Ceph BlueStore Performance on Latest Intel Server Platforms

Orlando Moreno
Performance Engineer, Intel Corporation
May 10, 2018
Legal Disclaimers

© 2017 Intel Corporation. Intel, the Intel logo, Xeon and Xeon logos 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.

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.

Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks/datacenter.

The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of 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.

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.

No computer system can be absolutely secure.

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.

Available on select Intel® processors. Requires an Intel® HT Technology-enabled system. Your performance varies depending on the specific hardware and software you use. Learn more by visiting http://www.intel.com/info/hyperthreading.
AGENDA

• Background
• Hardware and software configurations
• Performance overview
• Summary
BACKGROUND

Storage and Application Workloads

- SSD/NVMe Focus
- RBD/CephFS
- RadosGW – S3
Introducing Innovative NVMe*-Based Storage Solutions… for Today and the Future

Red Hat Ceph Storage* with Intel® Optane™ SSD DC P4800X combined with Intel® SSD DC P4500 delivers exceptional performance, lower latency, and reduced TCO.

1. Responsiveness defined as average read latency measured at Queue Depth 1 during 4k random write workload. Measured using FIO 2.15. Common configuration - Intel 2U Server System, OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Intel drives evaluated - Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P3700 1600GB. Samsung* drives evaluated – Samsung SSD PM1725a, Samsung SSD PM1725, Samsung PM963, Samsung PM953. Micron* drive evaluated – Micron 9100 PCIe* NVMe* SSD. Toshiba* drives evaluated – Toshiba ZD6300. Test – QD1 Random Read 4K latency, QD1 Random RW 4K 70% Read latency, QD1 Random Write 4K latency using FIO 2.15. *Other names and brands may be claimed as the property of others.
STORAGE EVOLUTION

Performance and Capacity for Every Need
PLATFORM EVOLUTION

Generation-to-Generation

Ceph Storage Performance
Small Block Workload

- Read
- Write

Ceph Storage Performance
Large Block Workload

- Read
- Write

Relative Performance to Intel®
Xeon® E5-2600 v4

Intel® Xeon® Scalable Processor

#redhat #rhsummit
BLUESTORE BACKEND

BlueStore is a new Ceph storage backend optimized for modern media

- key/value database (RocksDB) for metadata
- all data written directly to raw device(s)
- can combine HDD, SSD, NVMe, NVRAM
- ~2X faster than FileStore
- Better parallelism, efficiency on fast devices
- No double writes for data
- Performs well with very small journals
- Separate caching and data drives still recommended!
HARDWARE AND SOFTWARE CONFIGURATION

6-Node Disaggregated All-Flash Ceph Cluster

- Ceph 12.1.1-175 (Luminous rc) Bluestore
- 2x replication pool, 8192 PGs
- 1, 2, and 4 OSDs per NVMe SSD

6x Client Node
- Intel® Xeon® processor E5-2699 v4 @ 2.2GHz, 128GB mem
- 1x Single-Port 100GbE

6x Storage Node
- Intel® Xeon® Platinum 8176 Processor @ 2.1 GHz, 384GB Memory
- 1x Intel® Optane™ SSD DC P4800X 375G SSD as DB/WAL drive
- 4x 4.0TB Intel® SSD DC P4500 as data drive
- 2x Dual-Port 25GbE

#redhat #rhsummit
Determining Ceph RBD performance

Using Ceph Benchmarking Tool (CBT), FIO was run against several RBD volumes. Several metrics were collected:

- Aggregate IOPS and bandwidth
- Average and 99th percentile latency
- CPU utilization
BLUESTORE METADATA ON INTEL® OPTANE™ SSD

IOPS vs Latency
4KB Random Write, 60 libRBD Clients, 1-32 QD

99pct latency vs IO Depth
4KB Random Write, 60 libRBD Clients

*Lower is better
USING INTEL® OPTANE™ SSD FOR METADATA
RocksDB and WAL

Adding Intel® Optane™ SSD as metadata drive provides write latency improvements

• ~25% more IOPS with Optane for small block random writes
• ~50% increase in aggregate throughput (GB/s) for large (1MB) sequential writes
• Average latency decreases by up to 25%
• 2x lower long tail latency
SCALING STORAGE PERFORMANCE

Two vectors to scale Ceph performance:

- Co-locate multiple OSD processes on a NVMe device
- Add more NVMe devices per node
- Trade-offs for each method
OSD AND NVME SCALING

4KB Random Performance

2-4 OSDs/NVMe SSD and 4-6 NVMe SSDs per node are sweet spots
FUTURE WORK

RDMA in Ceph

- Default Ceph networking stack uses Async Messenger (TCP)
- Leverage RDMA to reduce CPU utilization and network layer latency
- Async Messenger compatible with RDMA (RoCE and iWARP)
- Functionally ready, but optimizations and testing on-going
SUMMARY

- Using Intel® Optane™ SSD DC P4800X combined with Intel® SSD DC P4500 for Ceph storage provides high performance, high capacity, and a more cost effective solution.
- Ceph Bluestore presents opportunities to utilize fast technology such as Intel® Optane™ SSD.
- On-going work to improve Ceph performance on NVMe and enable new technologies, such as RDMA.
THANK YOU

plus.google.com/+RedHat  facebook.com/redhatinc
linkedin.com/company/red-hat  twitter.com/RedHat
youtube.com/user/RedHatVideos
Pairing a new Intel developed controller, unique firmware innovations and disruptive capacities with Intel 3D NAND, the **cloud inspired** DC P4500 and DC P4600 have been **designed from the ground up** to deliver an all-PCIe* solution enabling highly efficient storage and caching to help data centers **do more per server, minimize service disruptions and efficiently manage at scale.**
STORAGE EVOLUTION

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
CEPH PARAMETERS

Global

- perf = true
- mutex.perf_counter = true
- throttler.perf_counter = false
- rbd cache = false
- rbd_cache_writethrough_until_flush = false
- rbd_op_threads = 2
- osd scrub load threshold = 0.01
- osd scrub min interval = 137438953472
- osd scrub max interval = 137438953472
- osd deep scrub interval = 137438953472
- osd max scrubs = 16

- log file = /var/log/ceph/$name.log
- log to syslog = false
- mon compact on trim = false
- osd pg bits = 8
- osd ppg bits = 8
- mon pg warn max object skew = 100000
- mon pg warn min per osd = 0
- mon pg warn max per osd = 32768

- osd_crush_chooseleaf_type = 0
CEPH PARAMETERS

OSD

osd_op_num_shards = 8
osd_op_num_threads_per_shard = 2
filestore_max_sync_interval = 1
filestore_op_threads = 10
filestore_queue_max_ops = 5000
filestore_queue_committing_max_ops = 5000
journal_max_write_entries = 1000
journal_queue_max_ops = 3000
objecter_inflight_ops = 102400
filestore_wbthrottle_enable = false
filestore_queue_max_bytes = 1048576000
filestore_queue_committing_max_bytes = 1048576000
journal_max_write_bytes = 1048576000
journal_queue_max_bytes = 1048576000
ms_dispatch_throttle_bytes = 1048576000
objecter_inflight_op_bytes = 1048576000
PERFORMANCE OVERVIEW

4KB Random Workload

IOPS vs Latency - 4OSD/NVMe
60 libRBD Clients, 1-128QD

- 319K Write IOPS @~6ms
- ~767K 70r/30w IOPS @~3ms
- 1.87M Read IOPS @~1ms
99TH PERCENTILE LATENCY

4KB Random Workload

IOPS vs Latency - 4OSD/NVMe
60 libRBD Clients, 1-128QD

- 47 ms @32 QD
- 24 ms @16 QD
- 1.16 ms @16 QD

#redhat #rhsummit
OSD AND NVME SCALING

4KB Random Read Performance

2-4 OSDs/NVMe SSD and 4-6 NVMe SSDs per node are sweet spots
OSD AND NVME SCALING

1MB Sequential Performance

**OSD & NVMe Scaling - Bandwidth**
1MB Sequential Write, 60 libRBD Clients

- Writes benefit from more OSDs
- Reads are bottlenecked by network

**OSD & NVMe Scaling - Bandwidth**
1MB Sequential Read, 60 libRBD Clients

#redhat #rhsummit
Sequential workloads are not CPU intensive