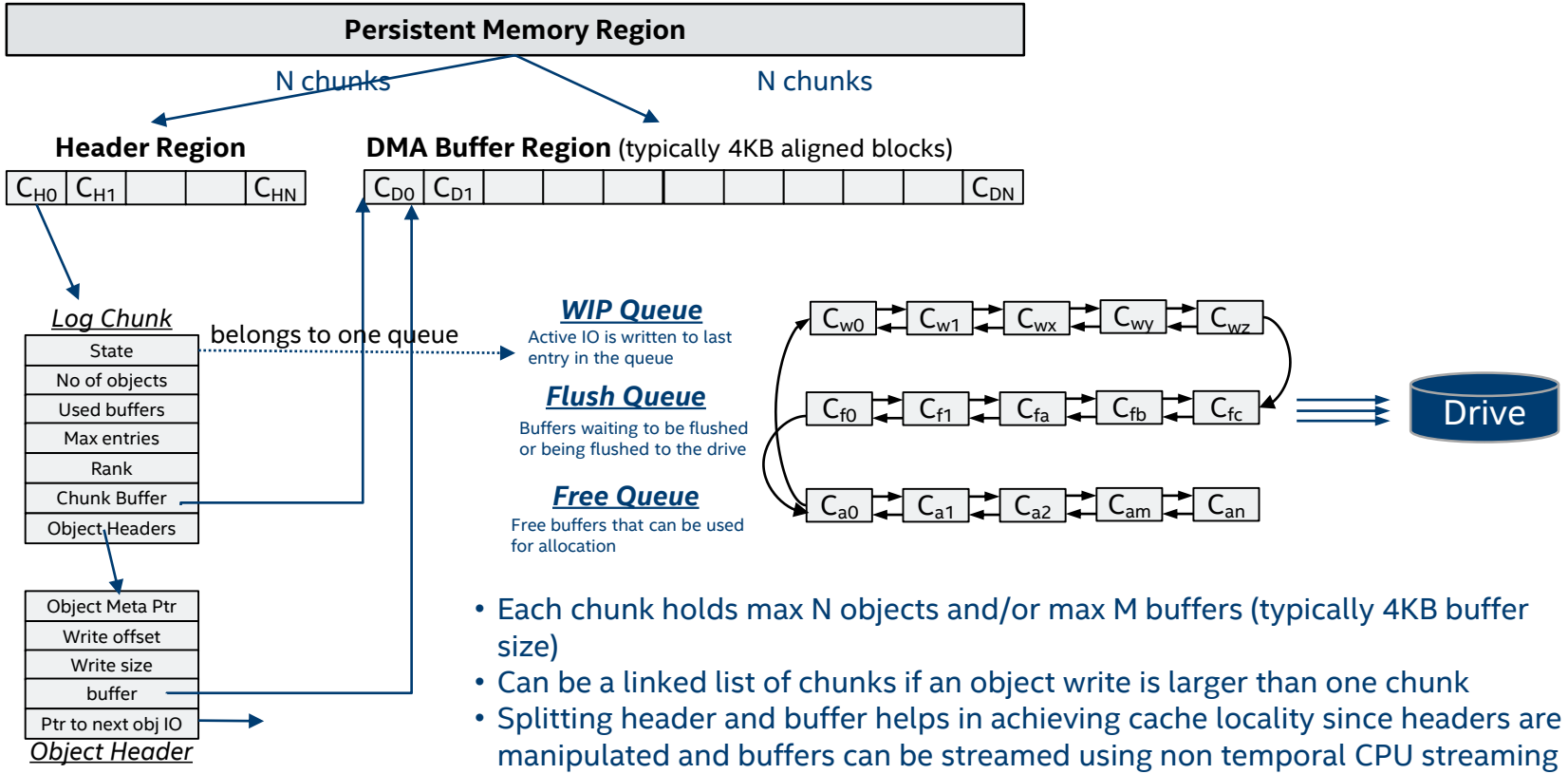
PMEM Metadata: Write Buffer & Lockless Design

Log File Format Details



- Locking is expensive
- Not persistent memory friendly – header and data combined
- Not DMA friendly (need 4K alignment)

PMEM Metadata: Write Buffer & Lockless Design



- Each chunk holds max N objects and/or max M buffers (typically 4KB buffer size)
- Can be a linked list of chunks if an object write is larger than one chunk
- Splitting header and buffer helps in achieving cache locality since headers are manipulated and buffers can be streamed using non temporal CPU streaming instructions

Mortimer Demo

Summary and Next Steps

- Mortimer **exploits NVMe-oF with distributed storage semantics** to deliver low latency open source software
- Mortimer uses **Intel Optane Persistent Memory** with optimized meta-data design in **app direct mode** to deliver low latency
- Mortimer uses **poll-mode, lockless design** to deliver an efficient storage solution to meet future computing trends
- **Mortimer to be open-sourced in 1H'21**
- **SNIA collaboration** to **extend NVMeoF protocol** for distributed storage

Thank you!!

Please take a moment
to rate this session.

Your feedback matters to us.