USING RECENTLY PUBLISHED CEPH REFERENCE ARCHITECTURES TO SELECT YOUR CEPH CONFIGURATION

Daniel Ferber

Open Source Software Defined Storage Technologist,
Intel Storage Group June 14, 2016

Ceph Days CERN

THE NEW CENTER OF POSSIBILITY

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.

Agenda for First Half of this Talk

- Inventory of Published Referenced Architectures from Red Hat and SUSE
- Walk through highlights of a soon to be published Intel and Red Hat Ceph Reference Architecture paper
- Introduce an Intel all-NVMe Ceph configuration benchmark for MySQL
- Show examples of Ceph solutions

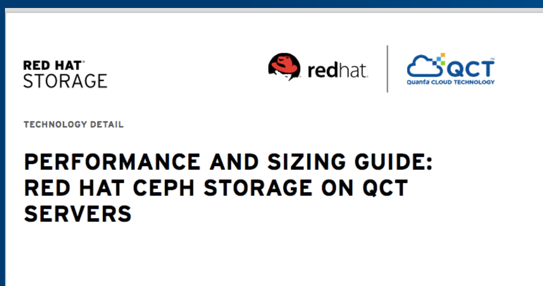
Dave Leone from Intel's SSD team will do second half of this presentation

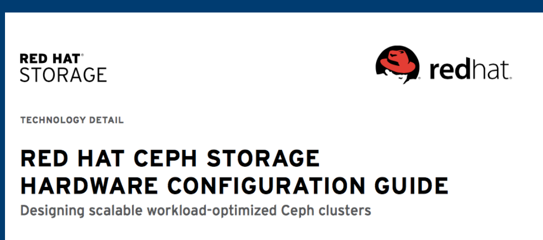
What Are Reference Architecture Key Components

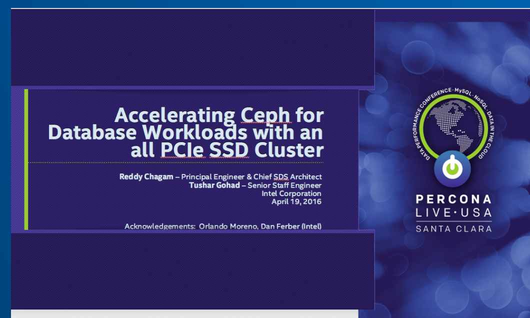
- Starts with workload (use case) and points to one or more resulting recommended configurations
- Configurations should be recipes that one can purchase and build
- Key related elements should be recommended
  - Replication versus EC, media types for storage, failure domains
- Ideally, performance data and tunings are supplied for the configurations

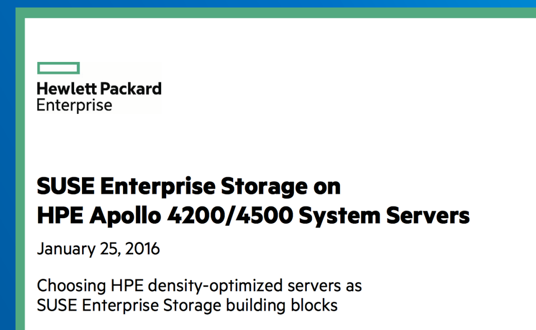
Tour of Existing Reference Architectures

Available Reference Architectures (recipes)









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

Available Reference Architectures (recipes)

- http://www.redhat.com/en/files/resources/en-rhst-cephstorage-supermicro-INC0270868_v2_0715.pdf
- http://www.qct.io/account/download/download?order_download_id=1065&dtype=Reference%20Architecture
- https://www.redhat.com/en/resources/red-hat-ceph-storage-hardware-configuration-guide
- https://www.percona.com/resources/videos/accelerating-ceph-database-workloads-all-pcie-ssd-cluster
- https://www.percona.com/resources/videos/mysql-cloud-head-head-performance-lab
- http://h20195.www2.hpe.com/v2/GetDocument.aspx?docname=4aa6-3911enw
- https://intelassetlibrary.tagcmd.com/#assets/gallery/11492083

A Brief Look at 3
of the Reference Architecture Documents

QCT AND RED HAT CEPH SOLUTION GUIDE

QCT CEPH PERFORMANCE AND SIZING GUIDE

- Target audience: Mid-size to large cloud and enterprise customers
- Showcases Intel based QCT solutions for multiple customer workloads
  - Introduces a three tier configuration and solution model:
    - IOPS Optimized, Throughput Optimized, Capacity Optimized
  - Specifies specific and orderable QCT solutions based on above classifications
  - Shows actual Ceph performance observed for the configurations

- Purchase fully configured solutions per above model from QCT
- Red Hat Ceph Storage Pre-Installed
- Red Hat Ceph Storage support included
- Datasheets and white papers at www.qct.io

QCT QxStor Red Hat Ceph Storage Edition Specification

	SMALL (500TB*)	MEDIUM (>1PB*)	LARGE (>2PB*)
Throughput optimized	16x RCT-200, each with D51PH-1ULH (1U) • 12x 8TB HDDs • 3x SSDs • 1x dual port 10GbE • 3x replica	6x RCT-400, each with T21P-4U/Dual (4U) • 2x 35x 8TB HDDs • 2x 2x PCIe NVMe SSDs • 2x single port 40GbE • 3x replica	11x RCT-400, with 11x T21P-4U/Dual (4U)
Cost/Capacity optimized		Nx RCC-400, each with T21P-4U/Dual • 2x 35x 8TB HDDs • 0x SSDs • 2x dual port 10GbE • Erasure Coding 4:2	
IOPS optimized	Future direction	Future direction	NA

SUPERMICRO PERFORMANCE AND SIZING GUIDE

- Target audience: Mid-size to large cloud and enterprise customers
- Showcases Intel based Supermicro solutions for multiple customer workloads
  - Introduces a three tier configuration and solution model:
    - IOPS Optimized, Throughput Optimized, Capacity Optimized
  - Specifies specific and orderable Supermicro solutions based on above classifications
  - Shows actual Ceph performance observed for the configurations

- Purchase fully configured solutions per above model from Supermicro
- Red Hat Ceph Storage Pre-Installed
- Red Hat Ceph Storage support included
- Datasheets and white papers at supermicro.com

TABLE 8. THROUGHPUT-OPTIMIZED SUPERMICRO SERVER CONFIGURATIONS.

	CEPH CLUSTER SIZE (USABLE CAPACITY)			
	STARTER (50 TB)	SMALL (500 TB)	MEDIUM (1 PB)	LARGE (2 PB)
OSD SERVER QUANTITY	• 4	• 32	• 63	• 125
PERFORMANCE (ESTIMATED)	• Read: 3,500 MB/s • Write: 1,200 MB/s	• Read: 28,000 MB/s • Write: 9,500 MB/s	• Read: 55,000 MB/s • Write: 19,000 MB/s	• Read: 110,000 MB/s • Write: 37,000 MB/s
SUPERMICRO SERVERS	SSG-2028R-OSD072 (w/ 4 TB HDDs), or SSG-F618H-OSD288 (w/ 4 TB HDDs): 1x E5-2620v3 64 GB RAM 12x 4T HDD 1x 800 GB PCIe			
NETWORKING	10 Gigabit Ethernet 10 Gigabit Ethernet Gigabit Ethernet			

INTEL CEPH SOLUTION GUIDE

INTEL SOLUTIONS FOR CEPH DEPLOYMENTS

- Target audience: Mid-size to large cloud and enterprise customers
- Showcases Intel based solutions for multiple customer workloads
  - Uses the three tier configuration and solution model:
    - IOPS Optimized, Throughput Optimized, Capacity Optimized
  - Contains Intel configurations and performance data
  - Contains a Yahoo case study

- Contains specific use case examples
- Adds a Good, Better, Best model for all SSD Ceph configurations
- Adds configuration and performance data for Intel* Cache Acceleration
- Overviews CeTune and VSM tools
- Datasheets and white papers at
  intelassetlibrary.tagcmd.com/#assets/gallery/11492083

CEPH STORAGE NODE - GOOD	
CPU	Intel® Xeon® Processor E5-2650v3
NIC	10GbE
Drives	1x 1.6 TB P3700
	12x 4 TB HDDs (1:12 ratio)
	(P3700 as Journal and caching)
Software	Intel CAS
	RSTe/MD4.3 (optional)

Table 3. Good Configuration

CEPH STORAGE NODE - BETTER	
CPU	Intel® Xeon® Processor E5-2690
Memory	128 GB
NIC	Dual 10GbE
Drives	1x 800 GB P3700
	4x 1.6 TB S3510
	(P3700 as Journal and caching)
Software	Intel CAS

Table 4. Better Configuration

CEPH STORAGE NODE - BEST	
CPU	Intel® Xeon® Processor E5-2699v3
Memory	>=128 GB
NIC	2x 40GbE
	4x dual 10GbE
Drives	4-6x 2 TB P3700

Table 5. Best Configuration

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

Quick Look at 3 Tables Inside the Intel and Red Hat Reference Architecture Document (to be published soon)

Generic Red Hat Ceph Reference Architecture Preview

https://www.redhat.com/en/resources/red-hat-ceph-storage-hardware-configuration-guide

TABLE 1. CEPH CLUSTER OPTIMIZATION CRITERIA.

OPTIMIZATION CRITERIA	PROPERTIES	EXAMPLE USES
IOPS-OPTIMIZED	• Lowest cost per IOPS • Highest IOPS • Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster)	• Typically block storage • 3x replication (HDD) or 2x replication (Intel SSD DC Series) • MySQL on OpenStack clouds
THROUGHPUT-OPTIMIZED	• Lowest cost per given unit of throughput • Highest throughput • Highest throughput per BTU • Highest throughput per watt • Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster)	• Block or object storage • 3x replication • Active performance storage for video, audio, and images • Streaming media
CAPACITY-OPTIMIZED	• Lowest cost per TB • Lowest BTU per TB • Lowest watt per TB • Meets minimum fault domain recommendation (single server is less than or equal to 15% of the cluster)	• Typically object storage • Erasure coding common for maximizing usable capacity • Object archive • Video, audio, and image object archive repositories

- IOPS optimized config is all NVME SSD
  - Typically block with replication
    Allows database work
  - Journals are NVME
  - Bluestore, when supported, will increase performance
- Throughout optimized is a balanced config
  - HDD storage with SSD journals
  - Block or object, with replication
- Capacity optimized typically all HDD storage
  - Object and EC

Intel and Red Hat Ceph Reference Architecture Preview

TABLE 2. BROAD SERVER SIZING TRENDS.

OPTIMIZATION CRITERIA	OPENSTACK STARTER (64 TB)	SMALL (250 TB)	MEDIUM (1 PB)	LARGE (2 PB)
IOPS-OPTIMIZED	• Servers with 2-4x PCIe/NVMe slots, or • Servers with 8-12x 2.5-inch SSD bays (SAS/SATA)		• Not typical	• Not typical
THROUGHPUT-OPTIMIZED	• Servers with 12-16x 3.5-inch drive bays		• Servers with 24-36x 3.5-inch drive bays	• Servers with 24-36x 3.5-inch drive bays
CAPACITY-OPTIMIZED				• Servers with 60-72x 3.5-inch drive bays

- IOPS optimized Ceph clusters are typically in the TB ranges
- Throughput clusters will likely move to 2.5" inch enclosures and all SSD over time
- Capacity optimized likely to favor 3.5" for HDD storage

https://www.redhat.com/en/resources/red-hat-ceph-storage-hardware-configuration-guide

Intel and Red Hat Ceph Reference Architecture Preview

TABLE 3. CONFIGURING INTEL SERVERS FOR RED HAT CEPH STORAGE.

OPTIMIZATION CRITERIA	OPENSTACK STARTER (100 TB)	SMALL (250 TB)	MEDIUM (1 PB)	LARGE (2 PB)
IOPS-OPTIMIZED	• Ceph RBD (block) pools • OSDs on 1-4 Intel SSD DC P3700 Series per server. Journals co-located on different partitions. • 1x Intel SSD DC P3700 per server: single-socket Intel Xeon Processor E5-2630v4 (10 cores) • 2x Intel SSD DC P3700 per server: dual-socket Intel Xeon Processor E5-2630v4 (20 cores) • 4x Intel SSD DC P3700 per server: dual-socket Intel Xeon Processor E5-2695v4 (36 cores) • Data protection: Replication (2x on SSD-based OSDs) with regular backups to the object storage pool • 2-4 OSDs per SSD or NVMe drive		• Not typical	• Not typical

- Specific recommended Intel processor and SSD models are now specified
- Intel processor recommendations depend on how many OSDs are used

Intel and Red Hat Ceph Reference Architecture

THROUGHPUT-OPTIMIZED	• Ceph RBD (block) or Ceph RGW (object) pools
	• OSDs on HDDs:
	  • Good: write journals on Intel SSD DC S3710 400TB drives, with a ratio of 4-5 HDDs to each SSD
	  • Better: write journals on Intel SSD DC P3700 800TB NVMe drives, with a ratio of 12-18 HDDs to each SSD
	• One CPU core-GHz per OSD. For example:
	  • 12 OSD/HDDs/server: single-socket Intel Xeon Processor E5-2620v4 (8 cores*2.1 GHz)
	  • 36 OSD/HDDs/server: dual-socket Intel Xeon Processor E5-2630v4 (20 cores*2.2 GHz)
	  • 60 OSD/HDDs/server: dual-socket Intel Xeon E5-2683v4 (32 cores*2.1 GHz)
	  • Data protection: Replication (read-intensive or mixed read/write) or erasure-coded (write-intensive)
	• High-bandwidth networking, greater than 10 GbE for servers with more than 12-16 drives

- Recommendations for specific Intel SSDs and journals, with two options
- Specific Intel processor recommendations, depending on how many OSDs

Intel and Red Hat Ceph Reference Architecture

OPTIMIZATION CRITERIA	OPENSTACK STARTER (100 TB)	SMALL (250 TB)	MEDIUM (1 PB)	LARGE (2 PB)
CAPACITY-OPTIMIZED	• Not typical	• Ceph RGW (object) pools • OSDs on HDDs. Write journals co-located on HDDs in separate partition. • One CPU core-GHz per OSD. See throughput-optimized section above for examples. • Data protection: Erasure-coded		

- No SSDs for capacity model
- Specific Intel processor recommendations are same as on previous throughput config recommendations, and are based on number of OSDs

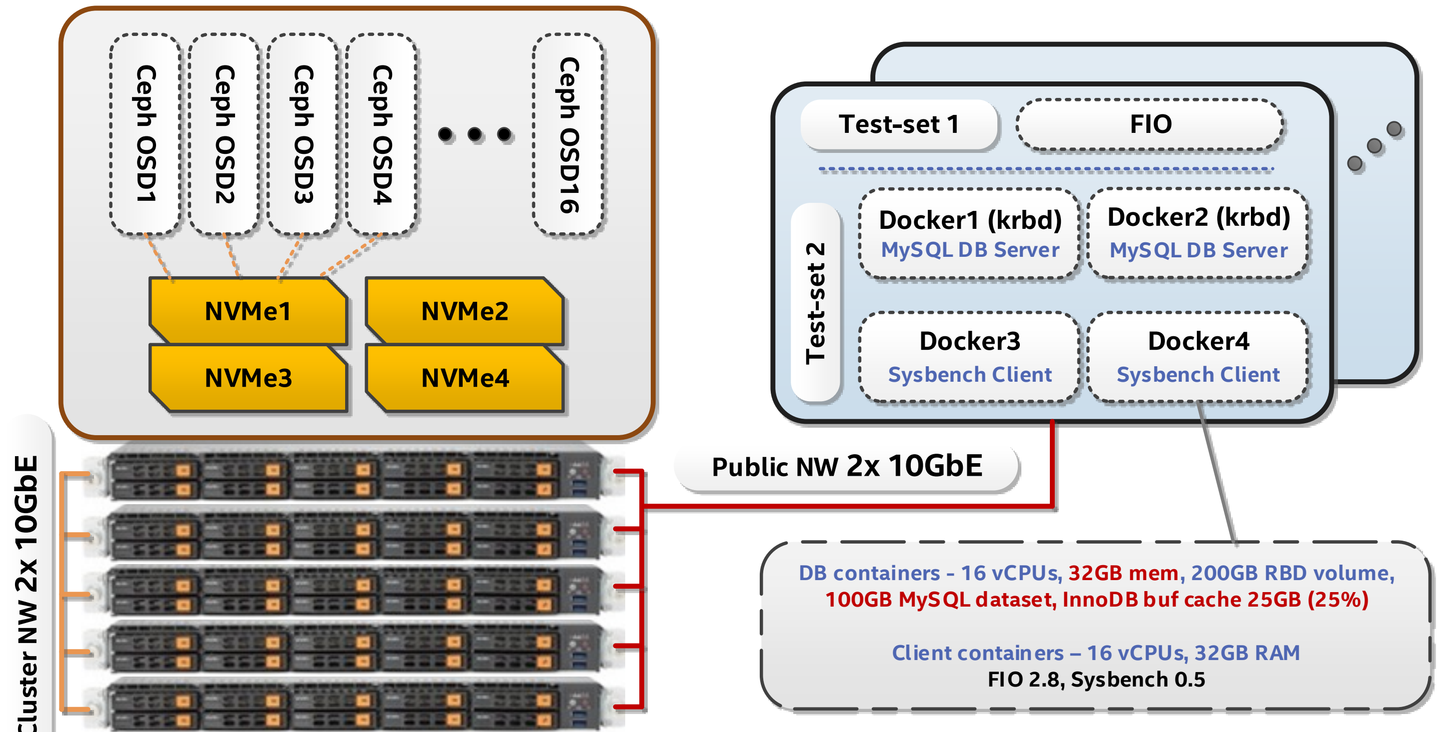
Intel all-NVMe SSD
Ceph Reference Architecture

Presented by Intel at Percona Live 2016

An “All-NVMe” high-density Ceph Cluster Configuration

5-Node all-NVMe Ceph Cluster
Dual-Xeon E5 2699v4@2.2GHz, 44C HT, 128GB DDR4
Centos 7.2, 3.10-327, Ceph v10.1.2, bluestore async

10x Client Systems + 1x Ceph MON
Dual-socket Xeon E5 2699v3@2.3GHz
36 Cores HT, 128GB DDR4

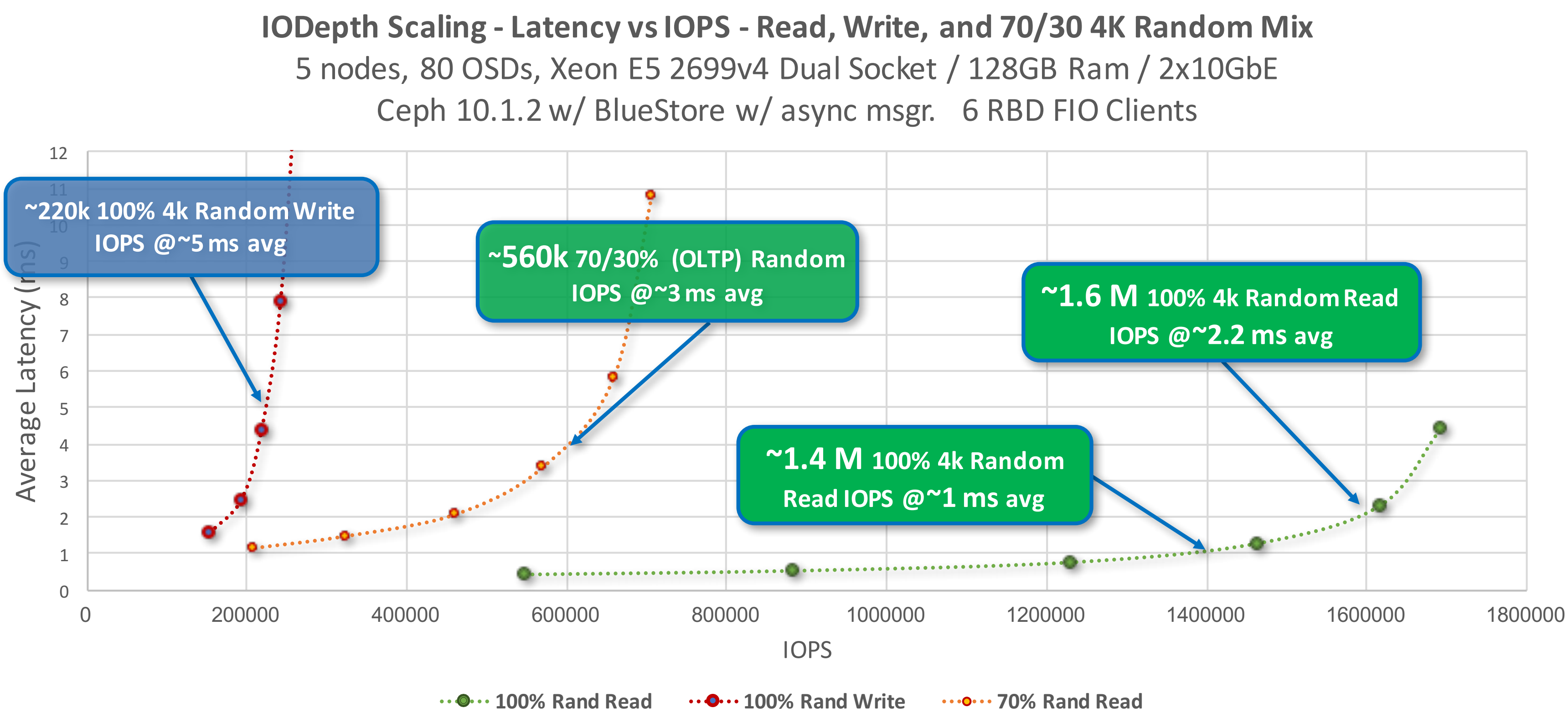


Supermicro 1028U-TN10RT+

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.

PERCONA LIVE

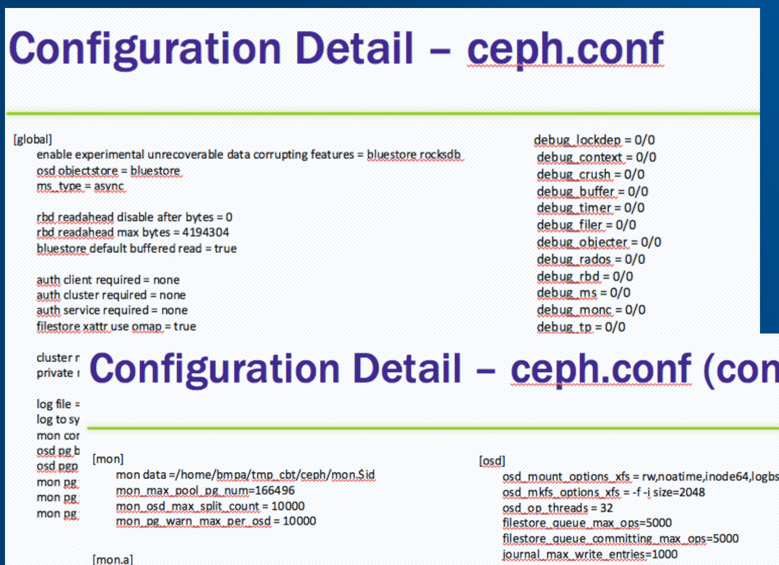
4K Random Read/Write Performance and Latency (Baseline FIO Test)



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.

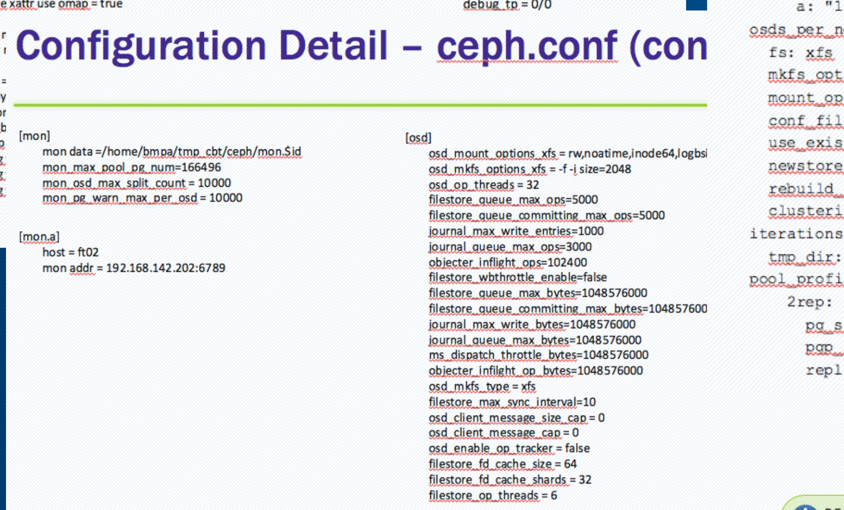
PERCONA LIVE

Tunings for the all-NVE Ceph Cluster



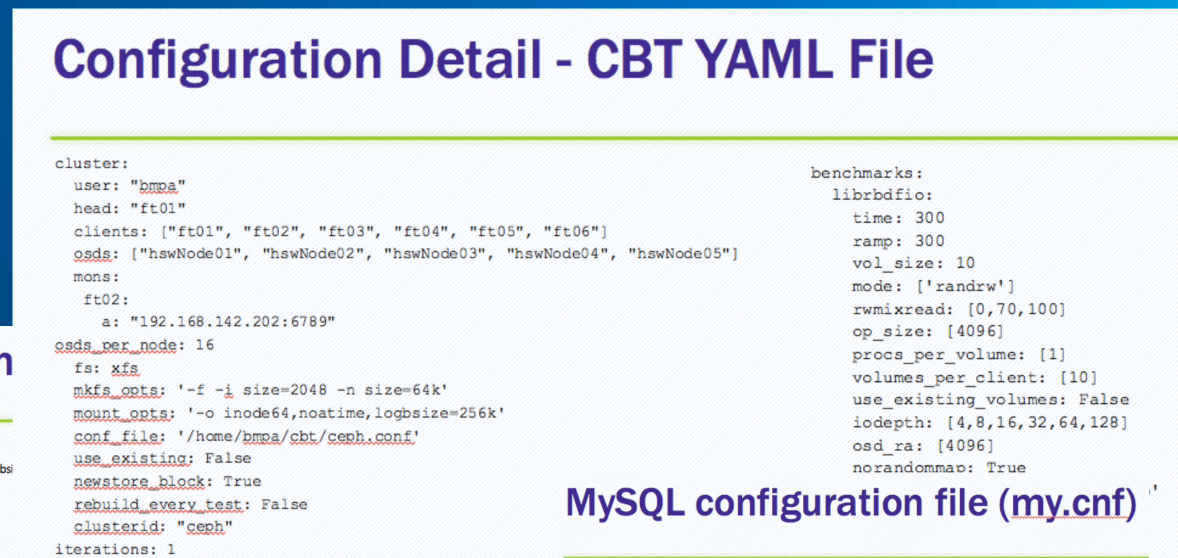
Configuration Detail – ceph.conf

```
[global]
    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 r
    private r

    log file =
    log to sy
    mon cor
    osd pg b
    osd pgp
    mon pg
    mon pg
    mon pg

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
```



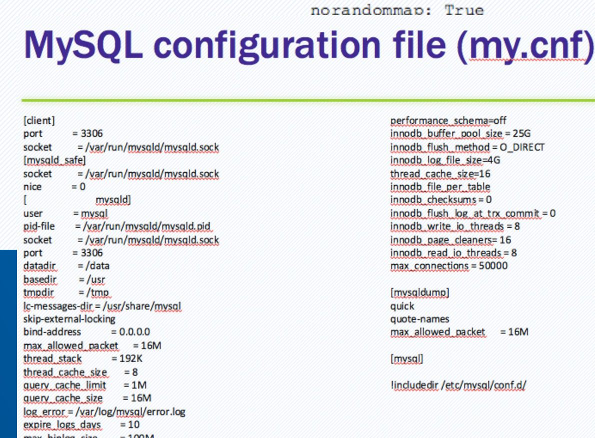
Configuration Detail – ceph.conf (con

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

[mon.a]
    host = ft02
    mon addr = 192.168.142.202:6789

[osd]
    osd_mount_options_xfs = rw,noatime,inode64,logbsi
    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=1000
    journal_queue_max_ops=3000
    objecter_inflight_ops=102400
    filestore_wbthrottle_enable=false
    filestore_queue_max_bytes=1048576000
    filestore_queue_committing_max_bytes=104857600
    journal_max_write_bytes=1048576000
    journal_queue_max_bytes=1048576000
    ms_dispatch_throttle_bytes=1048576000
    objecter_infilght_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
```



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 inode64,noatime,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:
  librbdfio:
    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
```



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/mysql/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/
```

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.

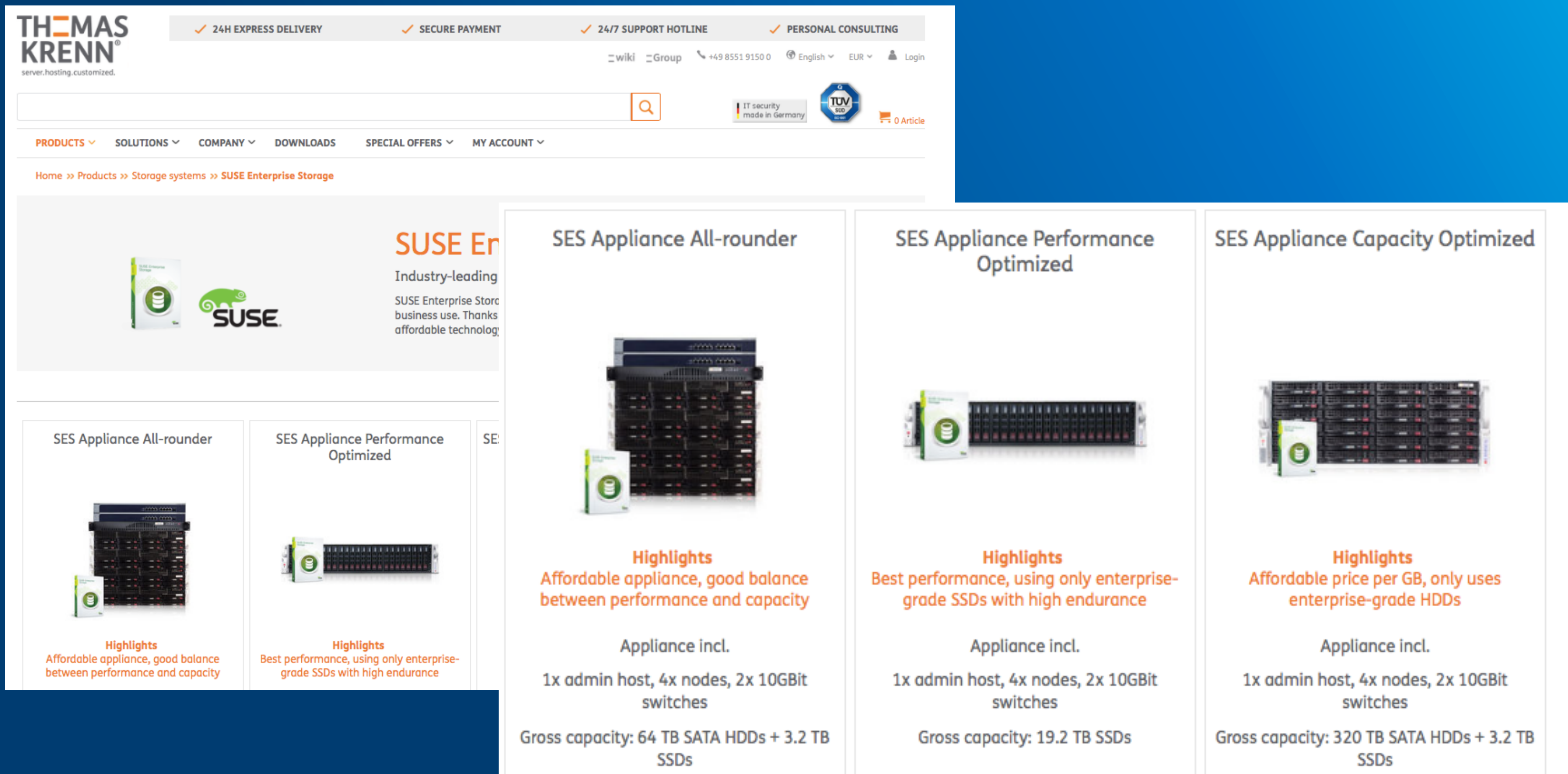
All NVMe Flash Ceph Storage – Summary

- Intel NVMe Flash storage works for low latency workloads
- Ceph makes a compelling case for database workloads
- 1.4 million random read IOPS is achievable in 5U with ~1ms latency today.
- Sysbench MySQL OLTP Performance numbers were good at 400k 70/30% OLTP QPS @~50 ms avg

- Using Xeon E5 v4 standard high-volume servers and Intel NVMe SSDs, one can now deploy a high performance Ceph cluster for database workloads
- Recipe and tunings for this solution are here:
  www.percona.com/live/data-performance-conference-2016/content/accelerating-ceph-database-workloads-all-pcie-ssd-cluster

Ceph Solutions Available
in addition to the
QCT, Supermicro, and HP Solutions Already Mentioned

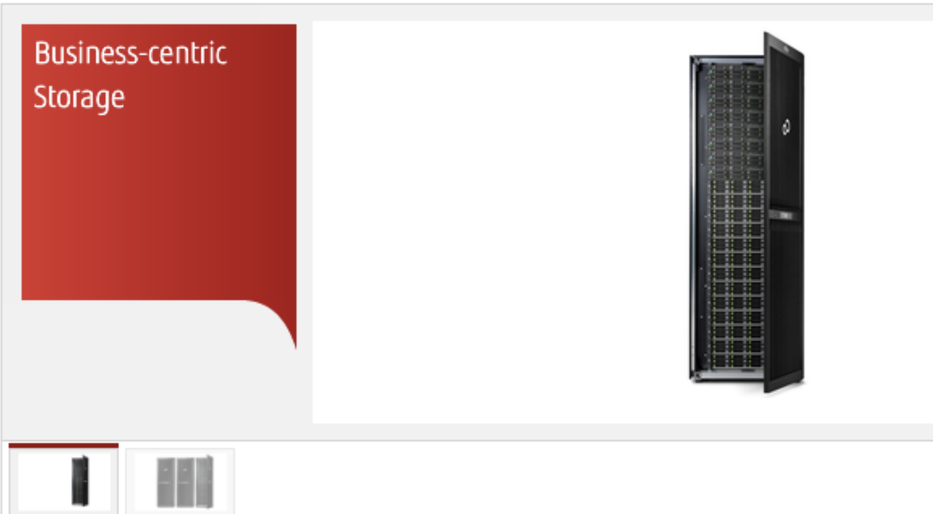
Thomas Krenn SUSE Enterprise Storage



https://www.thomas-krenn.com/en/products/storage-systems/suse-enterprise-storage.html

Fujistu Intel Based Ceph Appliance

FUJITSU Storage ETERNUS CD10000 S2



ETERNUS CD10000 S2 is a hyperscale, software-defined storage system designed to manage vast amounts of data. A configuration can start small and grow in line with the business. The architecture allows individual storage nodes to be added, exchanged and upgraded without downtime. Fujitsu integrates open source Ceph software in a complete and fully supported solution.

DARZ gains from Hyperscale storage system ETERNUS CD10000, to provide highly efficient offerings on Deutsche Börse Cloud Exchange (DBCE) marketplace

> **"Combining FUJITSU's technology with PROFI's skills and expertise has given us the quality, security and flexibility we need to join the DBCE marketplace."**
>
> **Lars Göbel,** Head of Sales and IT Operations, DARZ

http://www.fujitsu.com/global/products/computing/storage/eternus-cd/s2/

Ceph Reference Architectures Summary

Ceph Reference Architectures Summary

- The community has a growing number of good reference architectures
- Some point to specific hardware, others are generic
- Different workloads are catered for
- Some of the documents contain performance and tuning information
- Commercial support available for professional services and software support
- Intel will continue to work with its ISV and hardware systems partners on reference architectures
  - And continue Intel's Ceph development focused on Ceph performance

NEXT – A FOCUS ON NVM TECHNOLOGIES FOR TODAY'S AND TOMORROW'S CEPH

Dave Leone, Technical Marketing Engineer, Intel Corporation

June 2016

Solid State Drive (SSD) for Ceph today

Three Configurations for Ceph Storage Node

Standard/good (lowest cost)

NVMe/PCIe SSD for Journal + Caching, HDDs as OSD data drive

Example: 1 x Intel P3700 1.6TB as Journal and Cache + Intel CAS caching software, + 10 HDDs

Ceph storage node –Good	
CPU	Intel(R) Xeon(R) CPU E5-2650v3
Memory	64 GB
NIC	10GbE
Disks	1x 1.6TB P3700 + 10x 4TB HDDs (1:10 ratio) P3700 as Journal and caching
Caching software	Intel iCAS 3.0, option: RSTe/MD4.3

Better (higher cost, best TCO at the moment)

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

Example: 1 x Intel P3700 800GB + 4 x Intel S3510 1.6TB

Ceph Storage node –Better	
CPU	Intel(R) Xeon(R) CPU E5-2690
Memory	128 GB
NIC	Duel 10GbE
Disks	1x 800GB P3700 + 4x S3510 1.6TB

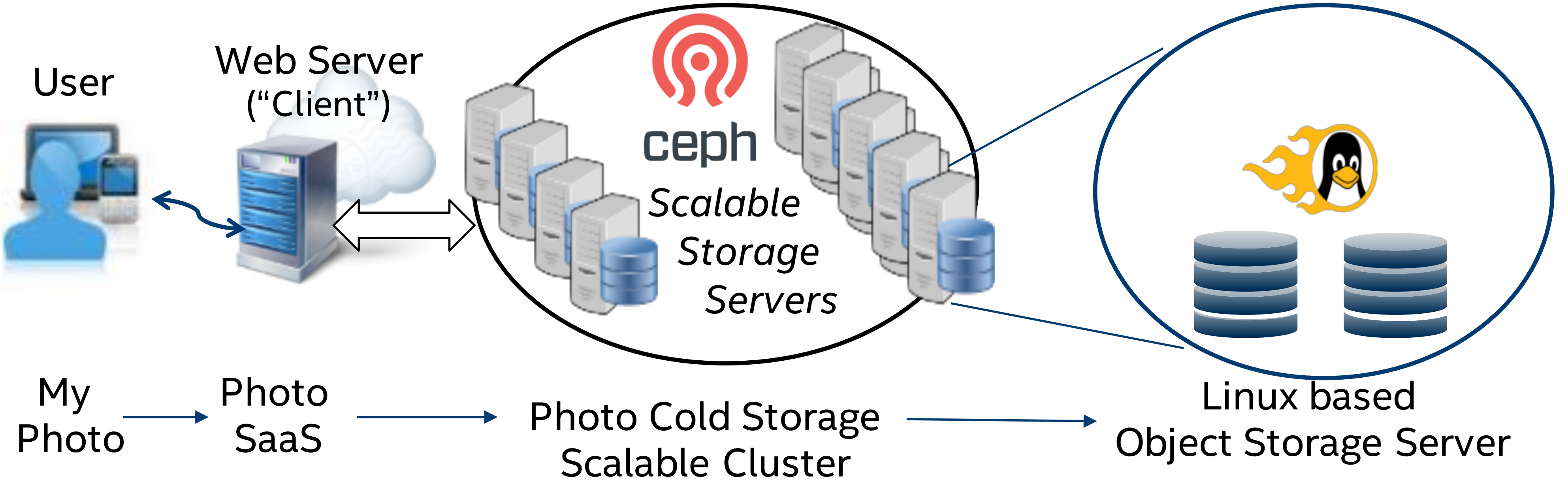
Best Performance ($$)

All NVMe/PCIe SSDs

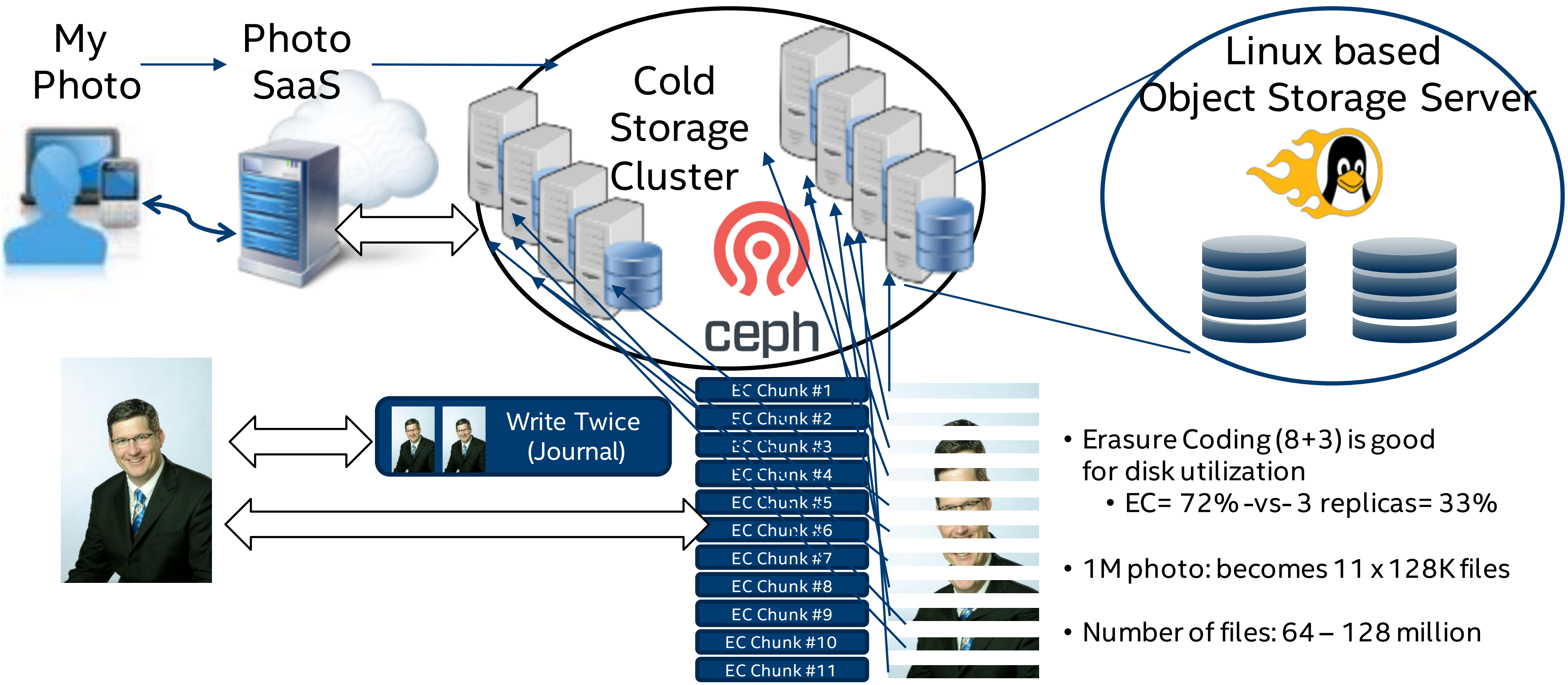
Example: 4 x Intel P3700 2TB SSDs

Ceph Storage node –Best	
CPU	Intel(R) Xeon(R) CPU E5-2699v3
Memory	>= 128 GB
NIC	2x 40GbE, 4x dual 10GbE
Disks	4 x P3700 2TB

Using Intel® NVMe SSDs to optimize Ceph* Software Defined Storage

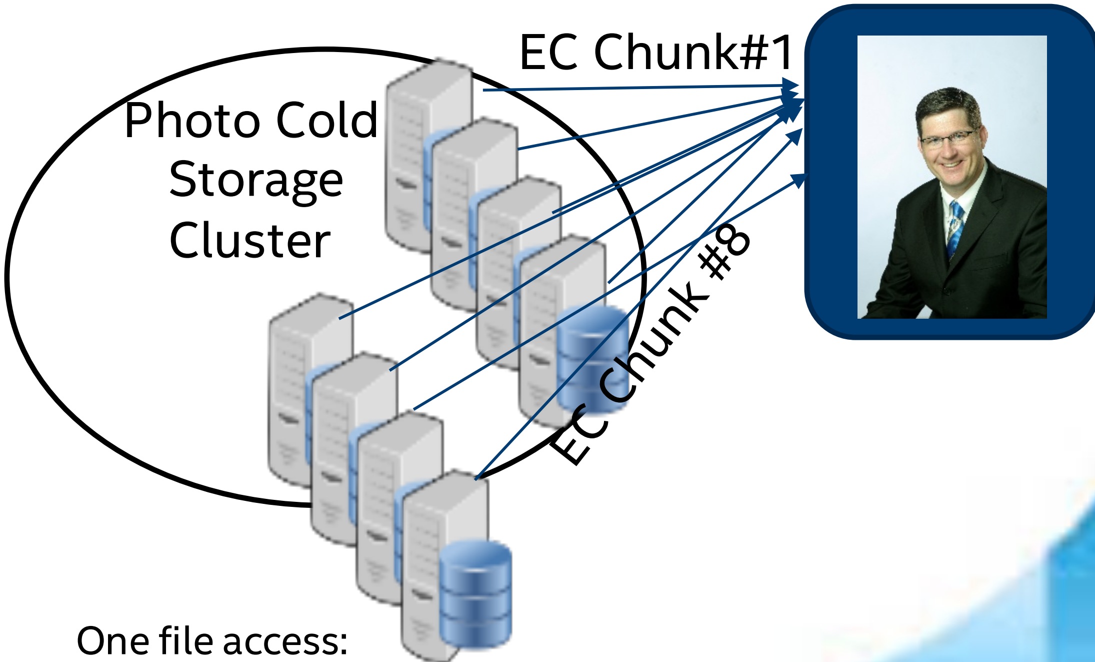


Ceph* Challenge #1: Huge Number of Small Files



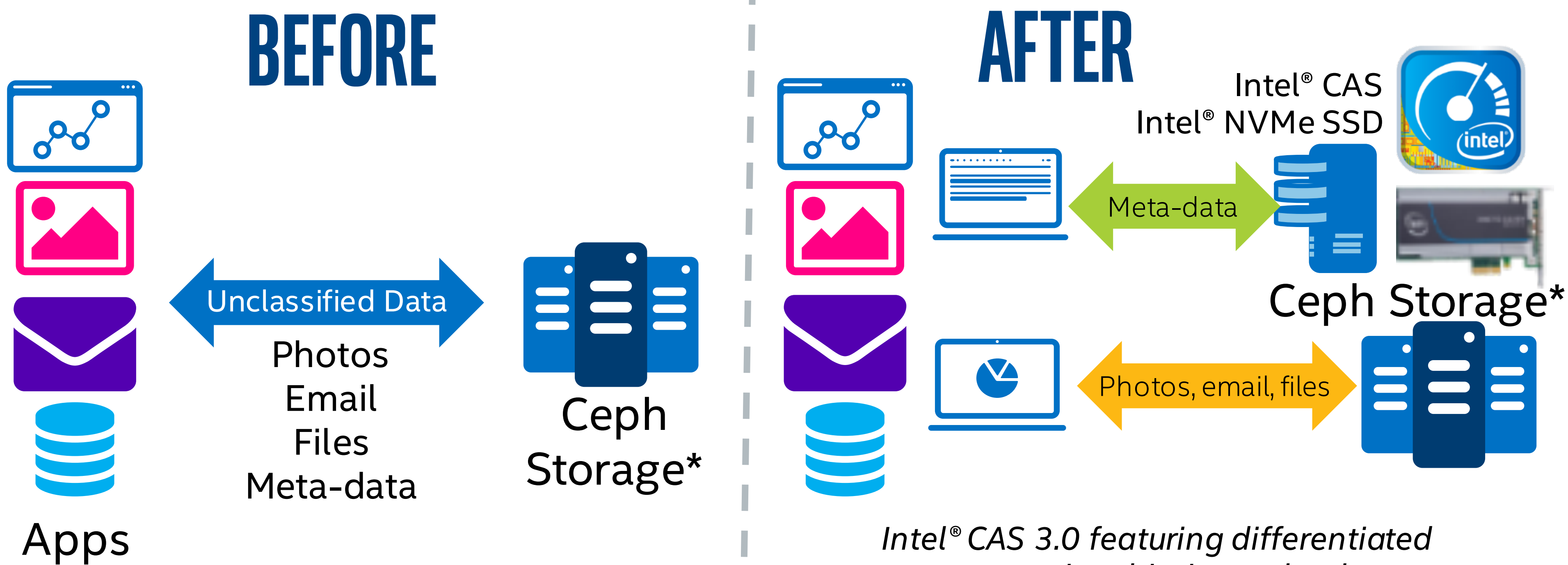
- Erasure Coding (8+3) is good for disk utilization
  - EC= 72% -vs- 3 replicas= 33%
- 1M photo: becomes 11 x 128K files
- Number of files: 64 – 128 million

Ceph* Challenge #2:
Long latency due to Erasure Code and meta-data lookups



One file access:
becomes 3-4 disk accesses

Solution to boost Ceph* performance using Intel CAS including DSS hinting



Intel® CAS 3.0 featuring differentiated storage services hinting technology

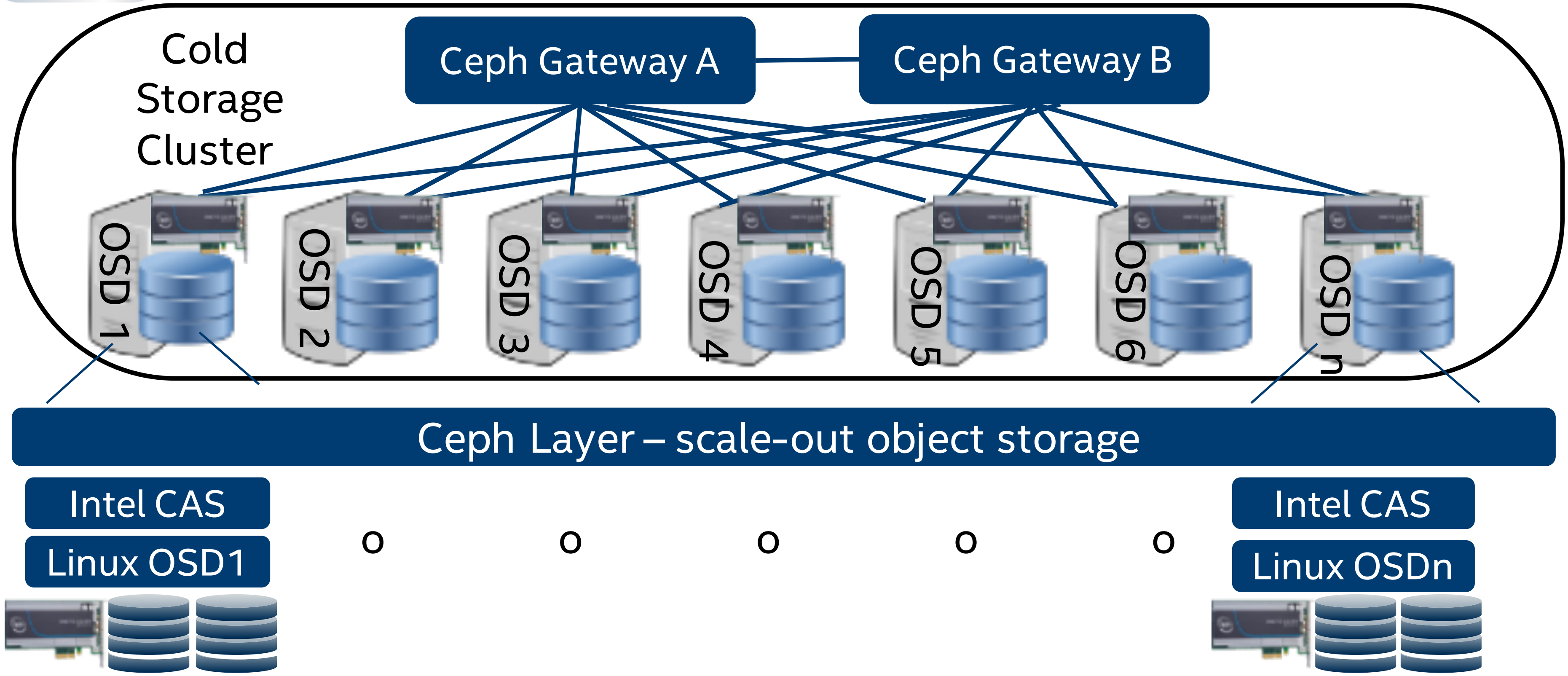
Benefits of classifying data types

I/O Classification Schema as implemented in Intel® CAS for Linux*

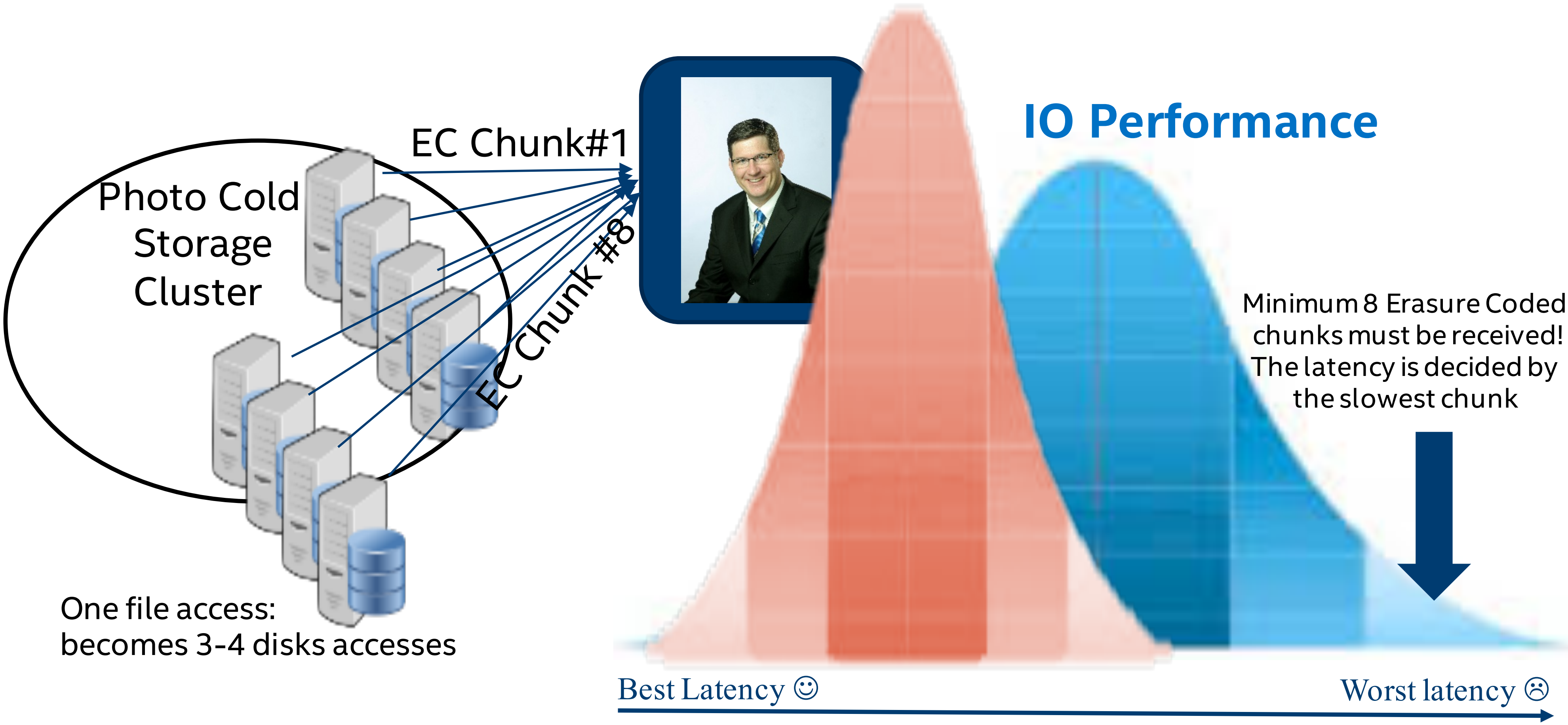
- Broadly applicable to Linux-based storage systems
- Intel CAS integrated Differentiated Storage Services (DSS) hinting, two elements:
  - Hint generation with patchless Meta-data tagging engine
  - Hint consumption by instrumenting the Intel Cache Acceleration SW to include the DSS I/O Classes (see the table on the right)
- Ability to selectively cache & evict based on block type & priority
  - Classifies I/O requests in software
  - Assigns policies to I/O classes
  - Enforces policies in the storage system
  - Evicts from cache based on priority
- Intel® CAS operates below the software stack at the Local filesystem block layer
  - No modification to Ceph*/Swift*/Lustre* stack required
- Benefits of this new approach:
  - End users can now uniquely identify the Meta-data and target only that data to the SSD cache
  - A very small cache tuned for best price-performance for a given workload

CAS I/O Classes
Unclassified
Meta-data (Superblock, Inode, IndirectBlk, Directory, etc)
<=4KiB
<=16KiB
<=64KiB
<=256KiB
<=1MiB
<=4MiB
<=16MiB
<=64MiB
<=256MiB
<=1GiB
>1GiB
O_DIRECT
Misc

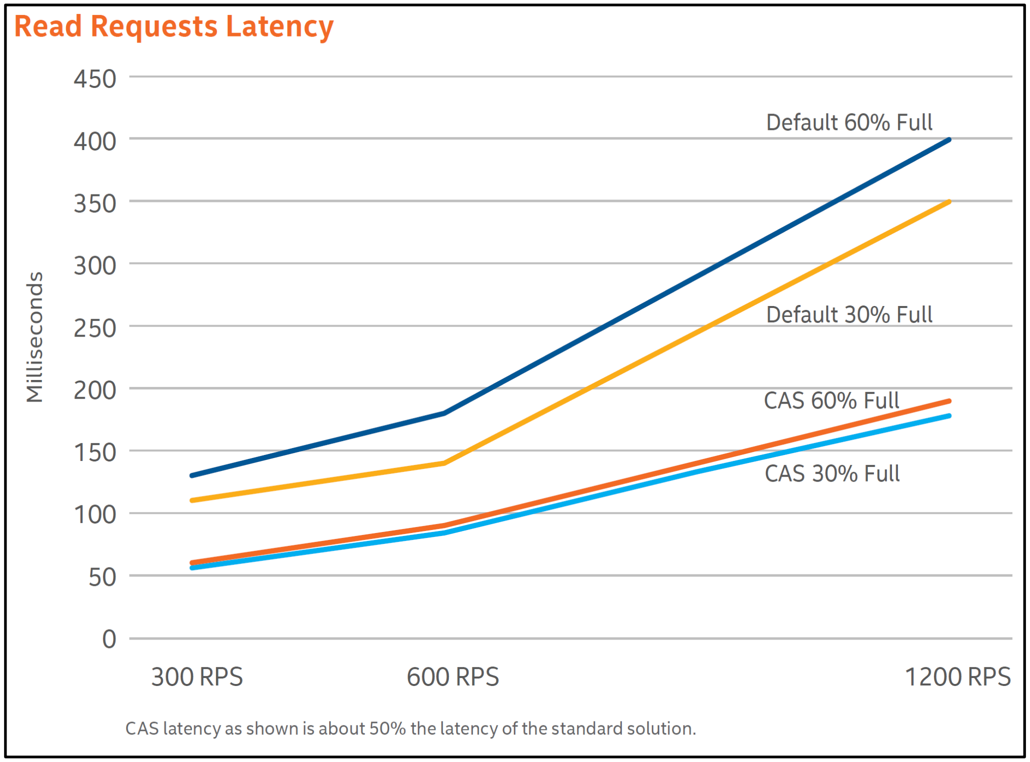
How caching is deployed to boost Ceph SDS

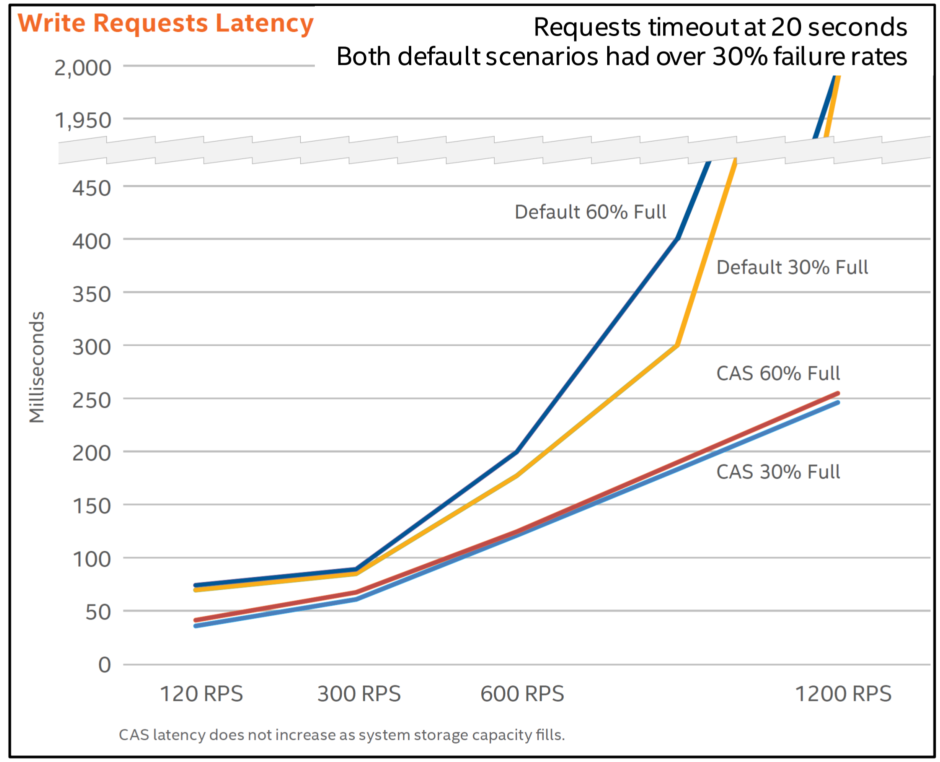


Benefit to latency distribution with metadata tagging



Yahoo* (Ceph* object) - Results

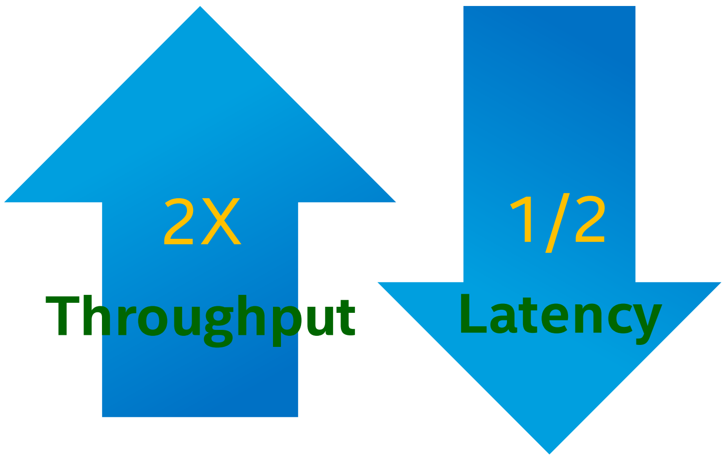




Results of Yahoo* internal benchmark testing - Ruiping Sun, Principal Architect, Yahoo*.

Hardware/Software Config: 8 OSD Nodes, each: HP ProLiant DL180 G6 ySPEC 39.5, 2x Xeon X5650 2.67GHz (HT enabled, total 12 cores, 24 threads), Intel 5520 IOH-36D B3 (Tylersburg), 48GB 1333MHz DDR3 (12x4GB PC3-10600 Samsung DDR3-1333 ECC Registered CL9 2Rx4), 10*8TB 7200 RPM SATA HDDs, 1*1.6TB Intel P3600 SSD (10GB journal per OSD, 1.5TB cache) *(CAS config only)*, 2*HP NC362i/Intel 82576 Gigabit NICs, 2*Intel 82599EB 10GbE NICs, RHEL 6.5 w/kernel 3.10.0-123.4.4.el7

Benefits for Ceph Storage* using Intel® NVMe SSDs with Intel® Cache Acceleration Software



- ***<5% NVMe SSD caching for 2X performance!***
- ***Intel Cache Acceleration Software available with license or as a bundle with Intel NVMe SSDs***

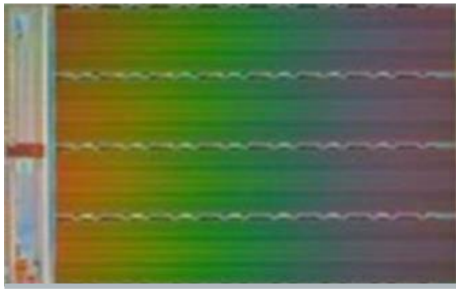
- ***To Learn More***
  - CAS Web Site
  - Ceph IDF 2015 Demo: https://www.youtube.com/watch?v=vtIlbxO4Zlk
  - Special Yahoo speaker IDF 2015: http://intelstudios.edgesuite.net//idf/2015/sf/aep/SSDS002/SSDS002.html
  - Intel Solutions for Ceph Deployments: http://www.intel.com/content/www/us/en/software/cache-acceleration-software-yahoo-brief.html
  - Intel Solutions for Ceph Deployments: http://intelassetlibrary.tagcmd.com/#assets/gallery/11492083

Get the Free 120-day Trial!
http://www.intel.com/cas

- ***Considerations for adoption***
  - Support RHEL, SLES, CentOS, ext4, ext3, xfs.
  - Intel will help to fine tune performance for your cloud workload
  - Have validated with Ceph Giant & Hammer. Currently testing Ceph Jewel, Lustre, Swift, and Hadoop.

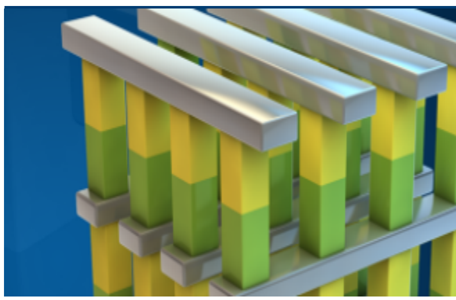
3D NAND and 3D XPoint™ for Ceph tomorrow

NAND Flash and 3D XPoint™ Technology for Ceph Tomorrow



3D MLC AND TLC NAND

BUILDING BLOCK ENABLING EXPANSION OF SSD INTO HDD SEGMENTS



3D XPOINT™

BUILDING BLOCKS FOR ULTRA HIGH PERFORMANCE STORAGE & MEMORY

3D XPoint™ TECHNOLOGY

In Pursuit of Large Memory Capacity … Word Access … Immediately Available …

Word (Cache Line)

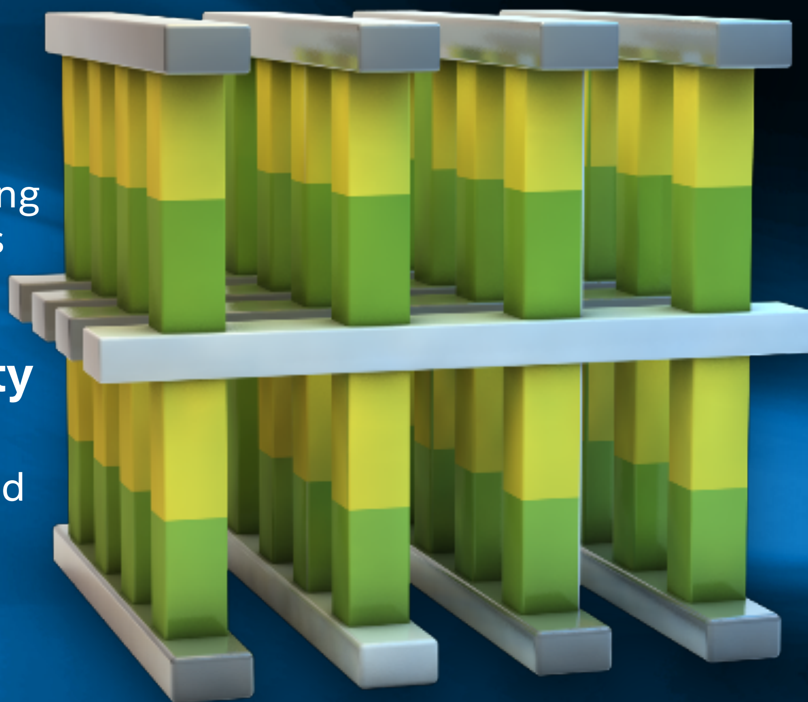
Crosspoint Structure

Selectors allow dense packing and individual access to bits

Large Memory Capacity

Crosspoint & Scalable

Memory layers can be stacked in a 3D manner



NVM Breakthrough Material Advances

Compatible switch and memory cell materials

Immediately Available

High Performance Cell and array architecture that can switch states 1000x faster than NAND

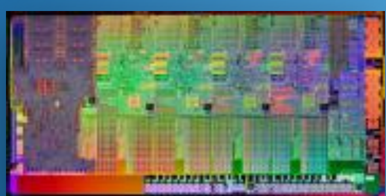
intel

3D XPoint™ TECHNOLOGY

Breaks the Memory Storage Barrier

STORAGE

SRAM
Latency: 1X
Size of Data: 1X



DRAM
Latency: ~10X
Size of Data: ~100X

3D XPoint ™ Memory Media
Latency: ~100X
Size of Data: ~1,000X

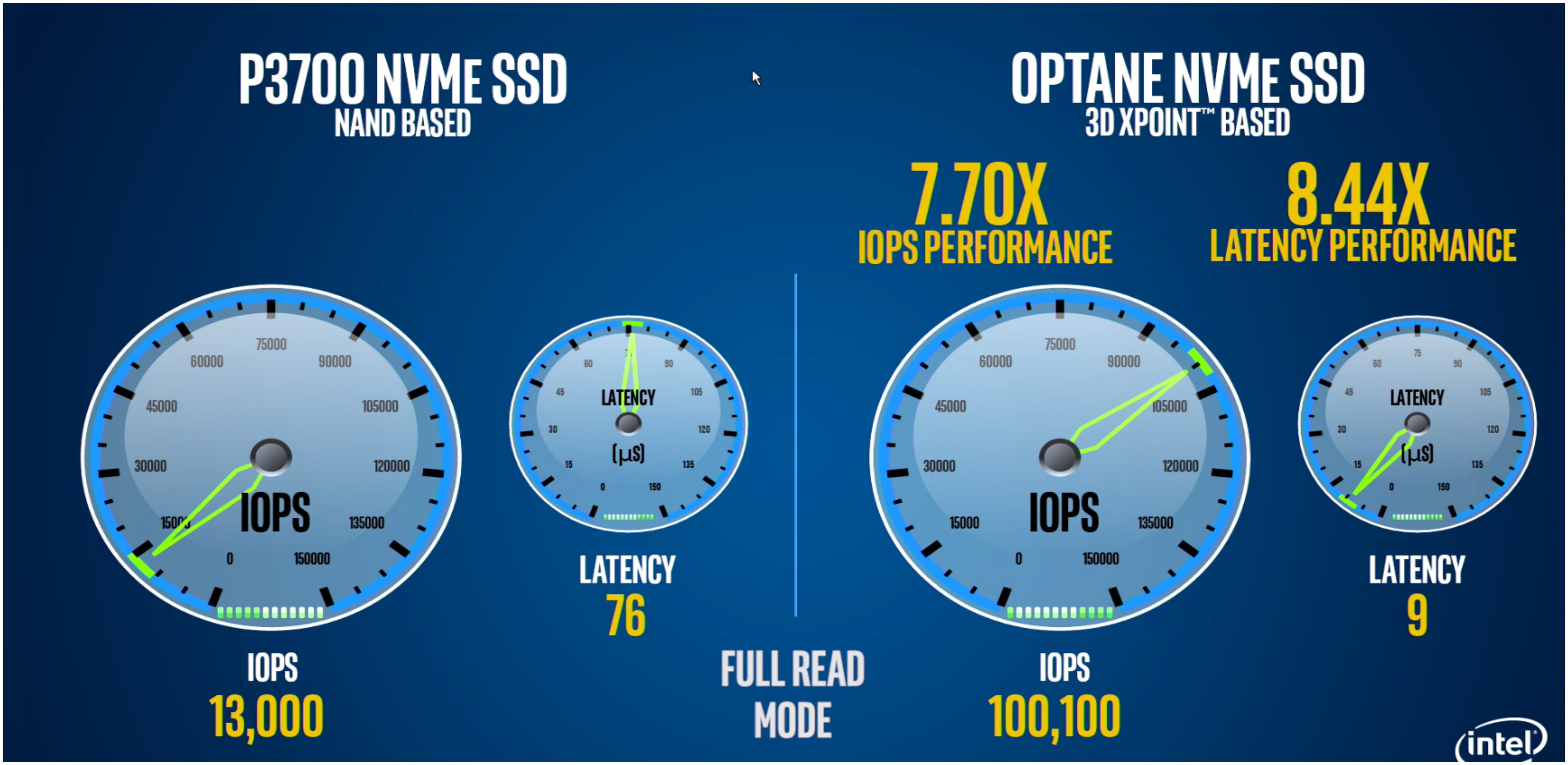
NAND SSD
Latency: ~100,000X
Size of Data: ~1,000X

HDD
Latency: ~10 MillionX
Size of Data: ~10,000X

MEMORY

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.

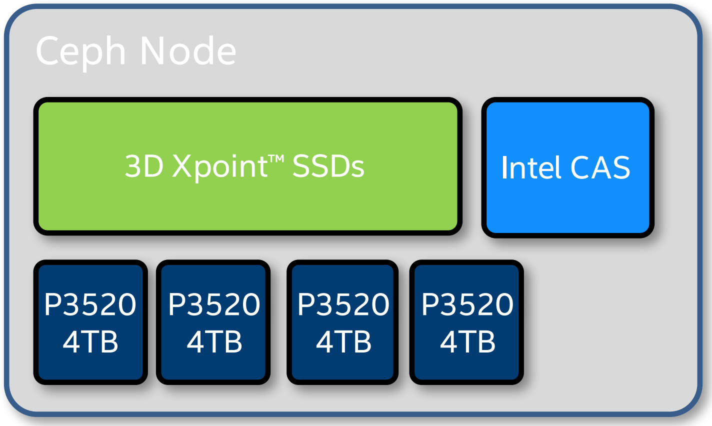
Intel® Optane™ (prototype) vs Intel® SSD DC P3700 Series at QD=1



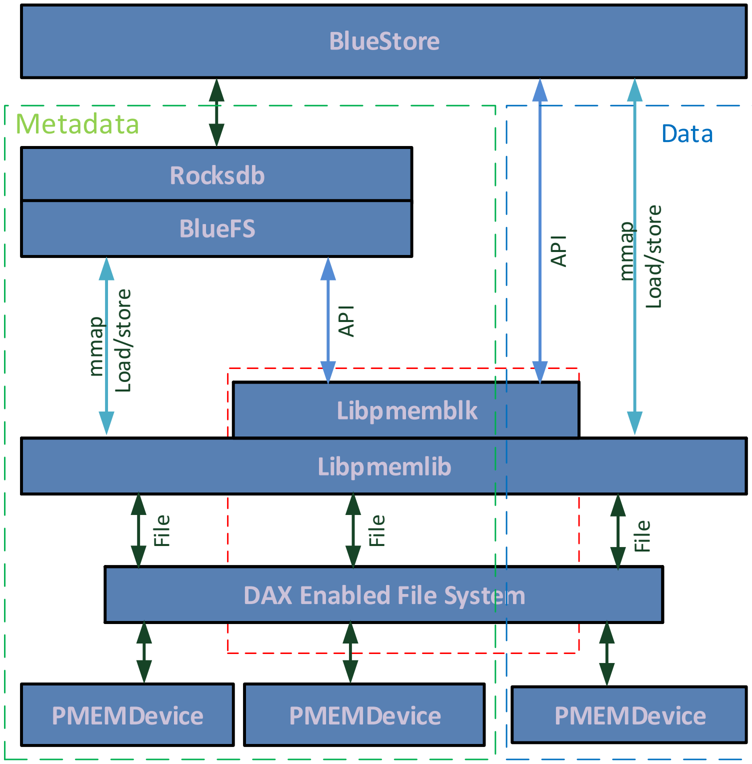
Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance. Server Configuration: 2x Intel® Xeon® E5 2690 v3 NVM Express* (NVMe) NAND based SSD: Intel P3700 800 GB, 3D Xpoint based SSD: Optane NVMe OS: Red Hat* 7.1

3D Xpoint & 3D NAND Solution Opportunities

- 3D XPoint as journaling and cache
- 3D NAND as primary storage



- 3D XPoint as Bluestore back end



Legal Notices and Disclaimers

Legal Information: Benchmark and Performance Claims Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

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.

Test and System Configurations: See Back up for details.

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

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the first quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "plans," "believes," "seeks," "estimates," "may," "will," "should" and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel's actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be important factors that could cause actual results to differ materially from the company's expectations. Demand for Intel's products is highly variable and could differ from expectations due to factors including changes in the business and economic conditions; consumer confidence or income levels; customer acceptance of Intel's and competitors' products; competitive and pricing pressures, including actions taken by competitors; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel's gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; and product manufacturing quality/yields. Variations in gross margin may also be caused by the timing of Intel product introductions and related expenses, including marketing expenses, and Intel's ability to respond quickly to technological developments and to introduce new features into existing products, which may result in restructuring and asset impairment charges. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Results may also be affected by the formal or informal imposition by countries of new or revised export and/or import and doing-business regulations, which could be changed without prior notice. Intel operates in highly competitive industries and its operations have high costs that are either fixed or difficult to reduce in the short term. The amount, timing and execution of Intel's stock repurchase program and dividend program could be affected by changes in Intel's priorities for the use of cash, such as operational spending, capital spending, acquisitions, and as a result of changes to Intel's cash flows and changes in tax laws. Product defects or errata (deviations from published specifications) may adversely impact our expenses, revenues and reputation. Intel's results could be affected by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. Intel's results may be affected by the timing of closing of acquisitions, divestitures and other significant transactions. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.

intel®