Building High Performance, High Capacity, Cost-Efficient All-Flash Cloud Storage System with Ceph

Yuan Zhou, yuan.zhou@intel.com, Senior Software Engineer
Jack Zhang, yuan.zhang@intel.com, Senior Enterprise Architect
Jian Zhang, jian.zhang@intel.com, Senior Software Engineer

Agenda

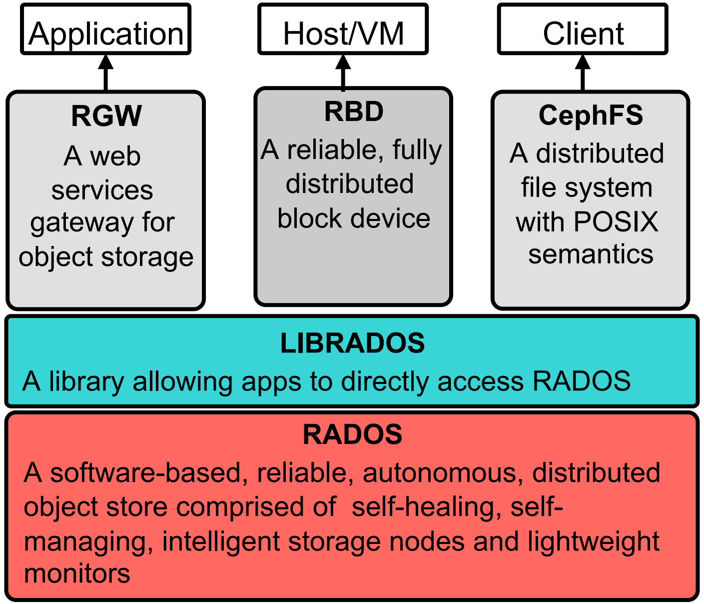
- Ceph* introduction
- Ceph all-flash configuration
- Breakthrough 1M IOPS Ceph cluster with SATA SSDs
- Ceph with 3D Xpoint™ and 3D NAND technologies
- Summary

*Other names and brands may be claimed as the property of others.

Ceph* introduction

Ceph is an open-source, massively scalable, software-defined storage system that provides object, block and file system storage in a single platform. It runs on commodity hardware—saving you costs and giving you flexibility—and because it's in the Linux* kernel, it's easy to consume.

- Object store (RADOSGW)
  - A bucket-based REST gateway
  - Compatible with S3 and swift
- File system (CEPH FS)
  - A POSIX-compliant distributed file system
  - Kernel client and FUSE
- Block device service (RBD)
  - OpenStack* native support
  - Kernel client and QEMU*/KVM driver



Three configurations for Ceph* storage node

- Standard/good
  - NVM Express* (NVMe)/PCI Express* (PCIe*) SSD for journal + caching, HDDs as OSD data drive
  - Example: 1 x Intel® SSD DC P3700 Series 1.6TB SSD as journal + Intel® Cache Acceleration Software (Intel® CAS) + 12 HDDs
- Better (best TCO, as in today's talk)
  - NVMe/PCIe SSD as journal + high capacity SATA* SSD for data drive
  - Example: 1 x Intel P3700 800GB + 4 x Intel S3510 1.6TB
- Best performance
  - All NVMe/PCIe SSDs
  - Example: 4 x Intel P3700 2TB SSDs

Ceph* storage node --Good	
CPU	Intel® Xeon® CPU E5-2650v3
Memory	64 GB
NIC	10GbE
Disks	1x 1.6TB P3700 + 16 x 4TB HDDs (1:12 ratio) P3700 as Journal and caching
Caching software	Intel CAS 3.0, option: Intel® Rapid Storage Technology enterprise/MD4.3

Ceph Cluster --Better	
CPU	Intel Xeon CPU E5-2690
Memory	128 GB
NIC	Duel 10GbE
Disks	1x 800GB P3700 + 4x S3510 1.6TB

Ceph Cluster --Best	
CPU	Intel Xeon CPU E5-2699v3
Memory	>= 128 GB
NIC	1x 40GbE, 4x 10GbE
Disks	4 x P3700 2TB

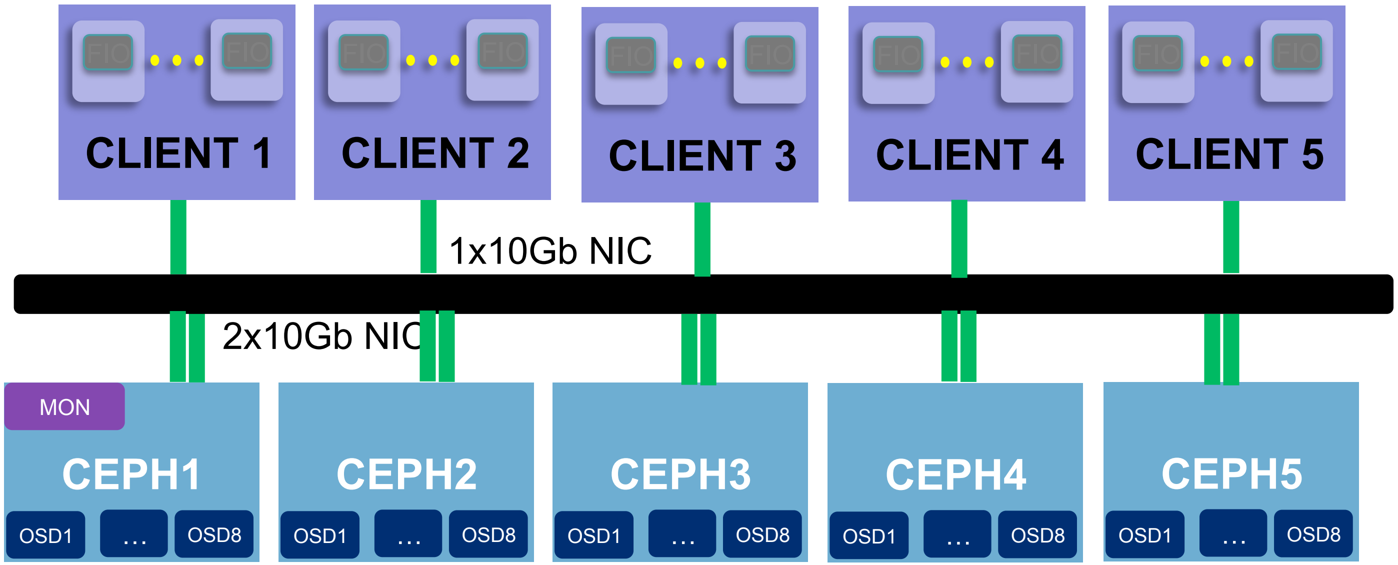
Standard/better/best designations reflect only levels of feature choices and are not intended as competitive comparisons.

Ceph* on all-flash array

- Storage providers are struggling to achieve the required high performance
  - There is a growing trend for cloud providers to adopt SSD
    - CSP who wants to build EBS alike service for their OpenStack* based public/ private cloud
- Strong demands to run enterprise applications
  - OLTP workloads running on Ceph
  - high performance multi-purpose Ceph cluster is a key advantage
  - Performance is still an important factor
- SSD price continue to decrease

Ceph* SATA all-flash array – Configuration w/ FileStore

Test Environment



5x Client Node

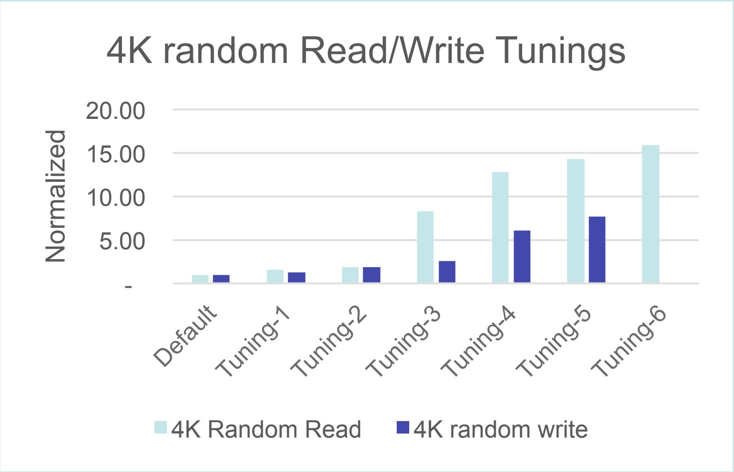
- Intel® Xeon™ processor E5-2699 v3 @ 2.3GHz, 64GB mem
- 10Gb NIC

5x Storage Node

- Intel Xeon processor E5-2699 v3 @ 2.3 GHz
- 128GB Memory
- 1x 1T HDD for OS
- 1x Intel® DC P3700 800G SSD for journal (U.2)
- 4x 1.6TB Intel® SSD DC S3510 as data drive
- 2 OSD instances one each S3510 SSD

Note: Refer to P27-P28 for detailed Ceph configurations.

Ceph* SATA all-flash array – Tuning and optimization efforts



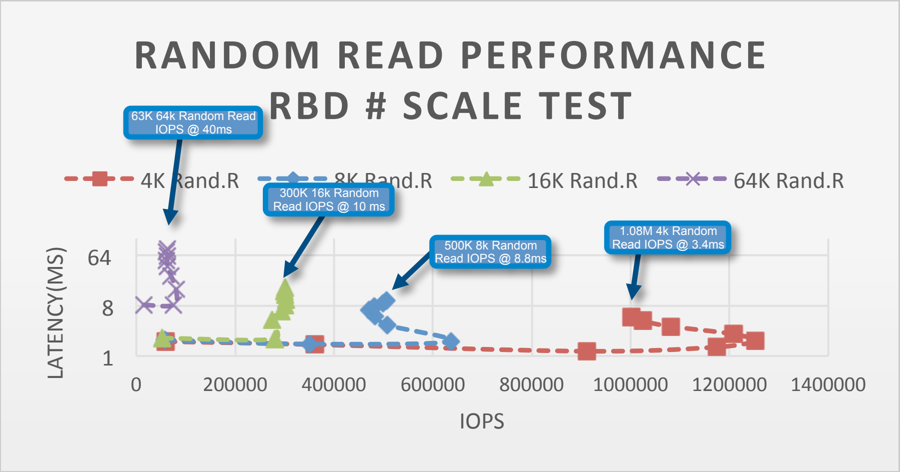
	4K Random Read Tunings	4K Random Write Tunings
Default	Single OSD	Single OSD
Tuning-1	2 OSD instances per SSD	2 OSD instances per SSD
Tuning-2	Tuning1 + debug=0	Tuning1+Debug 0
Tuning-3	Tuning2 + jemalloc	Tuning2+ op_tracker off, tuning fd cache
Tuning-4	Tuning3 + read_ahead_size=16	Tuning3+jemalloc
Tuning-5	Tuning4 + osd_op_thread=32	Tuning4 + Rocksdb to store omap
Tuning-6	Tuning5 + rbd_op_thread=4	N/A

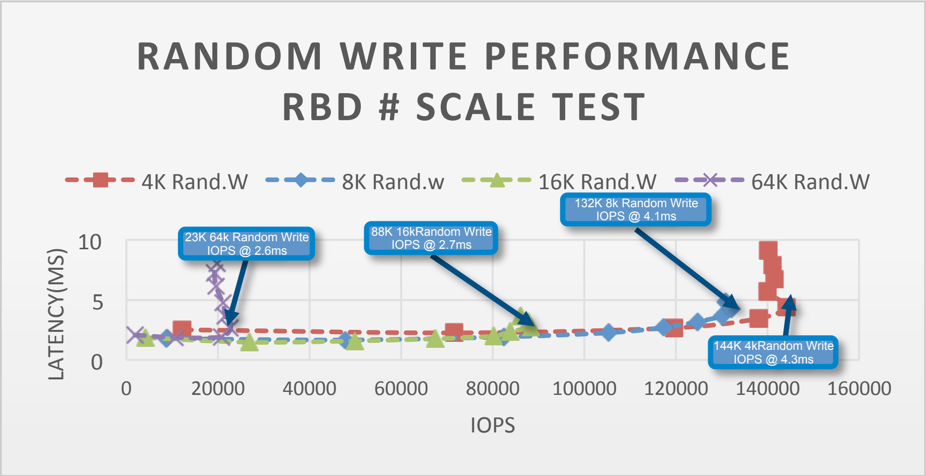
- Up to 16x performance improvement for 4K random read, peak throughput 1.08M IOPS
- Up to 7.6x performance improvement for 4K random write, 140K IOPS

*Refer to P27-28 for Optimization Notes.

Ceph* SATA all-flash array – Random read/write performance

- 1.08M IOPS for 4K random read, 144K IOPS for 4K random write with tunings and optimizations

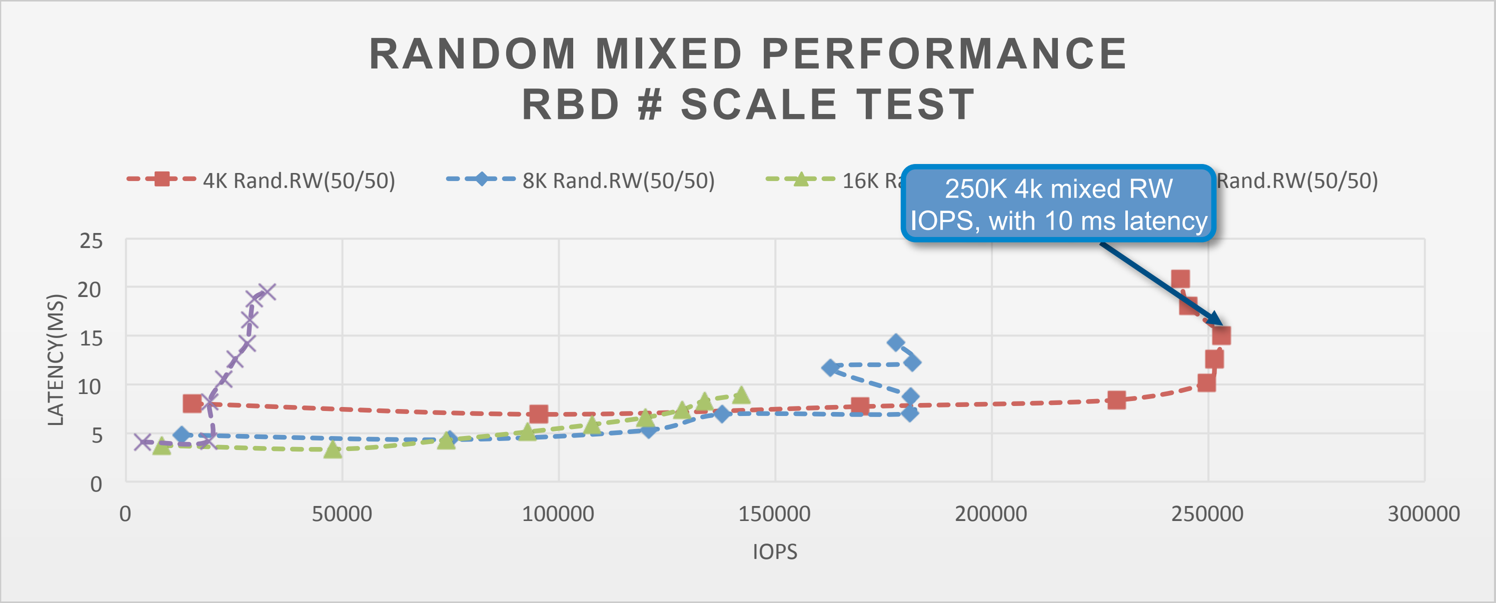




Excellent random read performance and Acceptable random write performance

*Refer to P27-28 for Optimization Notes.

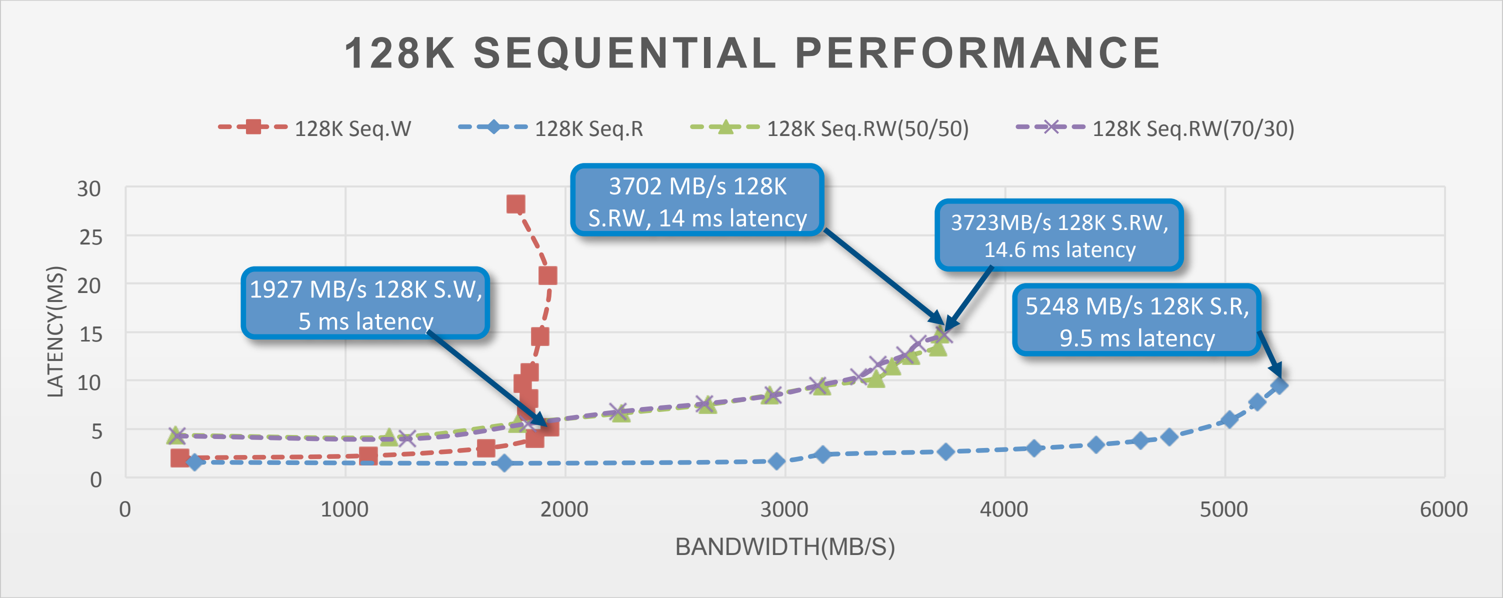
Ceph* SATA all-flash array – Mixed read/write performance



Random write has big impact on random read performance

*Refer to P27-28 for Optimization Notes.

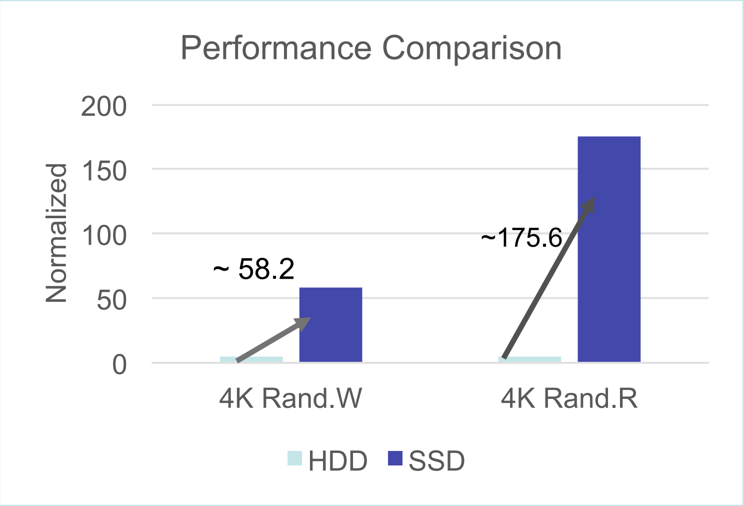
Ceph* SATA all-flash array – Sequential read/write performance



128 sequential read can hit 10 Gb NIC limitation

*Refer to P27-28 for Optimization Notes.

Ceph* SATA all-flash array – Ceph: SSD cluster vs. HDD cluster



- Both journal on PCI Express*/NVM Express* SSD
- 4K random write, need ~ 58x HDD Cluster (~ ***2320*** HDDs) to get same performance
- 4K random read, need ~ 175x HDD Cluster (~ ***7024*** HDDs) to get the same performance

Client Node

- 5 nodes with Intel® Xeon® processor E5-2699 v3 @ 2.30GHz, 64GB memory
- OS : Ubuntu* Trusty

Storage Node

- 5 nodes with Intel Xeon processor E5-2699 v3 @ 2.30GHz, 128GB memory
- Ceph Version : 9.2.0, OS : Ubuntu Trusty
- 1 x P3700 SSDs for Journal per node

Cluster difference:

SSD cluster : 4xS3510 1.6TB for OSD per node

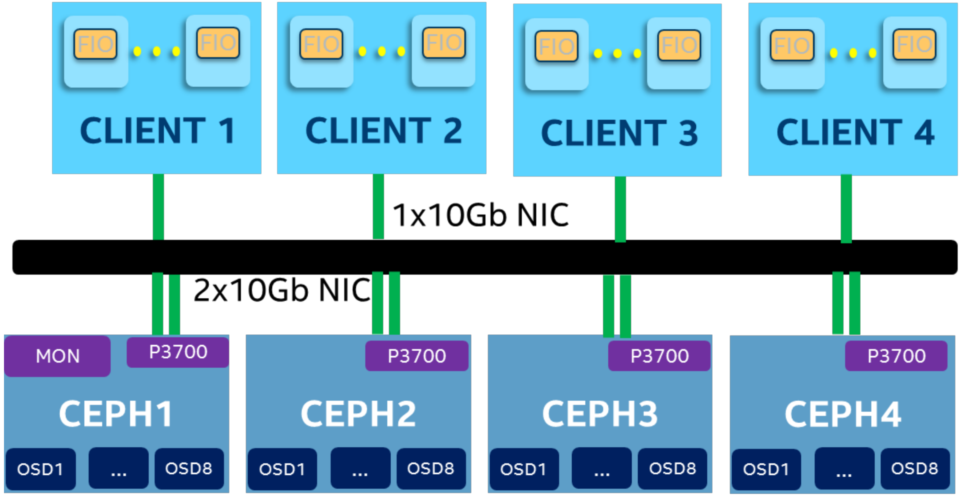
HDD cluster : 10 x STAT 7200RPM HDDs as OSD per node

ALL SSD Ceph helps provide excellent TCO (both Capx and Opex), not only in performance but also space, Power, Fail rate, etc.

*Refer to P27-28 for Optimization Notes.

Ceph* all-flash array performance with BlueStore

Test Environment



4x Client Node

- Intel® Xeon™ processor E5-2699 v3 @ 2.3GHz, 64GB mem
- 10Gb NIC

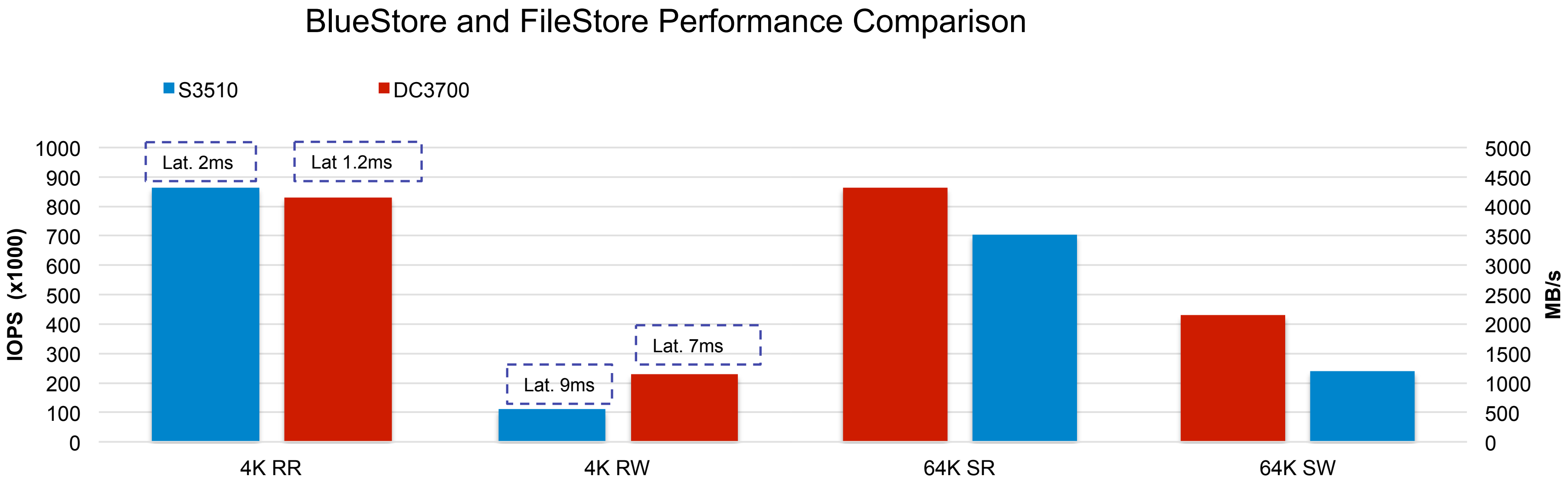
4x Storage Node

- Intel Xeon processor E5-2699 v3 @ 2.3 GHz
- 128GB Memory
- 1x 1T HDD for OS
- 1x Intel® DC P3700 2TB SSD for rocksdb WAL and database
- 4x 480GB Intel® SSD DC S3700 as data drive
- 2 OSD instances one each S3700 SSD

Software Configuration

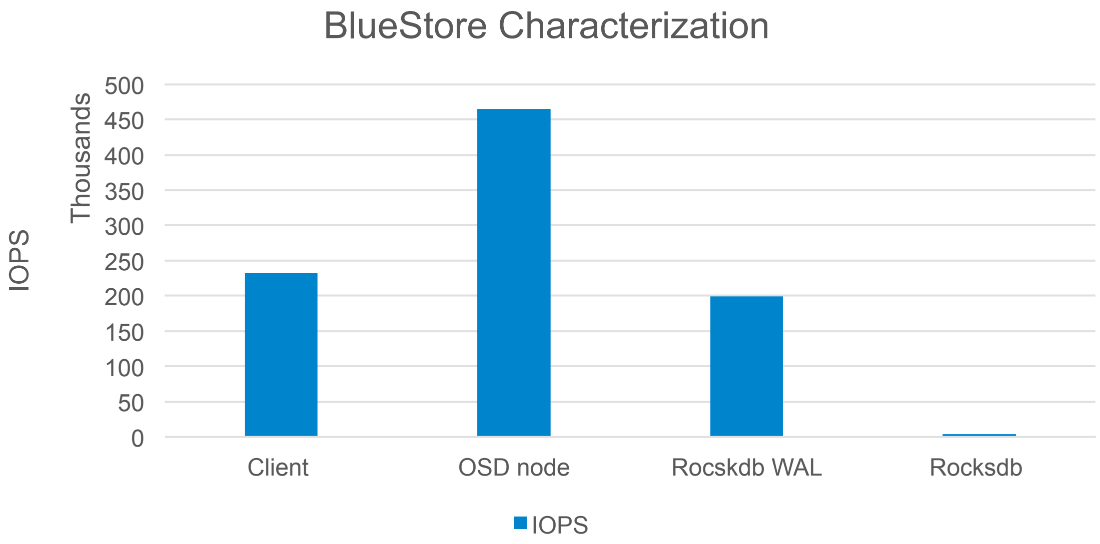
- Ceph 10.2.0

Ceph* all-flash array performance with BlueStore



- Significant performance increase for 4K random write and 64K sequential write
  - 2x improvement for 4K random write, 1.8x improvement for 64K sequential write
- The latency also becomes better for 4K random I/O.

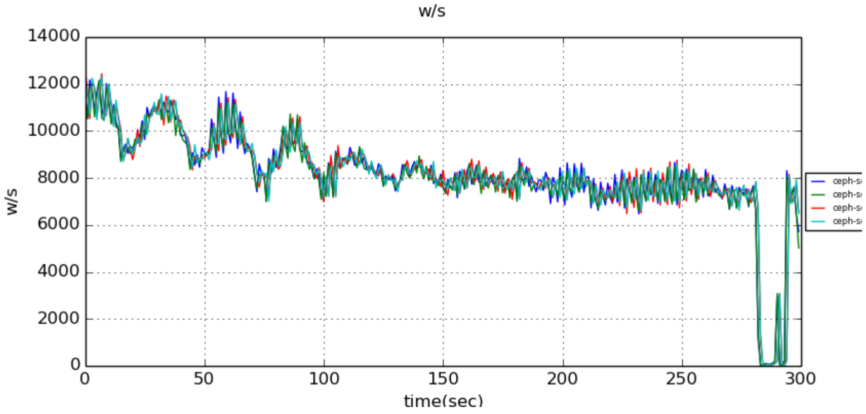
Ceph* all-flash array performance with BlueStore

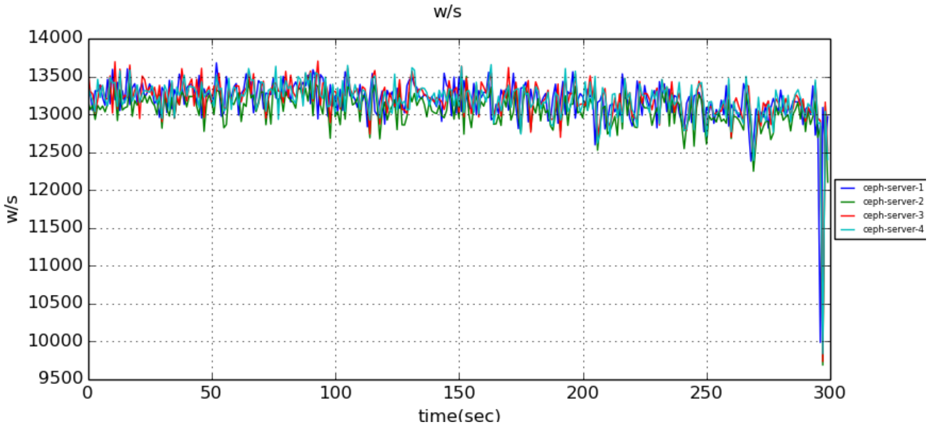


- High WAL bandwidth – need optimizations
  - 3.19x write amplification in total
- High latency - rocksdb?

Ceph* all-flash array performance with BlueStore

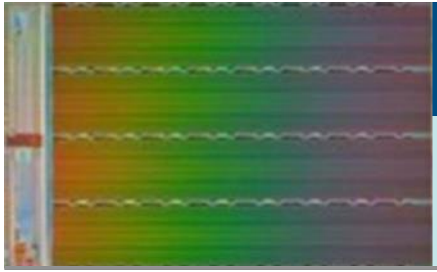
Rocksdb compaction overhead





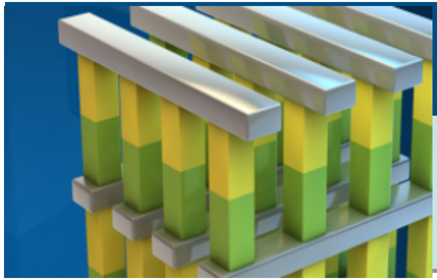
- Compaction will make throughput unstable – we can delay it with tunings, but cannot avoid it
- Still need rocksdb optimizations

Technology Driven: NVM Leadership



3D MLC and TLC NA

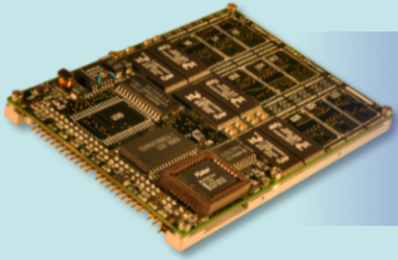
Building block enabling expansion of SSD into HDD segments



3D Xpoint™

Building blocks for ultra high performance storage & memory

Moore's Law Continues to Disrupt the Computing Industry



1992 12MB

First Intel® SSD for Commercial Usage

1,000,000x
the capacity while shrinking the form factor

2017 >10TB

U.2 SSD

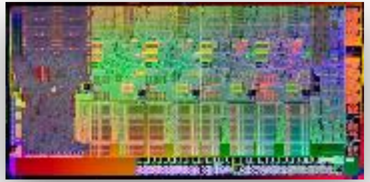
2014	2017	2018	2019
>6TB	>10TB	>30TB	1xxTB

Source: Intel projections on SSD capacity

3D Xpoint™ TECHNOLOGY

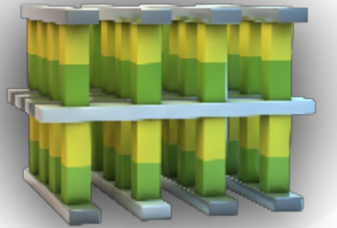
STORAGE

SRAM
Latency: 1X
Size of Data: 1X

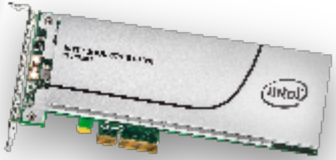


DRAM
Latency: ~10X
Size of Data: ~100X

3D XPoint™
Latency: ~100X
Size of Data: ~1,000X



NAND
Latency: ~100,000X
Size of Data: ~1,000X

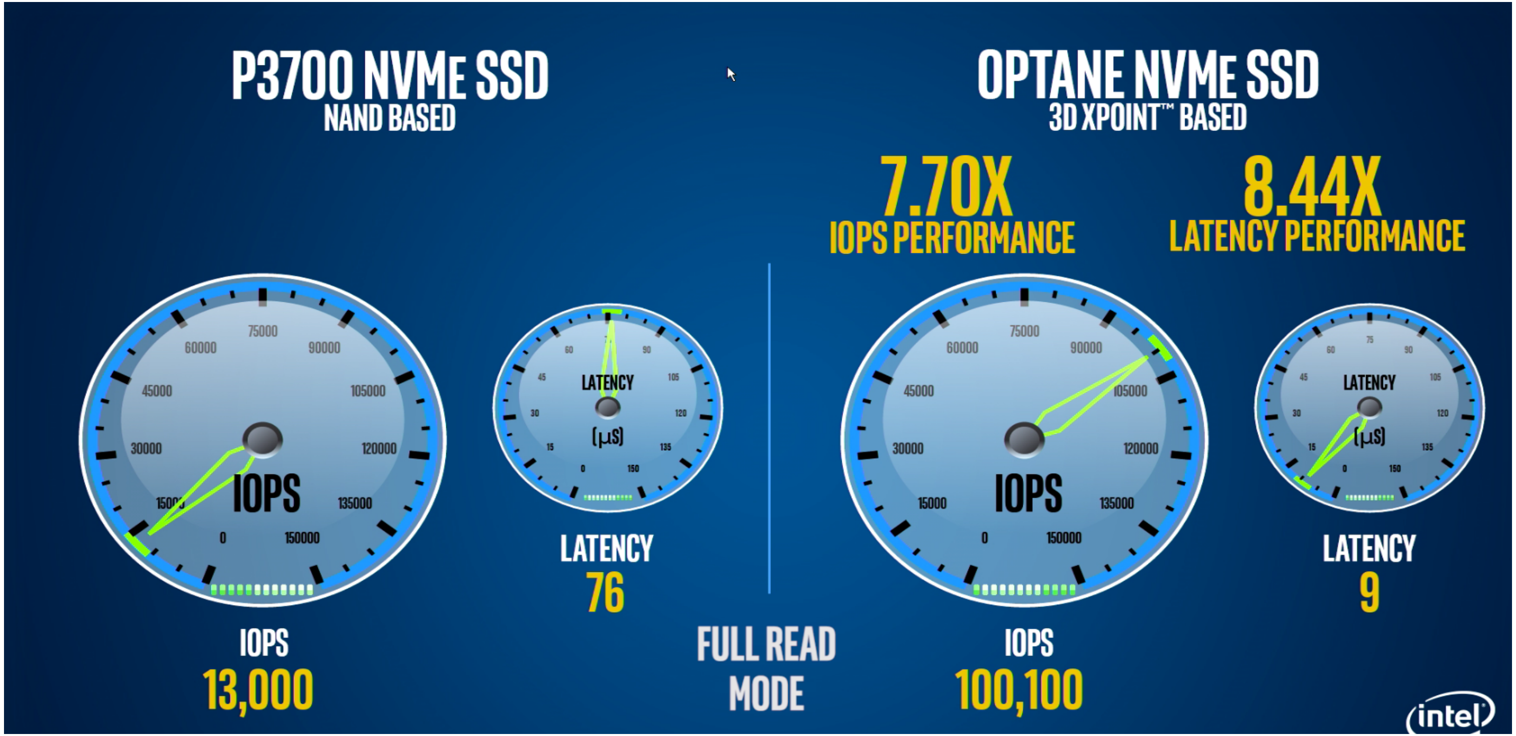


HDD
Latency: ~10 MillionX
Size of Data: ~10,000 X

MEMORY

Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

Intel® Optane™ storage (prototype) vs Intel® SSD DC P3700 Series at QD=1

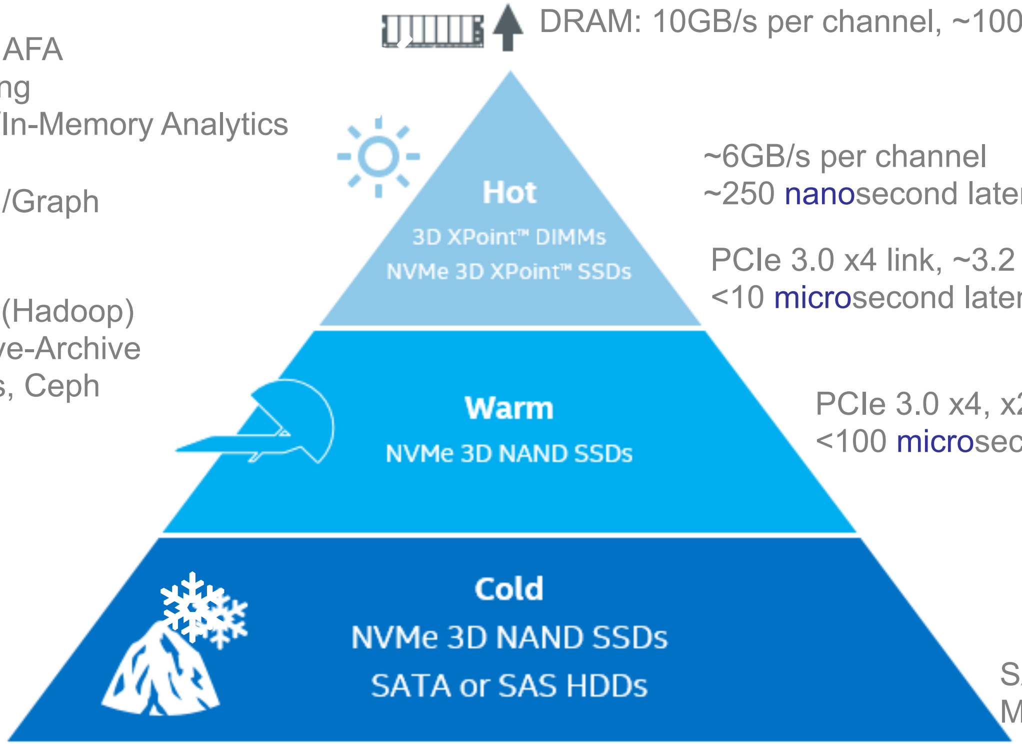


Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance. Server Configuration: 2x Intel® Xeon® E5 2690 v3 NVM Express* (NVMe) NAND based SSD: Intel P3700 800 GB, 3D Xpoint based SSD: Optane NVMe OS: Red Hat* 7.1

Storage Hierarchy Tomorrow

DRAM: 10GB/s per channel, ~100 nanosecond latency

- Server side and/or AFA
- Business Processing
- High Performance/In-Memory Analytics
- Scientific
- Cloud Web/Search/Graph



~6GB/s per channel
~250 nanosecond latency

PCIe 3.0 x4 link, ~3.2 GB/s
<10 microsecond latency

- Big Data Analytics (Hadoop)
- Object Store / Active-Archive
- Swift, lambert, hdfs, Ceph

PCIe 3.0 x4, x2 link
<100 microsecond latency

- Low cost archive

SATA 6Gbps
Minutes offline

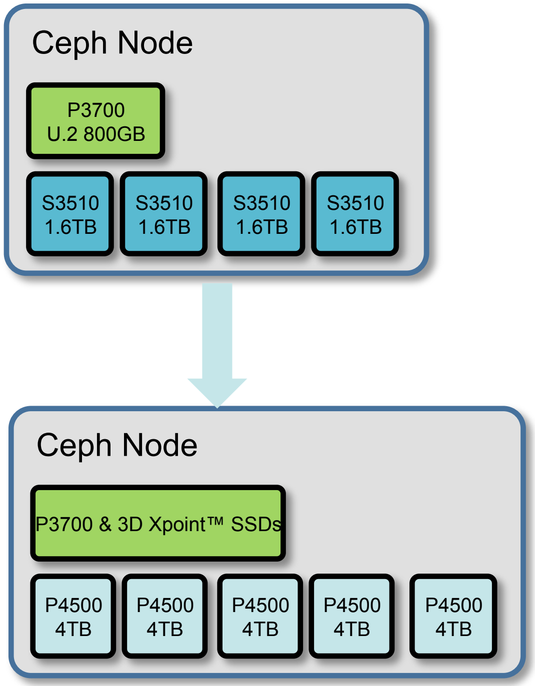
Comparisons between memory technologies based on in-market product specifications and internal Intel specifications.

3D XPoint™ & 3D NAND Enable High performance & cost effective solutions

Enterprise class, highly reliable, feature rich, and cost effective AFA solution:

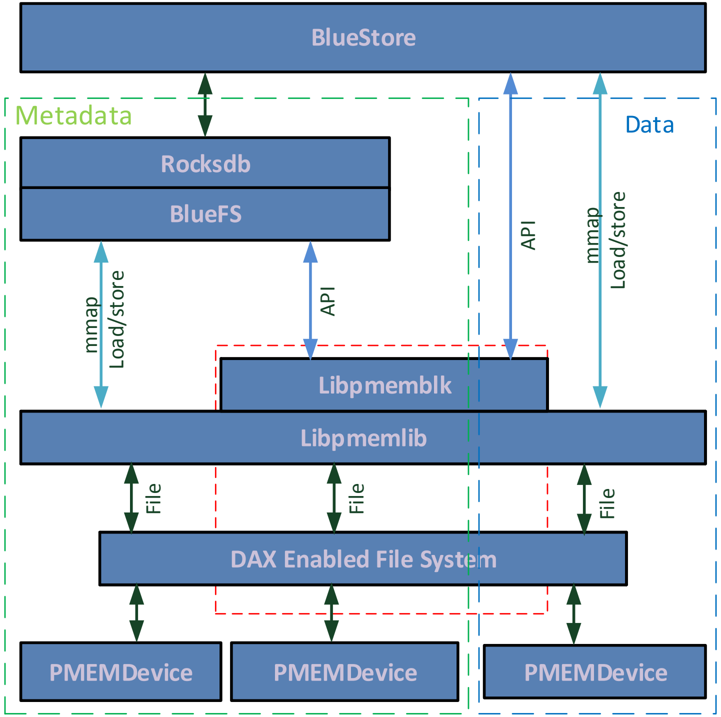
- Example:
- NVMe as Journal, 3D NAND TLC SSD as data store
  (performance) (capacity)

Enhance value through special software optimization on filestore and bluestore backend



3D Xpoint™ opportunities – Bluestore backend

- Three usages for PMEM device
  - Backend of bluestore: raw PMEM block device or file of dax-enabled FS
  - Backend of rocksdb: raw PMEM block device or file of dax-enabled FS
  - Backend of rocksdb's WAL: raw PMEM block device or file of DAX-enabled FS
- Two methods for accessing PMEM devices
  - libpmemblk
  - mmap + libpmemlib
    https://github.com/ceph/ceph/pull/8761



Summary

- Ceph is awesome!
- Strong demands for all-flash array ceph solutions
- SATA all-flash array Ceph cluster is capable of delivering over 1M IOPS with very low latency!
- Bluestore shows significant performance increase compared with filestore, but still needs to be improved
- Let's work together to make Ceph more efficient with all-flash array!

Legal notices and disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

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. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Intel, the Intel logo, 3D Xpoint, Intel Xeon, Intel Cache Acceleration Software and Intel Solid State Drive are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

Optimization Notice

Optimization Notice: 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.

Notice Revision #20110804

Backup

Testing Methodology

Storage interface

Use FIORBD as storage interface

Tool

- Use “dd” to prepare data for R/W tests
- Use fio (ioengine=libaio, direct=1) to generate 4 IO patterns: sequential write/read, random write/read
- Access Span: 60GB

Run rules

- Drop osds page caches (“1” > /proc/sys/vm/drop_caches)
- 100 secs for warm up, 600 secs for data collection
- Run 4KB/64KB tests under different # of rbds (1 to 120)

Ceph* All-Flash Tunings

= 0/0
debug journal = 0/0
debug mds_balancer = 0/0
debug mds = 0/0
mon_pg_warn_max_per_osd = 0/0
debug lockdep
auth = 0/0
debug auth
mds_log = 0/0
debug = 0/0
debug perfcounter = 0/0
debug = 0/0
debug
debug throttle = 0/0
debug client = 0/0
debug rgw = 0/0 migrator = 0/0
debug
debug finisher = 0/0
debug = 0/0
debug = 0/0
debug hadoop = 0/0
debug mds_locker = 0/0
debug
debug context = 0/0
debug osd = 0/0
debug bluestore = 0/0
debug objclass = 0/0

debug log = 0

debug filestore = 0/0 log_expire
debug
debug mds_log_expire = 0/0
debug = 0/0
debug rados = 0/0
debug
debug buffer = 0/0
asok = 0/0
debug objectcacher = 0/0
debug timer = 0/0
mutex_perf_counter = 0/0
mutex_perf_counter = True
rbd_cache = False
ms_crc_header = False
ms_crc_data = False
osd_pool_default_pgp_num
= 2
rbd_op_threads = 4
cephx require signatures = False
cephx sign messages = False
osd_pool_default_pg_num = 32768
throttler_perf_counter = False
auth_service_required = none
auth_cluster_required

auth_client_required
= none

osd_mount_options_xfs
osd_mount_options_xfs
rw,noatime,inode64,logbsize=256k,delaylog
xfs
filestore_queue_max_ops = 5000
osd_client_message_size_cap = 0
ms_dispatch_throttle_bytes = 1048576000
ms_dispatch_throttle_bytes = 1048576000
osd_mkfs_options_xfs
filestore_wbthrottle_enable = True
filestore_fd_cache_shards = 64
objecter_inflight_ops = 1024000
filestore_queue_committing_max_bytes = 1048576000
osd_op_num_threads_per_shard = 2
filestore_queue_max_bytes = 10485760000
osd_op_threads = 32
osd_op_num_shards = 16

filestore_op_threads = 16
osd_pg_object_context_cache_count = 10240
journal_queue_max_ops = 3000
journal_queue_max_bytes = 10485760000
journal_max_write_entries = 1000
filestore_queue_committing_max_ops = 5000
= 1048576000
osd_enable_op_tracker = False
filestore_fd_cache_size = 10240