Bridging the Gap Between NVMe SSD Performance and Scale Out Software

Anjaneya “Reddy” Chagam
Principal Engineer, Intel Corporation

Swaroop Dutta
Director, Product Management, Storage and Availability Business Unit, VMware

Murali Rajagopal
Storage Architect, VMware
INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained from your local Intel sales office or your distributor; please address requests to your local Intel sales office. Intel products are not intended for use in medical, life support, or safety applications where failure or malfunction of an Intel product can reasonably be expected to result in personal injury. Intel Corporation and its disclaimers of warranties are provided as is in this document. Intel Corporation makes no representations or warranties with respect to the accuracy or completeness of the content in this document and reserves the right to make changes to specifications and product descriptions at any time without notice. Intel assumes no liability whatsoever and Intel disclaims any express or implied warranty, including without limitation any implied warranties of merchantability or fitness for a particular purpose, with respect to instruction sets, hardware, software, product, or services described in this material and makes no commitment to update the information hereafter.

This list and/or this document may be subject to change without notice. Other vendors are listed by Intel as a convenience to Intel’s general customer base, but Intel does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices. This list and/or these devices may be subject to change without notice. Other names and brands may be claimed as the property of others. No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology is a security technology under development by Intel and requires for operation a computer system with Intel® Virtualization Technology, an Intel Trusted Execution Technology-enabled processor, chipset, BIOS, Authenticode Code Modules, and an Intel or other compatible measured virtual machine monitor. In addition, Intel Trusted Execution Technology requires the system to contain a TPMv1.2 as defined by the Trusted Computing Group and specific software for some uses. See http://www.intel.com/technology/security/ for more information.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor. * Other names and brands may be claimed as the property of others.

Intel Corporation or its subsidiaries in the United States and other countries.

Copyright © 2016, Intel Corporation. All rights reserved.
VMware Legal Disclaimer

- This presentation may contain product features that are currently under development.
- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
- Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.
- Technical feasibility and market demand will affect final delivery.
- Pricing and packaging for any new technologies or features discussed or presented have not been determined.
Agenda

- NVM Express (NVMe) Overview
- Ceph Scale Out
  - Introduction
  - NVMe use cases
  - Low latency workload performance
  - Plans
- VMware vSAN
  - Introduction
  - All Flash workloads
  - NVMe integration
  - Plans
NVM Express (NVMe)
Standardized interface for non-volatile memory, http://nvmexpress.org

What is NVMe?
NVMe Express’ (NVMe) is a standardized high performance software interface for PCI Express’ Solid State Drives

Built for SSDs
Architected from the ground up for SSDs to be efficient, scalable, and manageable

Developed to be lean
Streamlined protocol with new efficient queuing mechanism to scale for multi-core CPUs, low clock cycles per IO

Industry standard software and drivers that work out of the box

Performance: 1 GB/s per lane.. 4 GB/s, 8 GB/s, 16 GB/s per device..
Lower latency: Direct CPU connection
No host bus adapter (HBA): Lower power ~ 10W and cost ~ $15
Increased I/O opportunity: Up to 40 PCIe lanes per CPU socket
Form factor options: PCIe add-in-card, SFF-8639, M.2, SATA Express, BGA

Add-in-card
M.2
(SFF-8639)

Ready for next generation SSDs
New storage stack with low latency and small overhead to take full advantage of next generation NVM

Source: Intel. Other names and brands are property of their respective owners. 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.

2016 Storage Developer Conference. © Intel Corporation. All Rights Reserved.
Ceph with NVMe
Ceph - Introduction

- Open-source, object-based scale-out storage
- Object, Block and File in single unified storage cluster
- Highly durable, available – replication, erasure coding
- Runs on economical commodity hardware
- 10 years of hardening, vibrant community

- Scalability – CRUSH data placement, no single POF
- Replicates and re-balances dynamically
- Enterprise features – snapshots, cloning, mirroring
- Most popular block storage for OpenStack use cases
- Commercial support from Red Hat


© Intel Corporation. All Rights Reserved.
Ceph NVM Workloads

Storage Performance (IOPS, Throughput)

Storage Capacity (PB)

Higher

Lower

Block

Remote Disks
Boot Volumes
Test & Dev

NVM Focus

Object

Mobile Content Depot
Enterprise Dropbox
App Storage
Backup, Archive

VDI
CDN
Cloud DVR

Databases
HPC
BigData

Lower

Higher

CDN
Enterprise
Dropbox
Backup, Archive

Remote Disks
Boot Volumes
Test & Dev

BigData
Cloud DVR

Databases
HPC
Ceph and NVMe SSDs

Ceph Clients

Bare-metal

User

App

Flash

Cache

Kernel RBD

RADOS

Kernel

VM

Guest VM

App

Qemu/VirtIO

LIBRBD

RADOS

Hypervisor

Flash

Cache

Kernel RBD

RADOS

Host

Container

Container

App

Flash

Cache

Kernel RBD

RADOS

RADOS

IP Network

Ceph Cluster

Filestore Backend (Production)

OSD

CEPH NODE

Journal

Filestore

Flash

Filesystem

Cache

Flash

Bluestore Backend (Tech Preview)

OSD

CEPH NODE

Data

Metadata

RocksDB

BlueRocksEnv

BlueFS

Flash

Cache

Flash

Flash

Client-side Cache

Journal

Read Cache

OSD Data

* NVM – Non-volatile Memory

2016 Storage Developer Conference. © Intel Corporation. All Rights Reserved.
Configuration Options for Ceph Storage Node

Standard/good

NVM Express (NVMe)/PCI Express (PCIe) SSD for Journal + Caching, HDDs as OSD data drive

Example: 1x Intel P3700 1.6TB as Journal + Intel® Cache Acceleration Software (Intel® CAS) caching software + up to 16 HDDs

Better (best TCO, as of today)

NVMe/PCIe SSD as Journal + High capacity SATA SSD for data drive

Example: 1x Intel P3700 800GB + 6x Intel S3510 1.6TB

Best Performance

All NVMe/PCIe SSDs

Example: 4x Intel P3700 2TB SSDs

---

**Ceph storage node -- Good**

<table>
<thead>
<tr>
<th>CPU</th>
<th>Intel® Xeon® CPU E5-2650v4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>64 GB</td>
</tr>
<tr>
<td>NIC</td>
<td>10GbE</td>
</tr>
<tr>
<td>Disks</td>
<td>1x 1.6TB P3700 + 16 x 4TB HDDs (1:16 ratio) P3700 as Journal and caching</td>
</tr>
<tr>
<td>Caching software</td>
<td>Intel CAS 3.0, option: Intel® Rapid Storage Technology enterprise/MD4.3</td>
</tr>
</tbody>
</table>

**Ceph Cluster -- Better**

<table>
<thead>
<tr>
<th>CPU</th>
<th>Intel Xeon CPU E5-2690v4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>128 GB</td>
</tr>
<tr>
<td>NIC</td>
<td>Dual 10GbE</td>
</tr>
<tr>
<td>Disks</td>
<td>1x 800GB P3700 + 6x S3510 1.6TB</td>
</tr>
</tbody>
</table>

**Ceph Cluster -- Best**

<table>
<thead>
<tr>
<th>CPU</th>
<th>Intel Xeon CPU E5-2699v4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>&gt;= 128 GB</td>
</tr>
<tr>
<td>NIC</td>
<td>1x 40GbE, 4x 10GbE</td>
</tr>
<tr>
<td>Disks</td>
<td>4x P3700 2TB</td>
</tr>
</tbody>
</table>

Standard/Better/Best designations reflect only levels of feature choices and are not intended as competitive comparisons

2016 Storage Developer Conference. © Intel Corporation. All Rights Reserved.
Databases, NVMe SSDs and Ceph

- Why MySQL?
  - Leading open-source RDBMS
  - MySQL #4 workload on Openstack
    (#1-3 use databases too)
  - 70% Openstack apps use LAMP

- Why NVMe SSDs?
  - High throughput
  - Dependable Latency

- DBA-friendly Ceph feature-set
  - Shared, elastic storage pools
  - Snapshots (full and incremental) for easy backup
  - Copy-on-write cloning
  - Flexible volume resizing
  - Live Migration
  - Async Volume Mirroring
All-NVMe Ceph Cluster for MySQL Hosting

5-Node all-NVMe Ceph Cluster
Dual-Xeon E5 2699v4@2.2GHz, 44C HT, 128GB DDR4
RHEL7.2, 3.10-327, Ceph v10.2.0, bluestore async

10x Client Systems
Dual-socket Xeon E5 2699v3@2.3GHz
36 Cores HT, 128GB DDR4

Tests at cluster fill-level 82%
Hardware and Architectural Considerations

- **Compute**
  - Dual-socket Xeon E5v4 config for 4+ NVMes per storage node
  - Pin SoftIRQs for NVMe/NIC devices to it's associated NUMA node
  - Observed performance increases with higher core count, faster clock, and larger CPU cache
  - Xeon E5-2695v4 or better for 16 OSDs per node (ref: 5-core-GHZ/OSD)

- **Network**
  - Intel X520-T2 dual-10GbE
  - Separate public/cluster networks, split OSD subnets

- **Storage**
  - 1.6TB Intel P3700 NVMe SSDs for bluestore data and metadata
  - Latest Red Hat kernels drivers, supported Ceph SKUs such as Red Hat Ceph Storage (yielded us better performance)
  - Leverage Ceph cluster sizing and performance tuning guides available from Red Hat, Intel and partners (see references)

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Any difference in system hardware or software design or configuration may affect actual performance. See configuration slides in backup for details on software configuration and test benchmark parameters.

First Ceph cluster to break ~1.4 Million 4K random IOPS, ~1ms response time in 5U

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Any difference in system hardware or software design or configuration may affect actual performance. See configuration slides in backup for details on software configuration and test benchmark parameters.
Sysbench MySQL OLTP Performance

Sysbench Thread Scaling - Latency vs QPS – 100% read (Point SELECTs)
5 nodes, 80 OSDs, Xeon E5 2699v4 Dual Socket / 128GB Ram / 2x10GbE
Ceph 10.1.2 w/ BlueStore. 20 Docker-rbd Sysbench Clients (16vCPUs, 32GB)

~1.3 million QPS with 20 Sysbench clients, 8 Sysbench threads/client

~55000 QPS with 1 client
1 million QPS with 20 clients @ ~11 ms avg
2 Sysbench threads/client

InnoDB buf pool = 25%, SQL dataset = 100GB
Database page size = 16KB

2016 Storage Developer Conference. © Intel Corporation. All Rights Reserved.
Ceph - NVMe Focus Areas (2016-17)

- Ceph RBD NVMe SSD caching
- Data efficiency features – Compression and Dedupe
- Long tail latency optimization
- Ceph OSD optimizations to reduce CPU overhead
  - Data Plane Development Kit (DPDK) with user mode TCP/IP
  - Storage Performance Development Kit (SPDK) user mode NVMe
- BlueStore and PMEM integration
VMware VSAN with NVMe
Hyper-converged infrastructure

**HYPER-CONVERGED SOFTWARE**
Compute, storage and networking
Tightly integrated software stack

**INDUSTRY-STANDARD HARDWARE**
Convergence of physical storage on x86 hardware
Building-block approach
VMware VSAN: overview

- Software-Defined Storage
- Distributed, Scale-out Architecture
- Hyper-Converged Infrastructure
- Integrated with vSphere platform
- Policy Driven Control Plane
Tiered All-Flash and Hybrid Options

All-Flash
- SSD
- PCIe
- NVMe

- Writes cached first,
  Reads from capacity tier
- Capacity Tier
  Flash Devices
  Reads primarily from capacity tier

100K IOPS per Host
+ sub-millisecond latency

Hybrid
- SSD
- PCIe
- NVMe

- Read and Write Cache
- Capacity Tier
  SAS / NL-SAS / SATA

40K IOPS per Host
Current: VSAN All Flash

- **2 Tier Architecture:**
  - Tier 1 Caching: High performance, high endurance flash for caching writes
  - Tier 2 Data Persistence: Read intensive, low endurance drives for capacity

- **Space Efficiency:** 2X – 8X savings with Deduplication, Compression & Erasure Coding

- **Performance:** 4X IOPS of Hybrid VSAN; sub millisecond latency response times

- **Ideal Workloads:** Business Critical Applications (Exchange DAG), Transactional (OLTP, SQL), VDI

- **Customer Adoption:** Gaining significant momentum, aligned with enterprise adoption of flash, particularly NVMe
NVM Express (NVMe) – Market and Architecture is evolving

Enterprise SSD by Interface

By 2018, NVM Express projected to be > 70% of client SSD market

PCle projected as leading SSD interface in DC by 2018
Virtual SAN – NVMe - Benefits

Overview
- Non Volatile Memory Express (NVMe) is a highly optimized controller interface that significantly improves performance for enterprise workloads.
- NVMe devices provide increased performance over traditional SSDs.
  - Reduced latencies, significantly higher IOPS due to increased parallelism.
  - High endurance (3x), low power (30%).

Benefits
- Ideal for caching tier for All Flash configurations specifically for workloads that require high IOPS and low latencies.

NVMe Enablement
- NVMe devices are currently certified for VSAN Caching tier, specifically for All Flash configurations using the NVMe certification suite.
- Roadmap: Enhancing ESXi and VSAN storage stack to achieve close to raw device IOPS from NVMe (caching tier).
## Virtual SAN All-Flash – Intel NVMe Ready Node

<table>
<thead>
<tr>
<th>Components</th>
<th>Details</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SKU</td>
<td>VRN2208WAF8</td>
<td>1</td>
</tr>
<tr>
<td>ESXi Pre-Installed?</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>System</td>
<td>Intel® Server System R2208WTYYSR</td>
<td>2</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel® Xeon E5-2600 V4 (14 cores)</td>
<td>24</td>
</tr>
<tr>
<td>Memory</td>
<td>16GB DDR4 RDIMM</td>
<td>1</td>
</tr>
<tr>
<td>Caching Tier</td>
<td>Intel SSD DC P3700 Series SSDPE2MD400G4 (400 GB, 2.5-inch)</td>
<td>2</td>
</tr>
<tr>
<td>Capacity Tier</td>
<td>Intel® SSD DC S3510 Series SSDSC2BB012T6 (1.2 TB, 2.5-inch)</td>
<td>12</td>
</tr>
<tr>
<td>Controller</td>
<td>Intel RAID Controller RS3UC080</td>
<td>2</td>
</tr>
<tr>
<td>NIC</td>
<td>Intel Dual port 10Gb RJ45/SFP+</td>
<td>1</td>
</tr>
<tr>
<td>Boot Device</td>
<td>Intel SSD DC S3710 200GB</td>
<td>1</td>
</tr>
</tbody>
</table>
• Four Disk Groups with:
  • 4 x800GB Intel P3700 PCIe SSDs per host
  • 20x2TB Intel P3500 PCIe SSDs per host

• Provides 25TB of cache and 320TB Raw Storage

• Cost-Performance:
  • Over 7x Cost reduction per effective GB with Dedup/CMP
  • $0.25/GB – $1.86/GB
NVMe vs SAS SSD Performance

- Two Disk Groups with a total of:
  - 2x400GB Intel P3700 PCIe SSDs
  - 6x800 GB Intel S3500 SSDs
- 100GB working set per host
- Virtual SAN configured with Space efficiency features disabled

1.5x IOPS with NVMe vs SAS SSDs with 70/30 R:W ratios with ~1ms latency
What’s Coming Next?

Disclaimer: no feature or timeline commitments
Intel® Optane™ SSD Latency

Intel® Optane™ SSDs offer ~10x reduction in latency versus NAND SSD

The latency comparisons above include internal Intel estimates for the SSD NAND NVMe™ and SSD with Intel® Optane™ technology.
ESXi Application Performance Delivered by Intel® Optane™ is:
2.4x faster than NAND PCI Express*!

Software Optimizations may unleash even more performance and value!

VMware® ESXi 6.0 Update 1. Windows® Server VMs running 4kB 70/30 R/W QD8 using IOMeter. 4 workers per SSD device. SSDs used: NVM Express*: Intel® SSD Data Center P3700 Series (800 GB) achieving 66k IOPs (shown on slide 10), and Intel prototype SSD using Intel Optane Technology (shown here). SuperMicro® 2U SuperServer 2028U-TNR4T+. Dual Intel® Xeon® Processor E5-2699 v3 (45M Cache, 2.30 GHz). 192 GB DDR4 DRAM. Boot drive: Intel® SSD Data Center S3710 Series (256 GB). 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.
ESXi Storage Stack Enhancements for NVMe Performance Boost

Hardware:
- Intel® Xeon® E5-2687W v3 @3.10GHz (10 cores + HT)
- 64 GB RAM
- NVM Express® 1M IOPS @ 4K Reads

Software:
- vSphere® 6.0U2 vs. Future prototype
- 1 VM, 8 VCPU, Windows® 2012, 4 VMDK eager-zeroed
- IOMeter: 4K seq reads, 64 OIOs per worker, even distribution of workers to VMDK
Virtual SAN with Next-Generation Hardware (NVMe + NVDIMM)
Future: Evolving VSAN to exploit next-gen hardware

- **Performance Boost with Next Gen H/W**
  1. **High Speed NVMe**: Enable VSAN to use high speed, low latency NVMe for caching (2017)
  2. **ESXi Storage Stack Enhancements**: Achieve close to raw NVMe device IOPS (million IOPS)
  3. **NVDIMM (Metadata) + NVMe (Write Cache)**
  4. **RDMA over Ethernet**: To boost n/w transfers, reduce latencies & CPU utilization (2018)
vSphere NVMe Native Driver Stack
NVM Express Evolution & vSphere

- Queueing Interface
- NVM Command Set
- Admin Command Set
- End-to-end Protection (DIF/DIX)
- Security
- Physical Region Pages (PRPs)

NVMe 1.0 Released March 1, 2011
- General Scatter Gather Lists (SGLs)
- Multi-Path I/O & Namespace Sharing
- Reservations
- Autonomous Power Transitions During Idle

NVMe 1.1 Released October 11, 2012
- Implementation and Reporting Refinements
- Name Space Management
- Controller Memory Buffer
- Host Memory Buffer
- Ground Work for NVMe Management

NVMe 1.2 Released November 3, 2014
- General Scatter Gather Lists (SGLs)
- Multi-Path I/O & Namespace Sharing
- Reservations
- Autonomous Power Transitions During Idle
- Ground Work for NVMe Management

NVMe Over Fabric 1.0 Released June 5, 2016
- Defines extension to NVMe, for non PCI
- Primary focus on RDMA
- Compatible to FC-NVMe (INCITS 540)
- Host Memory Buffer
- Ground Work for NVMe Management

vSphere 5.5
- Introduce first async NVMe driver 1.0e
- Launch IOVP cert program for NVMe

vSphere 6.0
- Introduce first inbox NVMe driver
- Bring broader ecosystem support

vSphere (Future)
- Multiple name spaces, Queues
- NVMe Over Fabric
- vNVMe
- End-to-end NVMe Stack
Where to get more information?

• vSphere 5.5: Download VMware ESXi 5.5 Driver CD for NVM Express (NVMe) driver.

• vSphere 6.0: available as part of base image.
  – Also available for download VMware ESXi 5.5 nvme 1.2.0.27-4vmw NVMe Driver for PCI Express based Solid-State Drives

• NVMe Ecosystem:

• vSphere NVMe Open Source Driver to encourage ecosystem to innovate
  – https://github.com/vmware/nvme
THANK YOU
Ceph.conf (Best Config)

[osd]
  osd_mkfs_type = xfs
  osd_mount_options_xfs = rw,noatime,inode64,logbsize=256k,delaylog
  filestore_queue_max_ops = 5000
  osd_client_message_size_cap = 0
  objecter_inflight_op_bytes = 1048576000
  ms_dispatch_throttle_bytes = 1048576000
  osd_mkfs_options_xfs = -f -i size=2048
  filestore_wbthrottle_enable = True
  filestore_fd_cache_shards = 64
  objecter_inflight_ops = 1024000
  filestore_queue_committing_max_bytes = 1048576000
  osd_op_num_shards = 16
  osd_op_num_threads_per_shard = 2
  filestore_queue_max_bytes = 1048576000
  rbd_op_threads = 4
  filestore_max_sync_interval = 10
  filestore_op_threads = 16
  osd_pg_object_context_cache_count = 10240
  journal_queue_max_ops = 3000
  filestore_odsync_write = True
  journal_queue_max_bytes = 1048576000
  journal_max_write_entries = 1000
  filestore_queue_committing_max_ops = 5000
  journal_max_write_bytes = 1048576000
  osd_enable_op_tracker = False
  filestore_fd_cache_size = 10240
  osd_client_message_cap = 0
  journal_dynamic_throttle = True
  osd_enable_op_tracker = False

Default OSD shard and threads per shard tuning parameters appear to be well chosen. A 10% improvement in concurrent read performance may be gained by increasing the number of shards or threads per shard, though potentially at the expense of higher single operation write latency. This is especially true when these settings are configured to be significantly higher than default. Lowering the default values potentially can dramatically decrease concurrent read performance. The node used in this testing has 12 physical cores and it may be that simply matching the total number of shards/threads (across all OSDs) to the number of cores tends to produce the best overall results.

https://www.spinics.net/lists/ceph-users/attachments/pdfA9vNSS0XEF.pdf

© Intel Corporation. All Rights Reserved.
Configuration Detail – ceph.conf

[globals]

enable experimental unrecoverable data corrupting features = bluestore rocksdb
osd objectstore = bluestore
ms_type = async

rbd readahead disable after bytes = 0
rbd readahead max bytes = 4194304
bluestore default buffered read = true

auth client required = none
auth cluster required = none
auth service required = none
filestore xattr use omap = true

cluster network = 192.168.142.0/24, 192.168.143.0/24
private network = 192.168.144.0/24, 192.168.145.0/24

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

debug_lockdep = 0/0
debug_context = 0/0
debug_crush = 0/0
debug_buffer = 0/0
debug_timer = 0/0
debug_filer = 0/0
debug_objecter = 0/0
debug_rados = 0/0
debug_rbd = 0/0
debug_ms = 0/0
debug_monc = 0/0
debug_tp = 0/0
debug_auth = 0/0
debug_finisher = 0/0
debug_heartbeatmap = 0/0
debug_perfcounter = 0/0
debug_asok = 0/0
debug_throttle = 0/0
debug_mon = 0/0
debug_paxos = 0/0
debug_rgw = 0/0
perf = true
mutex_perf_counter = true
throttler_perf_counter = false
rbd cache = false
Configuration Detail – ceph.conf (continued)

```
[mon]
mon data=/home/bmps/tmp_cbt/ceph/mon.$id
mon_max_pool_pg_num=166496
mon_osd_max_split_count = 10000
mon_pg_warn_max_per_osd = 10000

[osd]
osd_mount_options_xfs = rw,noatime,inode64,logbsize=256k,delaylog
osd_mkfs_options_xfs = -f -i size=2048
osd_op_threads = 32
filestore_queue_max_ops=5000
filestore_queue_committing_max_ops=5000
journal_max_write_entries=10000
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
osd_mkfs_type = xfs
filestore_max_sync_interval=10
osd_client_message_size_cap = 0
osd_client_message_cap = 0
osd_enable_op_tracker = false
filestore_fd_cache_size = 64
filestore_fd_cache_shards = 32
filestore_op_threads = 6
```

```
[mon.a]
host = ft02
mon addr = 192.168.142.202:6789
```
Configuration Detail - CBT YAML File

cluster:
  user: "bmpa"
  head: "ft01"
  clients: ["ft01", "ft02", "ft03", "ft04", "ft05", "ft06"]
  osds: ["hswNode01", "hswNode02", "hswNode03", "hswNode04", "hswNode05"]
  mons:
    ft02:
      a: "192.168.142.202:6789"
  osds_per_node: 16
  fs: xfs
  mkfs_opts: '-f -i size=2048 -n size=64k'
  mount_opts: '-o inodetype,inoatime,logbsize=256k'
  conf_file: '/home/bmpa/cbt/ceph.conf'
  use_existing: False
  newstore_block: True
  rebuild_every_test: False
  clusterid: "ceph"
  iterations: 1
  tmp_dir: "/home/bmpa/tmp_cbt"

pool_profiles:
  2rep:
    pg_size: 8192
    pgp_size: 8192
    replication: 2

benchmarks:
  librbdio:
    time: 300
    ramp: 300
    vol_size: 10
    mode: ['randrw']
    rwmixread: [0,70,100]
    op_size: [4096]
    procs_per_volume: [1]
    volumes_per_client: [10]
    use_existing_volumes: False
    iodepth: [4,8,16,32,64,128]
    osd_ra: [4096]
    norandommap: True
    cmd_path: '/usr/local/bin/fio'
    pool_profile: '2rep'
    log_avg_msec: 250
MySQL configuration file (my.cnf)

[client]
port = 3306
socket = /var/run/mysqld/mysqld.sock

[mysqld_safe]
socket = /var/run/mysqld/mysqld.sock
nice = 0

[mysqld]
user = mysql
pid-file = /var/run/mysqld/mysqld.pid
socket = /var/run/mysqld/mysqld.sock
port = 3306
datadir = /data
basedir = /usr
tmpdir = /tmp
lc-messages-dir = /usr/share/mysql
skip-external-locking
bind-address = 0.0.0.0
max_allowed_packet = 16M
thread_stack = 192K
thread_cache_size = 8
query_cache_limit = 1M
query_cache_size = 16M
log_error = /var/log/mysqld/error.log
expire_logs_days = 10
max_binlog_size = 100M

performance_schema=off
innodb_buffer_pool_size = 25G
innodb_flush_method = O_DIRECT
innodb_log_file_size=4G
thread_cache_size=16
innodb_file_per_table
innodb_checksums = 0
innodb_flush_log_at_trx_commit = 0
innodb_write_io_threads = 8
innodb_page_cleaners = 16
innodb_read_io_threads = 8
max_connections = 50000

[mysqldump]
quick
quote-names
max_allowed_packet = 16M

[mysql]
!includedir /etc/mysql/conf.d/
**Sysbench commands**

**PREPARE**

```
```

**READ**

```
system --mysql-host=${host} --mysql-port=${mysql_port} --mysql-user=sbtest --mysql-password=sbtest --mysql-db=sbtest --mysql-engine=innodb --oltp-tables-count=32 --oltp-table-size=14000000 --test=/root/benchmarks/sysbench/sysbench/tests/db/oltp.lua --oltp-read-only=on --oltp-simple-ranges=0 --oltp-sum-ranges=0 --oltp-order-ranges=0 --oltp-distinct-ranges=0 --oltp-index-updates=0 --oltp-point-selects=10 --rand-type=uniform --num-threads=${threads} --report-interval=60 --warmup-time=400 --max-time=300 --max-requests=0 --percentile=99 run
```

**WRITE**

```
system --mysql-host=${host} --mysql-port=${mysql_port} --mysql-user=sbtest --mysql-password=sbtest --mysql-db=sbtest --mysql-engine=innodb --oltp-tables-count=32 --oltp-table-size=14000000 --test=/root/benchmarks/sysbench/sysbench/tests/db/oltp.lua --oltp-read-only=off --oltp-simple-ranges=0 --oltp-sum-ranges=0 --oltp-order-ranges=0 --oltp-distinct-ranges=0 --oltp-index-updates=100 --oltp-point-selects=0 --rand-type=uniform --num-threads=${threads} --report-interval=60 --warmup-time=400 --max-time=300 --max-requests=0 --percentile=99 run
```
Docker Commands

- **Database containers**
  - docker run -ti --privileged --volume /sys:/sys --volume /dev:/dev -d -p 2201:22 -p 13306:3306 --cpuset-cpus="1-16,36-43" -m 48G --oom-kill-disable --name database1 ubuntu:14.04.3_20160414-db /bin/bash

- **Client containers**
  - docker run -ti -p 3301:22 -d --name client1 ubuntu:14.04.3_20160414-sysbench /bin/bash
RBD Commands

- ceph osd pool create database 8192 8192
- rbd create --size 204800 vol1 --pool database --image-feature layering
- rbd snap create database/vol1@master
- rbd snap ls database/vol1
- rbd snap protect database/vol1@master
- rbd clone database/vol1@master database/vol2
- rbd feature disable database/vol2 exclusive-lock object-map fast-diff deep-flatten
- rbd flatten database/vol2
NVM Express

NVMe Development Timeline

- NVMe 1.0 – Mar-2011
  - Queueing Interface
  - Command Set
  - End-to-End Protection
  - Security
  - PRP's

- NVMe 1.1 – Oct-2012
  - Multi-Path IO
  - Namespace Sharing
  - Reservations
  - Autonomous Power Transition
  - Scatter Gather Lists

- NVMe 1.2 – Q4 2014
  - Host Memory Buffer
  - Replay Protected Memory
  - Active/Idle Power and RTD3
  - Temperature Thresholds
  - Namespace Management
  - Enhanced Status Reporting
  - Pass through support
  - Controller Memory Buffer
  - Firmware Update w/ no Reset

NVMe: CPU Efficient

Submission latency and CPU cycles reduced >50%*:

- NVMe: 2.8us, 9,100 cycles
- SAS: 6.0us, 19,500 cycles

* Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux kernel 3.13.
3D XPoint™ and Ceph

- First 3D XPoint Use Cases for Bluestore
  - Bluestore Backend, RocksDB Backend, RocksDB WAL

- Two methods for accessing PMEM devices
  - Raw PMEM blockdev (libpmemblk)
  - DAX-enabled FS (mmap + libpmemlib)
3D NAND - Ceph cost effective solution

- Enterprise class, highly reliable, feature rich, and cost effective AFA solution
  - NVMe SSD is today’s SSD, and 3D NAND or TLC SSD is today’s HDD
    - NVMe as Journal, high capacity SATA SSD or 3D NAND SSD as data store
    - Provide high performance, high capacity, a more cost effective solution
      - 1M 4K Random Read IOPS delivered by 5 Ceph nodes
        - Cost effective: 1000 HDD Ceph nodes (10K HDDs) to deliver same throughput
        - High capacity: 100TB in 5 nodes
  - with special software optimization on filestore and bluestore backend

2016 Storage Developer Conference. © Intel Corporation. All Rights Reserved.
Multi-partitioned NVMe SSDs

- High performance NVMe devices are capable of high parallelism at low latency
  - DC P3700 800GB Raw Performance: 460K read IOPS & 90K Write IOPS at QD=128
- High Resiliency of “Data Center” Class NVMe devices
  - At least 10 Drive writes per day
  - Power loss protection, full data path protection, device level telemetry
- By using multiple OSD partitions, Ceph performance scales linearly
  - Reduces lock contention within a single OSD process
  - Lower latency at all queue-depths, biggest impact to random reads
- Introduces the concept of multiple OSD’s on the same physical device
  - Conceptually similar crush map data placement rules as managing disks in an enclosure