Flash Memory Summit

# Increasing Ceph Performance Cost-Effectively with New Non-Volatile Technologies

Jian Zhang, Software Engineer Manager, jian.zhang@intel.com
Brien Porter, Senior Program Manager,  brien.porter@intel.com
Jack Zhang, Senior Enterprise Solution Architect,  yuan.zhang@intel.com

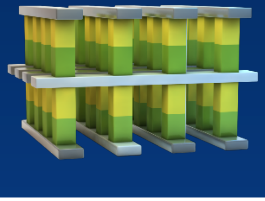
Santa Clara, CA
August 2017

# Agenda

- Ceph* with Intel® Non-Volatile Memory Technologies
- 2.8M IOPS Ceph* cluster with Intel® Optane™ SSDs + Intel® 3D TLC SSDs
- Ceph* Performance analysis on Intel® Optane™ SSDs based all-flash array
- Summary

# Intel® 3D NAND SSDs and OPTANE SSD Transform Storage

## Optimized STORAGE Solutions

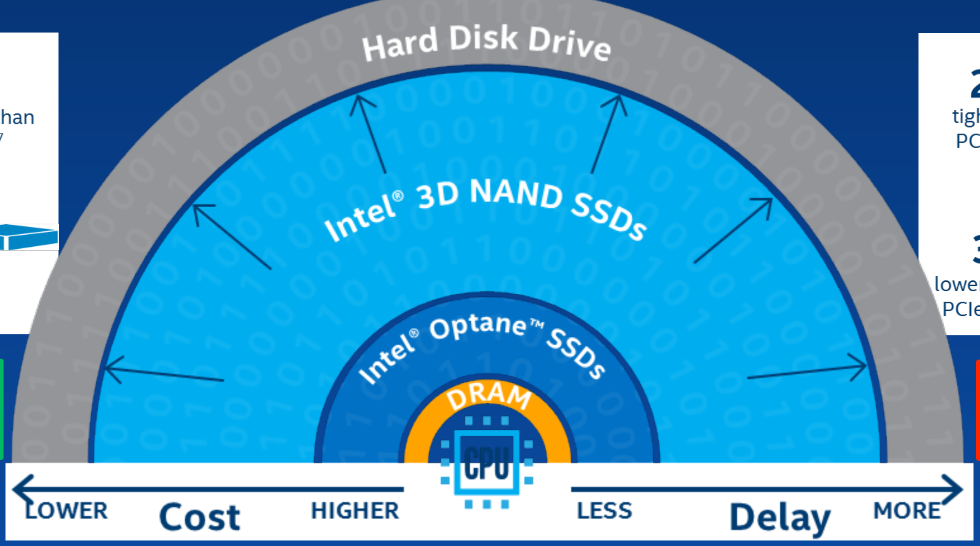


Up to **359x** more IOPS/$ than 10K HDD[6]

**>2X** higher endurance than 2D NAND SSDs[7]

Up to **217x** more IOPS/W than 10K HDD[6]

**More capacity** per rack unit[11]



Up to **200x** tighter QoS than PCIe NAND SSD

**>3X** higher endurance than PCIe NAND SSD

Up to **30%** lower power than PCIe ANND SSD

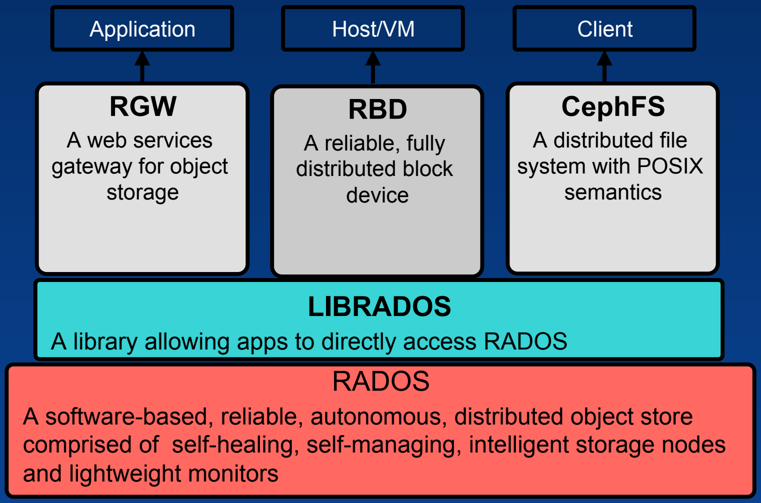
**More VMs, Same QoS** per rack

**Capacity** for Less

**Performance** for Less

Refer to appendix for footnotes

# A Brief Ceph Introduction



Application

Host/VM

Client

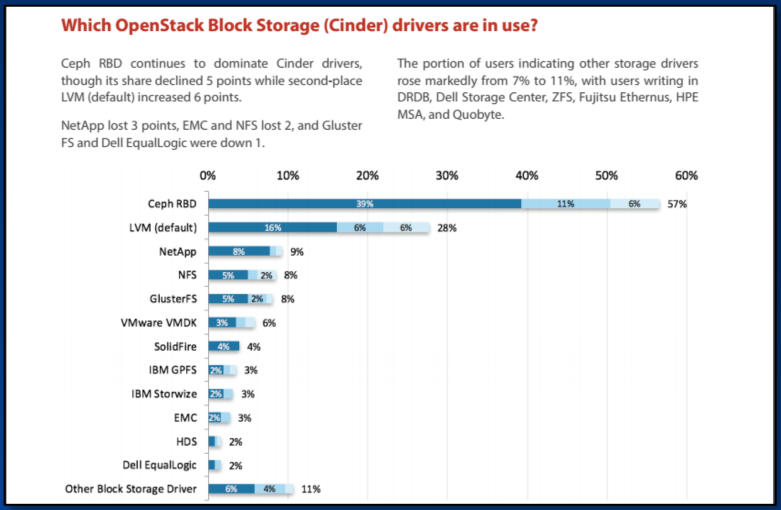
**RGW**
A web services gateway for object storage

**RBD**
A reliable, fully distributed block device

**CephFS**
A distributed file system with POSIX semantics

**LIBRADOS**
A library allowing apps to directly access RADOS

RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors



**Which OpenStack Block Storage (Cinder) drivers are in use?**

Ceph RBD continues to dominate Cinder drivers, though its share declined 5 points while second-place LVM (default) increased 6 points.

NetApp lost 3 points, EMC and NFS lost 2, and Gluster FS and Dell EqualLogic were down 1.

The portion of users indicating other storage drivers rose markedly from 7% to 11%, with users writing in DRDB, Dell Storage Center, ZFS, Fujitsu Ethernus, HPE MSA, and Quobyte.

- 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

References: http://ceph.com/ceph-storage, http://thenewstack.io/software-defined-storage-ceph-way,

# Who is using Ceph?

Telcom

CISCO

TIVIT synapsis

China unicom中国联通

中国电信 CHINA TELECOM

américa móvil

CSP/IPDC

YAHOO!

ebay

DigitalOcean

Ctrip

360

Letv

SHANDA 盛大集团

OEM/ODM

H3C

QCT Quanta CLOUD TECHNOLOGY

Hewlett Packard Enterprise

DELL

Enterprise, FSI, Healthcare, Retailers

AMX

Walmart

GE imagination at work

Adobe

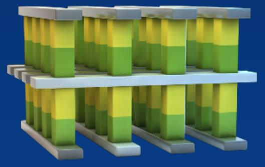
TARGET

Bloomberg

Searchable Examples: http://www.slideshare.net/inktank_ceph

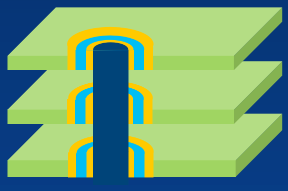
# Innovation for Cloud STORAGE : Intel® Optane™ + Intel® 3D NAND SSDs

- New Storage Infrastructure: enable high performance and cost effective storage:



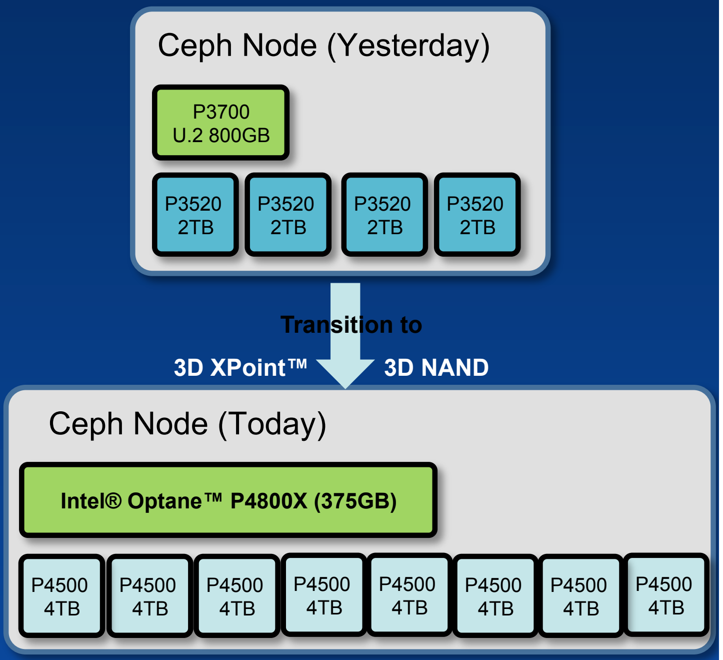
Journal/Log/Cache

+



Data

- Openstack/Ceph:
  - Intel Optane™ as Journal/Metadata/WAL (**Best** write performance, **Lowest** latency and **Best** QoS)
  - Intel 3D NAND TLC SSD as data store (cost effective storage)
  - **Best IOPS/$, IOPS/TB and TB/Rack**



Ceph Node (Yesterday)

P3700 U.2 800GB

P3520 2TB | P3520 2TB | P3520 2TB | P3520 2TB

**Transition to**

**3D XPoint™** **3D NAND**

Ceph Node (Today)

**Intel® Optane™ P4800X (375GB)**

P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB

# Suggested Configurations for Ceph Storage Node

- **Standard/good (baseline):**
- ***Use cases/Applications:*** *that need high capacity storage with high throughput performance*
  - **NVMe*/PCIe* SSD for Journal + Caching, HDDs as OSD data drive**
- **Better IOPS**
- ***Use cases/Applications:*** *that need higher performance especially for throughput,  IOPS and SLAs with medium storage capacity requirements*
  - **NVMe/PCIe SSD as Journal, High capacity SATA SSD for data drive**
- **Best Performance**
- ***Use cases/Applications****: that need highest performance (throughput  and IOPS) and low latency/QoS (Quality of Service).*
  - **All NVMe/PCIe SSDs**

| Ceph* storage node --Good | |
|---|---|
| CPU | Intel(R) Xeon(R) CPU E5-2650v4 |
| Memory | 64 GB |
| NIC | 10GbE |
| Disks | 1x 1.6TB P3700 + 12 x 4TB HDDs (1:12 ratio)<br>P3700 as Journal and caching |
| Caching software | Intel(R) CAS 3.0, option: Intel(R) RSTe/MD4.3 |

| Ceph* Storage node --Better | |
|---|---|
| CPU | Intel(R) Xeon(R) CPU E5-2690v4 |
| Memory | 128 GB |
| NIC | Duel 10GbE |
| Disks | 1x Intel(R) DC P3700(800G) + 4x Intel(R) DC S3510 1.6TB<br>Or 1xIntel P4800X (375GB) + 8x Intel® DC S3520 1.6TB |

| Ceph* Storage node --Best | |
|---|---|
| CPU | Intel(R) Xeon(R) CPU E5-2699v4 |
| Memory | >= 128 GB |
| NIC | 2x 40GbE, 4x dual 10GbE |
| Disks | 1xIntel P4800X (375GB) + 6x Intel® DC P4500 4TB |

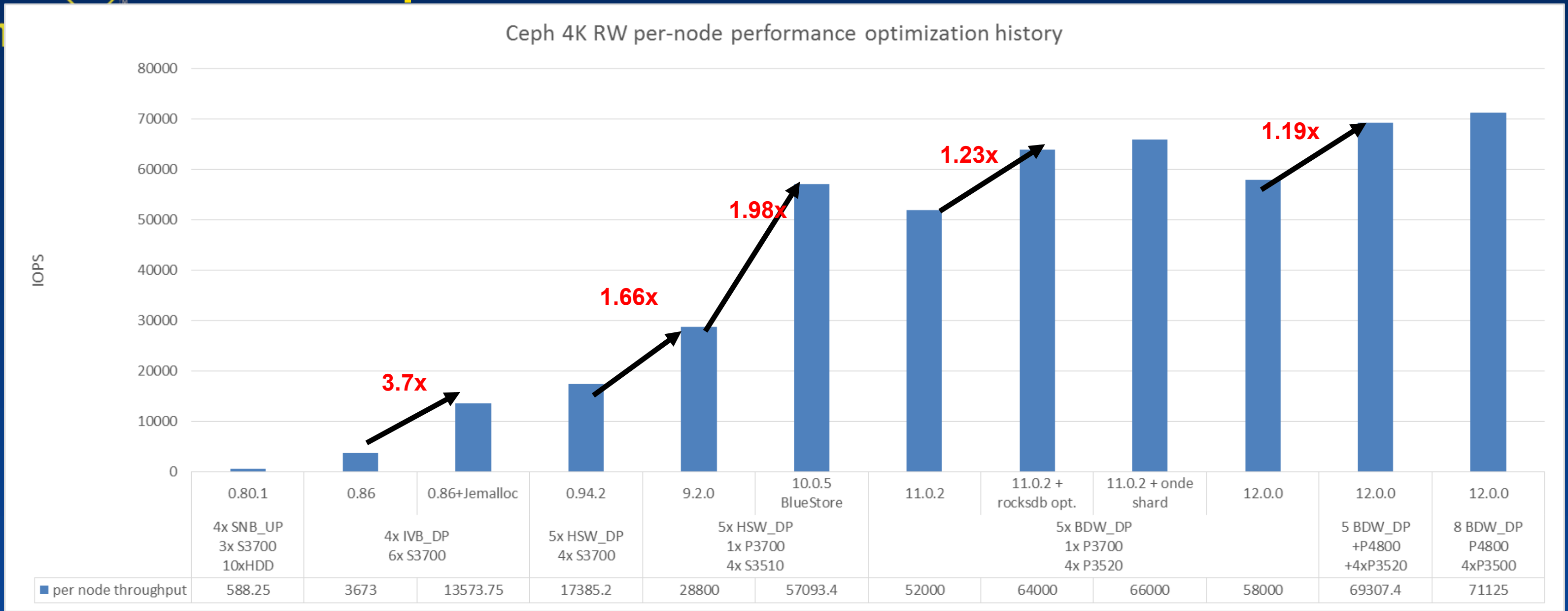
*Other names and brands may be claimed as the property of others. **More information at Ceph.com  (new RAs update soon!)**
http://tracker.ceph.com/projects/ceph/wiki/Tuning_for_All_Flash_Deployments

# Drivers for Ceph on All-Flash Arrays

- 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 Amazon EBS like services for their OpenStack* based public/private cloud
- Strong demands to run enterprise applications
  - OLTP workloads running on Ceph, tail latency is critical
  - high performance multi-purpose Ceph cluster is a key advantage
  - Performance is still an important factor
- SSD performance continue to increase while price continue to decrease

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

# Ceph Performance Trends with SSD



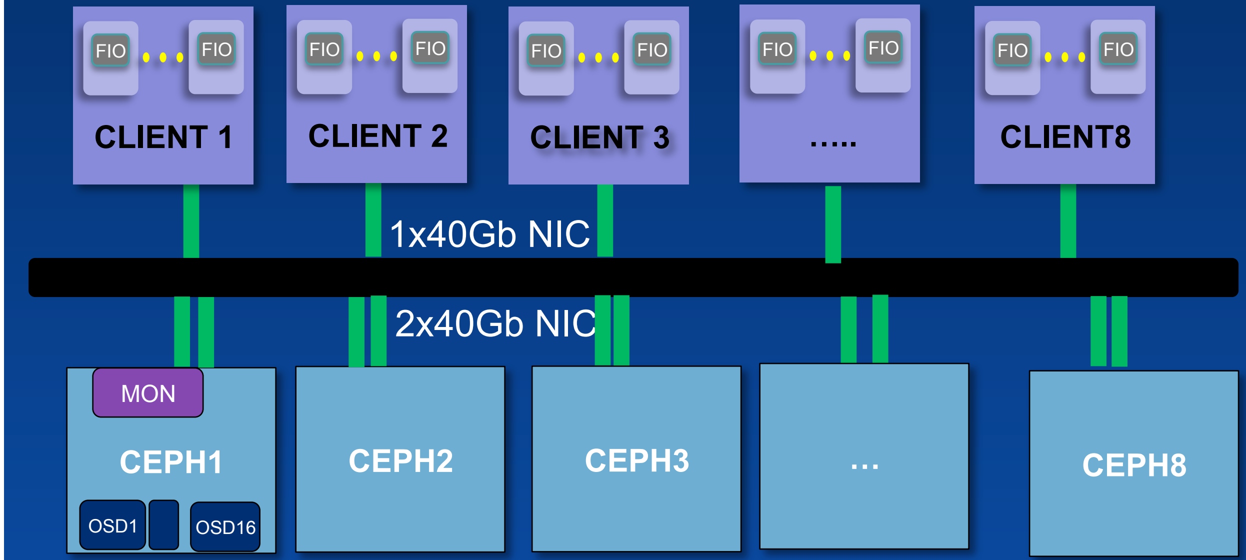
| | 0.80.1 | 0.86 | 0.86+Jemalloc | 0.94.2 | 9.2.0 | 10.0.5 BlueStore | 11.0.2 | 11.0.2 + rocksdb opt. | 11.0.2 + onde shard | 12.0.0 | 12.0.0 | 12.0.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4x SNB_UP<br>3x S3700<br>10xHDD | 4x IVB_DP<br>6x S3700 | | 5x HSW_DP<br>4x S3700 | 5x HSW_DP<br>1x P3700<br>4x S3510 | | 5x BDW_DP<br>1x P3700<br>4x P3520 | | | | 5 BDW_DP<br>+P4800<br>+4xP3520 | 8 BDW_DP<br>P4800<br>4xP3500 |
| ■ per node throughput | 588.25 | 3673 | 13573.75 | 17385.2 | 28800 | 57093.4 | 52000 | 64000 | 66000 | 58000 | 69307.4 | 71125 |

- 18x performance improvement in Ceph on all-flash array!

*Refer to backup section for detail cluster configuration

# Ceph All-Flash Optane Configuration

## Test Environment



Workloads
- Fio with librbd
- 20x 30 GB volumes each client
- 4 test cases: 4K random read & write; 64K Sequential read & write

8x Client Node
- Intel® Xeon™ processor E5-2699 v4 @ 2.3GHz, 64GB mem
- 1x X710 40Gb NIC

8x Storage Node
- Intel Xeon processor E5-2699 v4 @ 2.3 GHz
- 256GB Memory
- 1x 400G SSD for OS
- 1x Intel® DC P4800 375G SSD as WAL and DB
- 8x 2.0TB Intel® SSD DC P4500 as data drive
- 2 OSD instances one each P4500 SSD
- Ceph 12.0.0 with Ubuntu 16.10

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

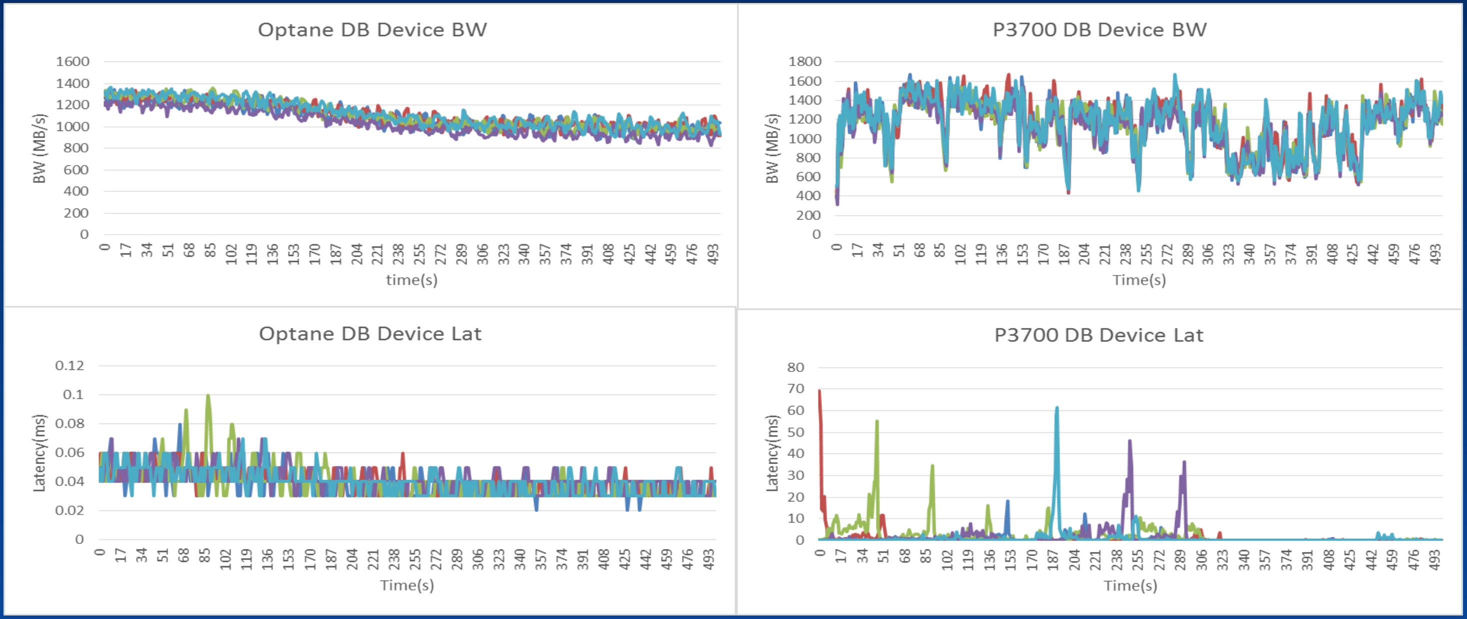
Flash Memory Summit

# Ceph Optane Performance Overview

| | Throughput | Latency (avg.) | 99.99% latency (ms) |
|---|---|---|---|
| 4K Random Read | 2876K IOPS | 0.9 ms | 2.25 |
| 4K Random Write | 610K IOPS | 4.0 ms | 25.435 |
| 64K Sequential Read | 27.5 GB/s | 7.6 ms | 13.744 |
| 64K Sequential Write | 13.2 GB/s | 11.9 ms | 215 |

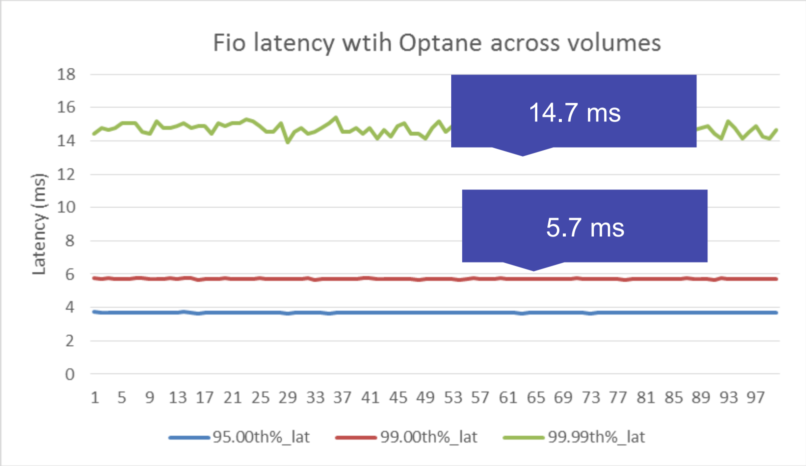
- Excellent performance on Optane cluster
  - random read & write hit CPU bottleneck

# Ceph Optane Performance Improvement



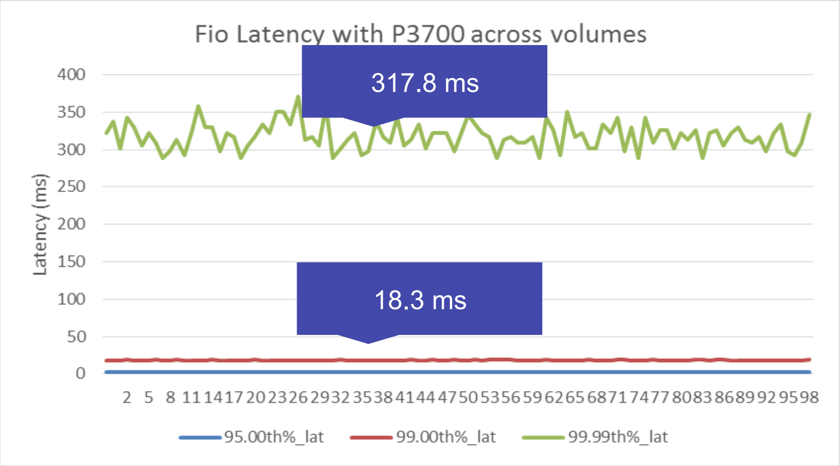
- The breakthrough high performance of Optane eliminated the WAL & rocksdb bottleneck
  - 1 P4800X or P3700 covers up to 8x P4500 data drivers as both WAL and rocksdb

# Ceph Optane Latency Improvement


Fio latency wtih Optane across volumes

14.7 ms

5.7 ms

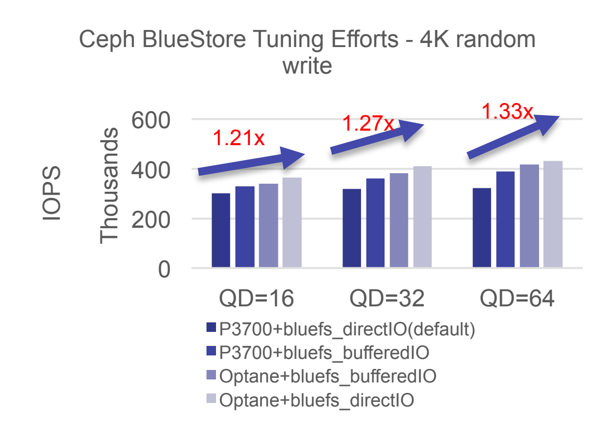

Fio Latency with P3700 across volumes

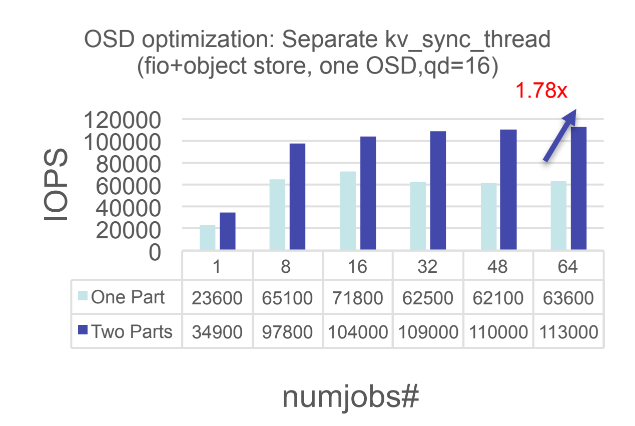
317.8 ms

18.3 ms

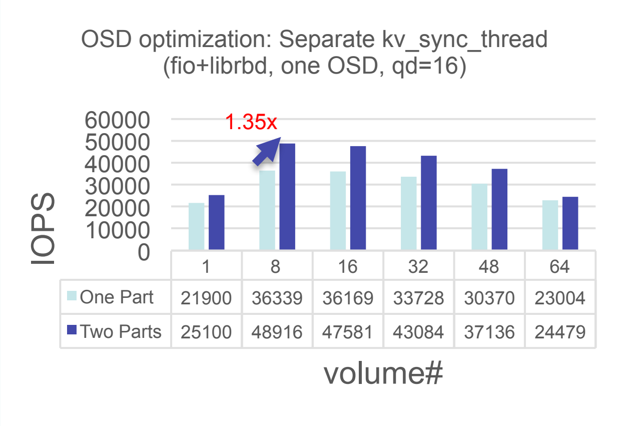
- Significant tail latency improvement with Optane
  - 20x latency reduction for 99.99% latency

# Ceph Performance Optimization on Optane





| | 1 | 8 | 16 | 32 | 48 | 64 |
|---|---|---|---|---|---|---|
| One Part | 23600 | 65100 | 71800 | 62500 | 62100 | 63600 |
| Two Parts | 34900 | 97800 | 104000 | 109000 | 110000 | 113000 |



| | 1 | 8 | 16 | 32 | 48 | 64 |
|---|---|---|---|---|---|---|
| One Part | 21900 | 36339 | 36169 | 33728 | 30370 | 23004 |
| Two Parts | 25100 | 48916 | 47581 | 43084 | 37136 | 24479 |

Optane Performance advantage over P3700 increased from 7% to 33% with tunings (bufferIO)

Optane Optimizations with separate kv_sync_thread

- Separate the thread to feed KV as much as possible. (#PR13943, merged)
- 1.77x performance boost with OSD side optimization on Optane, 1.3x with librbd interface
- Need to further optimize OSD layer

# Summary & Next

- Summary
  - Ceph* is awesome!
  - Strong demands for all-flash array Ceph* solutions
  - Optane based all-flash array Ceph* cluster is capable of delivering over 2.8M IOPS with very low latency!
  - Let’s work together to make Ceph* more efficient with all-flash array!
- Next
  - Improving Ceph network messenger performance with RDMA.
  - Ceph NVMeOF solutions
  - Client side cache on Optane with SQL workloads!

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

# Acknowledgements

- This is a joint effort
- Thanks for the contributions of Haodong, Tang, Jianpeng Ma of our Intel Shanghai development team

# Backup

# Ceph All Flash Tunings

```
[global]
 pid_path = /var/run/ceph
    auth_service_required = none
    auth_cluster_required = none
    auth_client_required = none
    mon_data = /var/lib/ceph/ceph.$id
    osd_pool_default_pg_num = 2048
    osd_pool_default_pgp_num = 2048
    osd_objectstore = bluestore
    public_network = 172.16.0.0/16
    cluster_network = 172.18.0.0/16
    enable experimental unrecoverable data corrupting features = *
    bluestore_bluefs = true
    bluestore_block_create = false
    bluestore_block_db_create = false
    bluestore_block_wal_create = false
    mon_allow_pool_delete = true
    bluestore_block_wal_separate = false
 debug objectcacher = 0/0
    debug paxos = 0/0
    debug journal = 0/0
    mutex_perf_counter = True
    rbd_op_threads = 4
    debug ms = 0/0
    debug mds = 0/0
    mon_pg_warn_max_per_osd = 10000
    debug lockdep = 0/0
    debug auth = 0/0
    ms_crc_data = False
    debug mon = 0/0
    debug perfcounter = 0/0
    perf = True
    debug monc = 0/0
    debug throttle = 0/0
    debug mds_migrator = 0/0
    debug mds_locker = 0/0
```

```
debug rgw = 0/0
   debug finisher = 0/0
   debug osd = 0/0
   debug mds_balancer = 0/0
   rocksdb_collect_extended_stats = True
   debug hadoop = 0/0
   debug client = 0/0
   debug zs = 0/0
   debug mds_log = 0/0
   debug context = 0/0
   rocksdb_perf = True
   debug bluestore = 0/0
   debug bluefs = 0/0
   debug objclass = 0/0
   debug objecter = 0/0
   debug log = 0
   ms_crc_header = False
   debug filer = 0/0
   debug rocksdb = 0/0
   rocksdb_collect_memory_stats = True
   debug mds_log_expire = 0/0
   debug crush = 0/0
   debug optracker = 0/0
   osd_pool_default_size = 2
   debug tp = 0/0
   cephx require signatures = False
   cephx sign messages = False
   debug rados = 0/0
   debug journaler = 0/0
   debug heartbeatmap = 0/0
   debug buffer = 0/0
   debug asok = 0/0
   debug rbd = 0/0
   rocksdb_collect_compaction_stats = False
   debug filestore = 0/0
   debug timer = 0/0
   rbd_cache = False
   throttler_perf_counter = False
```

```
▪     [mon]
▪        mon_data = /var/lib/ceph/mon.$id
▪        mon_max_pool_pg_num = 166496
▪        mon_osd_max_split_count = 10000
▪        mon_pg_warn_max_per_osd = 10000
▪     [osd]
▪        osd_data = /var/lib/ceph/mnt/osd-device-$id-data
▪        osd_mkfs_type = xfs
▪        osd_mount_options_xfs = rw,noatime,inode64,logbsize=256k
▪        bluestore_extent_map_shard_min_size = 50
▪        bluefs_buffered_io = true
▪        mon_osd_full_ratio = 0.97
▪        mon_osd_nearfull_ratio = 0.95
▪        bluestore_rocksdb_options =
      compression=kNoCompression,max_write_buffer_number=32,min_write_buffer_number
      _to_merge=2,recycle_log_file_num=32,compaction_style=kCompactionStyleLevel,write
      _buffer_size=67108864,target_file_size_base=67108864,max_background_compaction
      s=31,level0_file_num_compaction_trigger=8,level0_slowdown_writes_trigger=32,level0_
      stop_writes_trigger=64,num_levels=7,max_bytes_for_level_base=536870912,max_byte
      s_for_level_multiplier=8,compaction_threads=32,flusher_threads=8
▪        bluestore_min_alloc_size = 65536
▪        osd_op_num_threads_per_shard = 2
▪        osd_op_num_shards = 8
▪        bluestore_extent_map_shard_max_size = 200
▪        bluestore_extent_map_shard_target_size = 100
▪        bluestore_csum_type = none
▪        bluestore_max_bytes = 1073741824
▪        bluestore_wal_max_bytes = 2147483648
▪        bluestore_max_ops = 8192
▪        bluestore_wal_max_ops = 8192
▪
```

Flash Memory Summit

# Legal notices

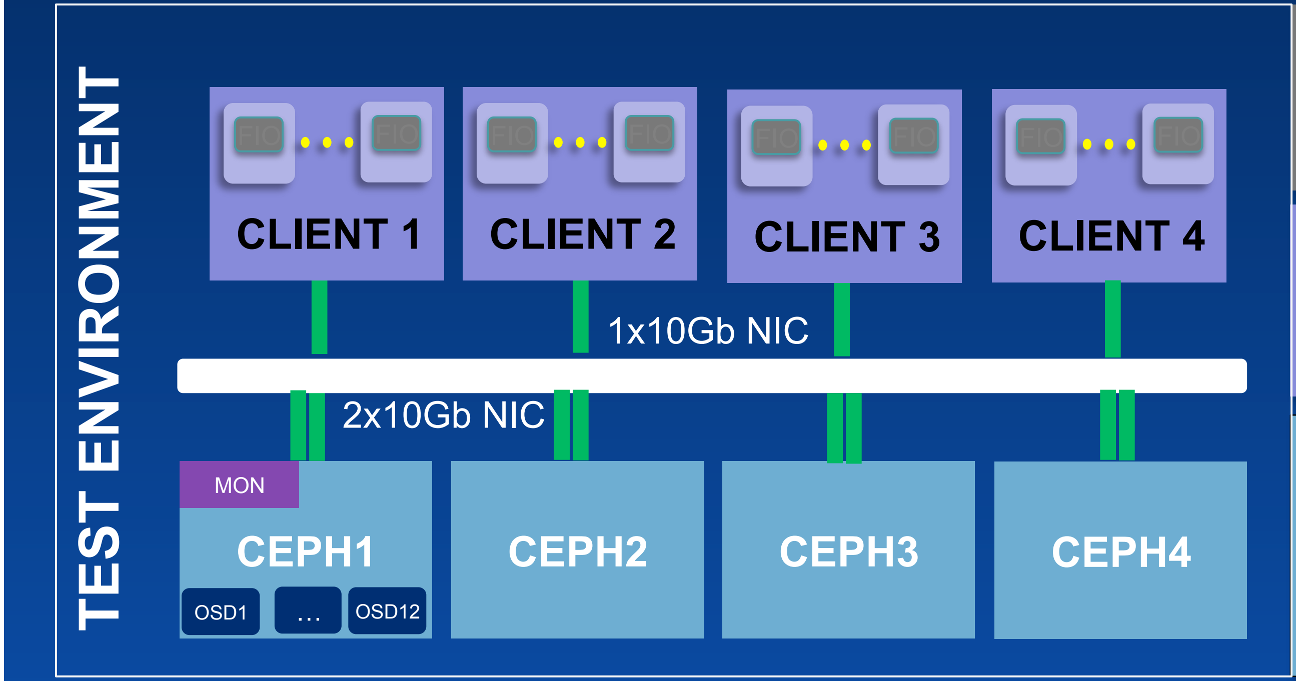
**FTC Optimization Notice**

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

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

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.

# Ceph* All Flash SATA configuration - IVB (E5 -2680 V2) + 6X S3700



TEST ENVIRONMENT

FIO • • • FIO

CLIENT 1

FIO • • • FIO

CLIENT 2

FIO • • • FIO

CLIENT 3

FIO • • • FIO

CLIENT 4

1x10Gb NIC

2x10Gb NIC

MON

CEPH1

OSD1 … OSD12

CEPH2

CEPH3

CEPH4

**WORKLOADS**

- Fio with librbd
- 20x 30 GB volumes each client
- 4 test cases: 4K random read & write; 64K Sequential read & write

**COMPUTE NODE**

2 nodes with Intel® Xeon™ processor x5570 @ 2.93GHz, 128GB mem
1 node with Intel Xeon processor E5 2680 @2.8GHz, 56GB mem

**STORAGE NODE**

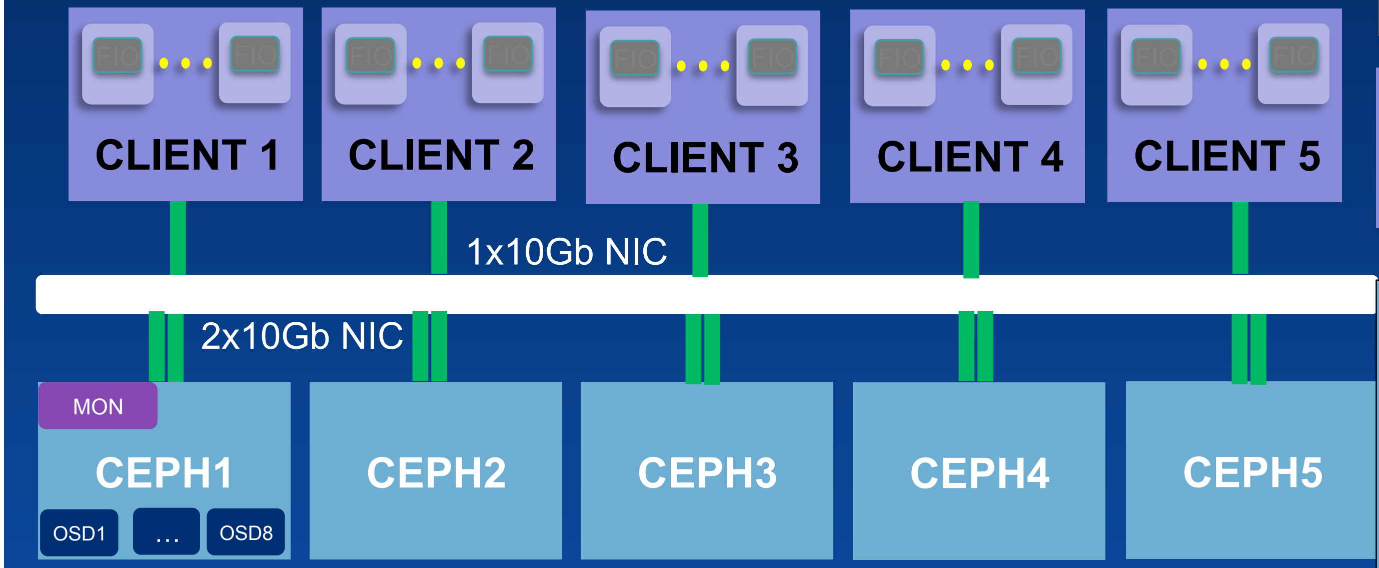
Intel Xeon processor E5-2680 v2
32GB Memory
1xSSD for OS
6x 200GB Intel® SSD DC S3700
2 OSD instances each Drive

*other names and brands may be claimed as the property of others.
Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

Flash Memory Summit

# Ceph* All Flash SATA configuration - HSW (E5 -2699 v3) + P3700 + S3510

## Test Environment



Workloads

- Fio with librbd
- 20x 30 GB volumes each client
- 4 test cases: 4K random read & write; 64K Sequential read & write

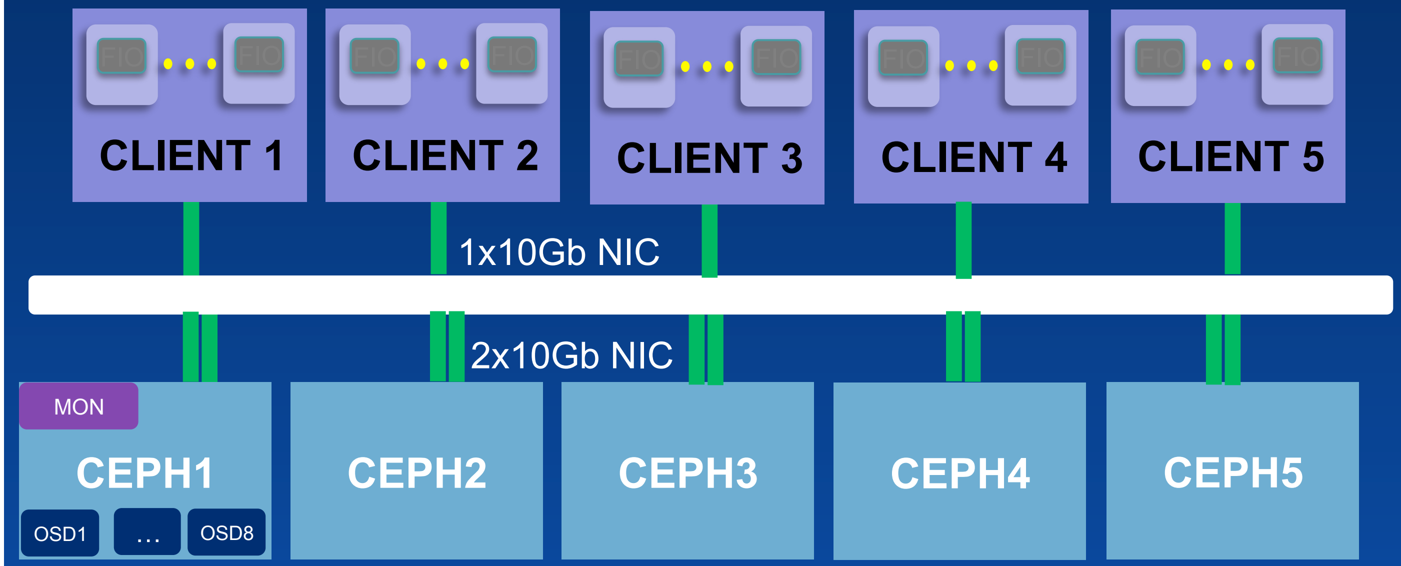
5x Client Node
Intel® Xeon® processor E5-2699 v3 @ 2.3GHz, 64GB memory
10Gb NIC

5x Storage Node
Intel Xeon processor E5-2699 v3 @ 2.3GHz, 64GB memory
1x Intel® DC P3700 800G SSD for Journal (U.2)
4x 1.6TB Intel® SSD DC S3510 as data drive
2 OSDs on single S3510 SSD

Flash Memory Summit

# Ceph* All Flash 3D NAND configuration - HSW (E5 -2699 v3) + P3700 + P3520

## Test Environment



Workloads

- Fio with librbd
- 20x 30 GB volumes each client
- 4 test cases: 4K random read & write; 64K Sequential read & write

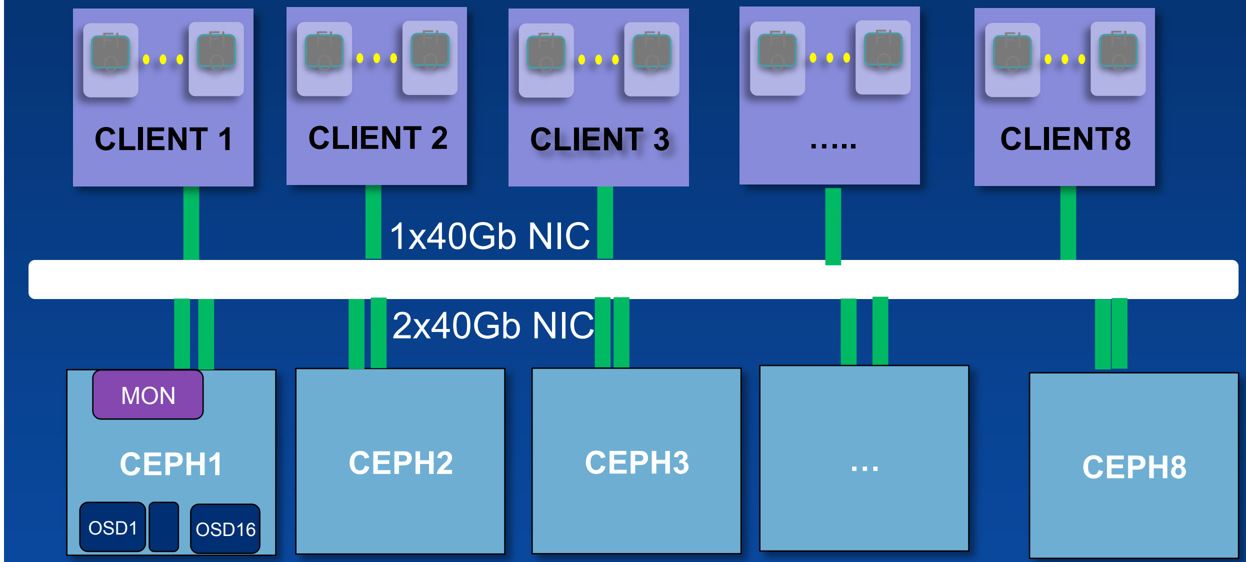
5x Client Node
Intel® Xeon™ processor E5-2699 v3 @ 2.3GHz, 64GB mem
10Gb NIC

5x Storage Node
Intel Xeon processor E5-2699 v3 @ 2.3 GHz
128GB Memory
1x 400G SSD for OS
1x Intel® DC P3700 800G SSD for journal (U.2)
4x 2.0TB Intel® SSD DC P3520 as data drive
2 OSD instances one each P3520 SSD

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

# Ceph* All Flash Optane configuration
# - BDW (E5-2699 v4) + Optane + P4500

## Test Environment



Workloads
- Fio with librbd
- 20x 30 GB volumes each client
- 4 test cases: 4K random read & write; 64K Sequential read & write

8x Client Node
- Intel® Xeon™ processor E5-2699 v4 @ 2.3GHz, 64GB mem
- 1x X710 40Gb NIC

8x Storage Node
- Intel Xeon processor E5-2699 v4 @ 2.3 GHz
- 256GB Memory
- 1x 400G SSD for OS
- 1x Intel® DC P4800 375G SSD as WAL and rocksdb
- 8x 2.0TB Intel® SSD DC P4500 as data drive
- 2 OSD instances one each P4500 SSD
- Ceph 12.0.0 with Ubuntu 14.01

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