Build High Performance, Cost effective Ceph All Flash Array Software Defined Storage Solutions with New Non-Volatile Memory Technologies

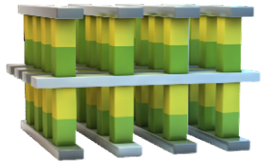
Jian Zhang, Software Engineer Manager, jian.zhang@intel.com
Brien Porter, Senior Program Manager, brien.porter@intel.com

Agenda

- Ceph* with Intel® Non-Volatile Memory Technologies
- Ceph AFA Reference Architectures Ceph* Performance analysis on Intel® Optane™ SSDs based all-flash array
- Bigdata Analytics on AFA
- Summary

Intel 3D NAND SSDs and OPTANE SSDs are Transforming Storage

Optimized Storage Performance

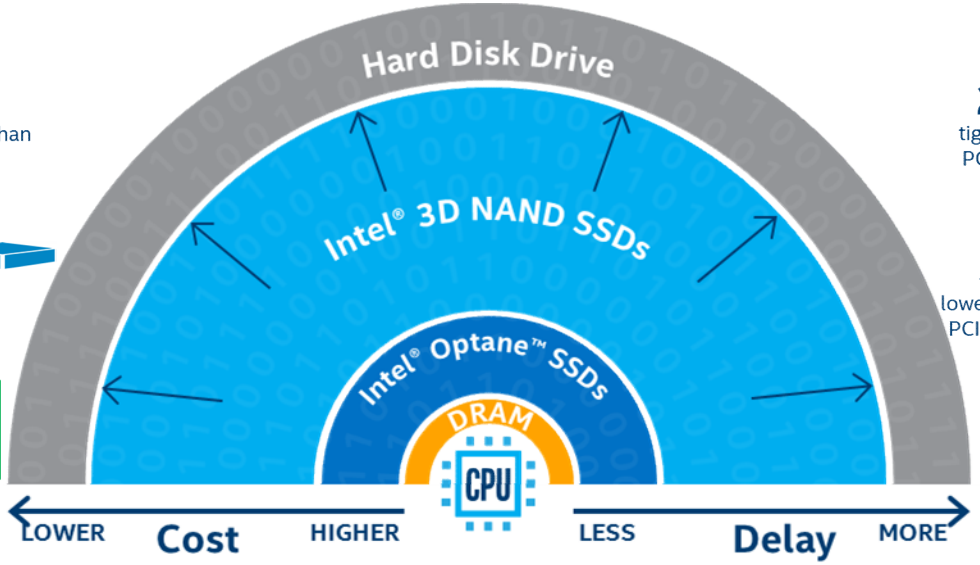


Up to **359x** more IOPS/$ than 10K HDD[6]

>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

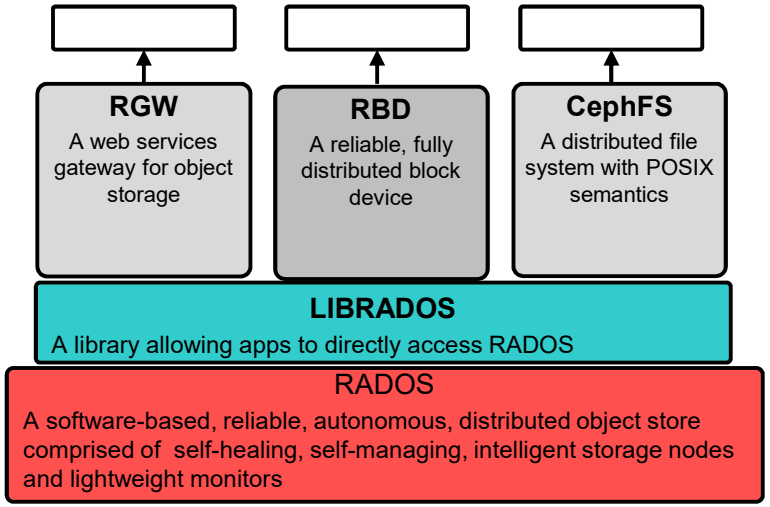
More VMs, Same QoS per rack

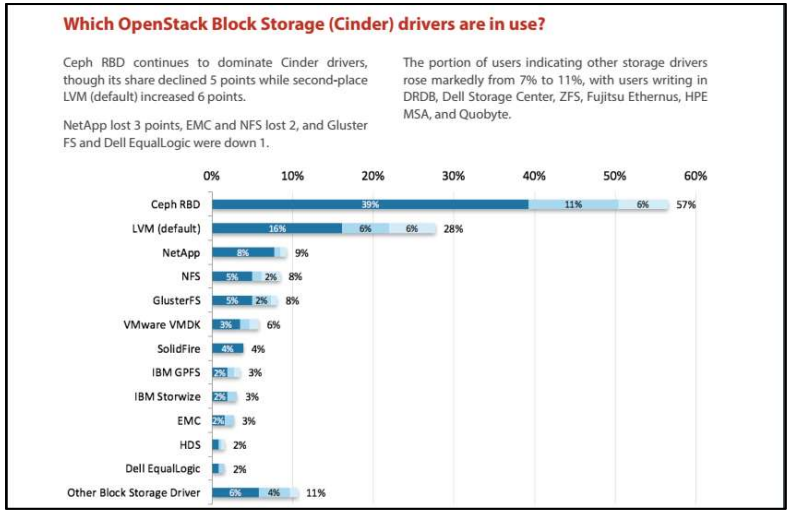
Capacity for Less

Performance for Less

Refer to appendix for footnotes

Ceph Introduction



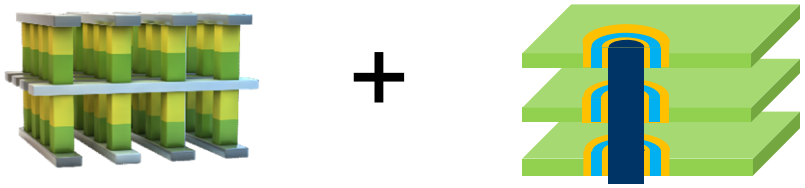


- Open-source, object-based scale-out storage
- Object, Block and File in single unified storage cluster
- Highly durable, available – replication, erasure coding
- Runs on economical commodity hardware
- Over 10 years of hardening, vibrant community

- 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

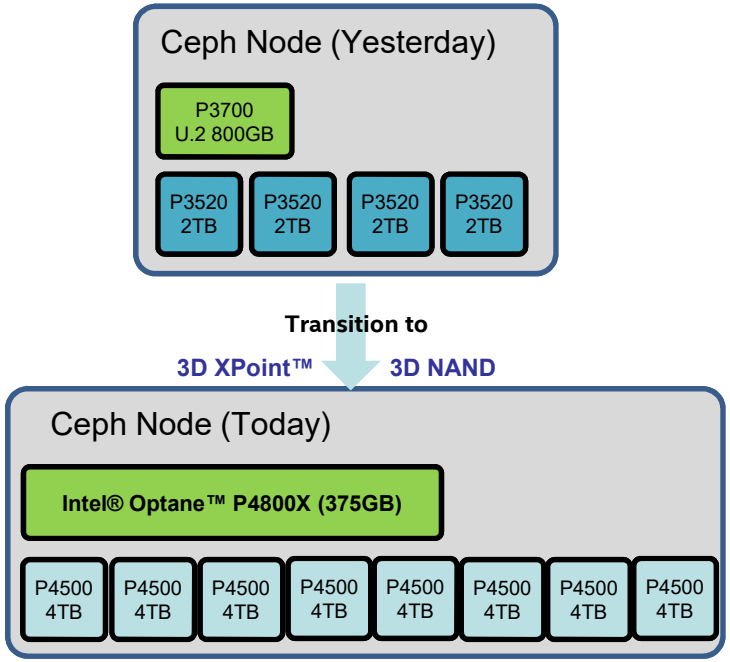
Innovation for Cloud Storage with Optane™ + Intel 3D NAND SSDs

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



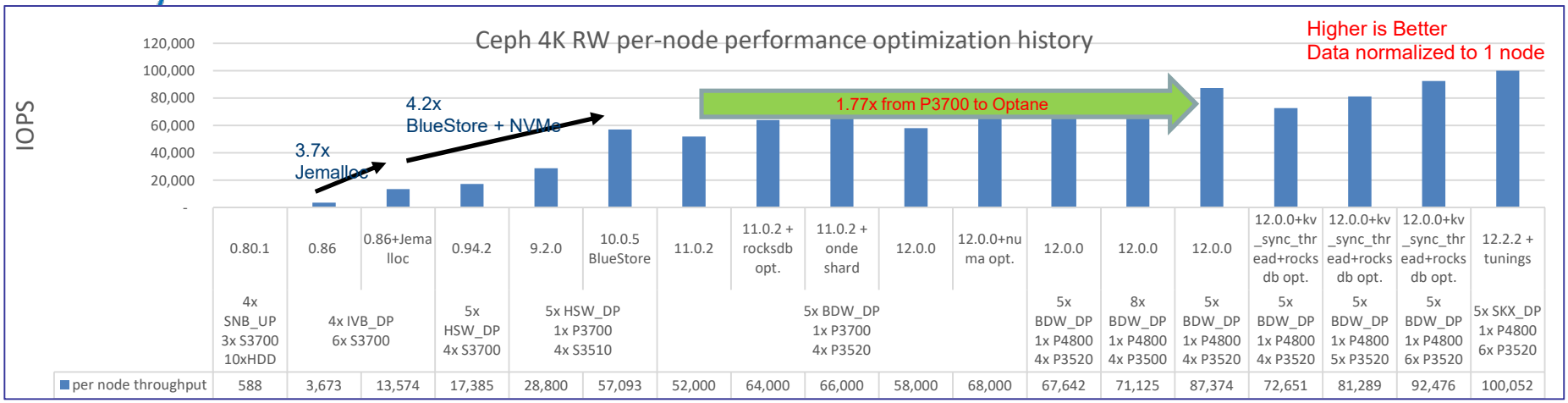
Journal/Log/Cache **Data**

- 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 NVMe based AFA performance

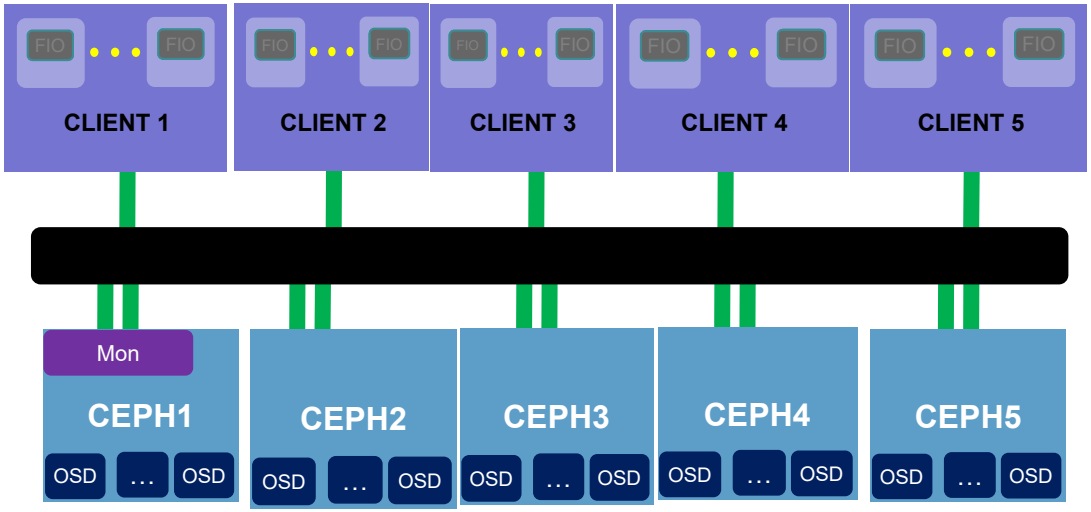
1.36x Ceph BlueStore Performance Improvement by introducing Optane!



- 1.31x performance improvement for P3700, 1.36x performance improvement for Optane
  - SW optimization demonstrated Optane performance advantages over P3700
- Still head rooms for performance improvement
  - 30% ~idle CPU, expect to get higher performance with more SSDs
- Performance refresh with 40Gb NICs
  - Performance with 40Gb NIC is the same as the performance with binding two 20Gb NICs

27x performance improvement in Ceph All Flash Array!

Ceph Cluster Configuration



Best configurations

Workloads
- Fio with librbd
- Ceph version 12.0.3(W/ two ky-sync-threads)/12.2.2
- 100x 30 GB volumes each client
- test case: 4K random write, 4k random read;64k sequential write, 64k sequential read

5x Client Node
- Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz / Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz, 384GB mem
- Mellanox 40G NIC

5x Storage Node
- Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz,384GB Memory
- 1x Intel Optane 375G SSD as db/wal drive
- 4x 2.0TB Intel® SSD DC P3520/2x 2.0TB Intel® SSD DC P3600 /4x 1.6TB Intel® SSD DC P3700/4x 2.0 TB Intel® SSD DC P3700 as data drive
- Mellanox 40G NIC

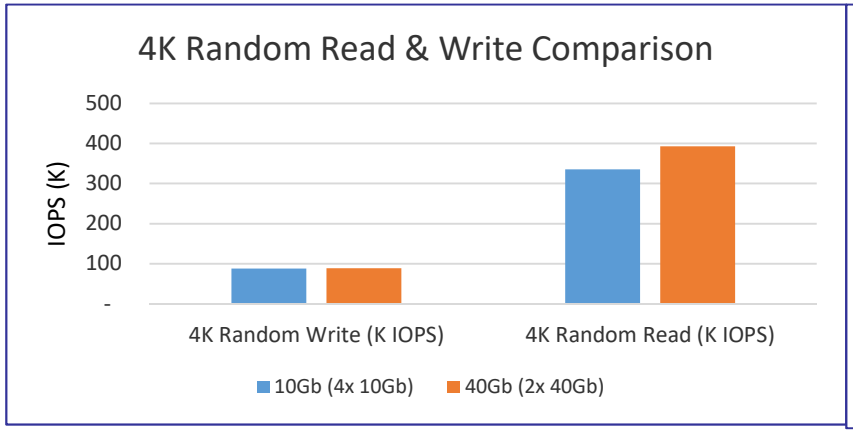
Ceph Storage Performance with OPTANE

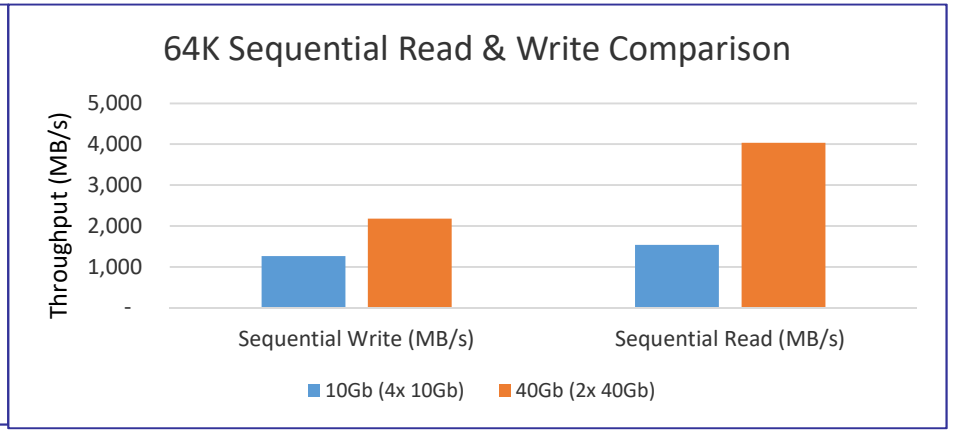
	Peak Performance	Avg. Latency (ms)	Avg. CPU %	IOPS/CPU
4K Random Write	500,259 IOPS	12.79	50	10005
4K Random Read	2,453,200 IOPS	5.36	60.87	40302
64K Sequential Read	21,949 MB/s	36.78	30.4	722
64K Sequential Write	8,714 MB/s	45.87	18.4	474

- Excellent performance on Optane cluster
  - 64K sequential read of these two configurations both hit the 40 GbE hardware limitation, bandwidth results of 64K sequential read are similar.
  - The 4K random read throughput is 2,453K IOPS with 5.36 ms average latency, while 4K random write throughput is 500K IOPS with 12.79 ms average latency.
    - No obvious bottleneck for random read & write, need future optimizations

Meltdown patch not applied

Ceph AFA Optimizations with 40Gb NICs

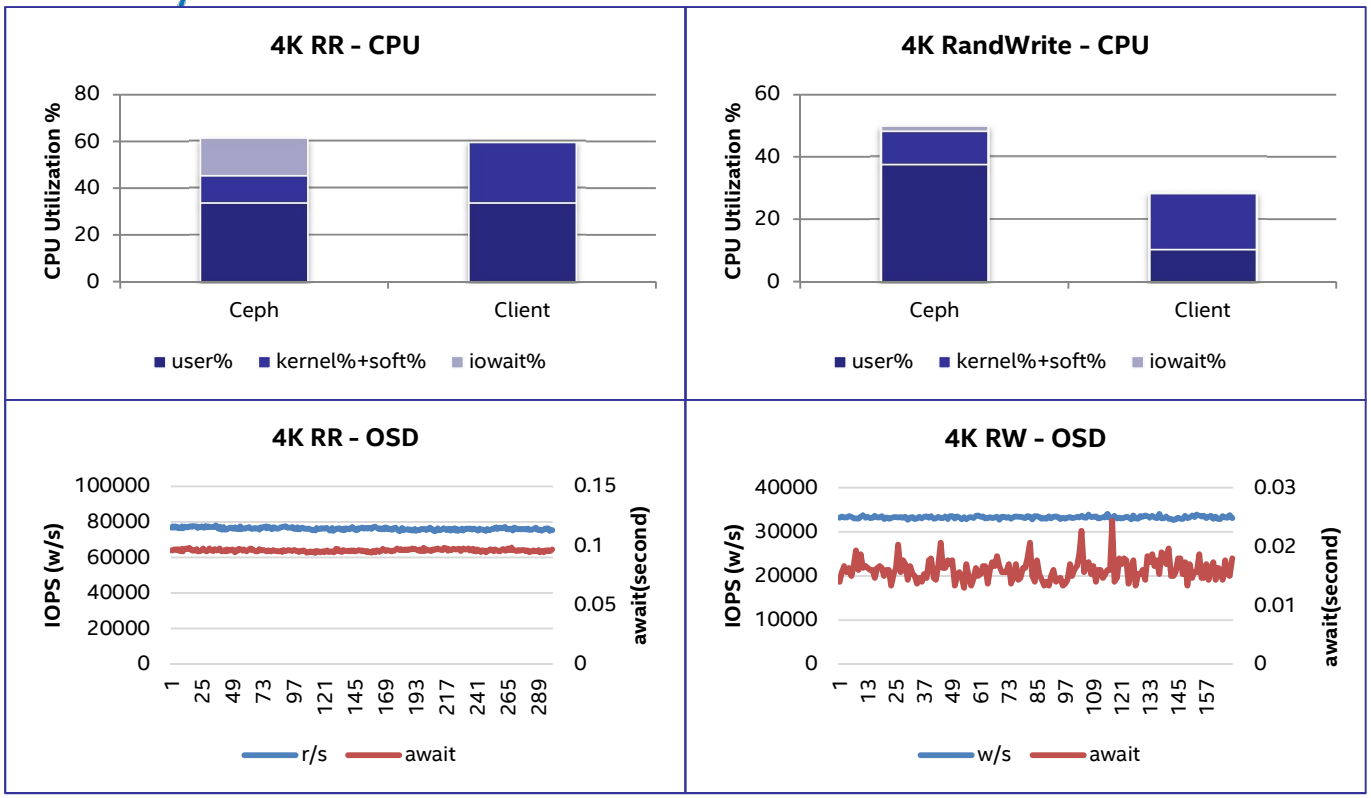




Performance improvement from 10Gb to 40Gb

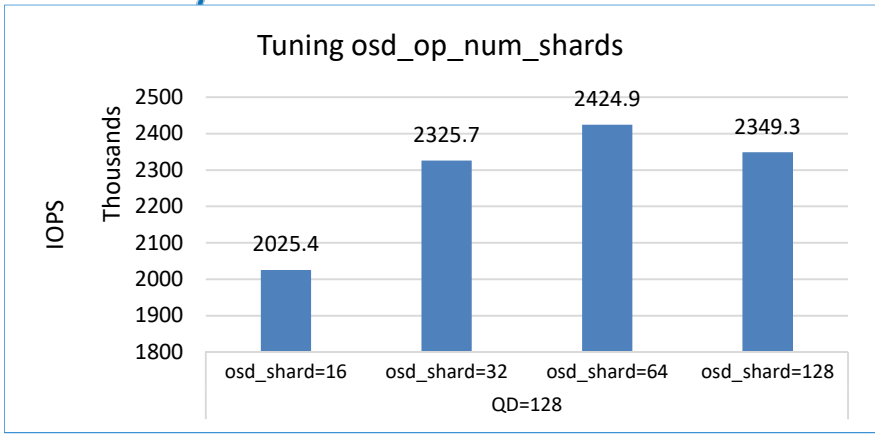
- 1.17x on 4K Random Read, 1.73x on Sequential Write, 2.63x on Sequential Read

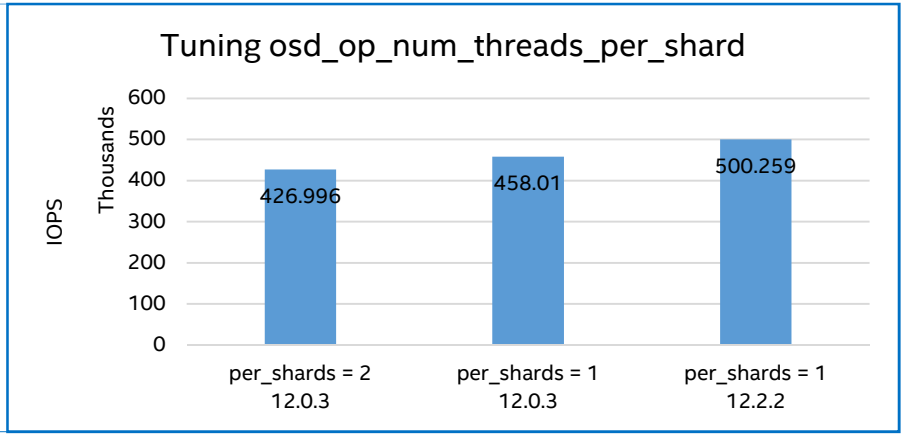
Ceph AFA system characterization



- No obvious bottlenecks
- Software stack requires further optimizations
- But where?

Performance Tuning and Optimizations

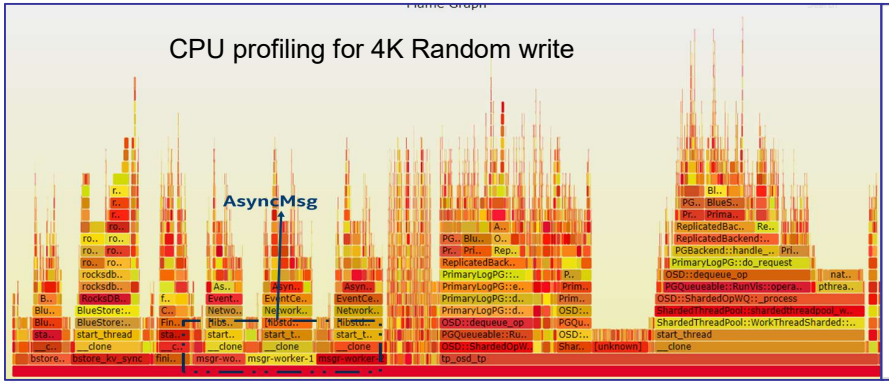


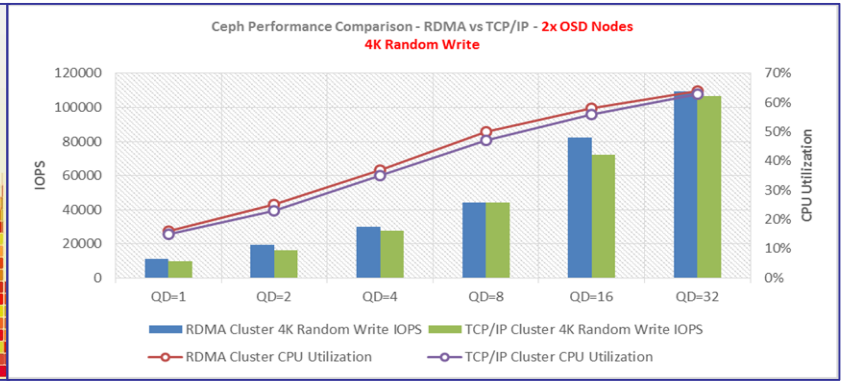


- 1.19x performance improvement after tuning osd_op_num_shards to 64,
- 1.17x performance improvement by tuning osd_op_num_threads_per_shard

Detail: Link

Looking Ahead: RDMA Optimizations

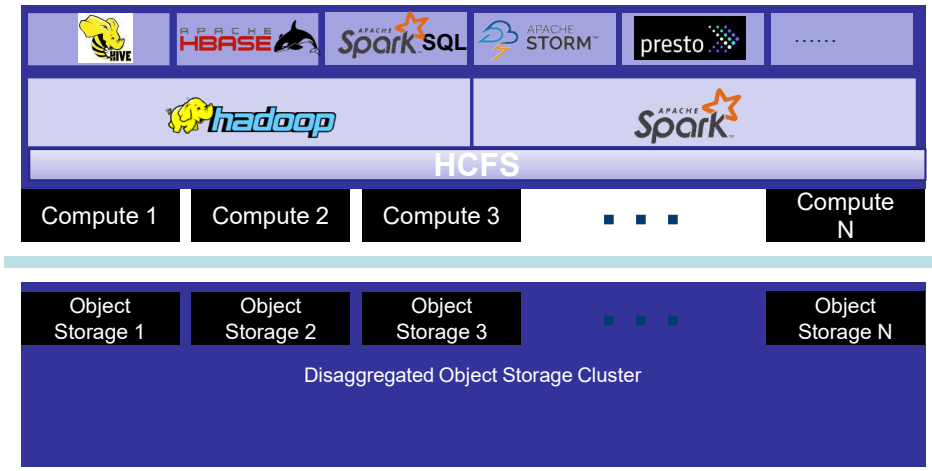




- Ceph networking layer consumes 20%+ CPU of the totally CPU used by Ceph in 4K random read workload.
- Ceph w/ iWARP delivers up to 17% 4K random write performance benefit than it w/ TCP/IP.
  - Patch was merged to upstream now.

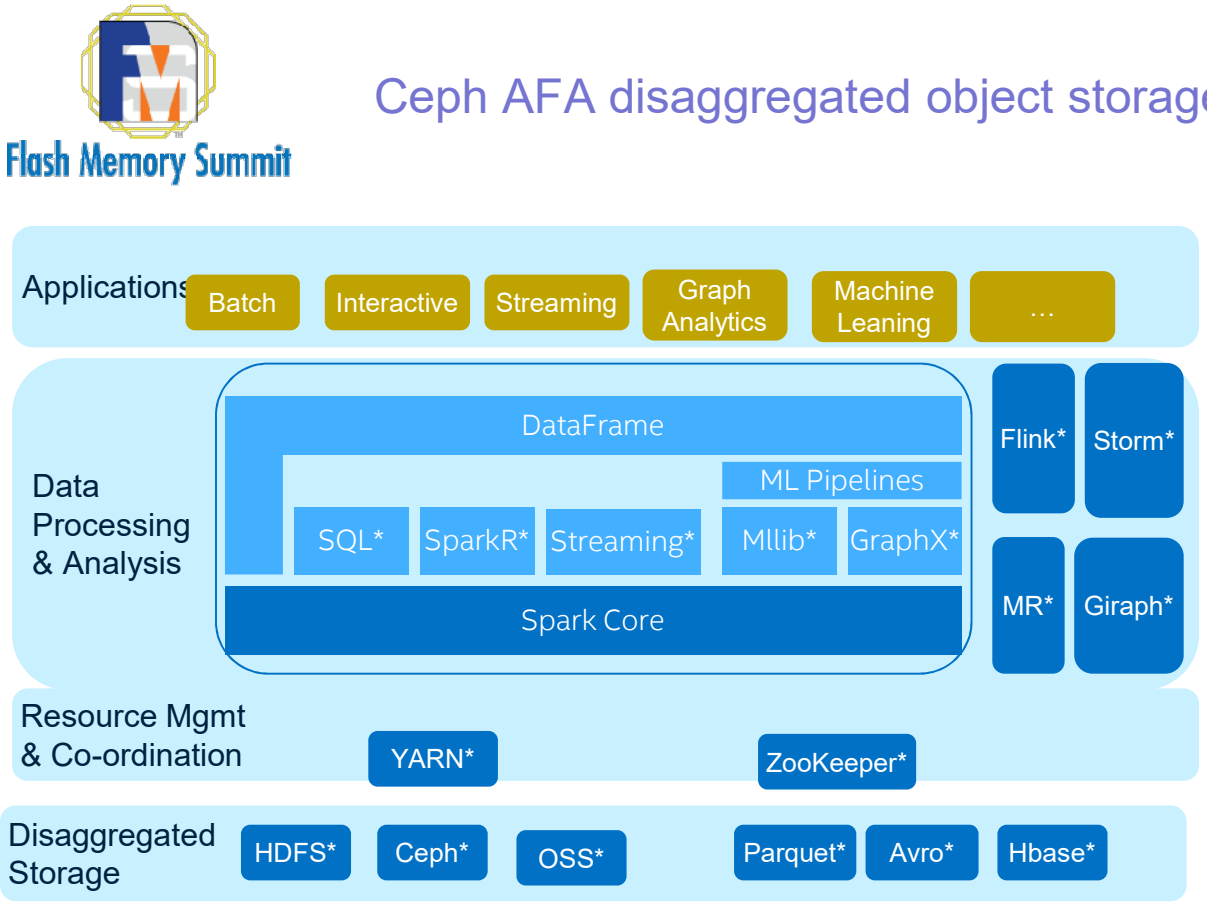
BigData Analytics on Ceph AFA

Disaggregating Object Storage



Hadoop Services
Virtual Machine
Container
Bare Metal

Object Storage Services
Co-located with gateway
Dynamic DNS or load balancer
Data protection via storage replication or erasure code
Storage tiering

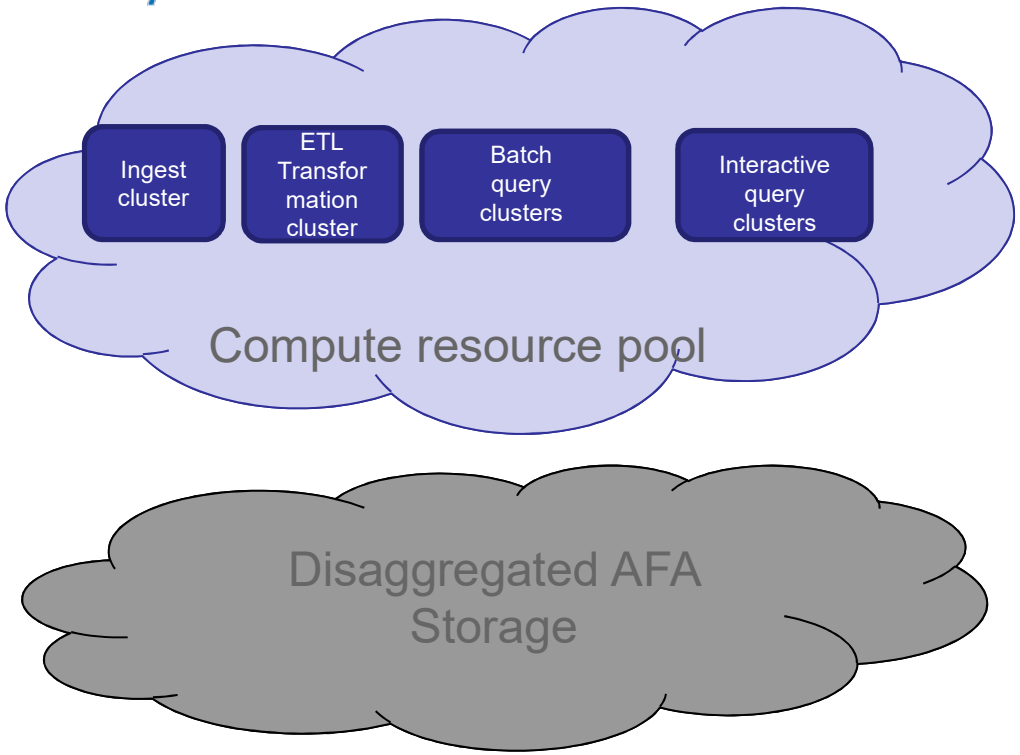


Ceph AFA disaggregated object storage for bigdata analytics

Hadoop Services orchestrations
- Virtual Machine
- Container
- Bare Metal

- Disaggregated Storage
- Remote HDFS based
- Object Storage systems
  - Co-located with gateway
  - Dynamic DNS or load balancer
  - Data protection via storage replication or erasure code
  - Storage tiering

BigData Usage Scenarios



Simple Read/Write

- DFSIO: TestDFSIO is the canonical example of a benchmark that attempts to measure the Storage's capacity for reading and writing bulk data.
- Terasort: a popular benchmark that measures the amount of time to sort one terabyte of randomly distributed data on a given computer system.

TPC-DS derived tests:

Batch ingestion

- Support collection of data from a variety of data sources in a consistent and repeatable manner designed to reduce data loss, improve traceability, increase availability, and increase timeliness.

Data Transformation

- ETL: Taking data as it is originally generated and transforming it to a format (Parquet, ORC) that more tuned for analytical workloads.

Batch Analytics

- To consistently executing analytical process to process large set of data.
- Leveraging 54 derived from TPC-DS * queries with intensive reads across objects in different buckets

Interactive Query

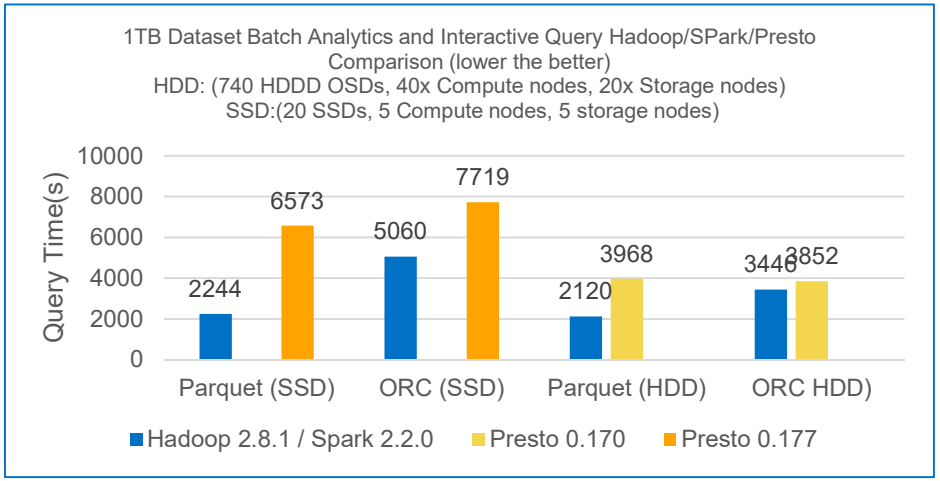
- This is very similar to the batch analytics workload, with the key distinction being required response time.

Streaming

- streaming data collection is the landing and aggregation of streaming data from messaging queues.

BigData on Object Storage Performance Overview

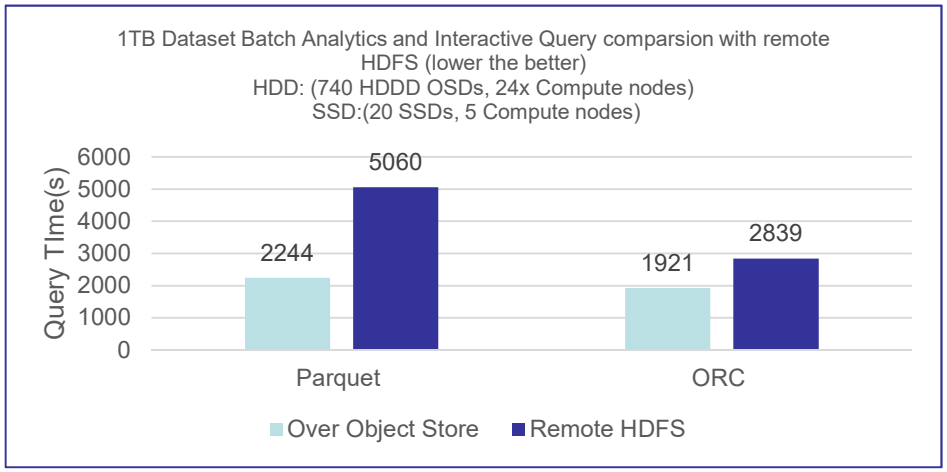
Batch Analytics



- Significant performance improvement from Hadoop 2.7.3/Spark 2.1.1 to Hadoop 2.8.1/Spark 2.2.0 (improvement in s3a)
- Batch analytics performance of 10-node Intel AFA is almost on-par with 60-node HDD cluster

Bigdata on Cloud vs. Remote HDFS

Batch Analytics



- On-par performance compared with remote HDFS
  - With optimizations, bigdata analytics on object storage is onpar with remote, especially on parquet format data
  - performance of s3a driver close to native dfsclient , and demonstrate compute and storage separate solution has a considerable performance compare with combination solution

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!
  - Significantly improved RDMA performance
  - RMDA performance is 24.7% higher than TCP/IP with optimizations (was 30% lower, based on softiwrap simulations).
- Bigdata analytics over disaggregated AFA storage demonstrated same functionality and close performance with HDFS solutions

…and in the next year QLC SSDs change the way Ceph users look at AFAs

Acknowledgement

- Thanks to the following Intel contributors: Yuan Zhou, Chendi Xue, Haodong Tang
- Bigdata analytics on Ceph data lake is an innovative solution co-developed by Intel Redhat, Quanta Cloud Technology (QCT)
- Special thanks to Kyle Bader, Karan Singh @Redhat, Marco Huang @ QCT for HDD based results

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_compactions=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_bytes_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
  -

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.