OPTIMIZING CEPH PERFORMANCE BY LEVERAGING INTEL® OPTANE™ AND 3D NAND TLC SSDS

Yuan Zhou, Software Engineer, yuan.zhou@intel.com

Jack Zhang, Storage Solution Architect, yuan.zhang@intel.com

Jun, 2017

Agenda

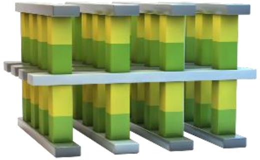
- Ceph* with Intel® Non-Volatile Memory Technologies
- Ceph* @ PRC
- 2.8M IOPS Ceph* cluster with Intel® Optane™ SSDs + Intel® 3D TLC SSDs
- Ceph* Performance analysis on Intel® Optane™ SSDs based all-flash array
- Summary

CEPH WITH INTEL® OPTANE™ AND INTEL® 3D TLC SSDS

INTEL® 3D NAND SSDS AND OPTANE SSD TRANSFORM STORAGE

Expand the reach of Intel® SSDs. Deliver disruptive value to the data center.

Optimized STORAGE Solutions



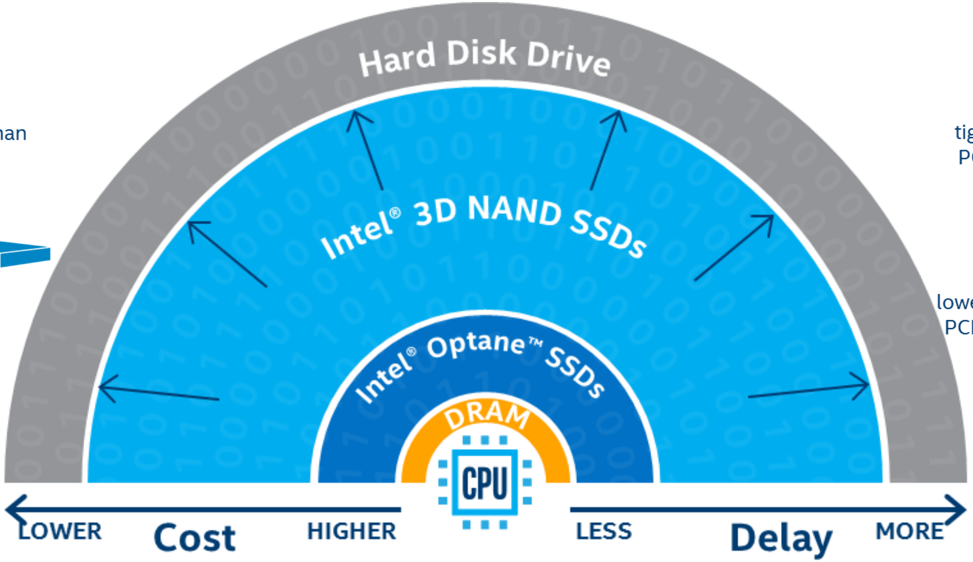
Up to
359x
more IOPS/$ than 10K HDD[6]

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

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

More capacity
per rack unit[11]

Capacity
for Less



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

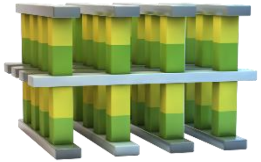
More VMs, Same QoS
per rack

Performance
for Less

Refer to appendix for footnotes

INNOVATION FOR CLOUD STORAGE: Intel® Optane™ + Intel® 3D NAND SSDs

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



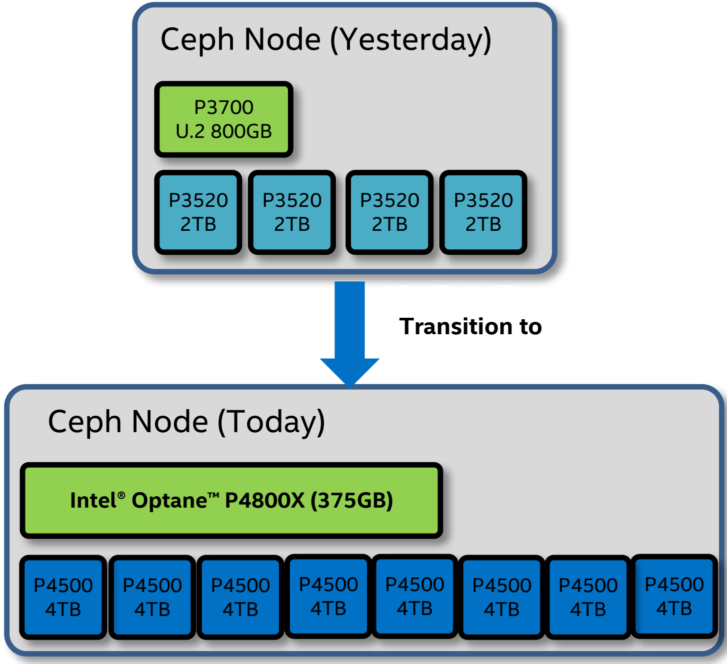
Journal/Log/Cache

+

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**



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**

More information at Ceph.com (new RAs update soon!)
http://tracker.ceph.com/projects/ceph/wiki/Tuning_for_All_Flash_Deployments

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) 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 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

Who is using Ceph?

Telcom

CSP/IPDC

OEM/ODM

Enterprise, FSI, Healthcare, Retailers

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

Ceph* @ PRC

Ceph* is very hot in PRC

Redevelopment based on the upstream code

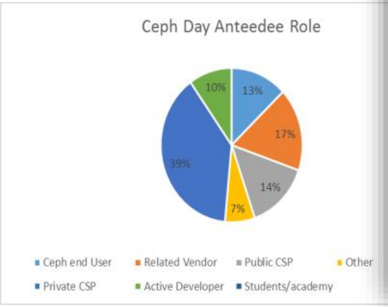
- More and more companies move to OpenSource storage solutions

Intel/Redhat held Three Ceph days at Beijing and Shanghai

- 1000+ attendees from 500+ companies
- Self media-reports and re-clippings
- Next one: Ceph Day Beijing on June, 10th

More and more PRC code contributors

- ZTE, XSKY, H3C, LETV, UnitedStack, AliYun, Ebay, EasyStack



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

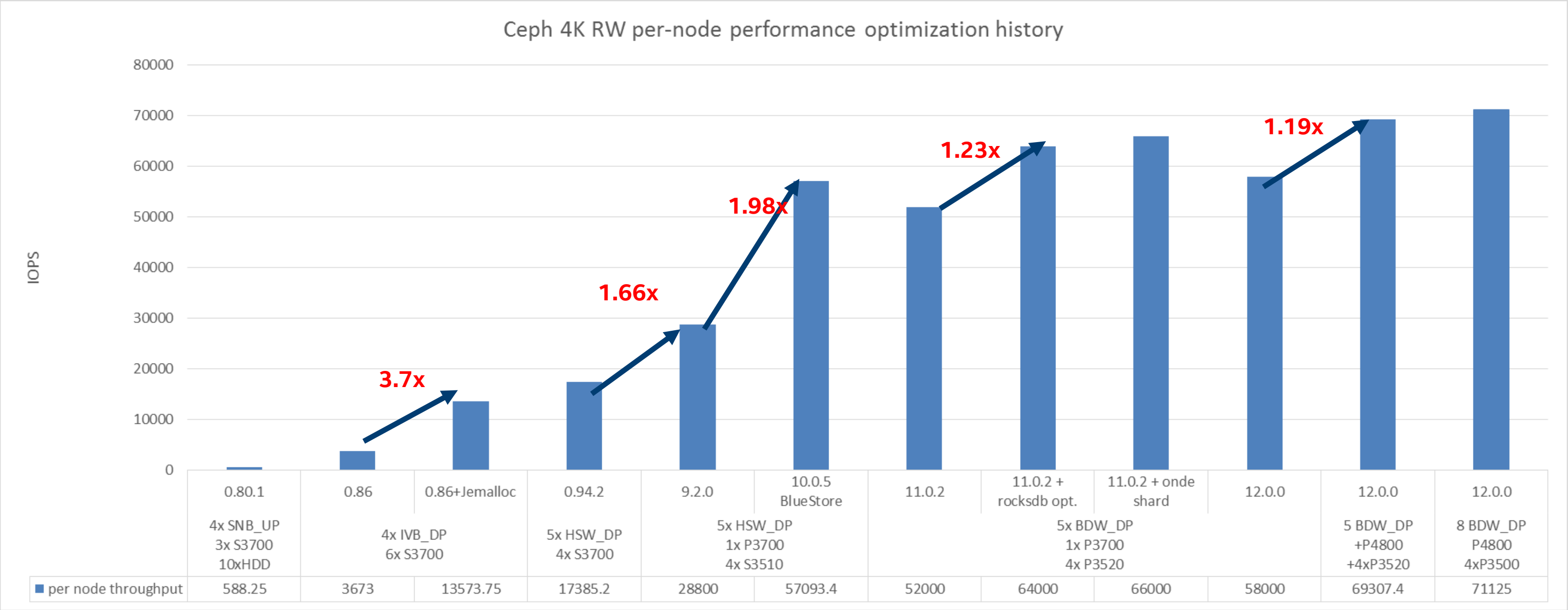
Strong demands to run enterprise applications

- OLTP workloads running on Ceph, tail latency is critical
- high performance multi-purpose Ceph cluster is a key advantage
- Performance is still an important factor

SSD price continue to decrease

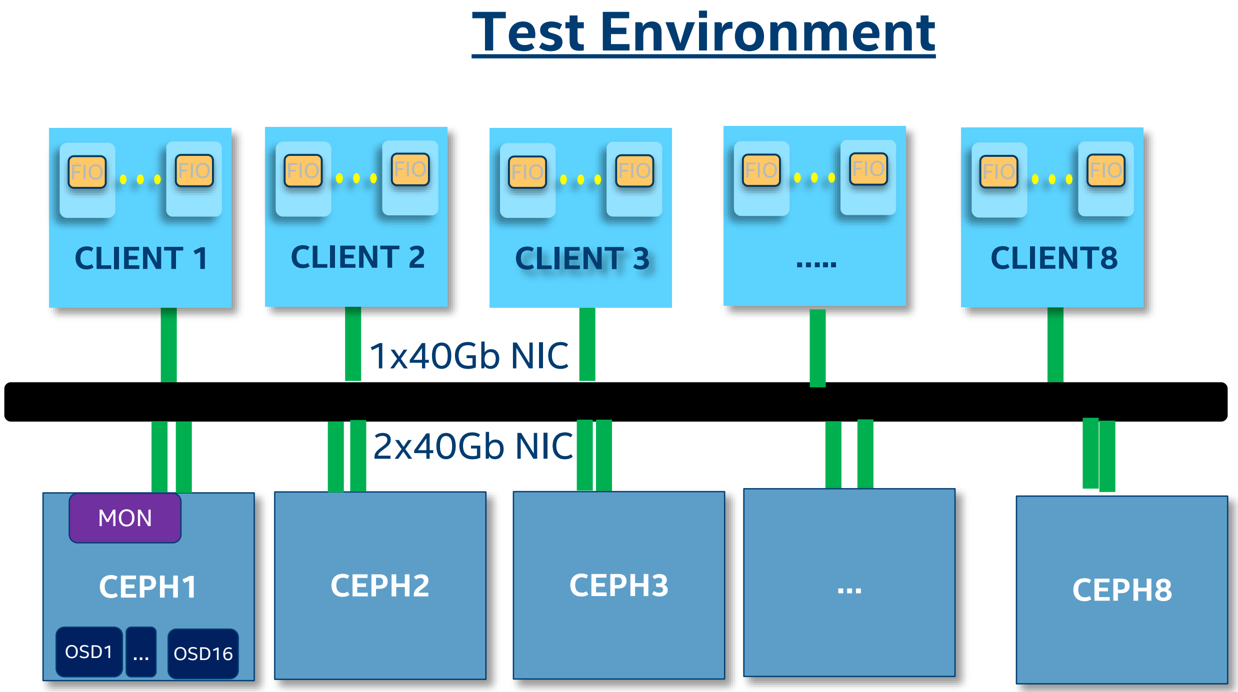
CEPH BLUESTORE PERFORMANCE ON INTEL® OPTANE™ BASED ALL-FLASH ARRAY

Ceph* performance trend with SSD – 4K Random Write



38x performance improvement in Ceph all-flash array!

Ceph* All Flash Optane configuration



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

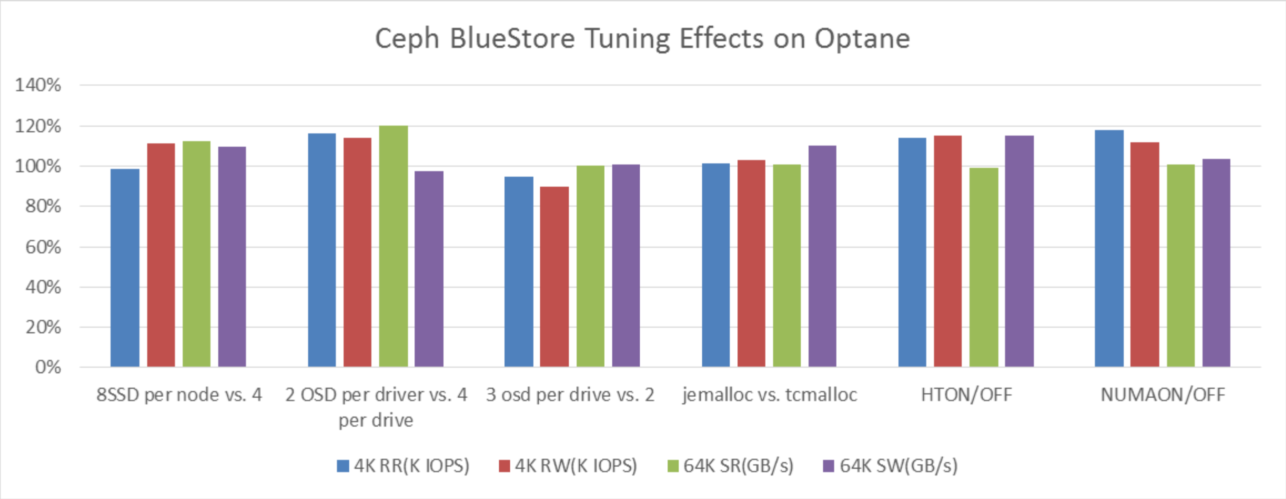
8x Storage Node
- Intel Xeon processor E5-2699 v4 @ 2.3 GHz
- 256GB Memory
- 1x 400G SSD for OS
- 1x Intel® DC P4800 375G SSD as WAL and DB
- 8x 2.0TB Intel® SSD DC P4500 as data drive
- 2 OSD instances one each P4500 SSD
- Ceph 12.0.0 with Ubuntu 16.10

Ceph* Optane Performance overview

	Throughput	Latency (avg.)	99.99% latency
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

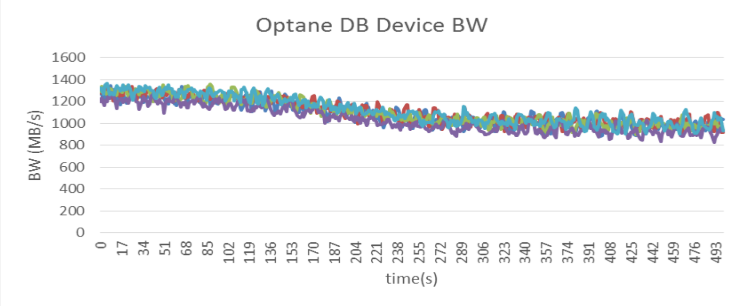
- Excellent performance on Optane cluster
  - random read & write hit CPU bottleneck

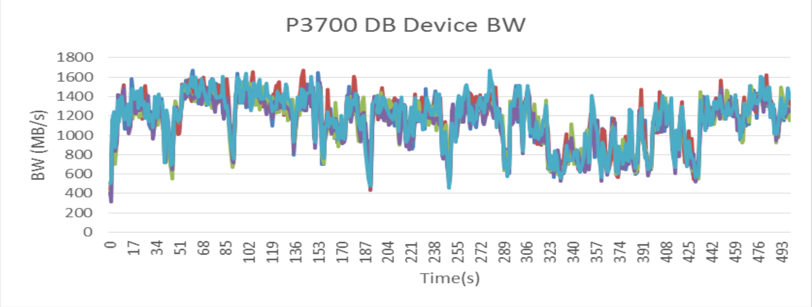
Ceph* Optane Performance – Tunings

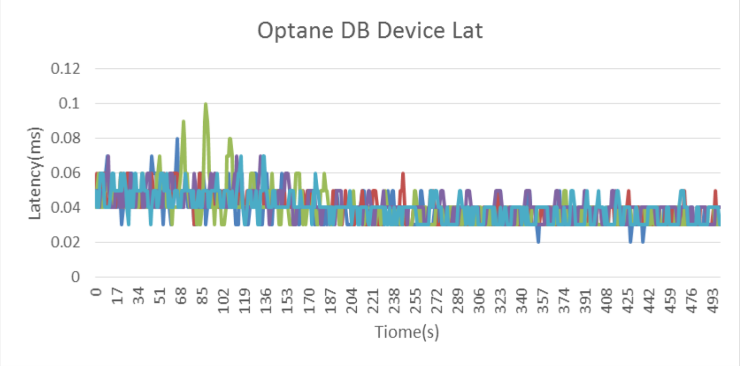


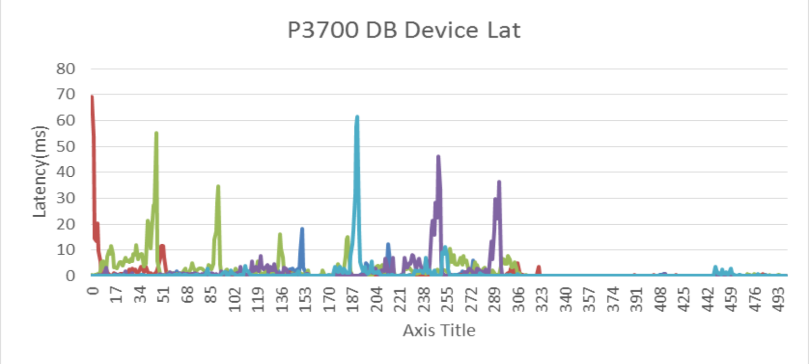
- Good Node Scalability, poor disk scalability for 4K block workloads (CPU throttled!)
- NUMA & HT helps a lot on the Performance
- Fine tune the # of OSD per node and Drive per Node.

Ceph* Optane Performance – Performance improvement



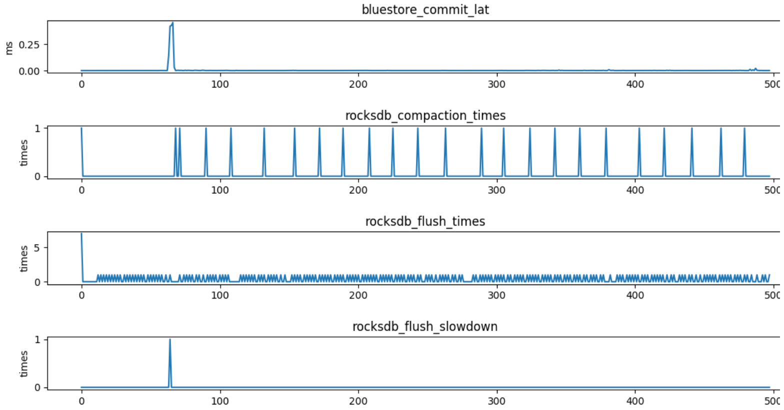


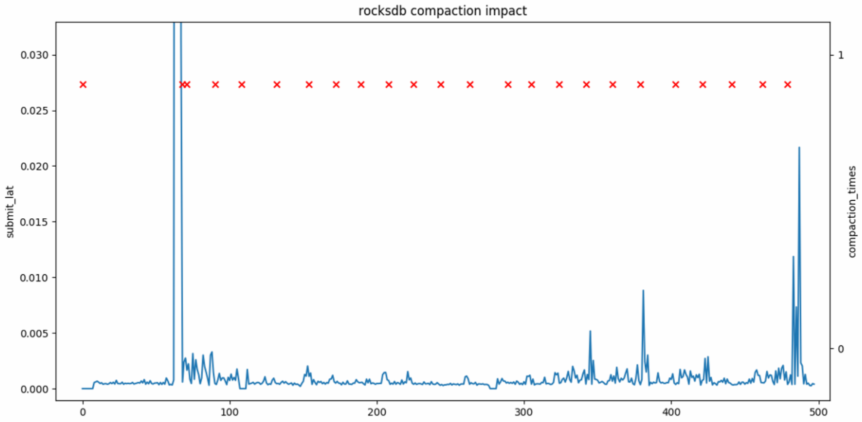




- The breakthrough high performance of Optane eliminated the WAL & rocksdb bottleneck
  - 1 P4800X or P3700 covers up to 8x P4500 data drivers as both WAL and rocksdb

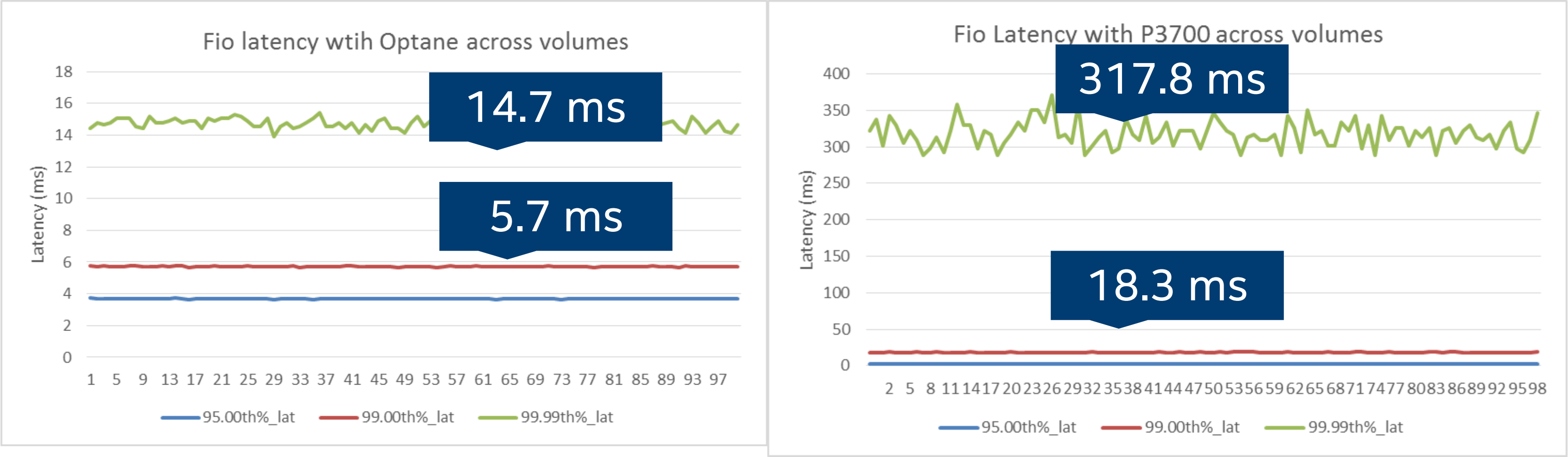
Ceph* Optane Performance – Rocksdb improvement





- Eliminate rocksdb write stall with Optane drive
  - Added event listener in rocksdb to provide real-time compaction&&flush information
  - Write stalls when flush or compaction can't keep up with the incoming write rate
  - rocksdb_compaction matches submit_transaction latency increase points

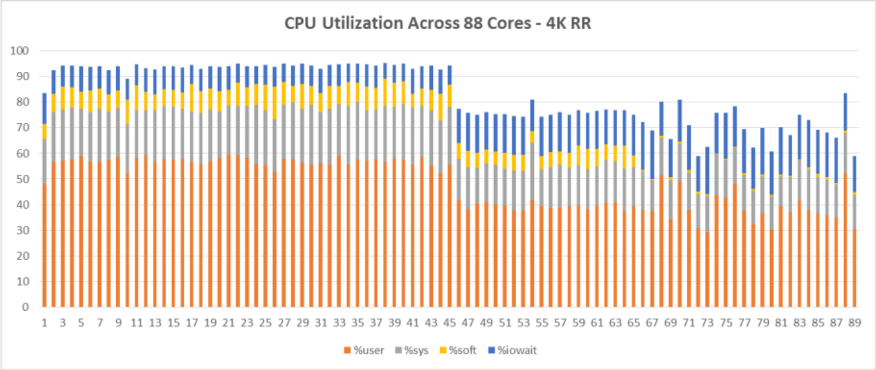
Ceph* Optane Performance – latency improvement

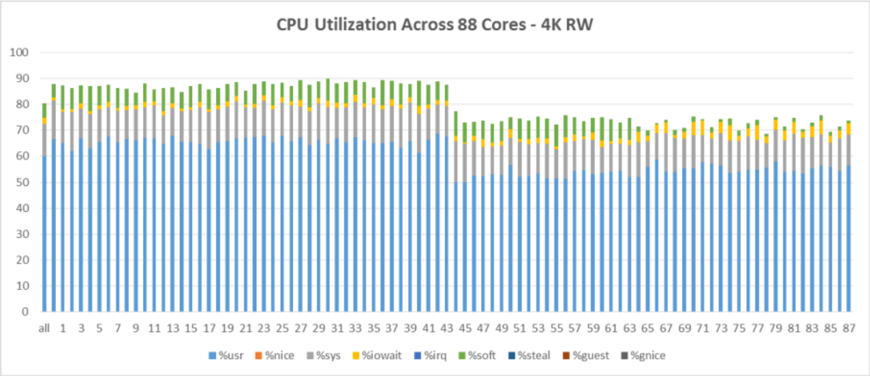


- Significant tail latency improvement with Optane
  - 20x latency reduction for 99.99% latency

CEPH PERFORMANCE ANALYSIS ON INTEL® OPTANE™ SSDS BASED ALL-FLASH ARRAY

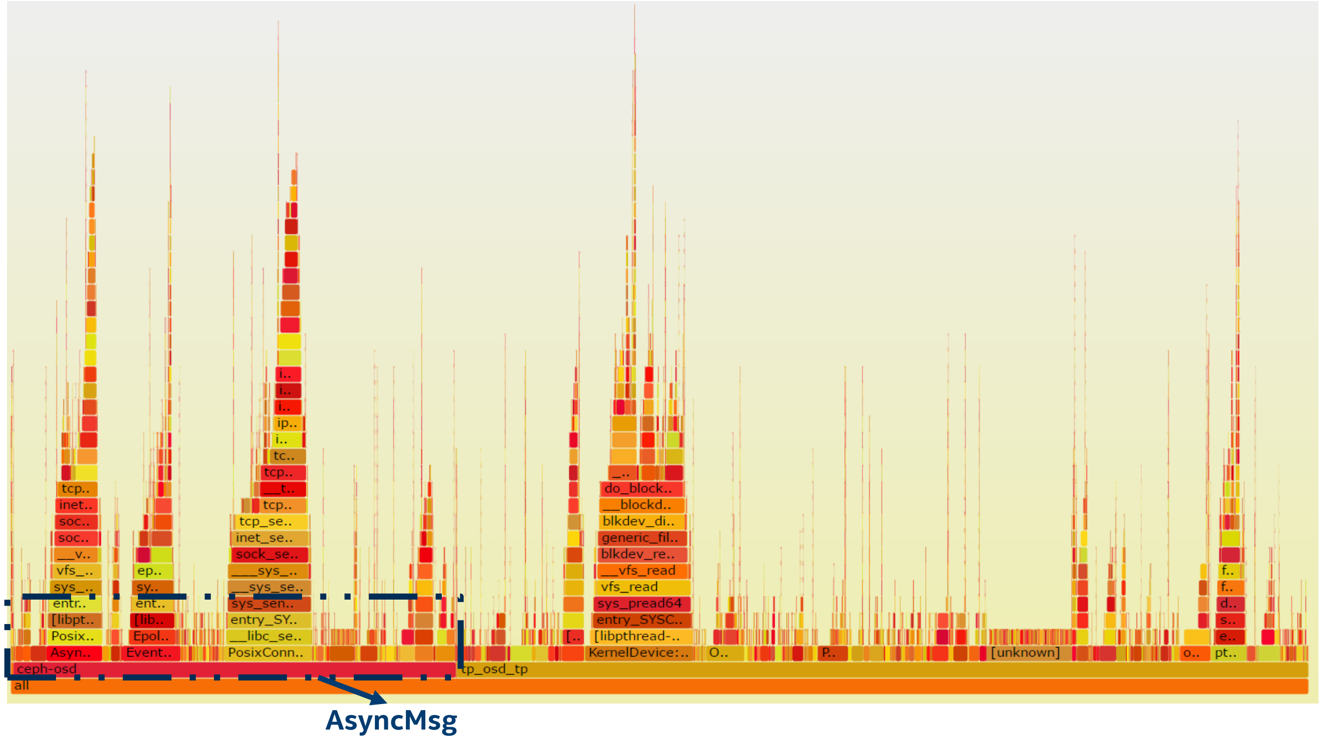
Ceph* Optane Performance Analysis - CPU utilizations





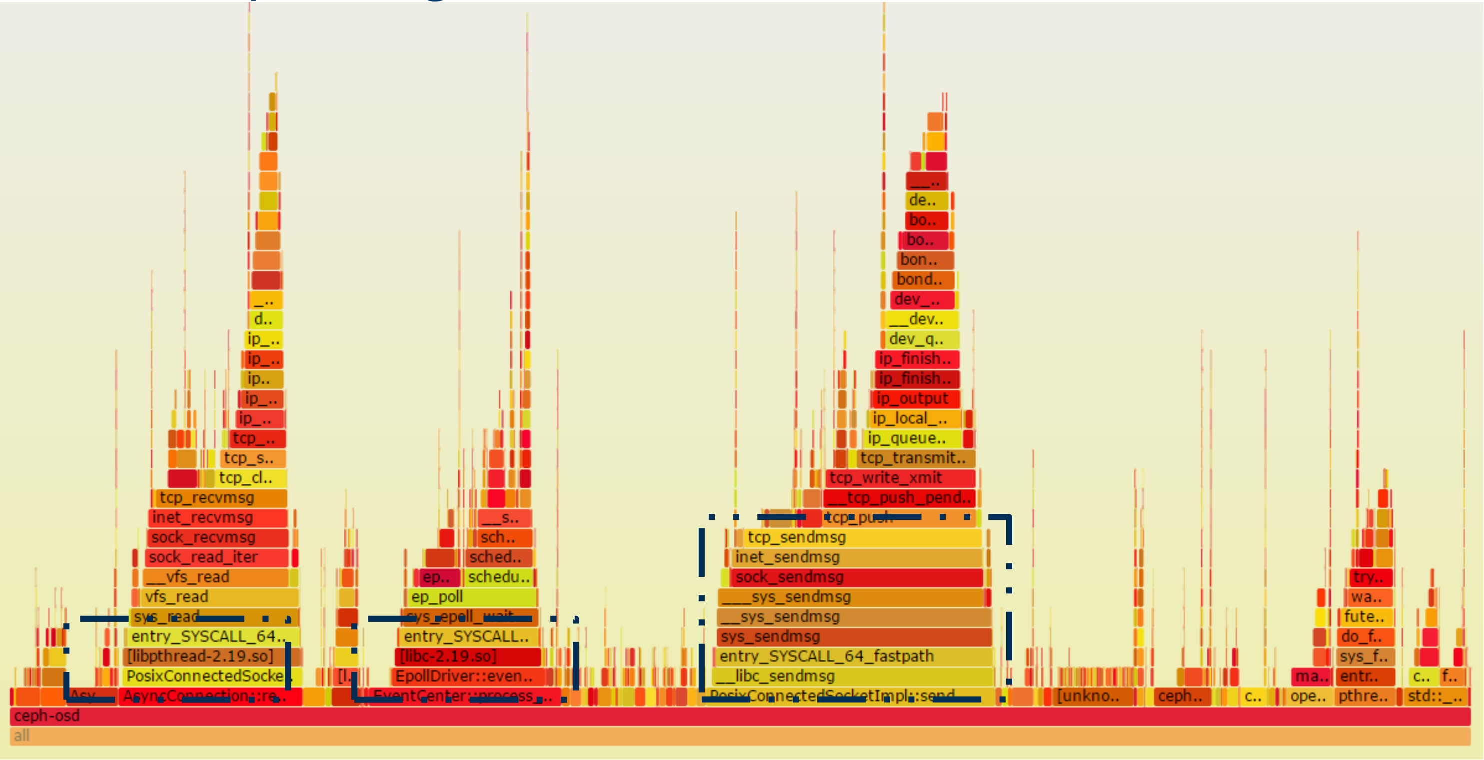
- Random Read & Write performance are throttled by CPU
  - Unbalanced CPU utilization caused by HT efficiency for random workloads
  - Limiting # of drive scalability for small block random workloads
- Need to optimize CPU utilization

Ceph* Optane Performance Analysis
- CPU profiling for 4K RR

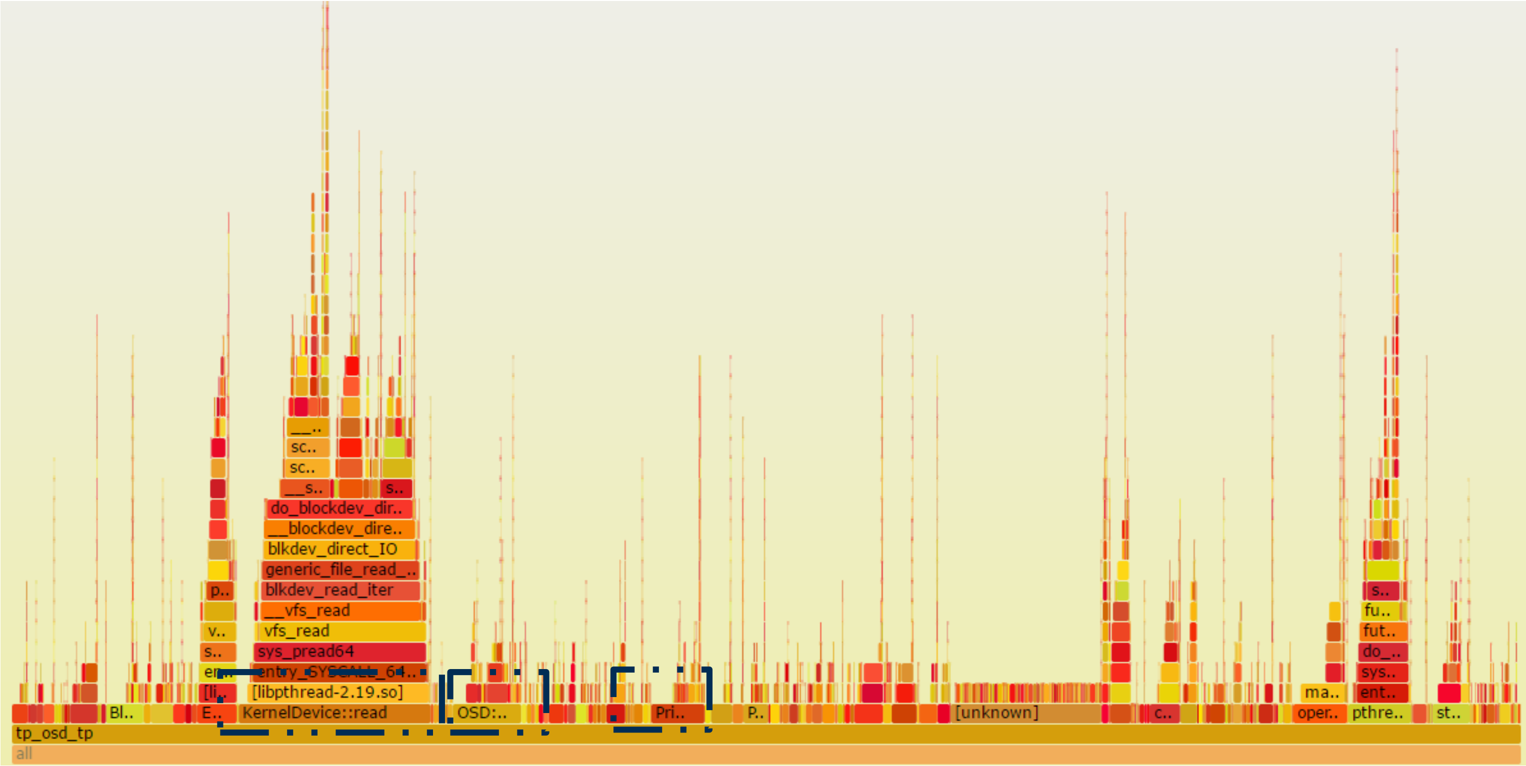


- Perf record for 30 second
- Ceph-osd: 34.1%, tp_osd_tp: 65.6%
- Heavy network messenger overhead

Ceph* Optane Performance Analysis - CPU profiling for 4K RR



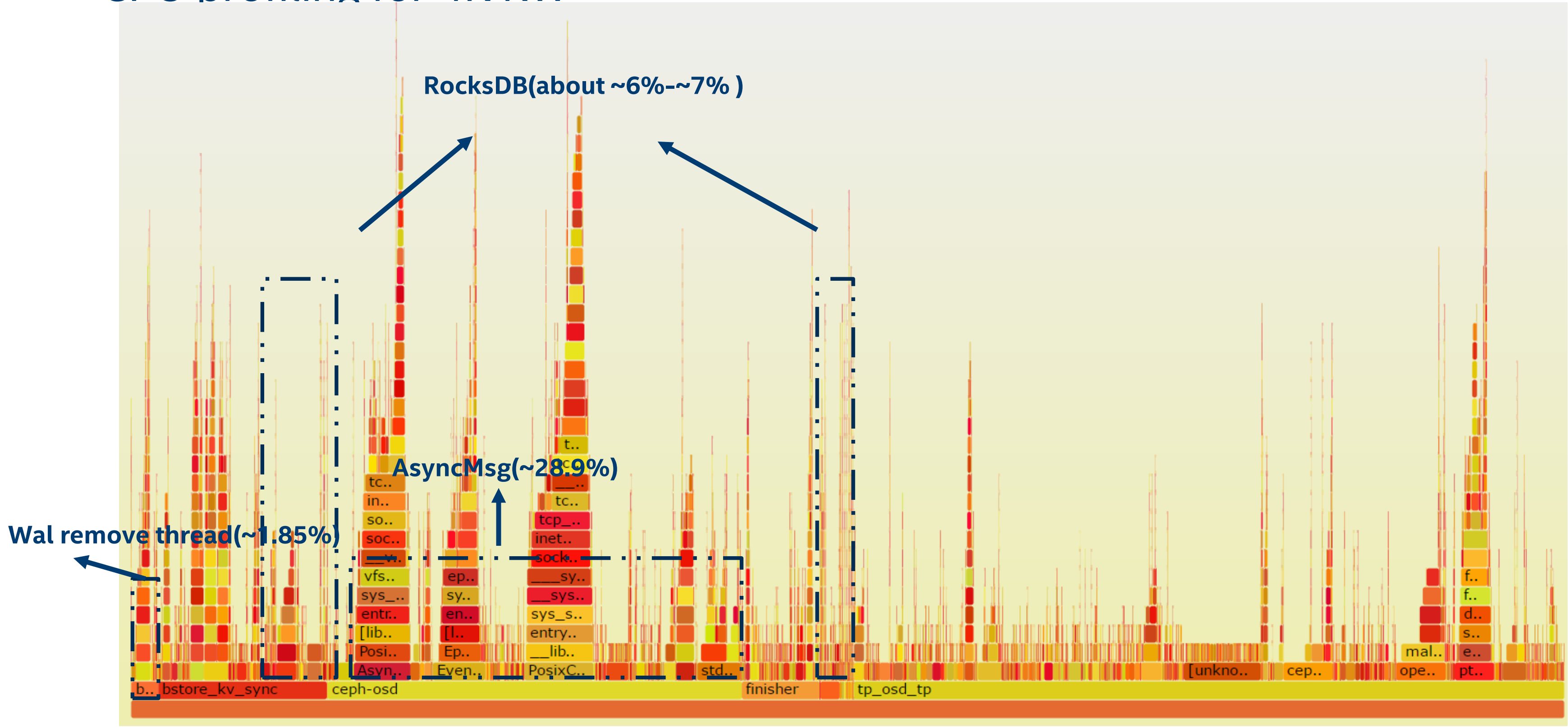
Ceph* Optane Performance Analysis
- CPU profiling for 4K RR



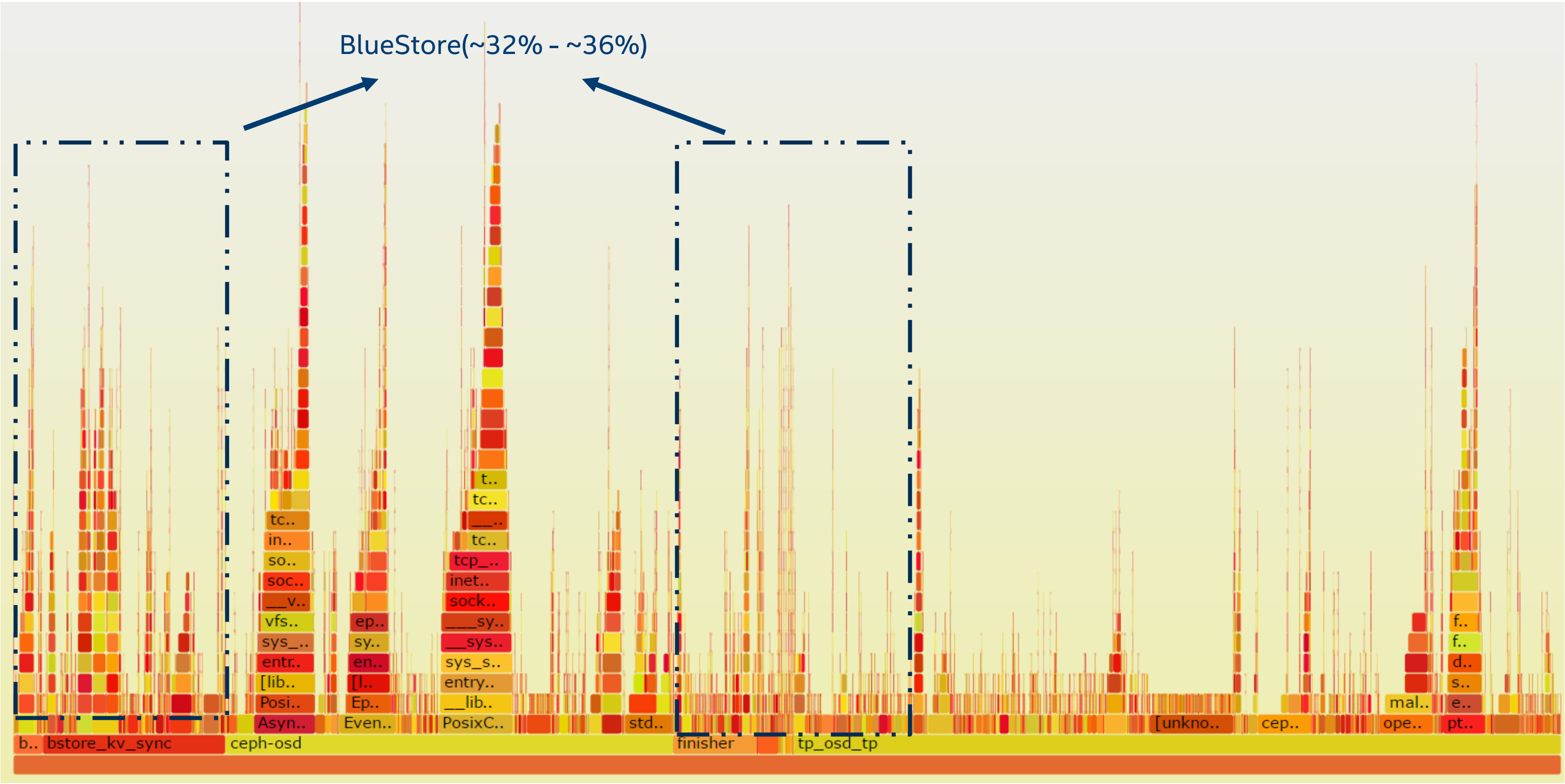
- The top three consumers of tp_osd_tp thread are
- KernelDevice::read,
- OSD::ShardedOpWQ::_process
- PrimaryLogPG::do_op.

- Perf record 30s.

Ceph* Optane Performance Analysis
- CPU profiling for 4K RW



Ceph* Optane Performance Analysis
- CPU profiling for 4K RW

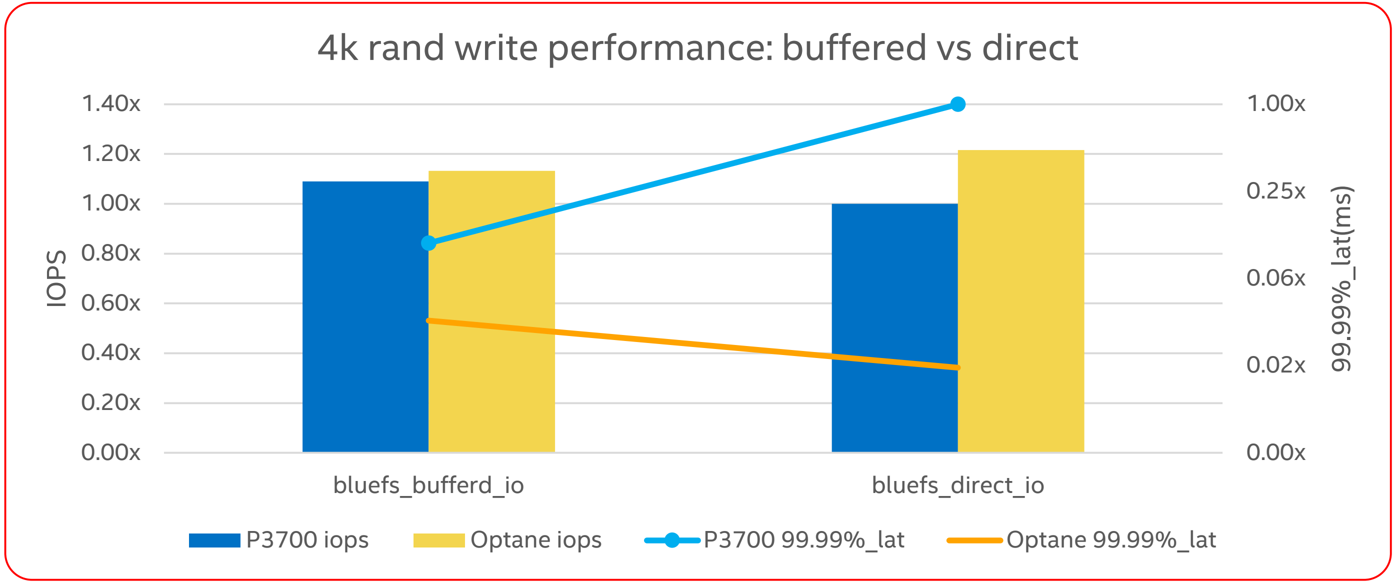


Ceph* Optane Performance Analysis - WAL tunings and optimizations

Tunings	4K RW IOPS	comments
Default: Baseline(NVMe as DB&& WAL drive)	340000	Separated DB&&WAL device
Tuning1: DB on NVMe && WAL on Ramdisk	360000	Move WAL to Ramdisk
Tuning2: Tuning1+disable rocksdb WAL	360000	RocksDB tuning
Tuning3: Tuning2+omit WAL in deferred write mode	410000	Don't write WAL in deferred write mode
Tuning4: Tuning1+write WAL but don't remove WAL from rocksdb	240000	Write WAL before write metadata into rocksdb, but will not clean WAL after write data to data device in deferred write mode
Tuning5: Tuning1+external WAL	380000	Write WAL to an external WAL device, and write its metadata to rocksdb

- Key takeaways:
  - rocksdb memTable overhead lead to 50000 IOPS difference. (tuning2 vs. tuning3)
  - If we don't clean bluestore WAL data in rocksdb, rocksdb overhead increase dramatically with the # of metadata increase:
  - Use an external WAL device to store WAL data, and just write WAL metadata into rocksdb!

Ceph* Optane Performance Analysis - Direct I/O vs Buffered I/O



- Key takeaways:
  - On P3700 setup, with bulefs_buffered_io, the performance improved 9%, 99.99% latency improved ~8x
  - However on Optane setup, we find bluefs_direct_io could improve the 10% performance, 99.99% latency also got improved ~1x

Ceph* Optane Performance Analysis - Metadata plane I/O pattern

DB

```
259,0   42      64     0.072669752 177116  A  WS 387830816 + 1032 <- (259,104) 1111072
259,0   86     134     0.075719488 177116  A  WS 387831840 + 1032 <- (259,104) 1112096
259,0   86     161     0.078582446 177116  A  WS 387832864 + 1040 <- (259,104) 1113120
259,0   24      64     0.081442440 177116  A  WS 387833896 + 1032 <- (259,104) 1114152
259,0   33      61     0.084122333 177116  A  WS 387834920 + 1032 <- (259,104) 1115176
259,0   77     103     0.086525275 177116  A  WS 387835944 + 1032 <- (259,104) 1116200
259,0   77     125     0.088835143 177116  A  WS 387836968 + 1040 <- (259,104) 1117224
259,0   77     147     0.089564736 177116  A  WS 387838000 + 432 <- (259,104) 1118256
259,0   77     158     0.092532258 177116  A   R 387838424 + 8 <- (259,104) 1118680
259,0   77     164     0.092561072 177116  A   R 387838128 + 8 <- (259,104) 1118384
259,0   77     170     0.092587360 177116  A   R 387838128 + 8 <- (259,104) 1118384
259,0   77     176     0.092693010 177116  A   R 387838128 + 304 <- (259,104) 1118384
259,0   77     186     0.093169176 177116  A  WS 387336240 + 8 <- (259,104) 616496
259,0   66     150     0.164727068 178265  A   R 387800744 + 16 <- (259,104) 1081000
259,0   76    3659     2.859572562 177118  A  WS 386732032 + 1032 <- (259,104) 12288
259,0   29    3039     2.864742636 177118  A  WS 386733056 + 1024 <- (259,104) 13312
259,0   29    3057     2.864780468 177118  A  WS 386736128 + 8 <- (259,104) 16384
259,0   39    3150     2.870420543 177118  A  WS 386736128 + 1032 <- (259,104) 16384
259,0   86    4688     2.876454314 177118  A  WS 386737152 + 1024 <- (259,104) 17408
259,0   86    4706     2.876482077 177118  A  WS 386740224 + 8 <- (259,104) 20480
259,0   86    4717     2.880833028 177118  A  WS 386740224 + 1032 <- (259,104) 20480
259,0   86    4744     2.885224128 177118  A  WS 386741248 + 1032 <- (259,104) 21504
259,0   74    4095     2.890598237 177118  A  WS 386742272 + 1032 <- (259,104) 22528
```

WAL

```
259,0   68    26    0.006772806 192607  A  WS 408268568 + 80 <- (259,105) 65304
259,0   68    37    0.007463665 192607  A  WS 408268640 + 112 <- (259,105) 65376
259,0   41    23    0.007970770 192607  A  WS 408268744 + 64 <- (259,105) 65480
259,0   41    29    0.008489614 192607  A  WS 408268800 + 64 <- (259,105) 65536
259,0   41    35    0.009196485 192607  A  WS 408268856 + 72 <- (259,105) 65592
259,0   41    41    0.009983880 192607  A  WS 408268920 + 88 <- (259,105) 65656
259,0   40    22    0.010737948 192607  A  WS 408269000 + 80 <- (259,105) 65736
259,0   36    33    0.011305476 192607  A  WS 408269072 + 64 <- (259,105) 65808
259,0   41    47    0.011651816 192607  A  WS 408269128 + 64 <- (259,105) 65864
259,0   85     9    0.012008326 192607  A  WS 408269184 + 56 <- (259,105) 65920
259,0   25     1    0.012191469 192607  A  WS 408269232 + 24 <- (259,105) 65968
259,0   25     7    0.012618032 192607  A  WS 408269248 + 24 <- (259,105) 65984
259,0   69     8    0.012783477 192607  A  WS 408269264 + 16 <- (259,105) 66000
259,0   69    15    0.012920923 192607  A  WS 408269272 + 8 <- (259,105) 66008
259,0   69    22    0.013206220 192607  A  WS 408269272 + 24 <- (259,105) 66008
259,0   69    29    0.014262027 192607  A  WS 408269288 + 104 <- (259,105) 66024
259,0   69    36    0.014881080 192607  A  WS 408269384 + 80 <- (259,105) 66120
259,0   69    43    0.015298038 192607  A  WS 408269456 + 48 <- (259,105) 66192
259,0   69    50    0.015556532 192607  A  WS 408269496 + 32 <- (259,105) 66232
259,0   69    57    0.015805013 192607  A  WS 408269520 + 24 <- (259,105) 66256
259,0   22    24    0.016778710 192607  A  WS 408269536 + 88 <- (259,105) 66272
259,0   22    31    0.017010918 192607  A  WS 408269616 + 16 <- (259,105) 66352
259,0   22    38    0.017258905 192607  A  WS 408269624 + 48 <- (259,105) 66360
259,0   66    28    0.017474505 192607  A  WS 408269664 + 40 <- (259,105) 66400
259,0   66    34    0.017693495 192607  A  WS 408269696 + 40 <- (259,105) 66432
259,0   66    40    0.017847418 192607  A  WS 408269728 + 16 <- (259,105) 66464
259,0   66    46    0.017962023 192607  A  WS 408269736 + 8 <- (259,105) 66472
259,0   66    52    0.018059066 192607  A  WS 408269736 + 32 <- (259,105) 66472
259,0   66    58    0.019135273 192607  A  WS 408269760 + 128 <- (259,105) 66496
259,0   66    64    0.019674721 192607  A  WS 408269880 + 48 <- (259,105) 66616
259,0   66    70    0.020015913 192607  A  WS 408269920 + 56 <- (259,105) 66656
259,0   66    76    0.020357403 192607  A  WS 408269968 + 64 <- (259,105) 66704
259,0   66    82    0.020627556 192607  A  WS 408270024 + 16 <- (259,105) 66760
259,0   80    36    0.020807658 192607  A  WS 408270032 + 16 <- (259,105) 66768
259,0   80    42    0.020986979 192607  A  WS 408270040 + 24 <- (259,105) 66776
259,0   69    67    0.021134715 192607  A  WS 408270056 + 8 <- (259,105) 66792
```

- Key takeaways:
  - Sync writes!
  - Large sequential write w/ offset-overlapping on WAL/DB device
  - Small disk reads on DB device due to metadata lookup missing on writes
  - BlueFS will save tail part of previous unaligned writes and merge with current writes
  - Medium-sized requests due to db transaction batch in BlueStore

SUMMARY

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

- Client side cache on Optane with SQL workloads!

Acknowledgements

This is a joint team work.

Thanks for the contributions of Haodong, Tang, Jianpeng Ma.

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=6710886
4,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_b
ytes_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.