SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2016

Hyper Converged Cache Storage Infrastructure For Cloud

Chendi, Xue <chendi.xue@intel.com>
Yuanhui, Xu <yuanhui.xu@intel.com>
Yuan, Zhou <yuan.zhou@intel.com>
Jian, Zhang <jian.zhang@intel.com>

Intel APAC R&D

Agenda

- Introduction
- Hyper Converged Storage
- Hyper Converged Cache Architecture
  - Overview
  - Design details
  - Performance overview
- Hyper Converged Cache with 3D XPoint™ technology
- Summary

Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Introduction

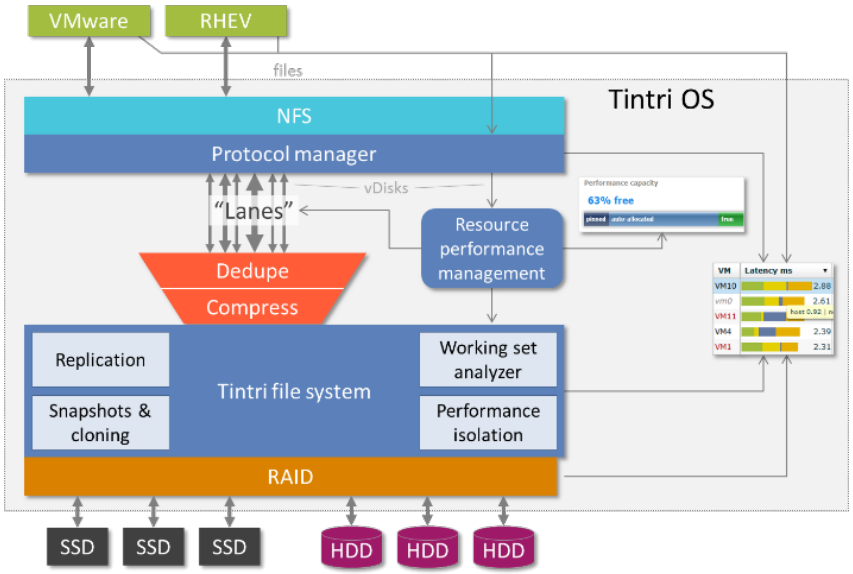
- Intel® Cloud and Bigdata Engineering Team
- Deliver optimized open source cloud and Bigdata solutions on Intel® platforms
  - Open source leadership @Spark*, Hadoop*, OpenStack*, Ceph* etc.
- Working closely with community and end customers
- Bridging advanced research and real-world applications

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

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Storage

- **Hyper-converged Infrastructure and Hyper-converged storage**
  - "Converged systems are essentially pooled systems comprising the four essential datacenter components – servers, storage, networks, and management software." [1]
  - Hyper-converged infrastructure pushes storage change.



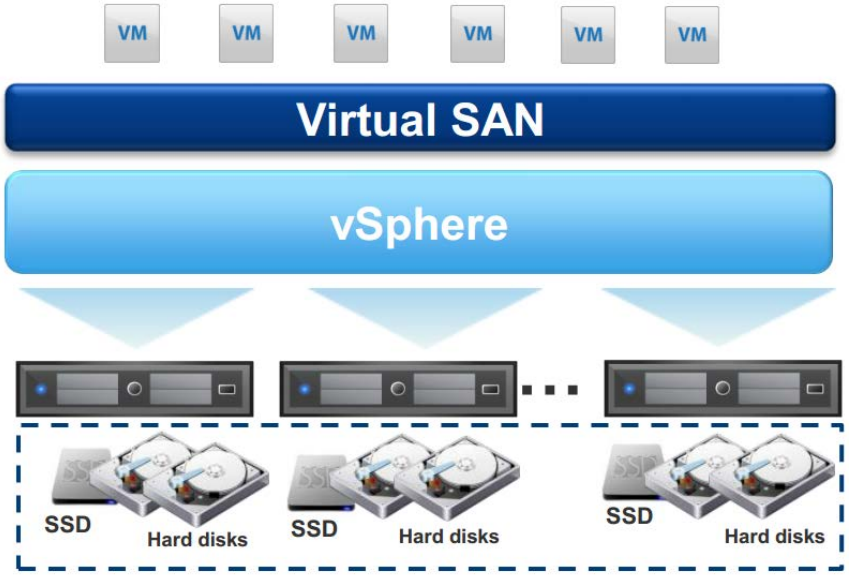
[1] http://idc-cema.com/eng/trendspotter/62716-hyper-convergence-when-converged-systems-grow-up
[PICTURE Source] http://blogs.vmware.com/virtualblocks/2015/05/29/20-common-vsan-questions/

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Storage

- **Managing VMs and not storage**
  - All storage actions are taken on a per virtual machine basis rather than having to understand LUNs, RAID groups, storage interfaces, etc.



[PICTURE Source] http://www.storagenewsletter.com/rubriques/software/tintri-os-3-2-global-center-2-0-and-syncvm-available/

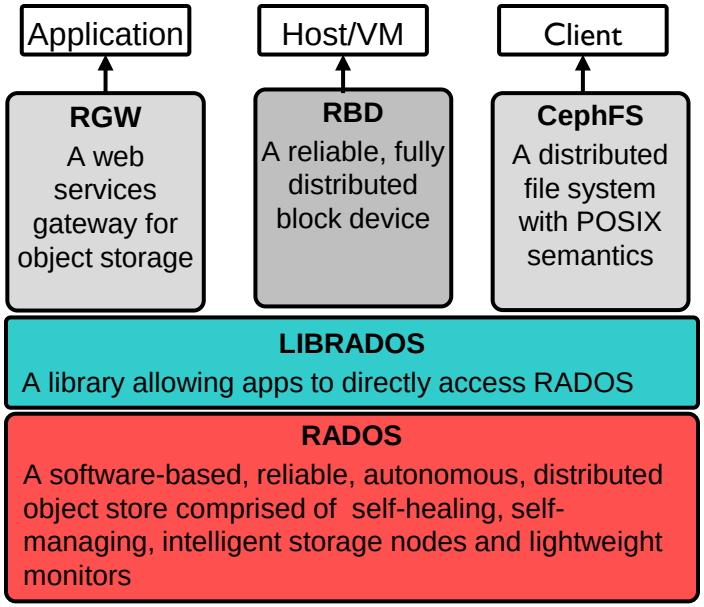
Intel does not control or audit third-party info or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Ceph*: OpenStack* de fecto storage backend[1]

Ceph* is an open-source, massively scalable, software-defined storage system that provides object, block and file system storage in a single platform. It runs on commodity hardware—saving you costs and giving you flexibility—and because it's in the Linux* kernel, it's easy to consume.

- Object store (RADOSGW)
  - A bucket-based REST gateway
  - Compatible with S3 and swift
- File system (CEPHFS)
  - A POSIX-compliant distributed file system
  - Kernel client and FUSE
- Block device service (RBD)
  - OpenStack* native support
  - Kernel client and QEMU*/KVM driver



[1] https://www.openstack.org/summit/openstack-summit-hong-kong-2013/session-videos/presentation/ceph-the-de-facto-storage-backend-for-openstack

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

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Gap on OpenStack* Storage

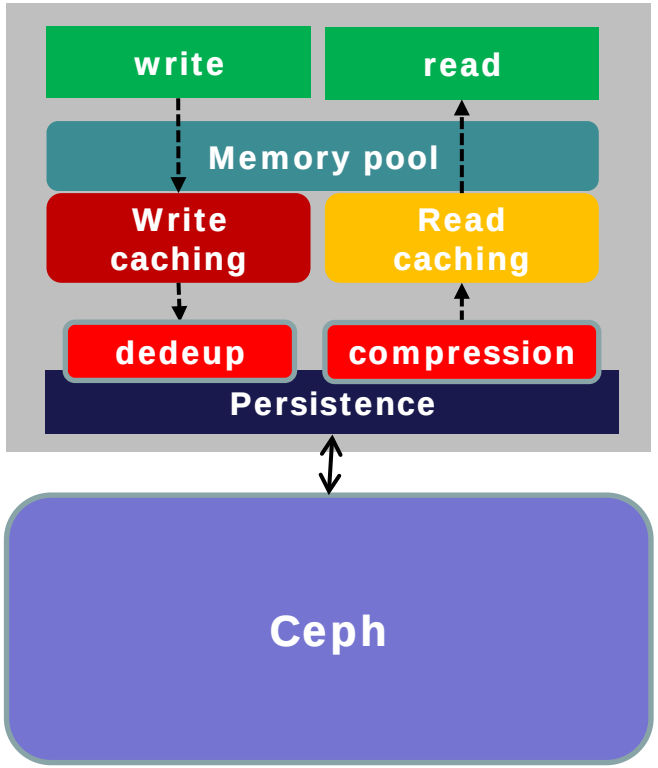
- A strong demands for SSD caching in Ceph* cluster
- Ceph* SSD caching performance has gaps
  - Cache tiering, Flashcache/bCache not work well
- OpenStack* storage lacks a caching layer

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

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Overview

- Building a hyper-converged cache solutions for the cloud
  - **Started with** Ceph*
  - **Block cache**, object cache, file cache
- Extensible Framework
  - Pluggable design/cache policies
  - General caching interfaces: Memcached like API
  - Support third-party caching software
- Advanced data services:
  - Compression, deduplication, QOS
- Value added feature for future SCM device



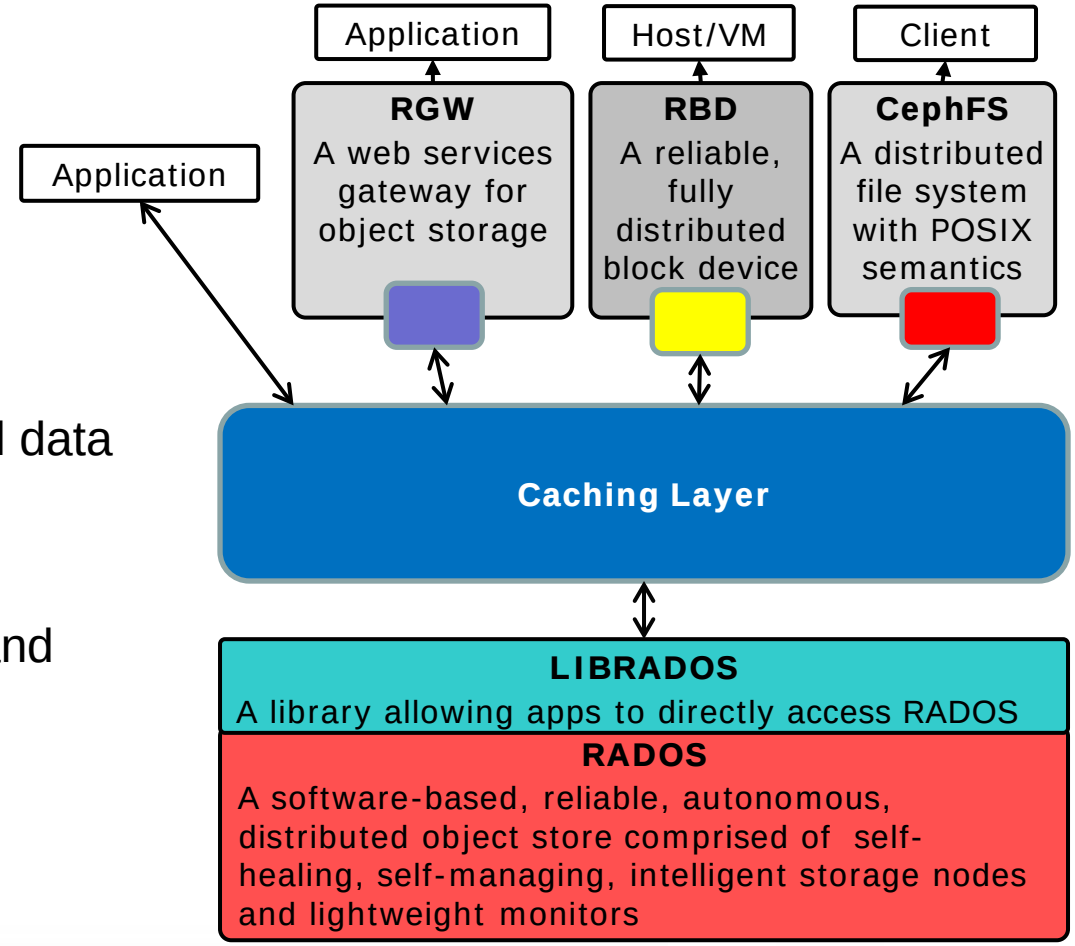
*Other names and brands may be claimed as the property of others.

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: different adapters

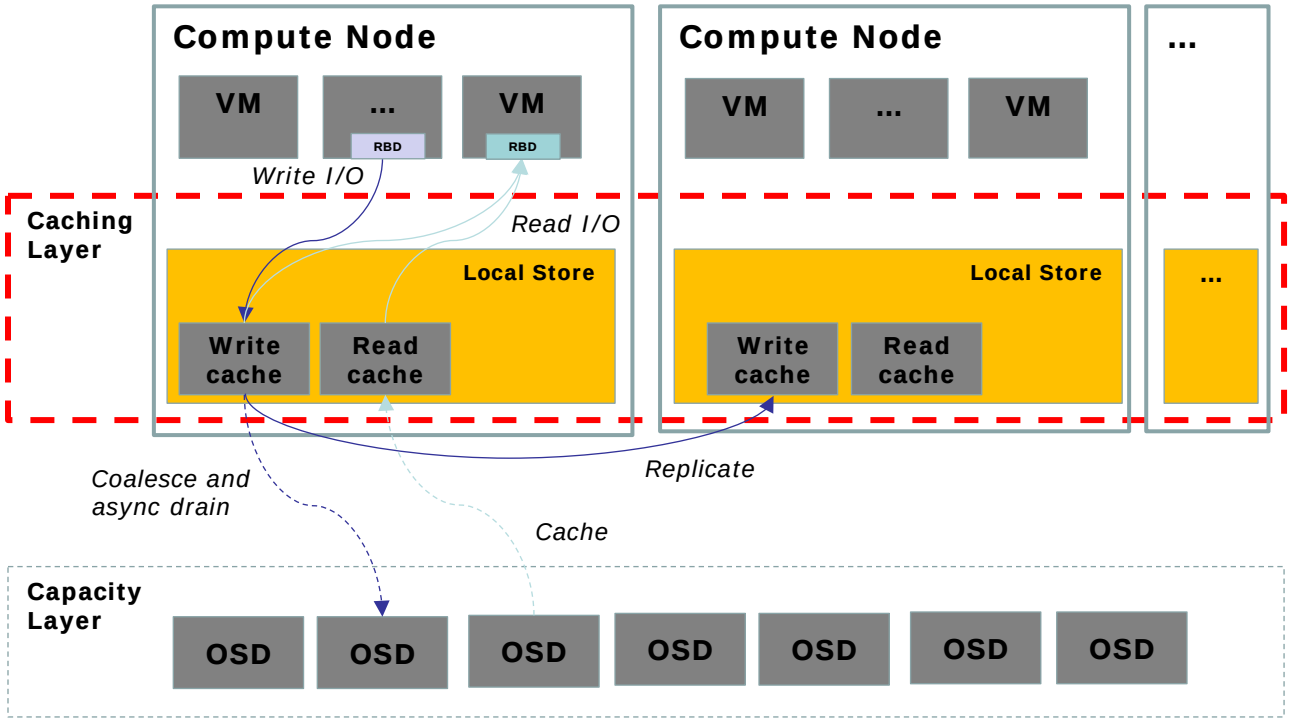
- **RBD**:
  - Hooks on librbd
  - caching for small writes
- **RGW**:
  - Caching over http
  - For metadata and small data
- **CephFS**:
  - Extend POSIX API
  - Caching for metadata and small writes



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

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Design Details
block cache details(1)

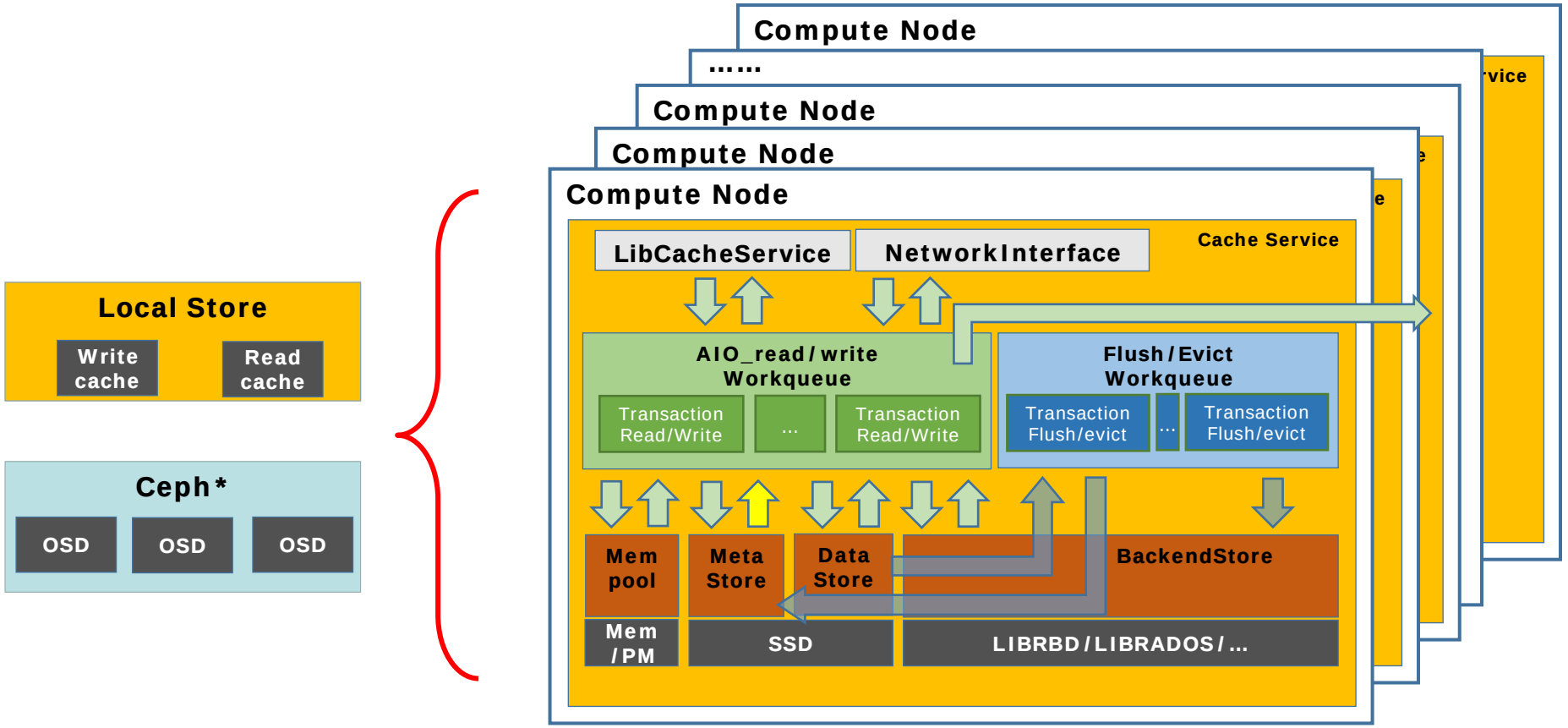


- Hyper-converged deployment
- Also, support deduped read cache and persistent write cache for VM scenario.

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Design Details
block cache details(2)



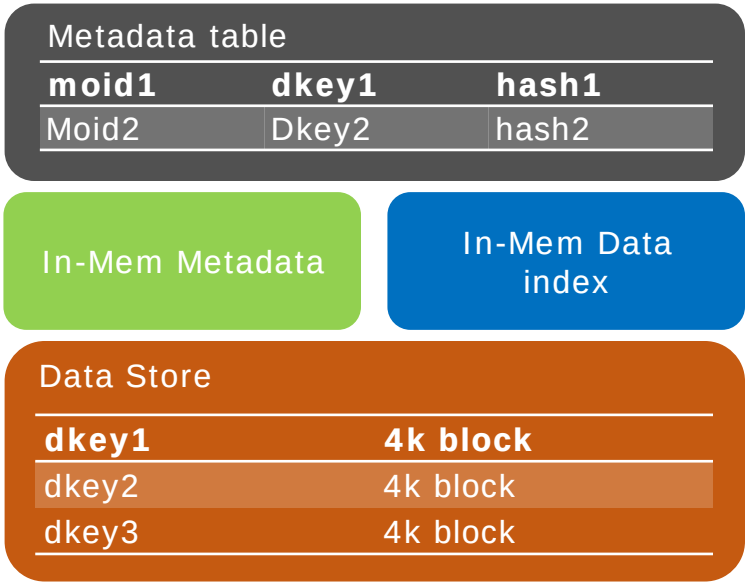
- Transactional read/write support
- Differential service for each RBD

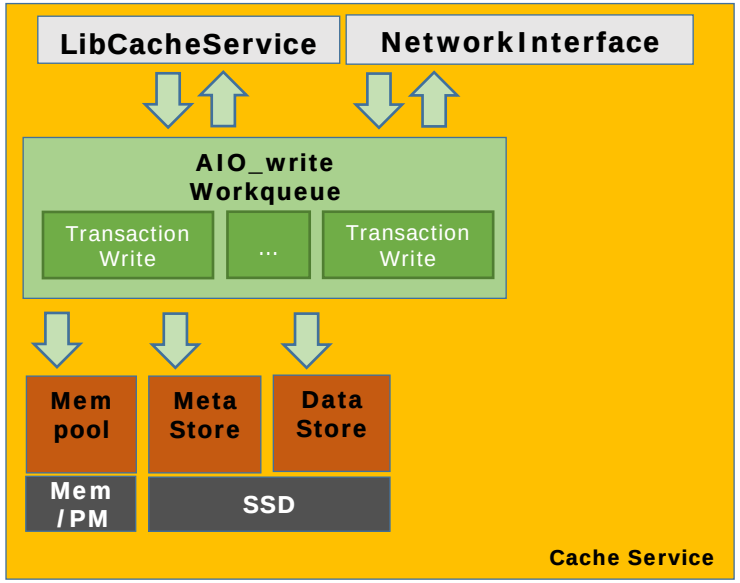
*Other names and brands may be claimed as the property of others.

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Design Details
block cache details(2)



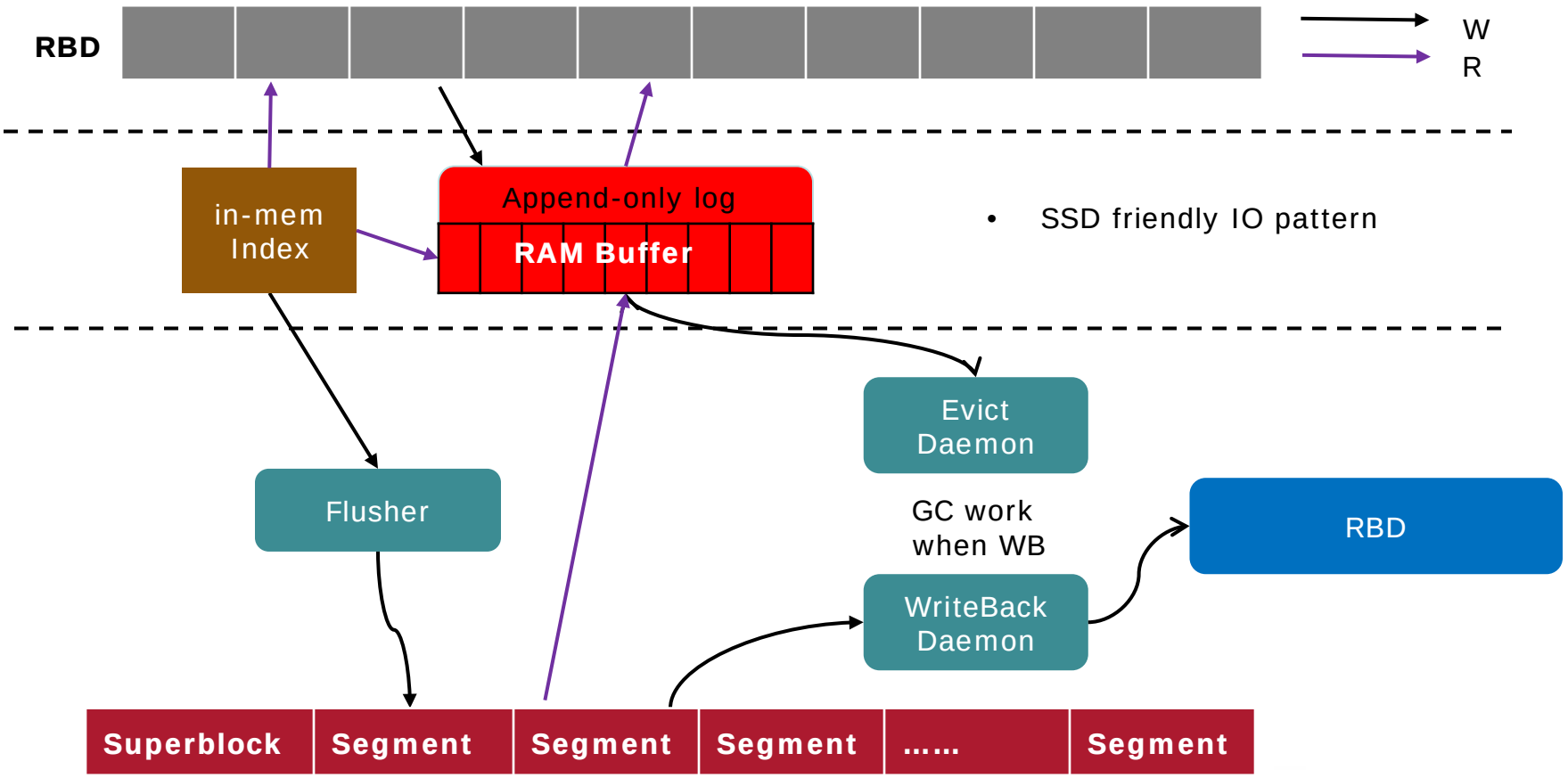


- Write cache is using log appending.
- On each write request, persistent the data into free slots on SSD, and update the metadata table
- if it's in the read cache, will also invalidate that entry

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

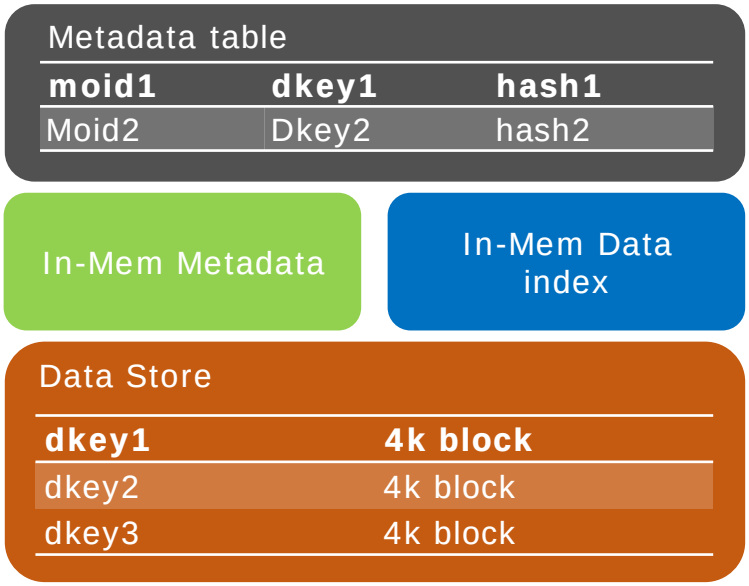
Hyper Converged Cache: Data Store

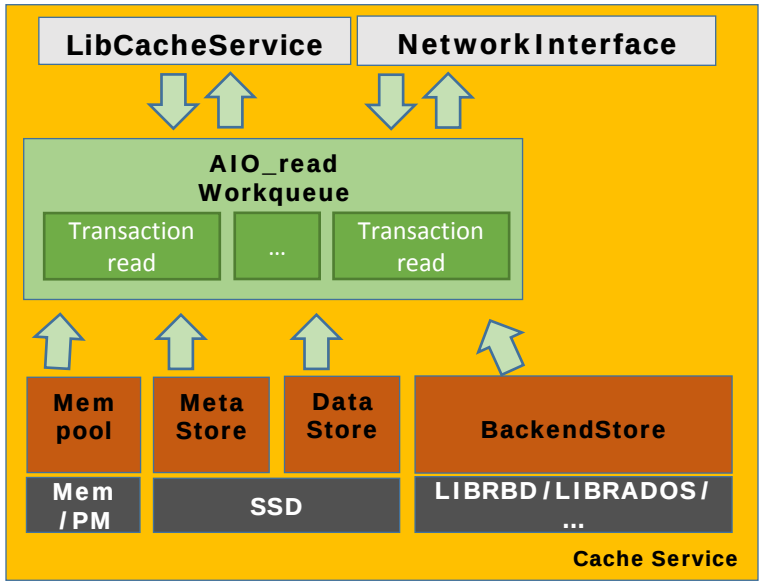


SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Read Cache

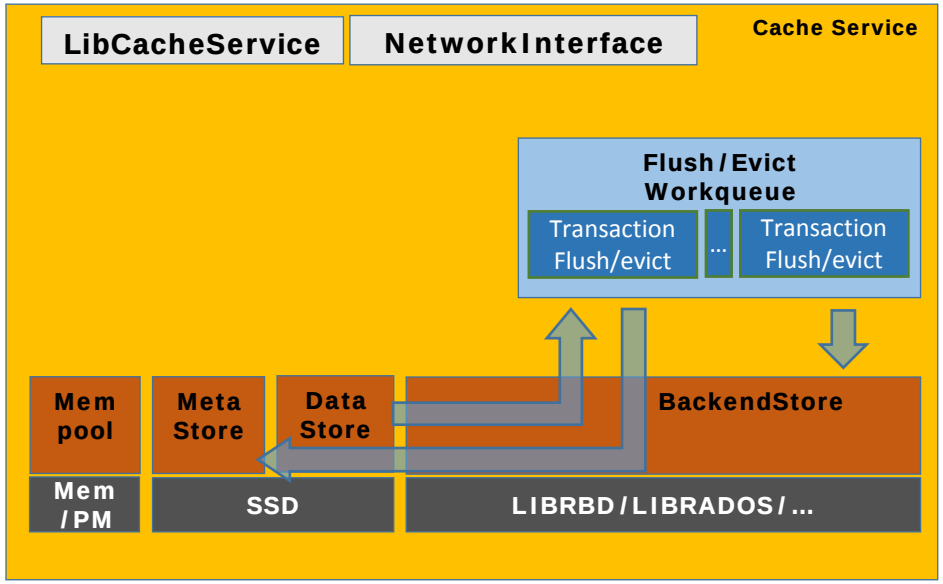




- Read cache is CAS (content-addressable storage) and stores hash/value combinations on SSD or flash storage.
- On each read request, look up hash in the metadata table first
- If miss, then go to look up in the write-cache
- Go to Ceph cluster if miss again

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Flush & Evict



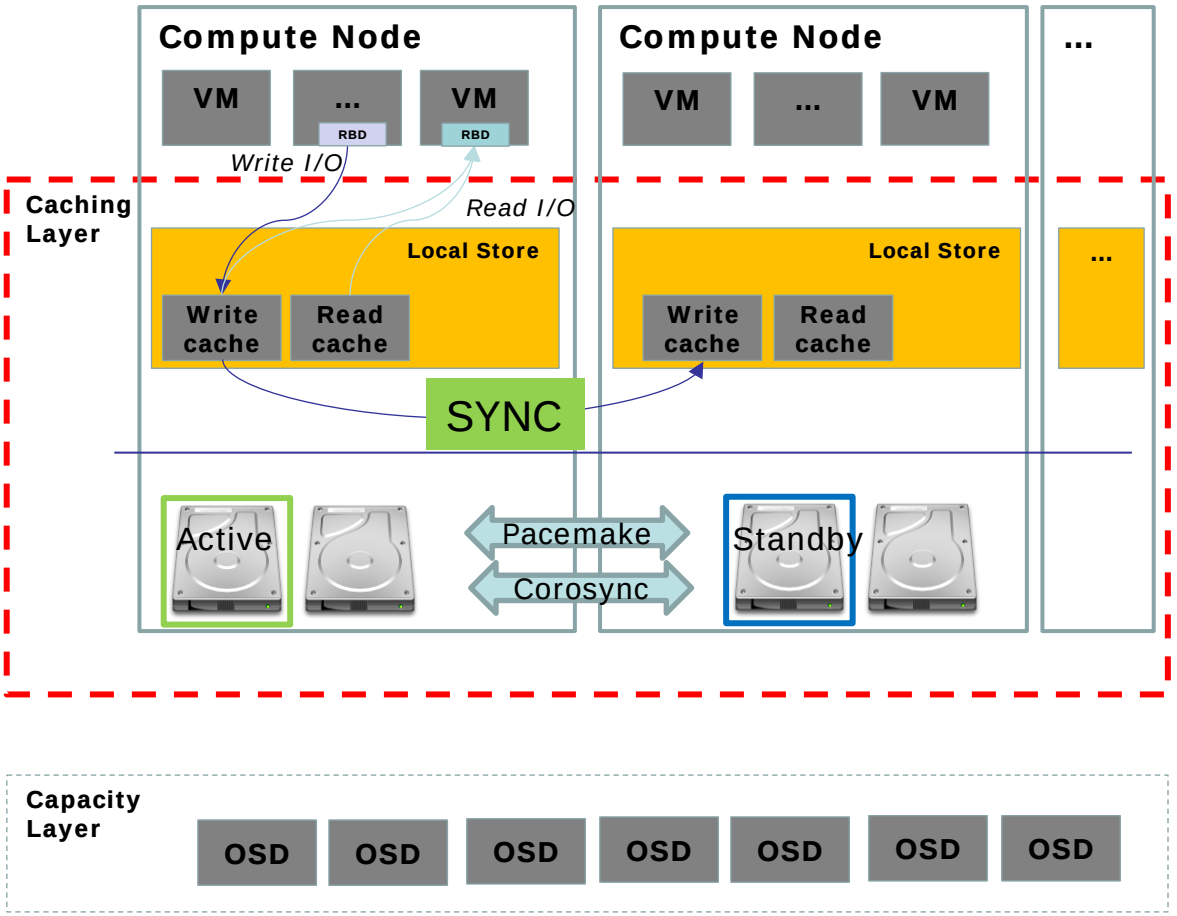
- Cache Service will automatically flush the cached contents to Ceph cluster as the cache_ratio reaches certain value.
- Based on LRU, the hot data will be kept in cache

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Failover & Recovery

- Master/Slave architecture
  - Two hosts are required in order to provide physical redundancy
- The cache layer will run into read-only state if master fails
  - All cached writes will be flushed to Ceph
  - All writes will be written to Ceph directly
  - Also can cache writes if only single copy of cache is acceptable.
- Pacemaker* + corosync* to handle system availability

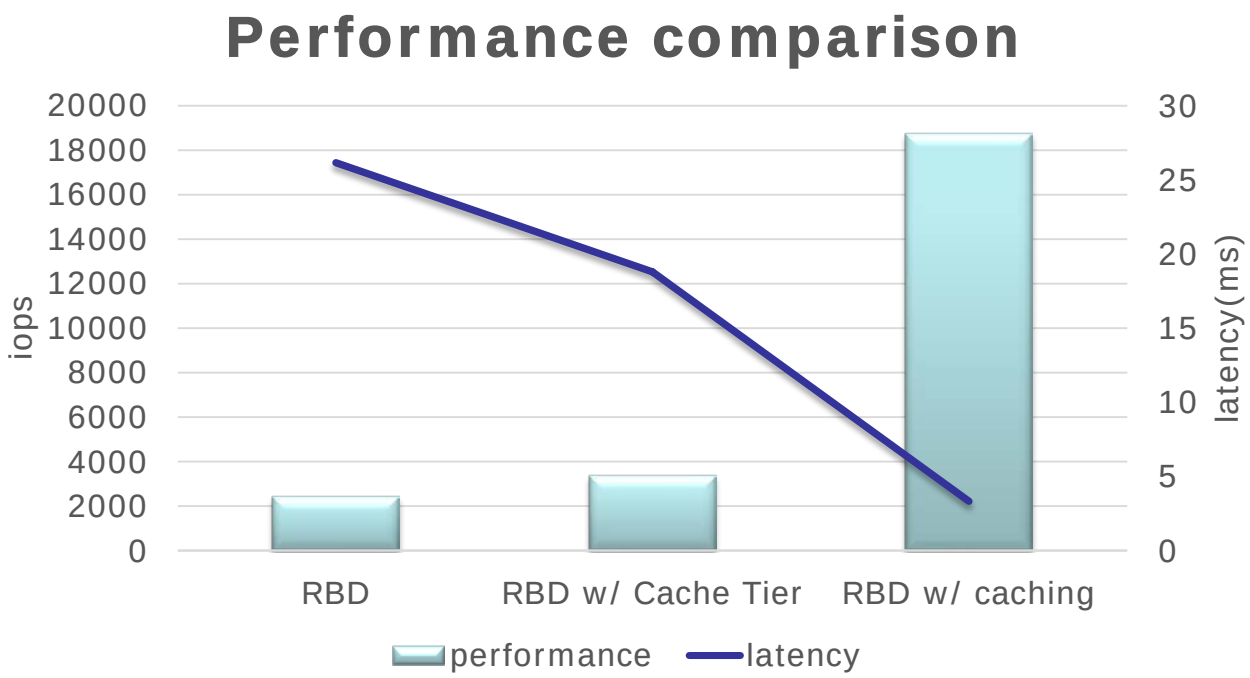


*Other names and brands may be claimed as the property of others.

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache: Performance
Overview



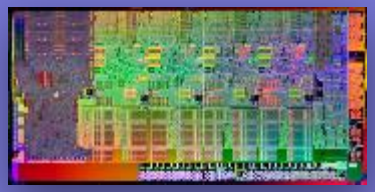
- Hyper converged cache is able to provide ~7x performance improvements w/ zipf 4k randwrite, the latency also decreased ~92%.
- Comparing with cache tier, the performance improved ~5x, the code path is much simpler.

Performance numbers are Intel Internal estimates
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

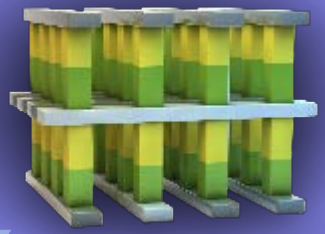
3D XPoint™ Technology

SRAM
Latency: 1X
Size of Data: 1X

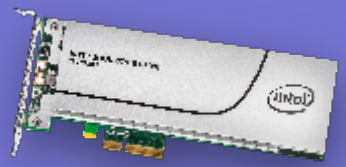


DRAM
Latency: ~10X
Size of Data: ~100X

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



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



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

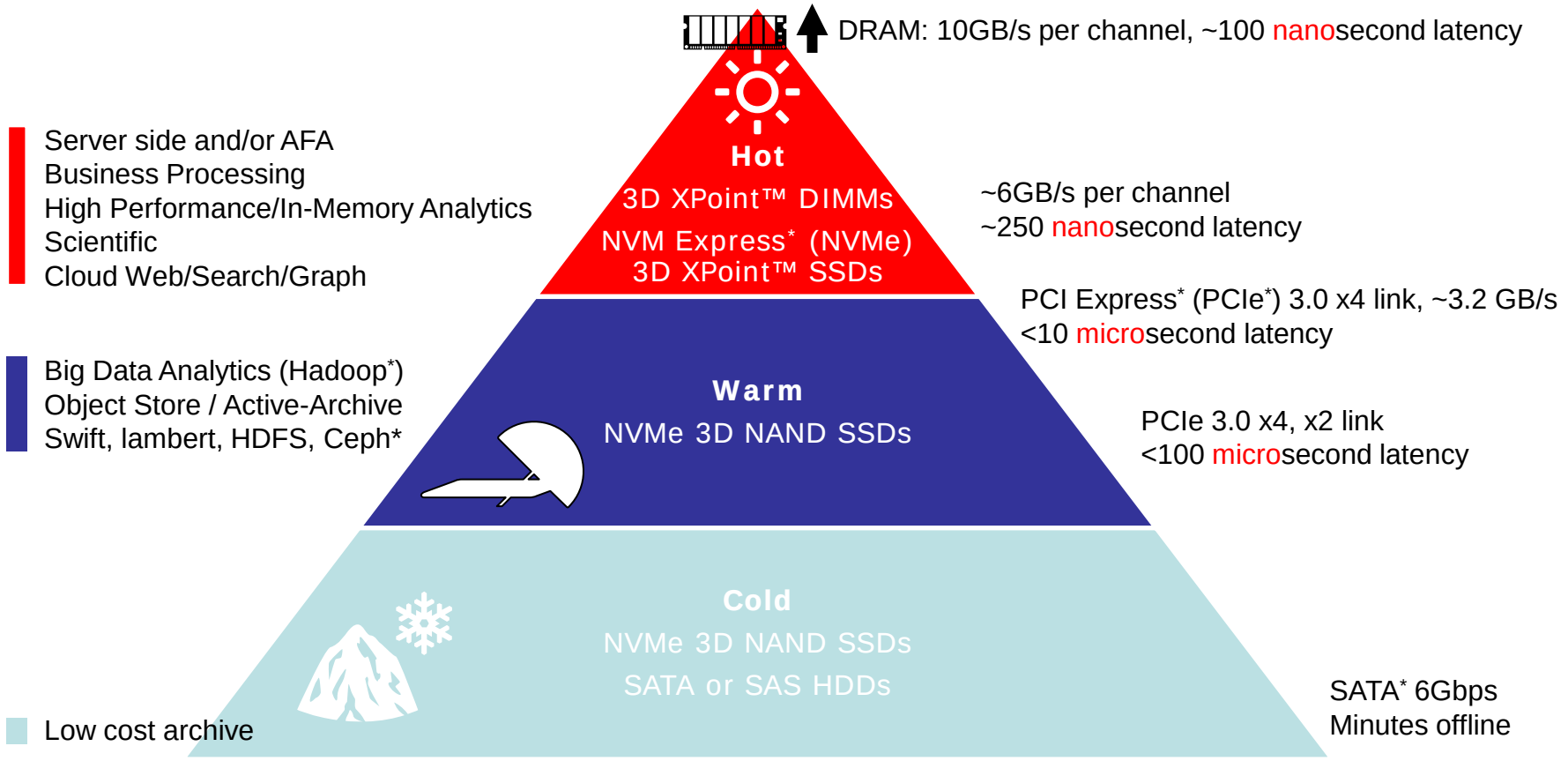
MEMORY

STORAGE

Performance numbers are Intel Internal estimates
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries
[1] http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160810_K21_Zhang_Zhang_Zhou.pdf
Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

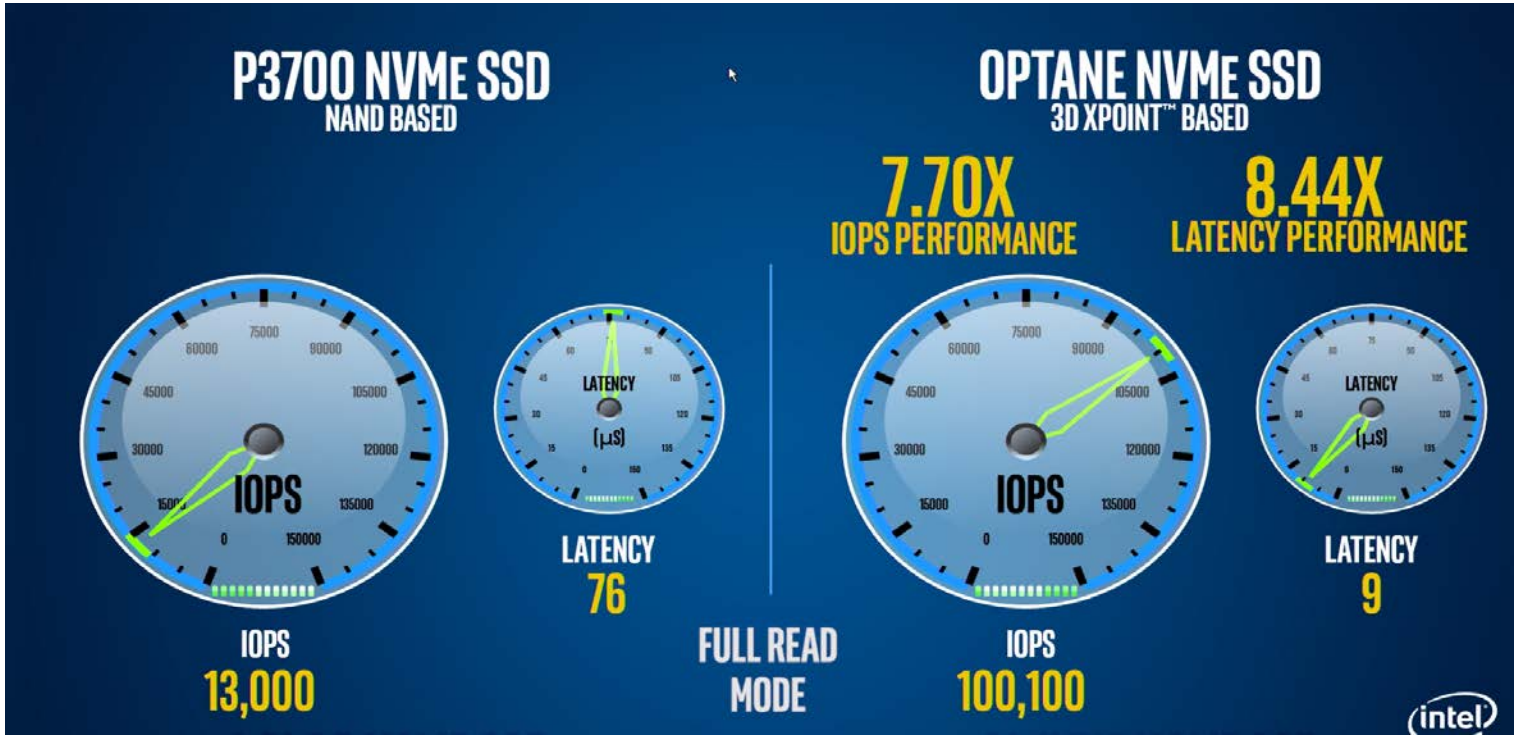
Storage Hierarchy Tomorrow



Performance numbers are Intel Internal estimates
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries
[1] http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160810_K21_Zhang_Zhang_Zhou.pdf
Comparisons between memory technologies based on in-market product specifications and internal Intel specifications.

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

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

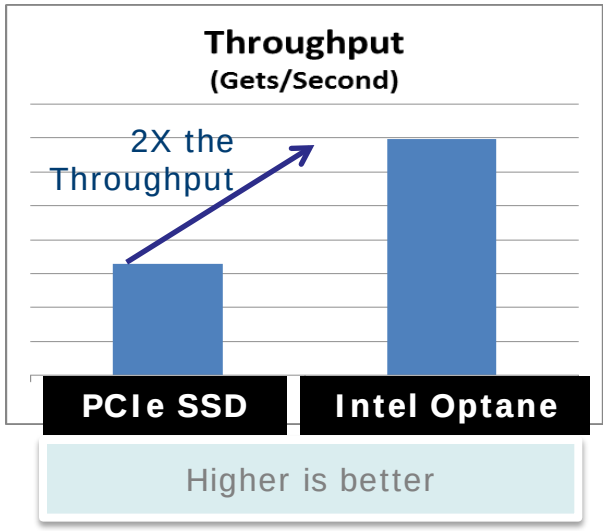


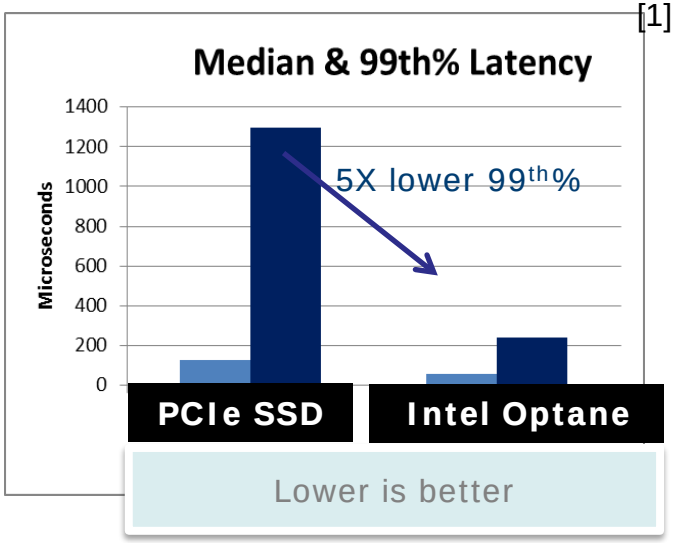
Performance numbers are Intel Internal estimates
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries
[1] http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160810_K21_Zhang_Zhang_Zhou.pdf

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

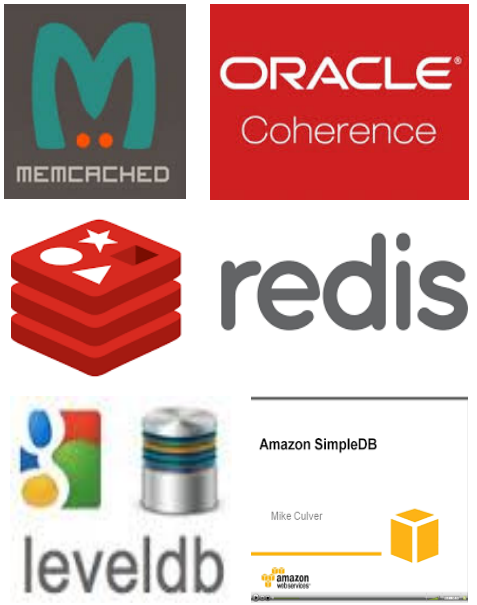
2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Intel® Optane™ shows significant performance improvement over PCIe SSD for RocksDB* Key/Value cloud benchmark*





[1]



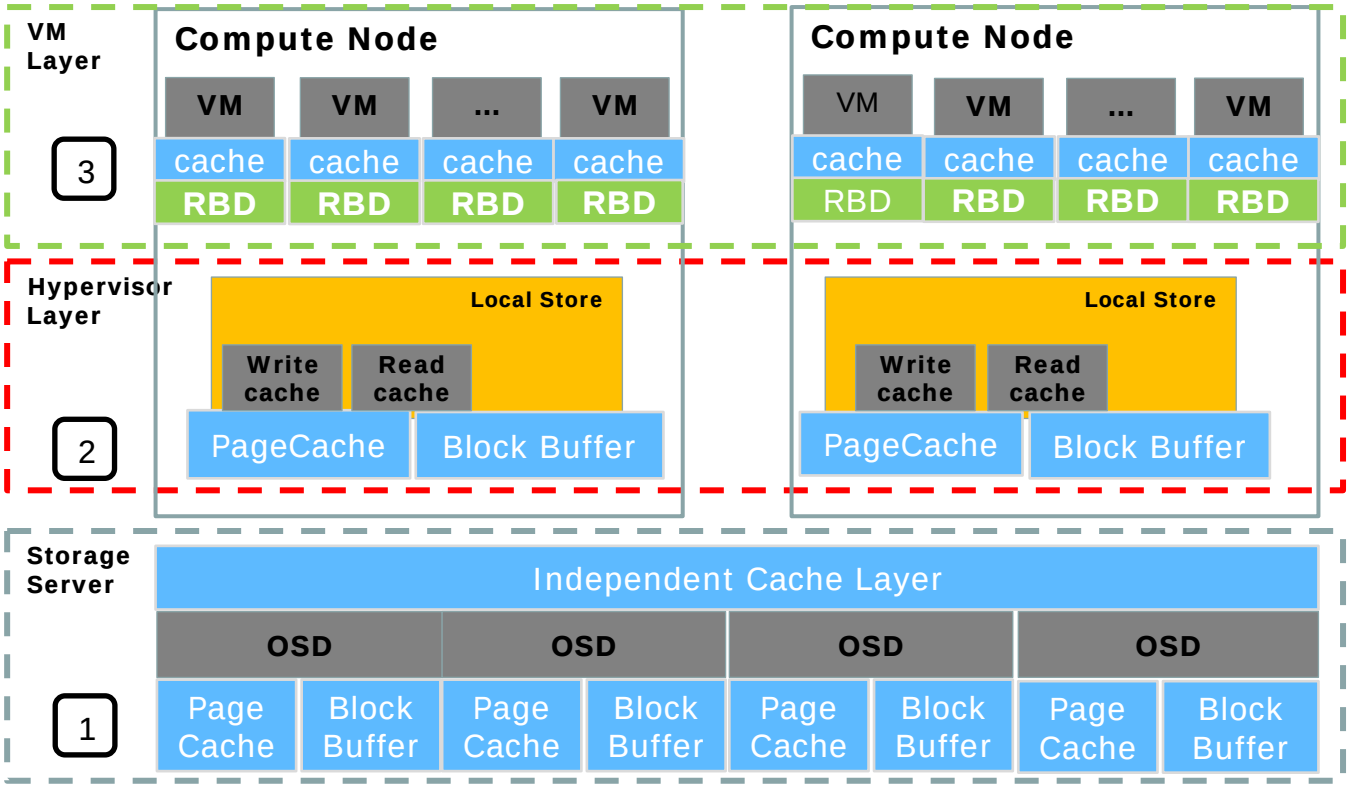
Performance numbers are Intel Internal estimates
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries

[1] http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160810_K21_Zhang_Zhang_Zhou.pdf

*Benchmarked on early prototype samples, 2S Haswell/Broadwell Xeon platform single server. Data produced without any tuning. We expect performance to improve with tuning.
*Other names and brands may be claimed as the property of others.

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Hyper Converged Cache with 3D XPointTM technology



- Using Intel® Optane™ device as block buffer cache device.
- Using Intel® Optane™ device as page caching device.
- Using 3D XPoint™ device as OS L2 memory?

Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Summary

- Hyper Converged Cache provides ~6x performance improvements, w/ ~92% latency reduce.
- With the emerging new media like 3D-XPoint™, the caching benefit will be more higher
- Next step:
  - Tests on objects and filesystem

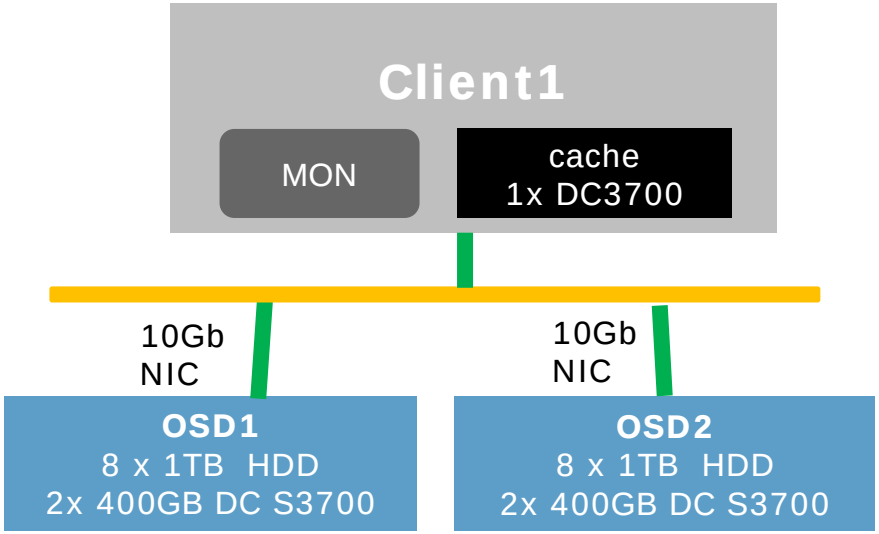
Performance numbers are Intel Internal estimates
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Intel and Intel logos are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Backup

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

H/W Configuration



Client Cluster	
CPU	Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.80GHz
Memory	96 GB
NIC	10Gb
Disks	1 HDD for OS 400G SSD for cache
Ceph Cluster	
CPU	OSD: Intel(R) Xeon(R) CPU E31280 @ 3.50GHz
Memory	32 GB
NIC	10GbE
Disks	2 x 400 GB SSD (Journal) 8 x 1TB HDD (Storage)

- 2 hosts Ceph cluster each host has 8 x 1TB HDD as OSDs and 2x Intel® DC S3700 SSD journal
- 1 Client with 1x 400GB Intel® DC S3700 SSD as cache device

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

S/W Configuration

- Ceph* version : 10.2.2 (Jewel)
- Replica size : 2
  - Data pool : 16 OSDs. 2 SSDs for journal, 8 OSDs on each node
  - OSD Size : 1TB * 8
  - Journal Size : 40G * 8
  - Cache: 1 x 400G Intel® DC S3700
  - FIO volume size: 10G
- Cetune test benchmark
  - fio + librbd

Cetune: https://github.com/01org/cetune

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

SDC 16

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Testing Configuration

- Test cases:
  - Operation: 4K random write with fio (zipf=1.2)
- Detail case:
  - Cache size < volume size (w/ zipf)
    - w/o flush & evict: cache size 10G.
    - w/ flush w/o evict: cache size 10G.
    - w/ flush & evict: cache size 10G.
  - Hot data = volume size * zipf1.2(5%), runtime = 4 hours
- Caching Parameters:
  - object_size=4096
  - cache_flush_queue_depth=256
  - cache_ratio_max=0.7
  - cache_ratio_health=0.5
  - cache_dirty_ratio_min=0.1
  - cache_dirty_ratio_max=0.95
  - cache_flush_interval=3
  - cache_evict_interval=5
  - Runtime: Base: 200s ramp up, 14400s run
  - DataStoreDev=/dev/sde
  - cache_total_size=10G
  - cacheservice_threads_num=128
  - agent_threads_num=32

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.

Legal Notices and Disclaimers

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].

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.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

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.

Intel, the Intel logo, Xeon, 3D-XPoint™ are trademarks of Intel Corporation in the U.S. and/or other countries.

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

2016 Storage Developer Conference. © 2016 Intel Corp. All Rights Reserved.