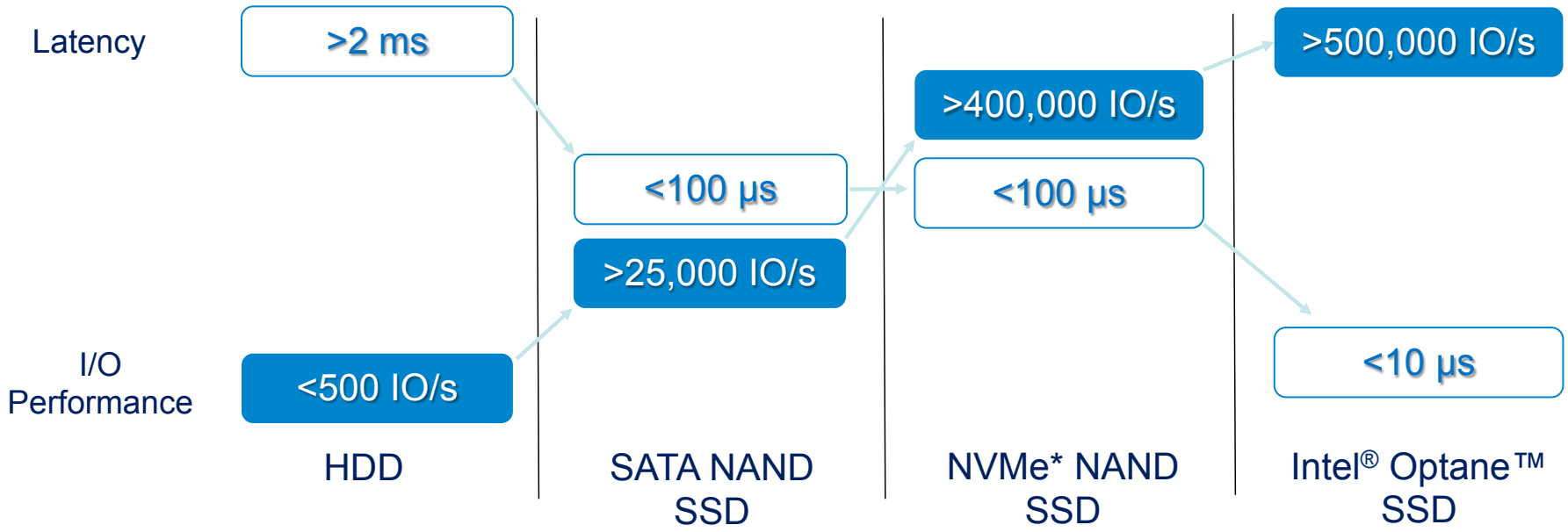
Ceph Optimizations for NVMe

Chunmei Liu, Intel Corporation

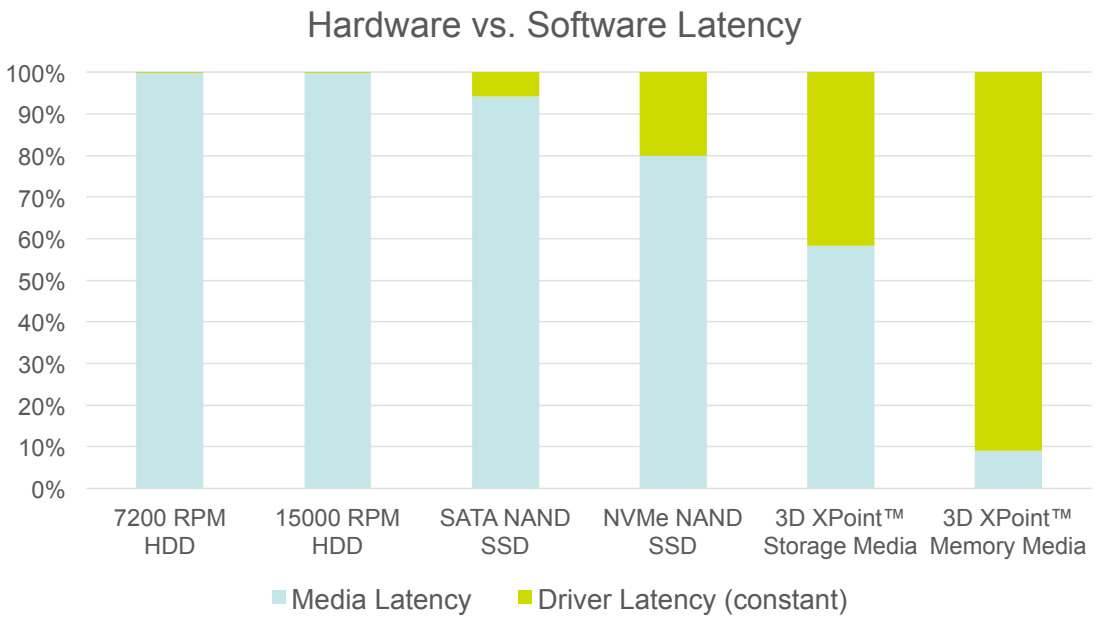
Contributions: Tushar Gohad, Xiaoyan Li, Ganesh Mahalingam, Yingxin Cheng, Mahati Chamarthy

Table of Contents

Software is the bottleneck



The Problem: Software has become the bottleneck



- Historical storage media: no issues
- 3D XPoint™ media approaches DRAM
- Cycles spent on negating old media inefficiencies are now wasted

Ceph Introduction

- Open-source, object-based scale-out distribute storage system
- Software-defined, hardware-agnostic – runs on commodity hardware
- Object, Block and File support in a unified storage cluster
- Highly durable, available – replication, erasure coding
- Replicates and re-balances dynamically

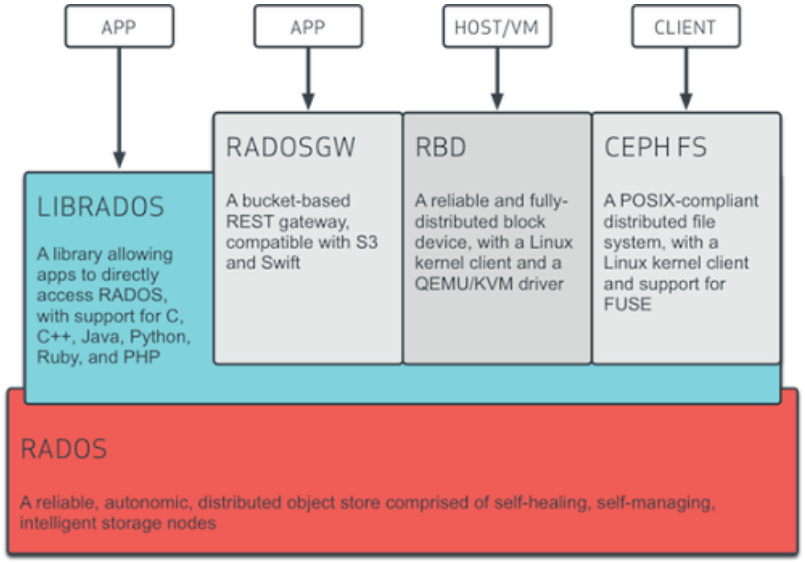
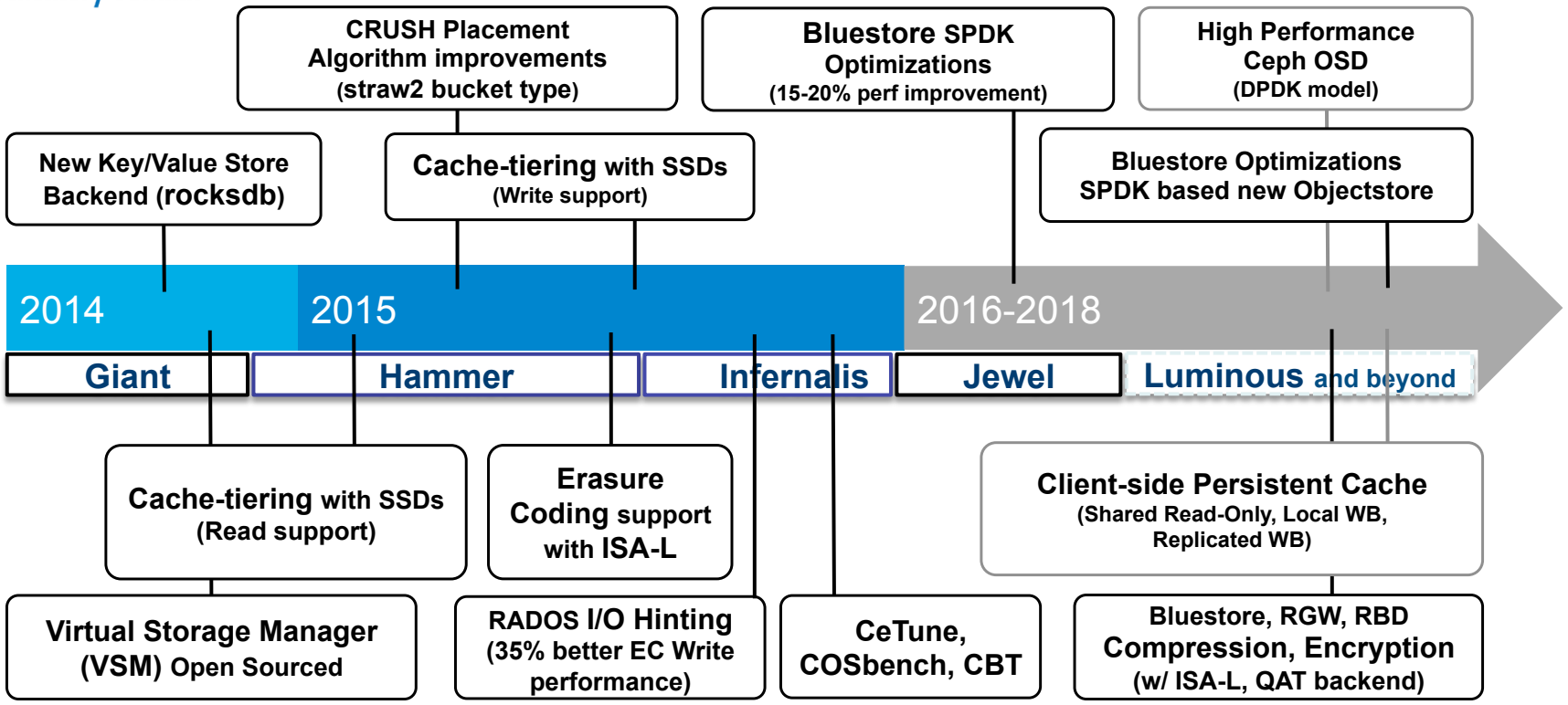
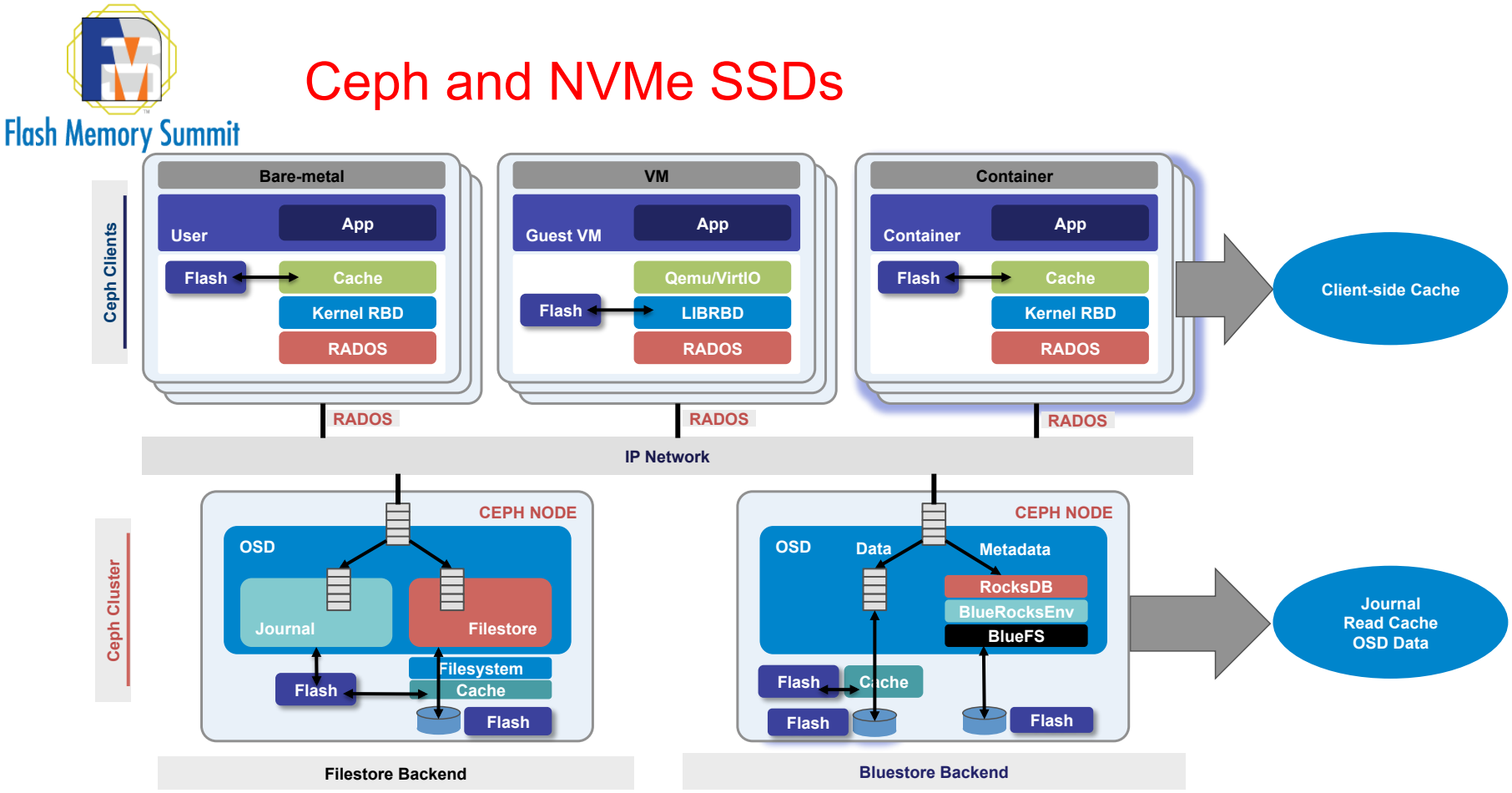


Image source: http://ceph.com/ceph-storage

Intel's Ceph Contribution Timeline





* NVM – Non-volatile Memory

State of Ceph Performance (All-NVMe Backends)

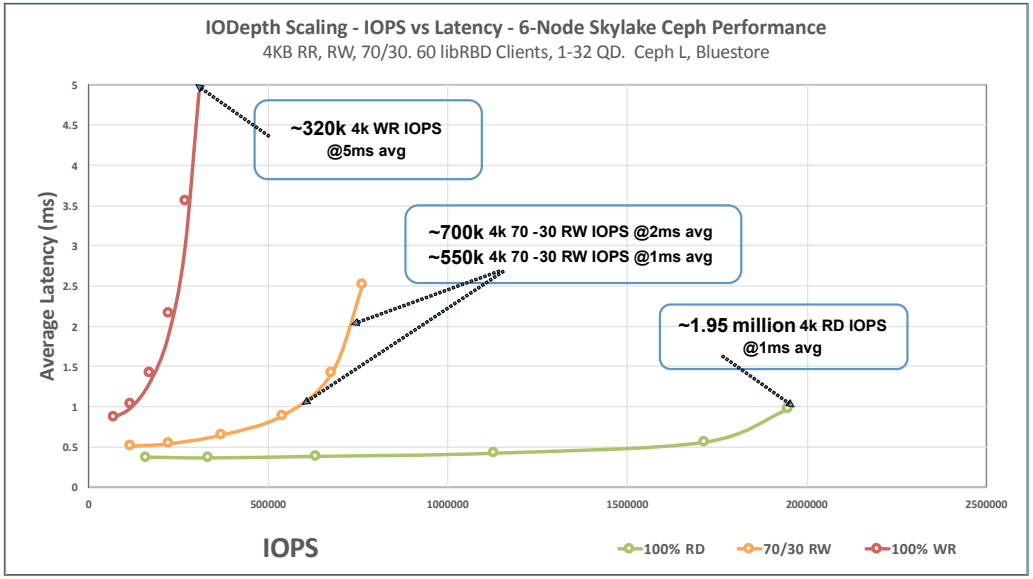
Flash Memory Summit

6x Ceph Nodes

- Intel Xeon Platinum 8176 Processor @ 2.1 GHz, 384GB
- 1x Intel P4800X 375G SSD as DB/ WAL drive
- 4x 4.0TB Intel® SSD DC P4500 as data drives
- 2x Dual-Port Mellanox 25Gb
- Ceph 12.1.1-175 (Luminous rc) Bluestore
- 2x Replication Pool

6x Client Nodes

- Intel® Xeon™ processor E5-2699 v4 @ 2.2GHz, 128GB
- Mellanox 100GbE



	Platform (Spec)	Ceph (Measured)
4K Random Read IOPS per Node	645K[1] * 4 = **3.1M**	**1.95m/6 =~325K@1ms**
4K Random Write IOPS per Node	48K[1] * 4 = **692K**	**(320K * 2) / 6 = ~107K**

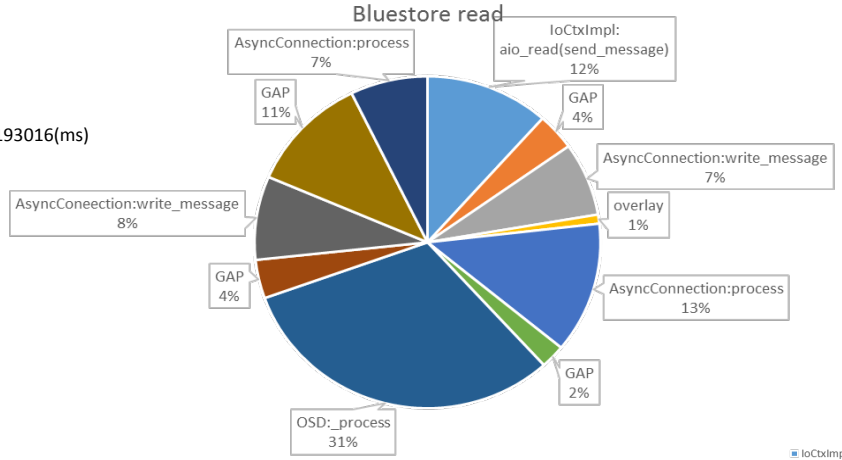
1 - https://www.intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds/dc-p4500-series/dc-p4500-4tb-2-5inch-3d1.html

Ceph performance bottleneck

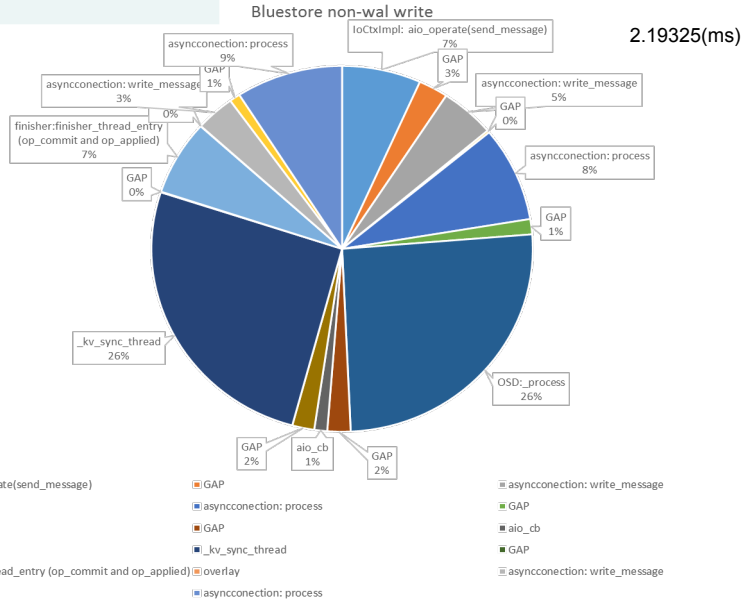
OSD component	IO write (2.2ms)	IO read (~1.2ms)
Messenger (network) thread	~ 10%	~20%
OSD process thread	~30%	~30%
Bluestore thread	~30%	
Pglock (locks)	Tens to hundreds us and to ms	
librbd	25% -30% of overall latency	

Test environment: bluestore use pcie-nvme as bluestore disk and Key-Value. 1osd, 1 mon and benchmark on the 1 server. Read/write 1 request simultaneously.

1.193016(ms)



2.19325(ms)



Intel Storage Software Ingredients

Intel® Intelligent Storage Acceleration Library (Intel® ISA-L)

storage-domain algorithms optimized from the silicon up

OS agnostic, forward- and backward-compatible: across entire Intel processor line, Atom® to Xeon®

Enhances Performance for data integrity (CRC), security/encryption, data protection (EC/RAID), and compression

Storage Performance Development Kit (SPDK)

drivers and libraries to optimize NVM Express* (NVMe) and NVMe over Fabrics (NVMe-oF)

Software Ingredients for Next-Gen Media
lockless, efficient components that scale to millions of IOs per second per core

User-space Polled-Mode Architecture
open source, BSD licensed for commercial or open source projects

Intel Network Stack Software Ingredients

Intel® Data Plane Developer Kit (Intel® DPDK)

libraries and drivers that accelerate packet processing

Enhance Performance for User Space network stack
Receive and Send Packets Within the Minimum Number of CPU Cycles

User-space Polled-Mode Architecture
open source, BSD licensed for commercial or open source projects

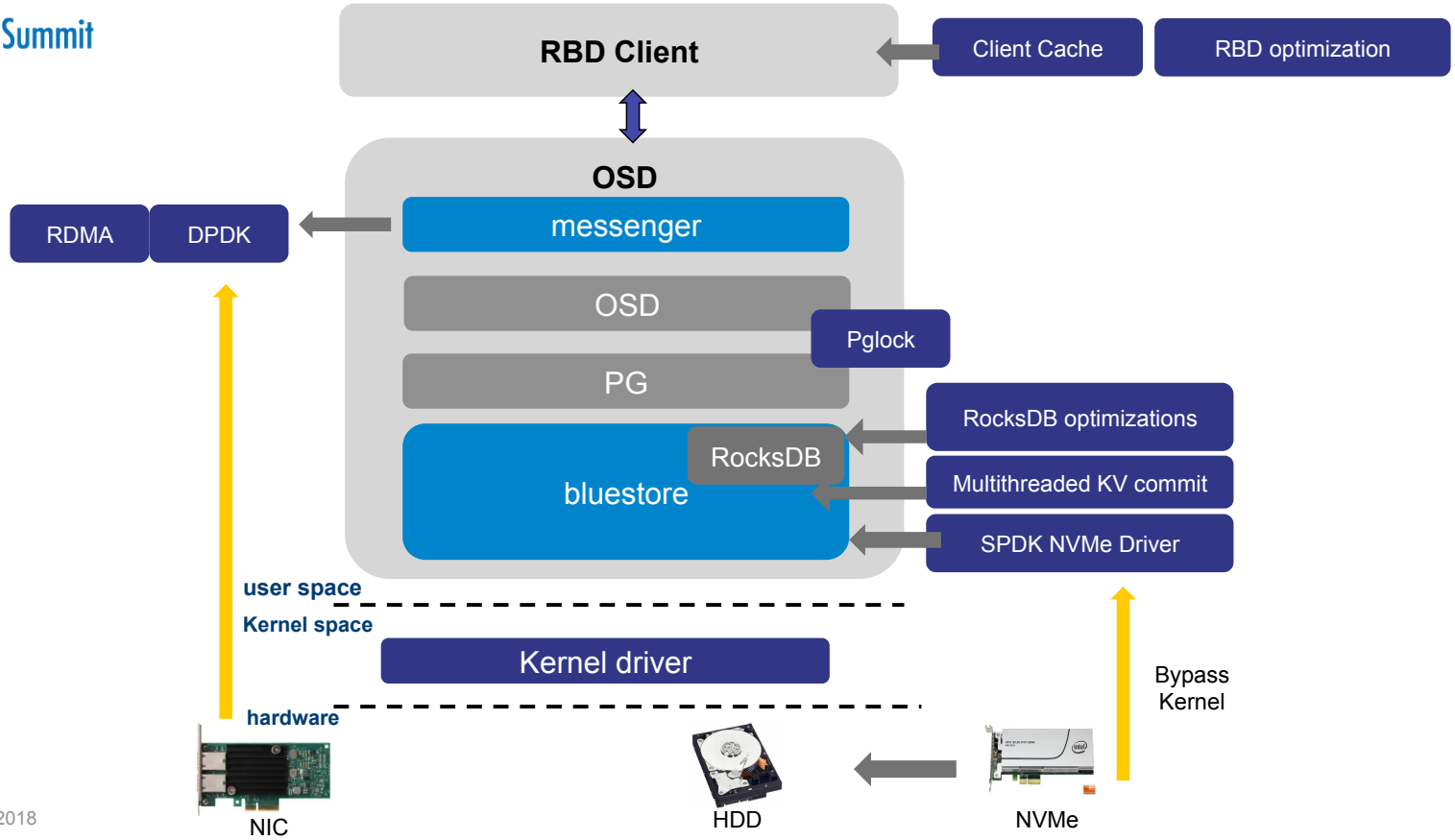
Remote Direct Memory Access(RDMA)

memory-to-memory data communication

Software Ingredients for parallel computer clusters
permits high-throughput, low-latency networking

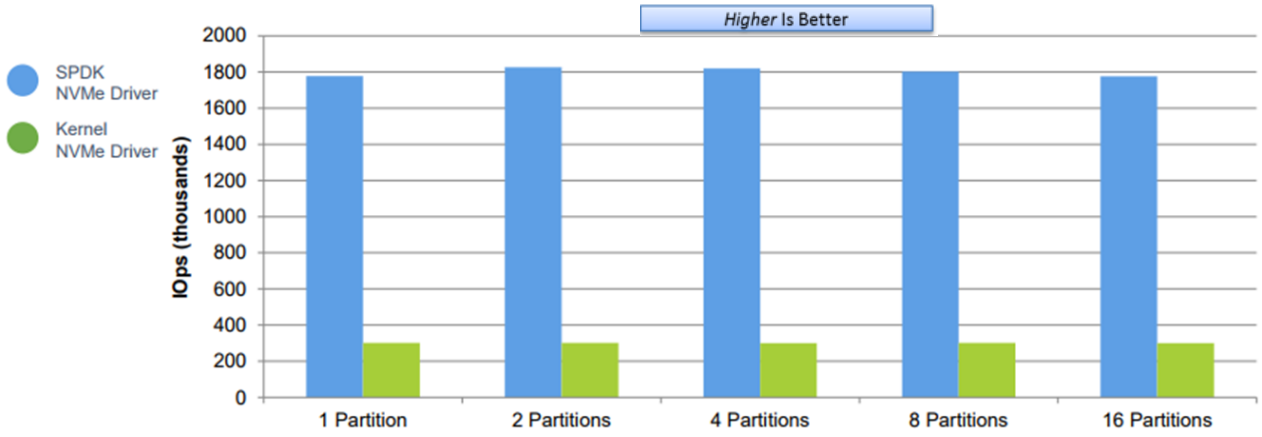
User-space Polled-Mode Architecture
open source, BSD licensed for commercial or open source projects

Ceph Software Stack - Layering



SPDK NVMe Driver

- NVMe driver used in BlueStore.
  - User space NVMe drivers provided by SPDK to accelerate Ios on NVMe SSDs.



*Up to 6X more IOPS/core for NVME vs. Linux Kernel

Ceph +SPDK

From Ziye yang test: SPDK（9322c258084c6abdeefe00067f8b310a6e0d9a5a）
ceph version 11.1.0-6369-g4e657c7 (4e657c7206cd1d36a1ad052e657cc6690be5d1b8)
Single image + fio job =1, iodepth-256

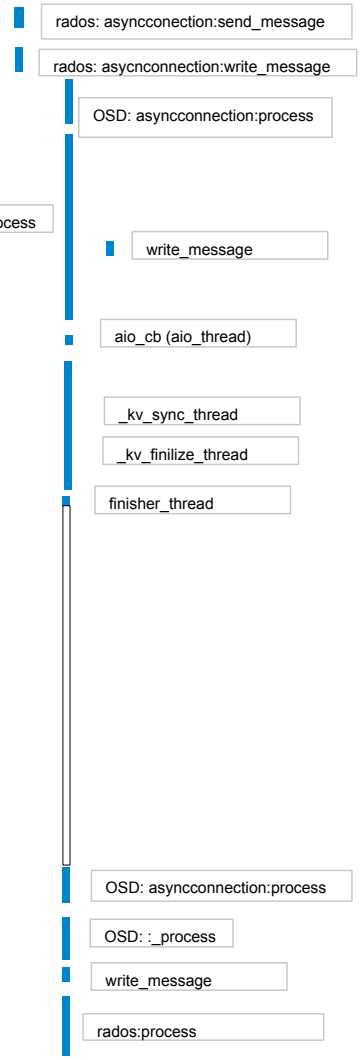
- SPDK NVMe driver alone can’t bring obvious benefit to Ceph

SPDK （df46c41a4ca2edabd642a73b25e54dcb173cf976）
Single image+ io job=1, iodepth-256

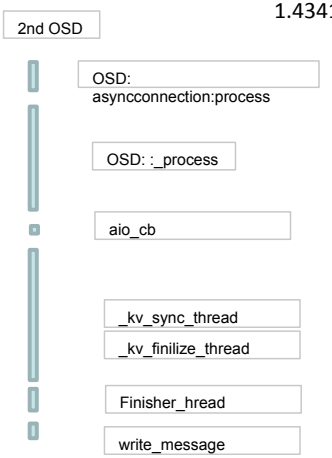
Bluestore – Write Datapath

Write

3.931488(ms)



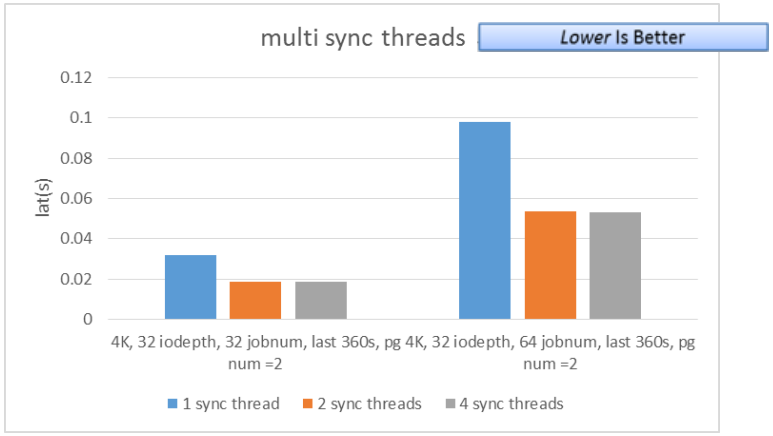
1.434139(ms)



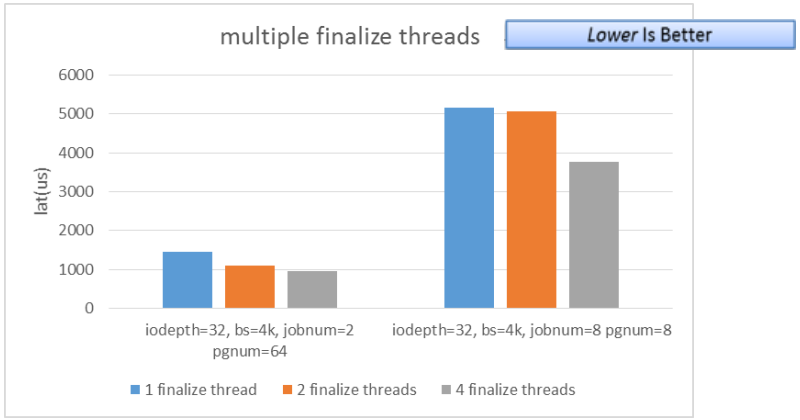
* Non-wal such as write to new blob, write align; wal-overlap write

Multi-kv-threads in bluestore

- Bluestore threads
  - One txc_aio_finish thread handle all ShardWQ threads non WAL aio write
  - One deferred_aio_finish thread handle all ShardWQ threads WAL(deferred) aio write
  - One kv_sync_thread handle rocksDB transaction sync
  - One kv_finalize_thread handle transactions deferred_aio submit
  - One finisher thread handle client reply, this is set by configure file can be changed.
- Add multiple kv_sync_threads
  - Test fio+bluestore
  - Depends on parameter configuration
    - Some case better some case worse

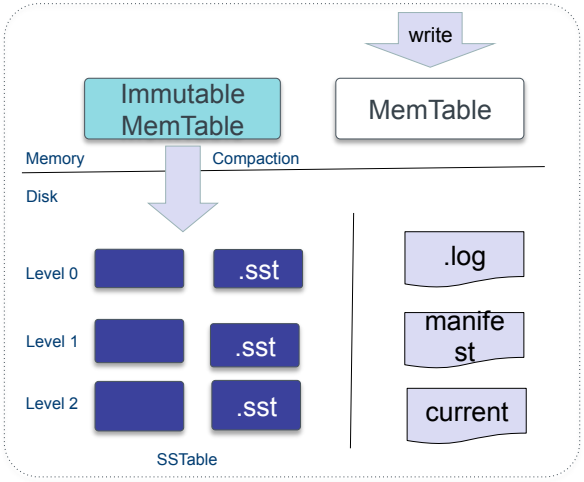


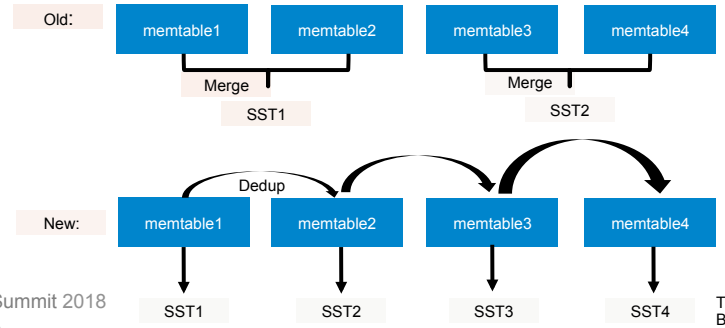
- Add multiple kv_finalize_threads
  - Test fio+bluestore
  - Depends on parameter configuration
    - Some case better some case worse

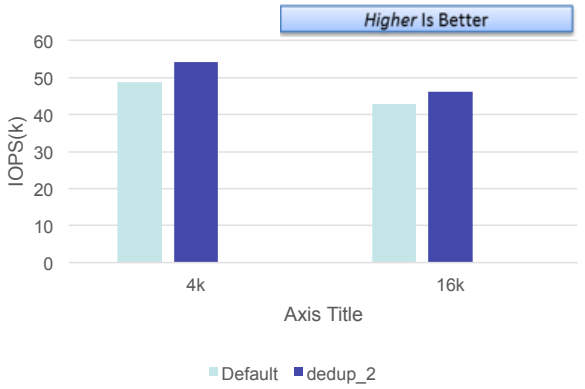


RocksDB optimization

- A key-value database, improved by Facebook.
- Based on LSM (Log-Structure merge Tree).
- Key words:
  - Active MemTable
  - Immutable MemTable
  - SST file
  - LOG
- Random writes 4k/16k.
- Add a flush style: to delete duplicated entries recursively.



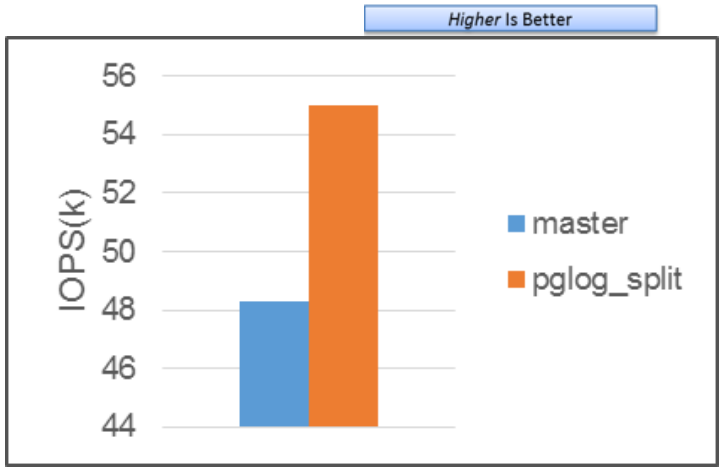


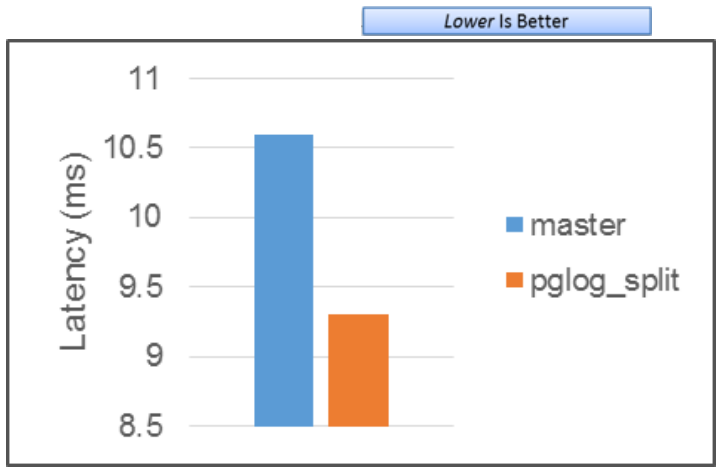


This means that any key that is repeatedly updated or any key that is quickly deleted will never leave the WAL. Both write/read performance in rocksdb is improved. Write can improve up to 15%, read can improve up to 38%. Bluestore IO performance is improved little.

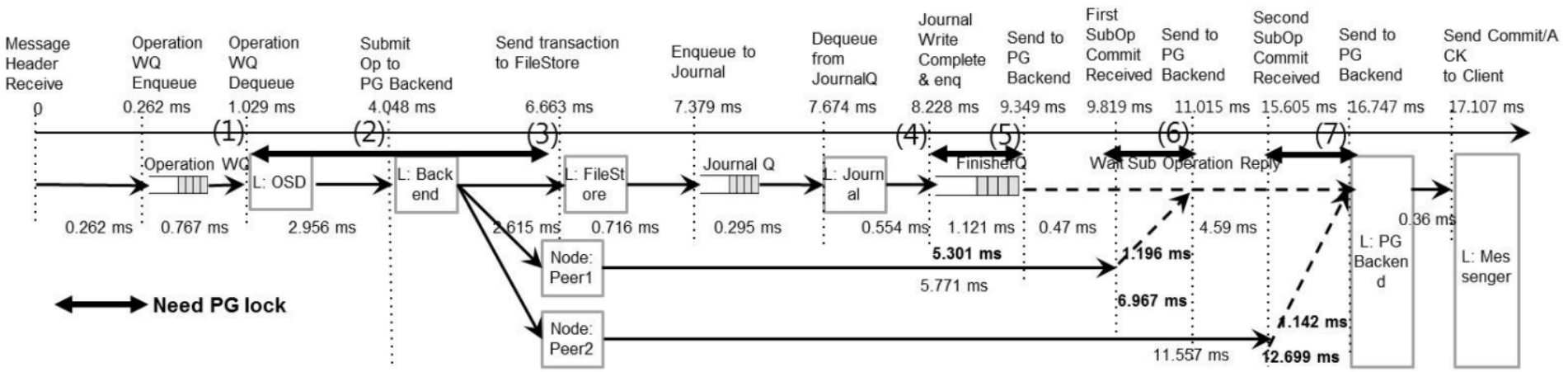
RocksDB optimization (cont.)

- Pglog split
  - Move pglog out from RocksDB, and put it into raw disk.





Pglock expense



∗ Ceph latency analysis for write path

Pglock expense

- Cost for acquiring pglock

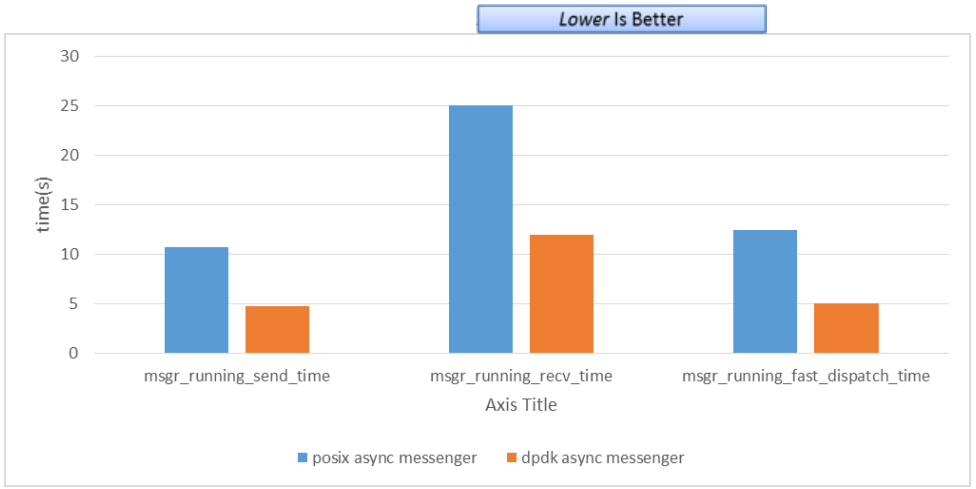
Threads * shard num	In ShardWQ (us)	Get pglock (us)
2_5	138.42	10.64
2_64	41.74	35.89

Cpu cores :22 Processors:87 OSD: 1, Mon:1, MGR:1 on same server Pool size: 1 Rados bench –p rbd –b 4096 –t 128 10 write PG num: 64 bluestore

- Evaluate pglock influence in OSD performance

Ceph Messenger with DPDK

- DPDK user space network Stack(zero copy) used in Messenger (network)
  - Fio rbd bs 4k io depth 32 time120 job4 rw



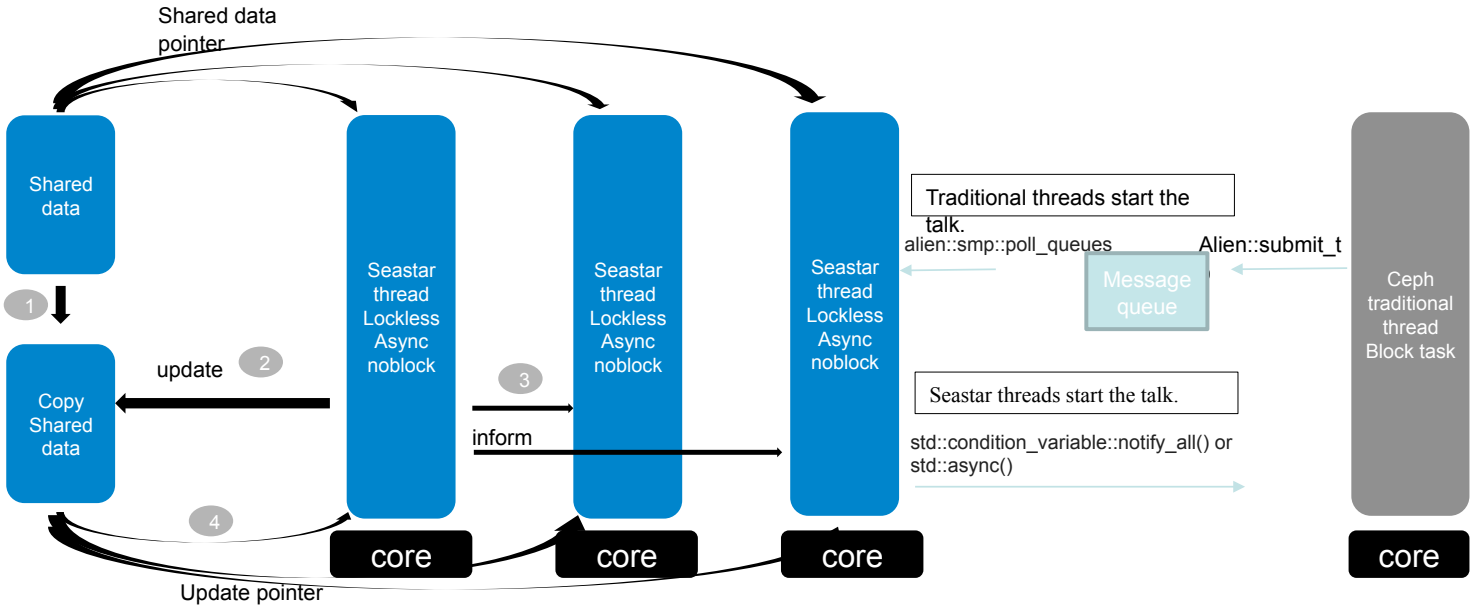
RBD Client Bottlenecks

- RBD worker thread pool size limited to 1
  - Race conditions recently discovered forced this change. WIP to remove this limitation
- Resource contention between librbd workers
  - RBD cacher has a giant global lock – WIP to redesign the client cache
  - The ThreadPool has a global lock – WIP on async RBD client
- Per-OSD session lock
  - The fewer OSDs you have, the higher the probability for IO contention
- Single threaded finisher
  - All AIO completions are fired from a single thread -- so even if you are pumping data to the OSDs using 8 threads, you are only getting serialized completions.

OSD Refactor

- Local performance improvement not cause obvious benefit in ceph
  - Many queues and threads switch in an IO request loop
  - Many locks for synchronize between threads
  - Synchronous and asynchronous mixed process
- Ceph community think about other framework--Seastar
  - Shared-nothing design: Seastar uses a shared-nothing model that shards all requests onto individual cores.
  - High-performance networking: Seastar offers a choice of network stack, including conventional Linux networking for ease of development, DPDK for fast user-space networking on Linux, and native networking on OSv.
  - Futures and promises: An advanced new model for concurrent applications that offers C++ programmers both high performance and the ability to create comprehensible, testable high-quality code.
  - Message passing: A design for sharing information between CPU cores without time-consuming locking
- OSD Refactor
  - Based on Seastar asynchronous programming framework, all operations will be asynchronous
  - Lockless, no any block in Seastar threads

OSD Refactor Framework



Thank you!

Questions?

Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. **No computer system 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® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

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.

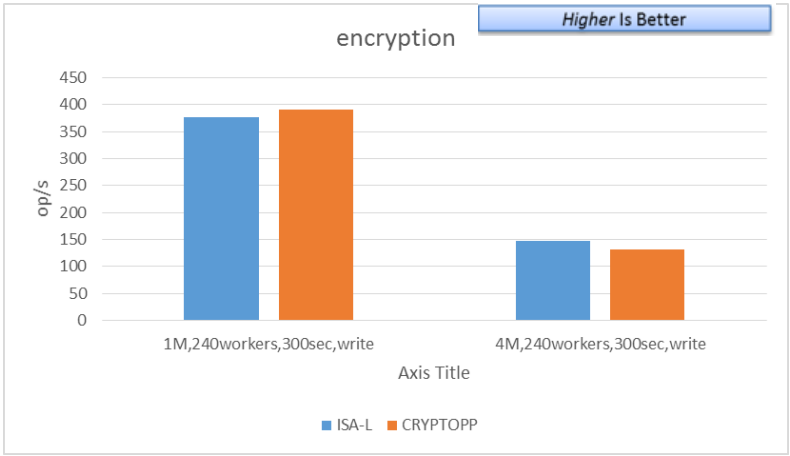
The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.

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.

Back Up

Ceph with ISA-L and QAT

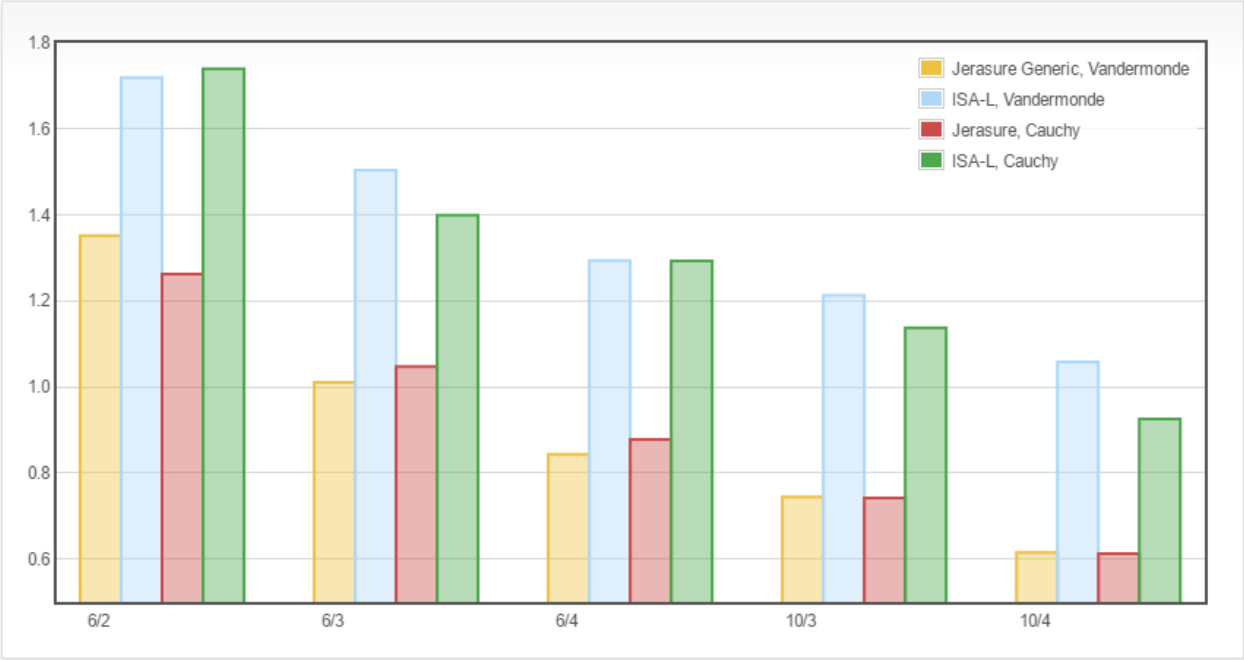
- Erasure coding
  - ISA-L offload support for Reed-Soloman codes
  - Supported since Hammer
- Compression
  - BlueStore
    - ISA-L offload for zlib compression supported in upstream master
    - QAT offload for zlib compression
- Encryption
  - BlueStore
    - ISA-L offloads for RADOS GW encryption in upstream master
    - QAT offload for RADOS GW encryption



* When object file is bigger than 4M, ISA-L gets better performance

Ceph Erasure Coding Performance (Single OSD)

Encode Operation – Reed-Soloman Codes



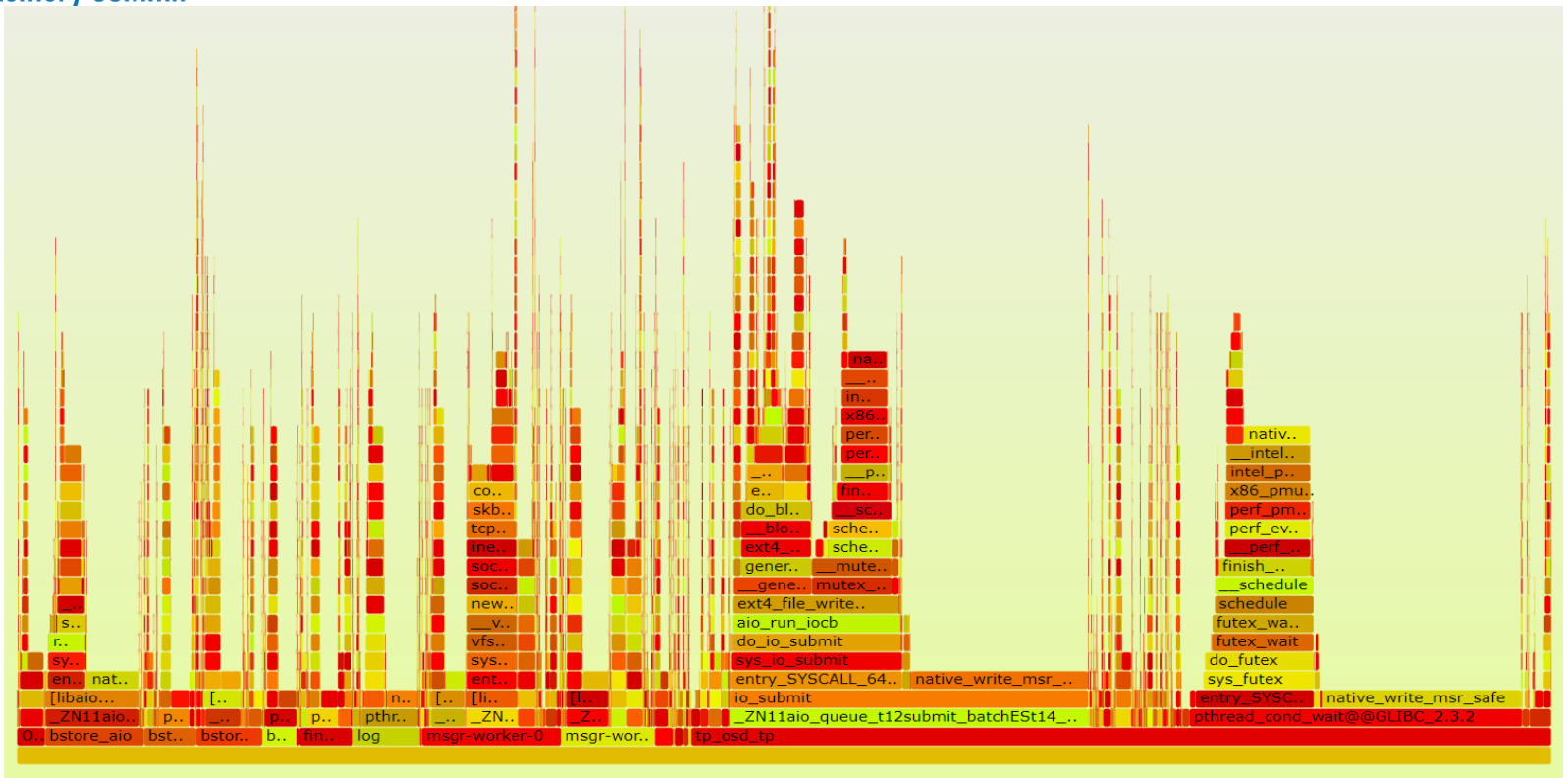
encode: Y = GB/s, X = K/M

Source as of August 2016: Intel internal **measurements** with Ceph Jewel 10.2.x on dual E5-2699 v4 (22C, 2.3GHz, 145W), HT & Turbo Enabled, Fedora 22 64 bit, kernel 4.1.3, 2 x DH8955 adaptor, DDR4-128GB

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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. Any difference in system hardware or software design or configuration may affect actual performance. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. **For more information go to** http://www.intel.com/performance

ISA-L Encode is up to 40% Faster than alternatives on Xeon-E5v4

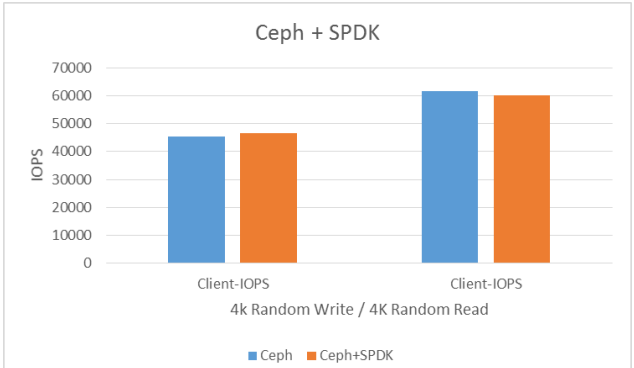
OSD rados benchmark flame gragh

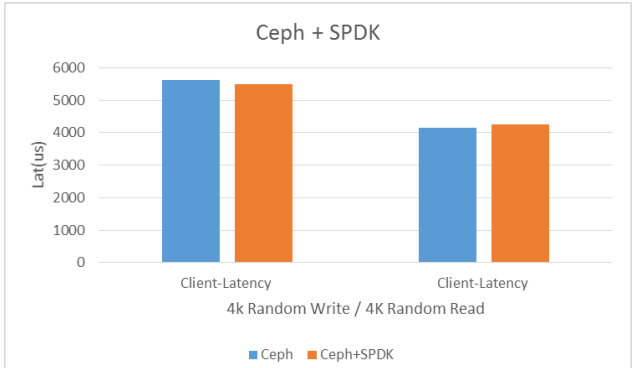


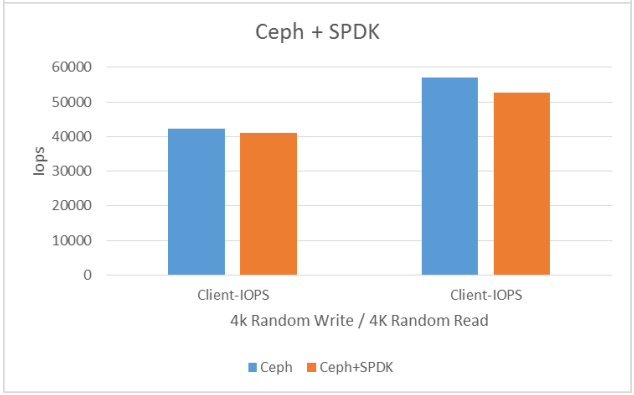
sudo perf record -p `pidof ceph-osd` -F 99 --call-graph dwarf -- sleep 60
./bin/rados -p rbd bench 30 write
rados bench -p rbd -b 4096 -t 60 60 write

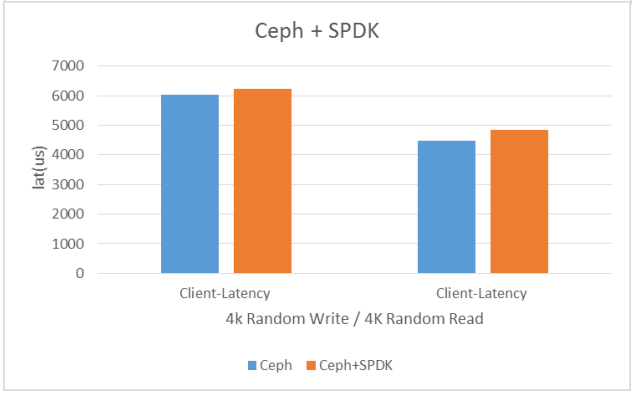
Ceph +SPDK

From Ziye yang test: SPDK（9322c258084c6abdeefe00067f8b310a6e0d9a5a）
ceph version 11.1.0-6369-g4e657c7 (4e657c7206cd1d36a1ad052e657cc6690be5d1b8)
Single image + fio job =1, iodepth-256





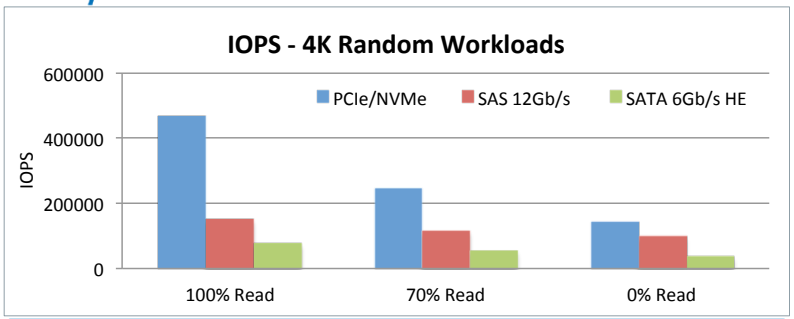




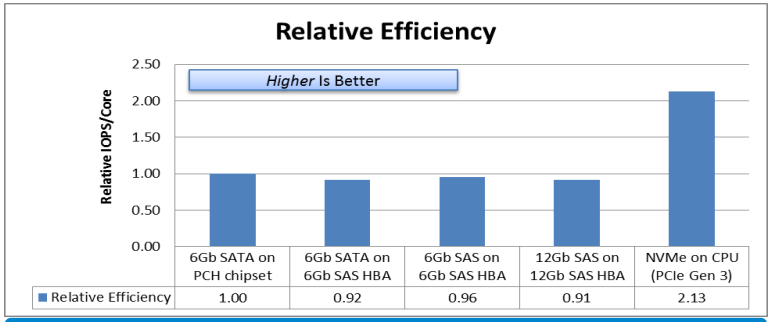
- SPDK can’t bring obvious benefit in Ceph

SPDK （df46c41a4ca2edabd642a73b25e54dcb173cf976）
Single image+ io job=1, iodepth-256

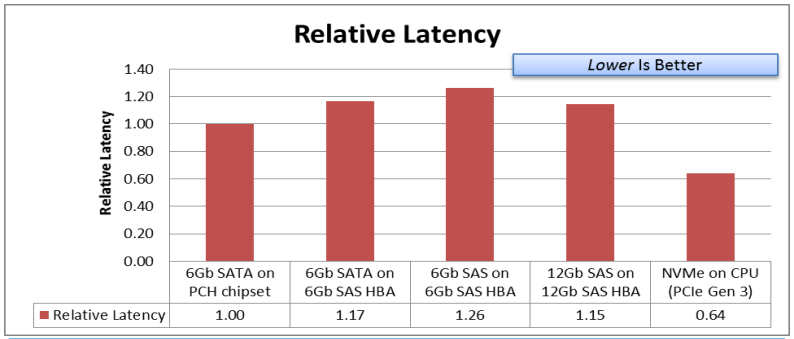
NVMe: Best-in-Class IOPS, Lower/Consistent Latency



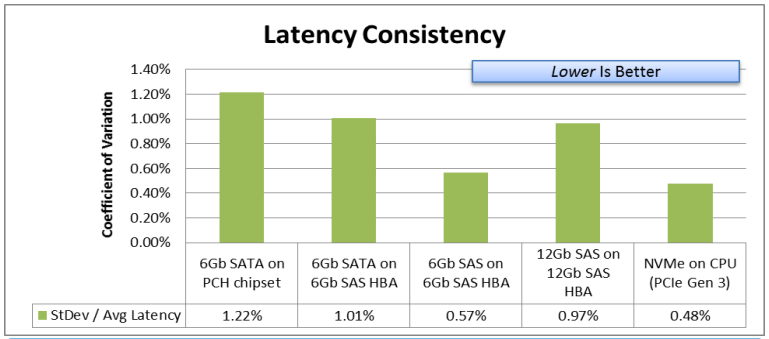
3x better IOPS vs SAS 12Gbps



For the same #CPU cycles, NVMe delivers over 2X the IOPs of SAS!



Lowest Latency of Standard Storage Interfaces



Gen3 NVMe has 2 to 3x better Latency Consistency vs SAS

Test and System Configurations: PCI Express® (PCIe®)/NVM Express® (NVMe) Measurements made on Intel® Core™ i7-3770S system @ 3.1GHz and 4GB Mem running Windows® Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by IOmeter® tool. SAS Measurements from HGST Ultrastar® SSD800M/1000M (SAS), SATA S3700 Series. For more complete information about performance and benchmark results, visit http://www.intel.com/performance. Source: Intel Internal Testing.