INTEL NVM TECHNOLOGY AND SOLUTION EVOLUTIONS

Benny Ni, Strategic Business Development Manager, NSG, Intel
Ping She, Strategic Planner, NSG, Intel

NOTICES AND DISCLAIMERS

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 product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

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. For more complete information visit http://www.intel.com/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

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.

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.

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.

AGENDA

- Intel SSD Update – Benny
- Caching with OCF bdev – Ping
- Intel® Volume Management Device – Ping

New in SPDK v19.01

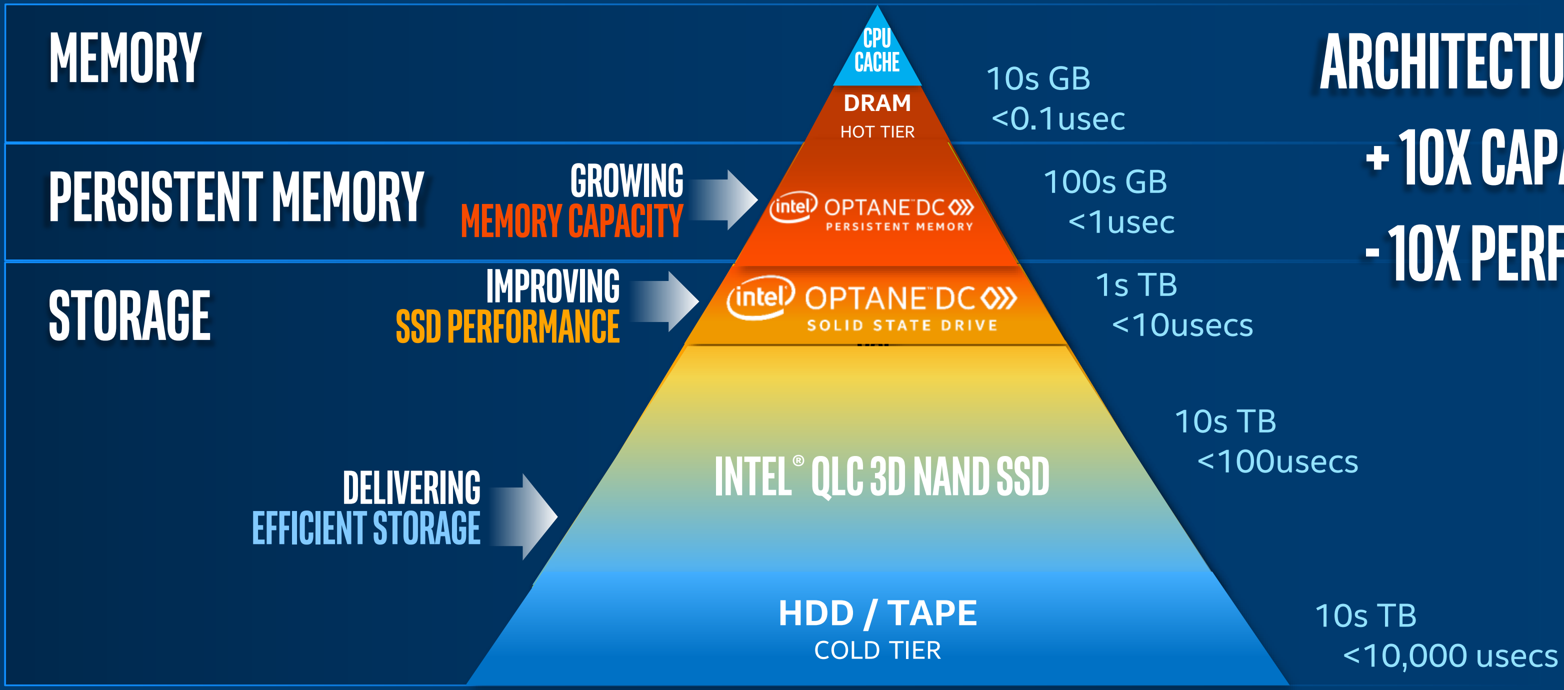
New in SPDK v19.07

INTEL SSD UPDATE

CONNECTED PLATFORM

MEMORY & STORAGE

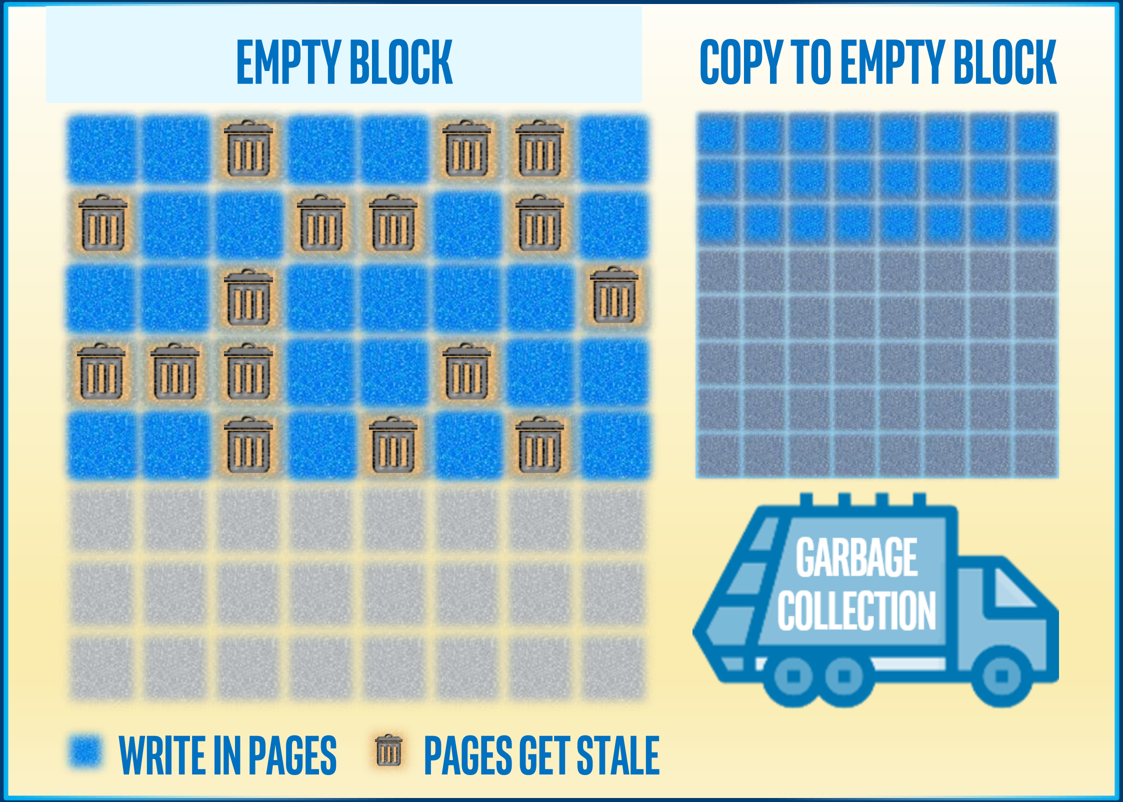
INTEL® OPTANE™ TECHNOLOGY AND INTEL® QLC 3D NAND ARE REVOLUTIONIZING THE MEMORY & STORAGE HIERARCHY



ARCHITECTURE GUIDELINE:

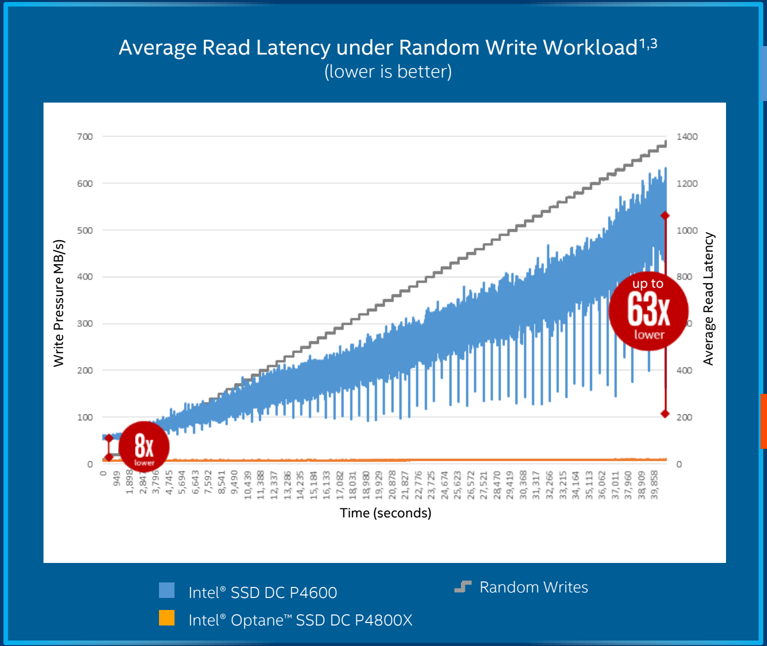
+ 10X CAPACITY
- 10X PERFORMANCE

EVERY NAND SSD IS SLOWED BY GARBAGE COLLECTION

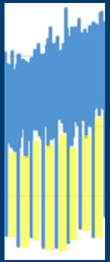


Permission to use truck icon made by monkik from www.flaticon.com with reference

INTEL® OPTANE™ DC TECHNOLOGY ELIMINATES GARBAGE COLLECTION



NAND: INCONSISTENT PERFORMANCE



INTEL® OPTANE™ DC SSD: CONSISTENT PERFORMANCE

[1] Source – Intel-tested: Response Time refers to average read latency measured at queue depth 1 during 4k random write workload. Measured using FIO 3.1. Common Configuration - Intel 2U Server System, OS CentOS 7.5, kernel 4.17.6-1.el7.x86_64, CPU 2 x Intel® Xeon® 6154 Gold @ 3.0GHz (18 cores), RAM 256GB DDR4 @ 2666MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P4600 1.6TB. Latency – Average read latency measured at QD1 during 4K Random Write operations using FIO 3.1. Intel Microcode: 0x2000043; System BIOS: 00.01.0013; ME Firmware: 04.00.04.294; BMC Firmware: 1.43.91f76955; FRUSDR: 1.43. SSDs tested were commercially available at time of test. The benchmark results may need to be revised as additional testing is conducted. Performance results are based on testing as of July 24, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. 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. For more complete information visit www.intel.com/benchmarks.

Permission to use truck icon made by monkik from www.flaticon.com with reference

INTEL® OPTANE™ DC TECHNOLOGY

ACCELERATE SOLUTIONS INNOVATIONS

STORAGE

RDMA/Replication
Oracle Exadata*

SDS
Ceph*

INFRASTRUCTURE

Memory
VMWare ESXi*
MSFT Hyper-V*
KVM[1]
Redis Labs
Memcached
VDI

DATABASE

Caching/Persistence
SAP HANA*
MS-SQL[1]
Oracle Exadata*
Aerospike*
Redis*
RocksDB*

AI / ANALYTICS

Real Time Analytics
SAS

Machine Learning Analytics
Apache Spark*[1]

HPC

Scratch & I/O Nodes
HPC Flex Memory

COMMS

Content Delivery Network (CDN)
Comms SP custom

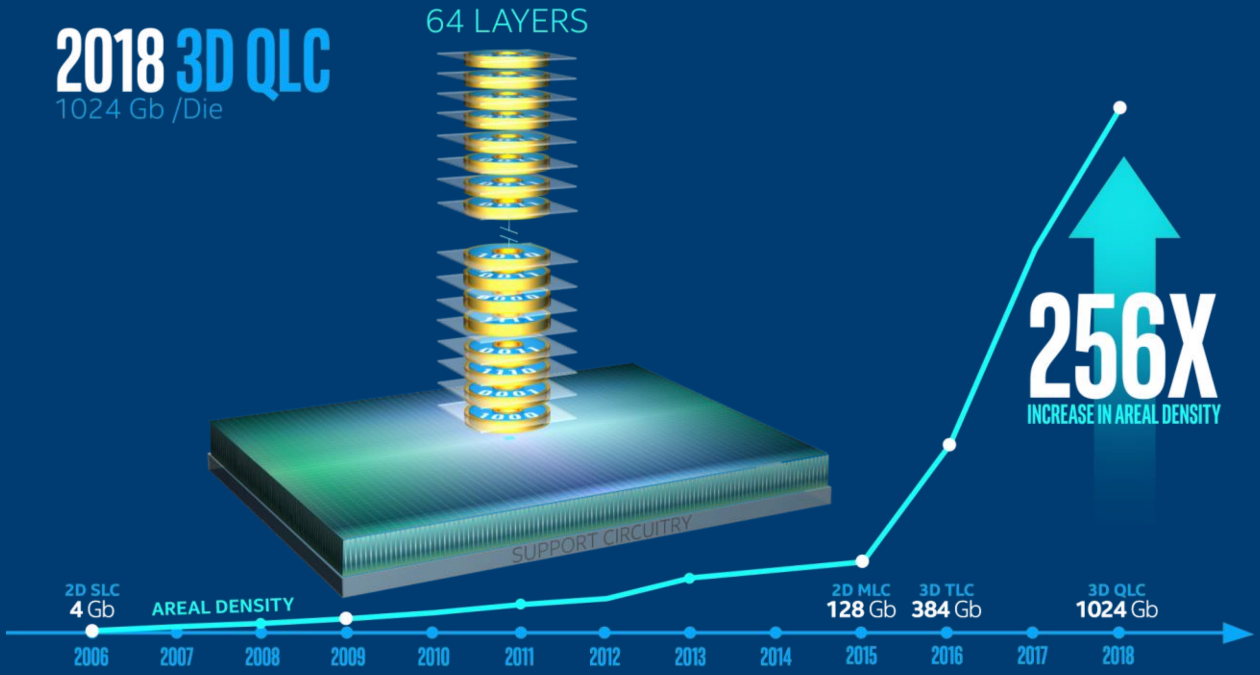
Hyper-Converged (HCI)

STORAGE	MEMORY
VMware vSAN*	VMWare ESXi*
Microsoft S2D*	MSFT Hyper-V*
Nutanix*	KVM*
Cisco HyperFlex*	

intel OPTANE DC PERSISTENT MEMORY
intel OPTANE DC SOLID STATE DRIVE
Greatest Affinity

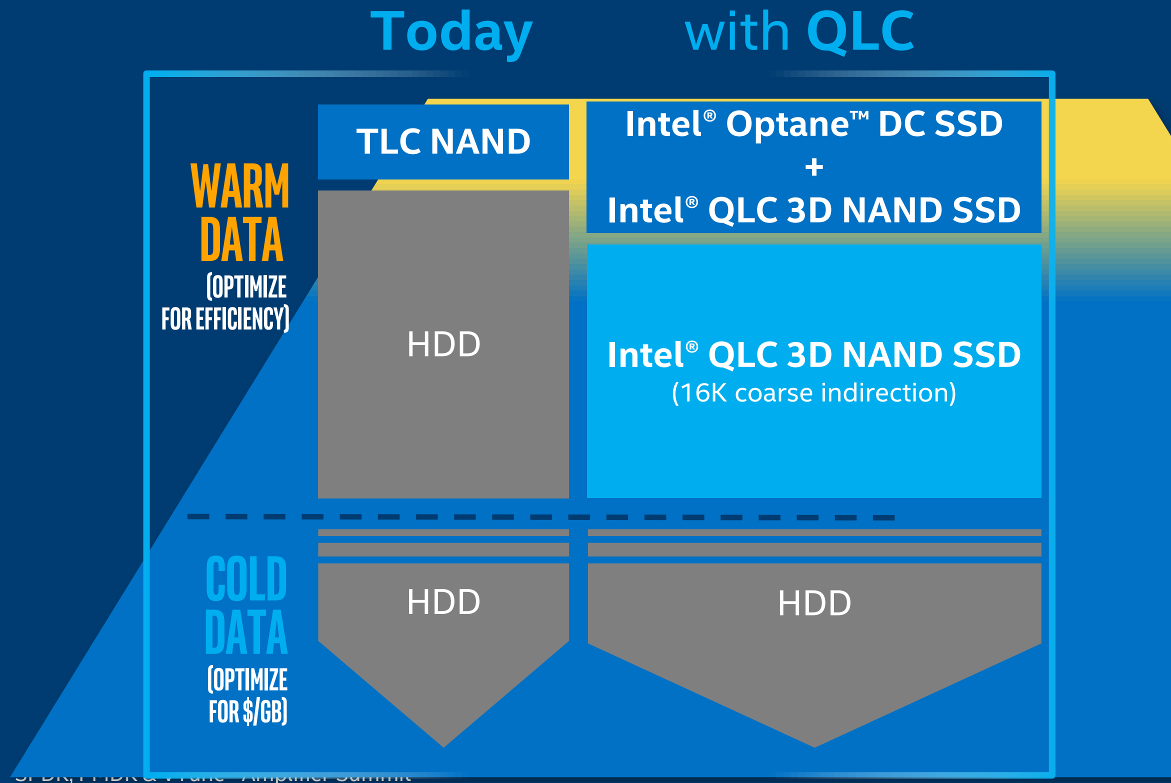
[1] Note: Intel® Optane™ DC Persistent Memory is primary, however NSG is also targeting Intel® Optane™ DC SSD P4800X for this workload. SSD-based workload examples are included in this presentation.
*Other names and brands may be claimed as the property of others

QLC - SCALING FASTER THAN MOORE'S LAW



INTEL® QLC TECHNOLOGY

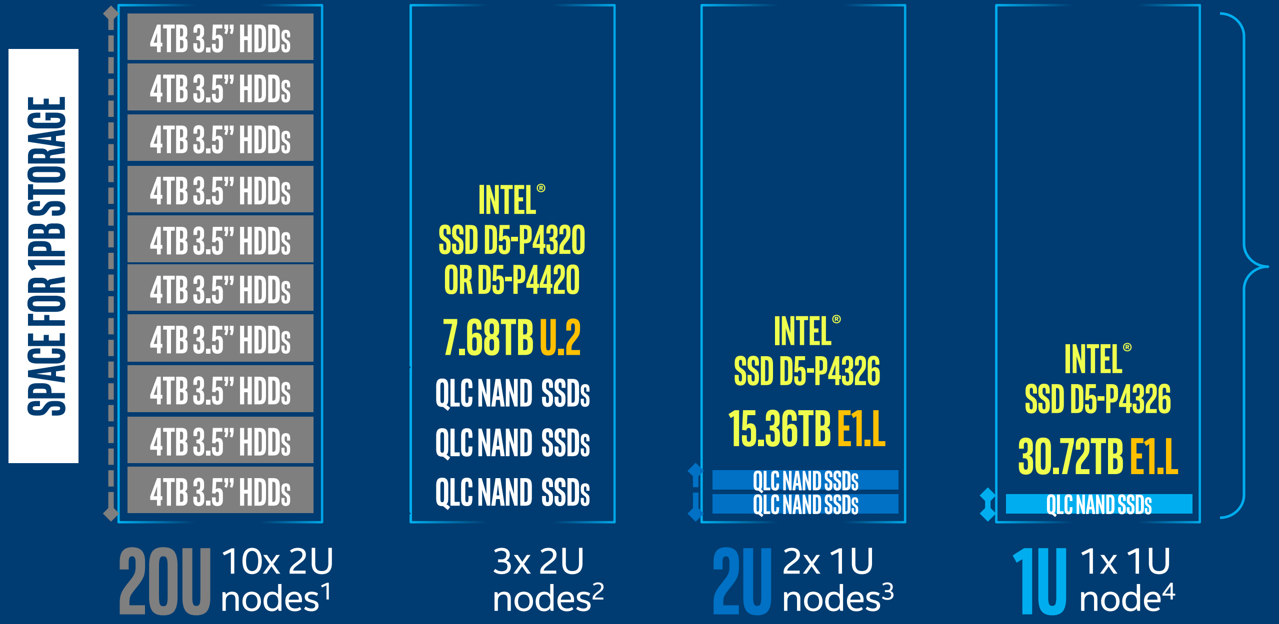
ACCELERATING HDD DISPLACEMENT IN WARM STORAGE



- BETTER PERFORMANCE
- BETTER SCALABILITY
- HIGHER QUALITY
- LESS WEIGHT
- LESS POWER
- LESS SPACE
- BETTER TCO

INTEL® QLC TECHNOLOGY

OFFERING GREATER SCALABILITY AND MAKING IMPOSSIBLE POSSIBLE



Enabling ~1PB in 1U delivers up to 20x reduction of warm storage footprint

INTEL® OPTANE™ AND INTEL® QLC 3D NAND

ACCELERATE STORAGE ARCHITECTURE EVOLUTION

	WORKING DATA	CAPACITY DATA
DATA CENTER/ CLOUD		
	Fast Access	*Large Capacity*

ACCELERATE CEPH*

WITH COMBINATION OF INTEL® OPTANE™ SSDS + INTEL® QLC NAND SSDS

	WAS		IS
CPU	Intel Xeon Gold inside	CPU	Intel Xeon Gold inside
	—	ROCKSDB*/ WAL	Intel® Optane™ DC SSD 1x Intel® Optane™ SSD DC P4800X **750GB**
CAPACITY/ ROCKSDB*/ WAL	Intel® 3D NAND DC SSD 4x Intel® SSD DC P4510	CAPACITY	Intel® QLC NAND Technology 4x Intel® SSD DC P4320

up to **50%** Lower P99 latency[1]

about the **SAME** Storage cost[2]

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. For more complete information visit www.intel.com/benchmarks. Performance results are based on testing as of October, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

[1] Source – Intel tested: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. NVMe configuration overview : Intel® Xeon® Gold 6142 Processor, Intel® SSD DC P4510, BIOS: 00.01.0013; ME: .00.04.294; BMC: 1.43.91f76955; Intel® Optane™ SSD config:: identical with exception of Intel™ Optane® SS DC P4800X for cache/RocksDB/WAL See detailed configurations in Appendix D.

[2] Source – Intel: NAND SSD pricing is estimated as of 10/19/2018 and subject to change.

BAIDU CLOUD AI SOLUTION

INTEL® XEON® PROCESSORS + INTEL® OPTANE™ SSDS +INTEL® QLC 3D NAND SSDS

≡ Gigabit

TOPICS COMPANIES TOP 10

AI FINTECH

Baidu and Intel reveal new collaborations in artificial intelligence

By BEN MOUNCER • Sep 04, 2018, 5:00PM

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

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.

For additional information visit https://newsroom.intel.com/news/baidu-cloud-collaborates-intel-ai/#gs.qsZoQIRc and https://www.gigabitmagazine.com/ai/baidu-and-intel-reveal-new-collaborations-artificial-intelligence

CACHING WITH OCF BDEV

Intel CAS vs Open CAS vs OCF

- Intel CAS (Caching Acceleration Solution)

Validated and supported product, shipping for 5 years. Includes CAS Linux and Windows.

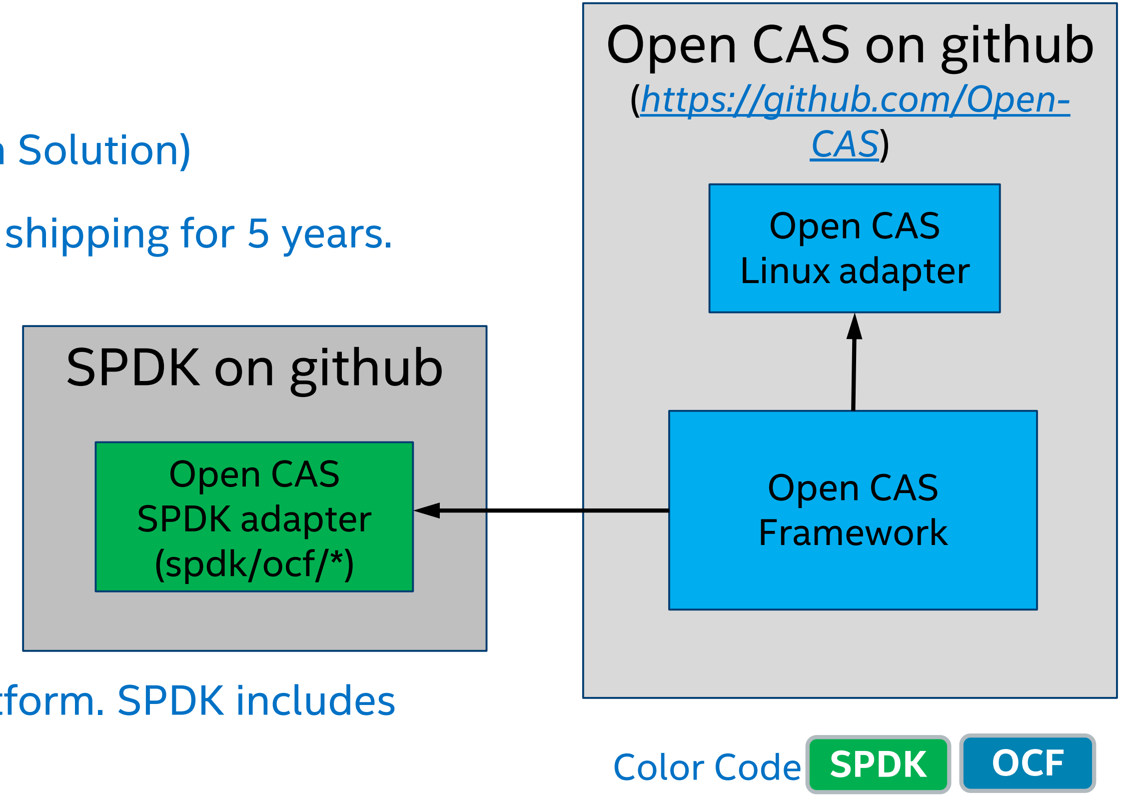
- Open CAS

Open source version of

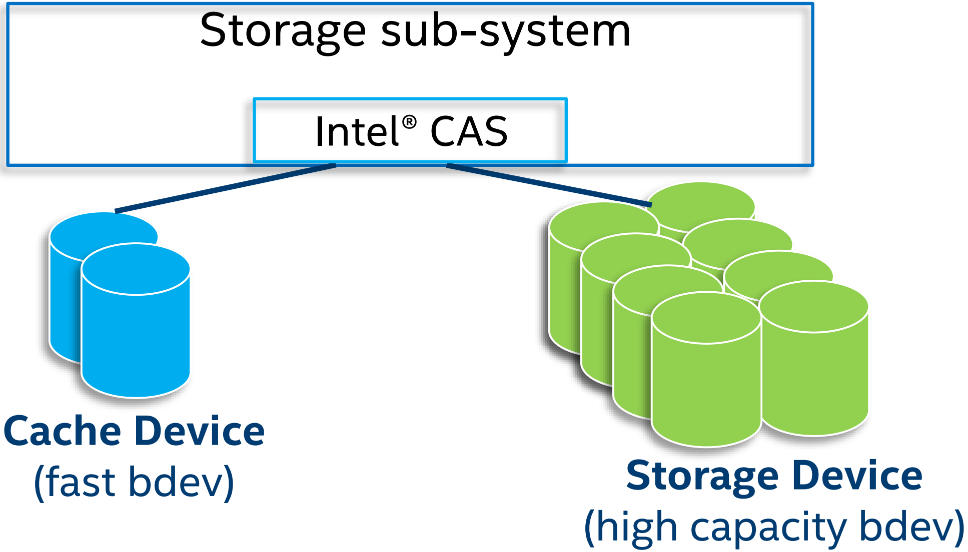
Intel CAS Linux.

- OCF (Open CAS Framework)

Cache engine. Independent of platform. SPDK includes OCF as a submodule.



What is Intel® CAS?



Ex: Intel® Optane™ Device + TLC SSD

Ex: Intel® Optane™ Device + QLC SSD

Ex: TLC SSD + HDDs

Intel CAS Specialty #1

Classify IOs: file system metadata, file and directory name, IO size, data lifetime

Intel CAS Specialty #2

Many modes: write-thru, write back, write around, write only, pinning

Intel® CAS Typical Usages

Application Acceleration

MYSQL* UP TO 5.1X
AS FAST
W/CAS + INTEL® OPTANE™ S

OIL SEARCH* USING CAS UP TO 5X

IKEA* UP TO 2X
AINS

Virtual Machine Acceleration

ALIYUN* VM PERFORMANCE 6.1X
FASTER
W/ CAS & INTEL® OPTANE™

Cloud Storage Acceleration

CEPH* READS UP TO 4.9X
FASTER
WITH CAS + INTEL® OPTANE™ SSD

4.8X
FASTER
WITH CAS + INTEL® OPTANE™ SSD

3X
THROUGHPUT INCREASE

WHITEPAPER

A JOINT PUBLICATION BY INTEL AND TWITTER

Data Center
Hadoop Performance

Boosting Hadoop* Performance and Cost Efficiency with Caching, Fast SSDs, and More Compute

Through experimentation and collaboration, Twitter discovers that increasing the core density of its Hadoop* clusters by 6X would result in 30 percent lower TCO and up to 50 percent faster runtimes[1]

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. See appendix for configuration information. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

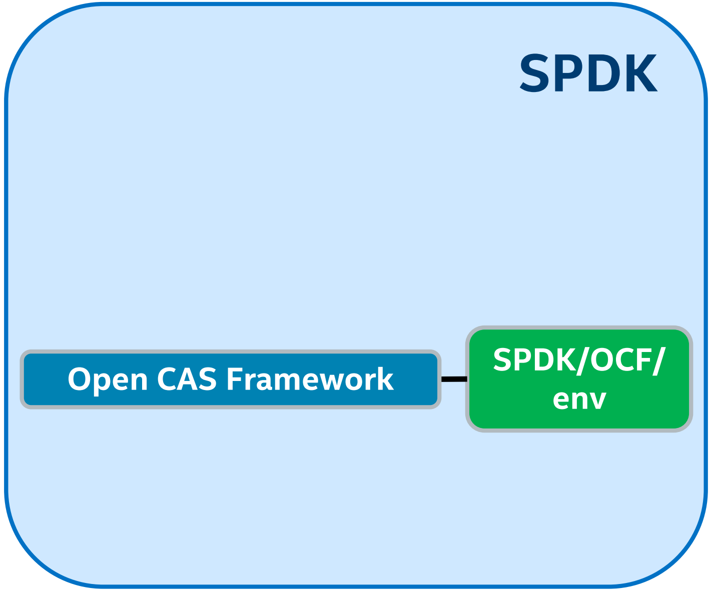
OCF Integration to SPDK - Environment

spdk/lib/bdev/ocf/env/*

OCF env is platform dependent,
enabling it to work in different environments

Requires implementation of:

- Memory allocators
- Logging
- Atomic variables
- Others...

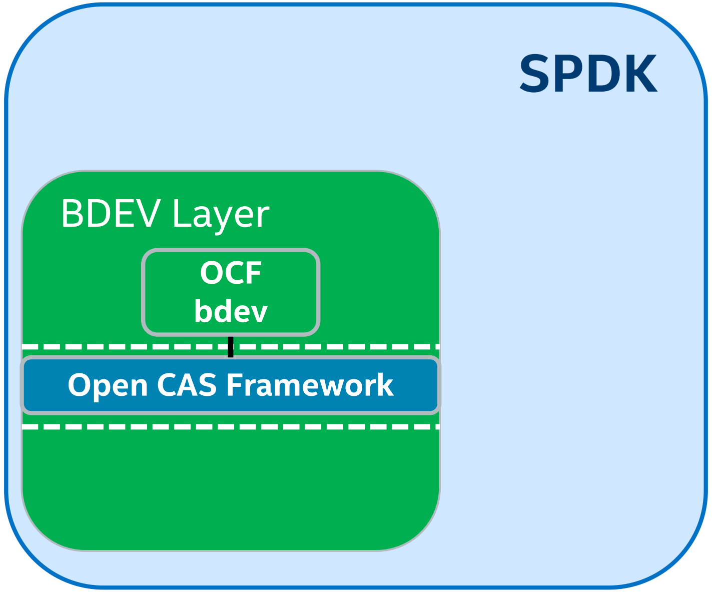


OCF Integration to SPDK - Top Adapter

spdk/lib/bdev/ocf/vbdev_ocf.c

SPDK bdev layer IO operations are passed to cache and core devices using OCF bdev

OCF abstracts access to core device via cache.

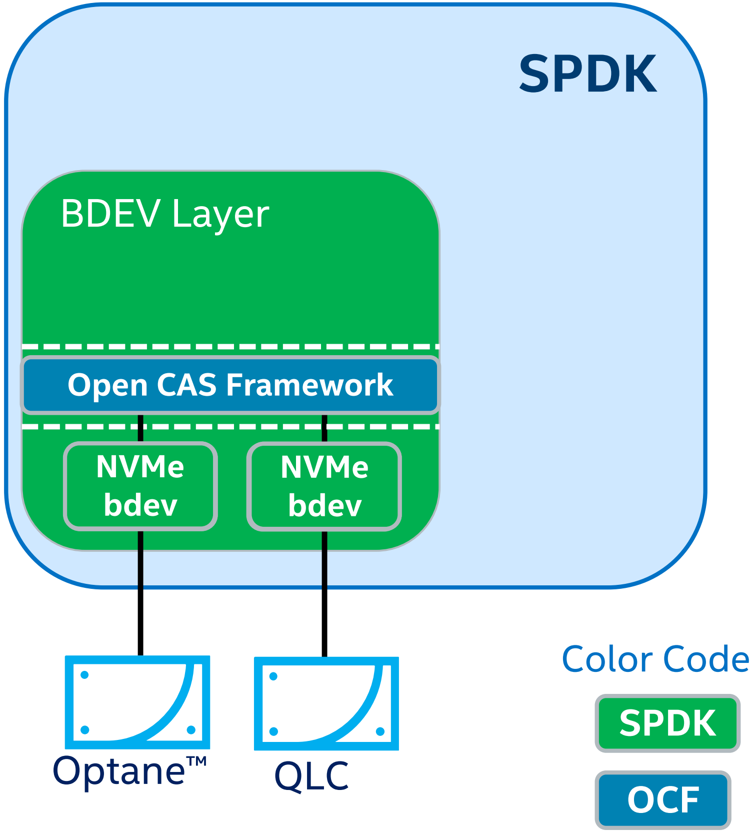


OCF Integration to SPDK - Bottom Adapter

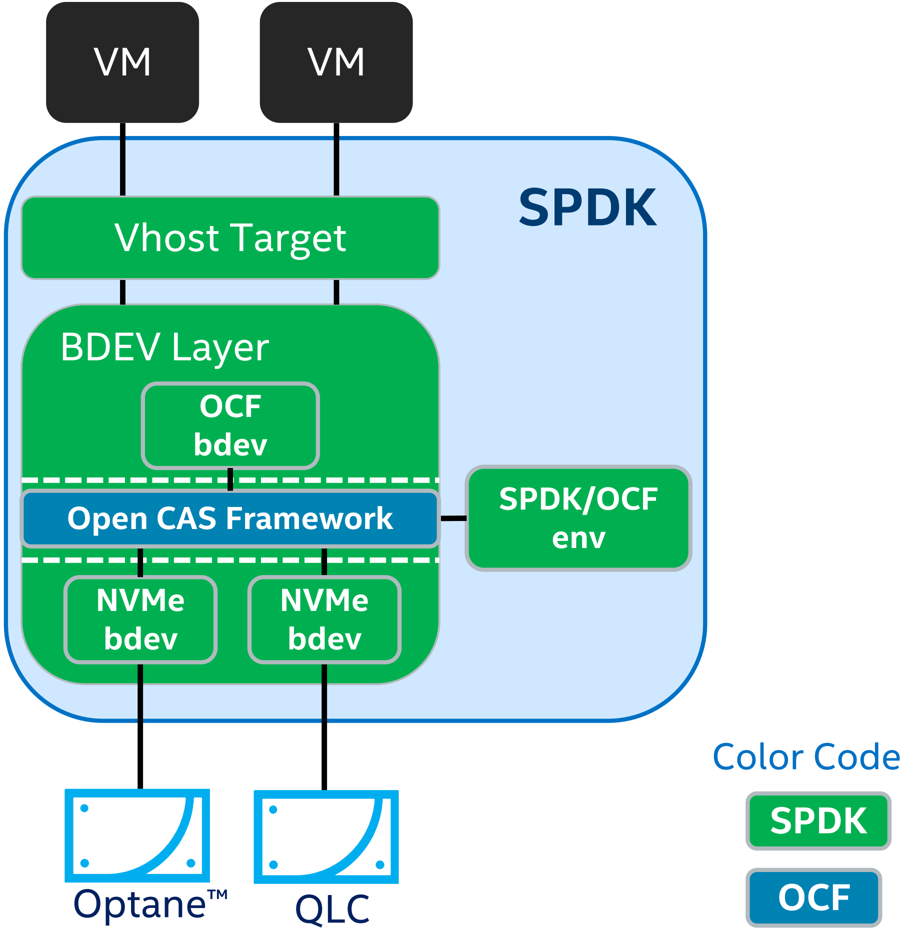
spdk/lib/bdev/ocf/volume.c

To access different types of storage using common NVMe bdev interface. Ex:

- Cache device
- Core device. i.e. big capacity storage device to be cached



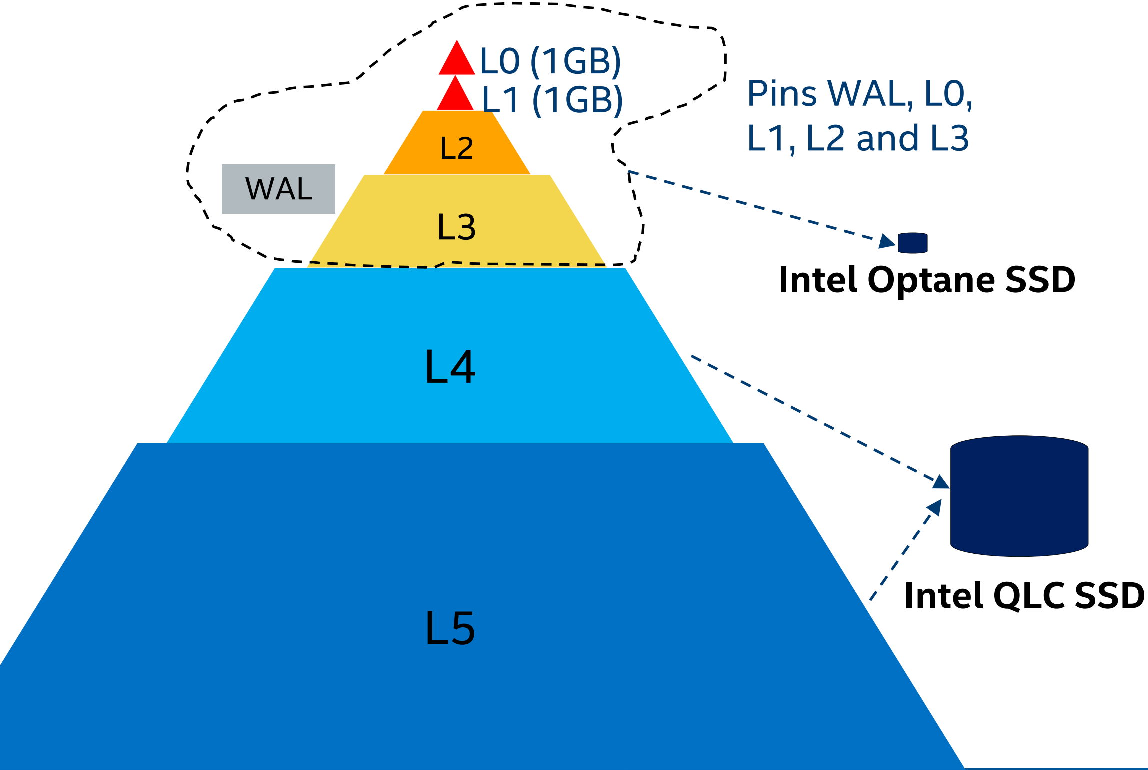
OCF Integration to SPDK - Summary



OCF – Outlook

- More Optane+TLC/QLC/PLC
  - Read caching, write buffering, tiering, pinning
  - ex: RocksDB O+Q use case (next page)
- More IO Classification
  - ex: data lifetime, process ID
- More integration to open source
  - ex: Open Stack, Ceph

O+Q for RocksDB



	Before (TLC)	After (O+QLC)	O+QLC vs TLC
Caching	NA	1x 375GB P4800 (Used 118GB)	Only 2% cache
Storage	1x 8TB P4510	1x 7.68TB P4320	
Write BW (MB/s)	30	40	**30% better**
99.99% QoS(ms)	30	14	**1x better**
Endur-ance (EDWPD)	1.7	2.7	**60% better**

RocksDB system config

The database has 6B keys (key size 32B, value size 1024B) and total 6 levels.

Our tests were performed on Fedora 25 (kernel 4.13.16) and RocksDB v 5.17.2.
We use fillseq to prep a database. Once the database is ready, we run readwhilewriting to update keys. Test stops after ~700M updates.

CPU Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz, 2 sockets, 22 cores, memory 256GB, BIOS Version: SE5C610.86B.01.01.0016.033120161139. Release Date: 03/31/2016,

To prep database:

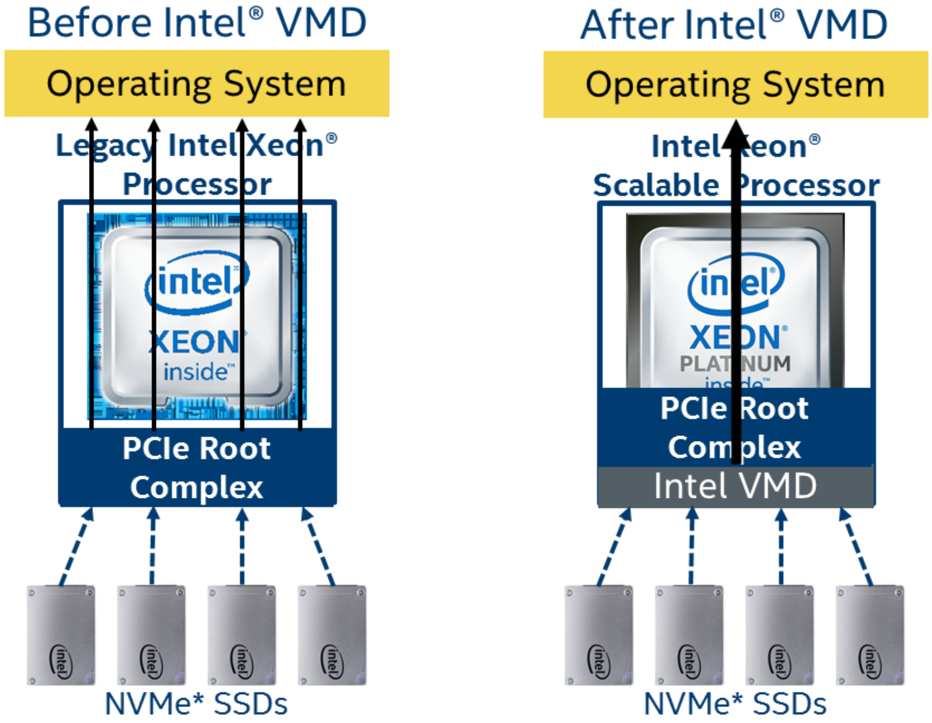
```
# db_bench --db=/mnt/rocksdb \
    --num_levels=6 \
    --key_size=${KEY_SIZE} --value_size=${VALUE_SIZE} \
    --block_size=4096 \
    --cache_size=$((8 * GiB)) --cache_numshardbits=6 \
    --compression_type=none --compression_ratio=0.5 \
    --hard_rate_limit=2 --rate_limit_delay_max_milliseconds=1000000 \
    --write_buffer_size=$((1024 * MiB)) --max_write_buffer_number=4 \
    --target_file_size_base=$((128 * MiB)) --max_bytes_for_level_base=$((1024 * MiB))
\
    --max_bytes_for_level_multiplier=10 \
    --sync=0 --verify_checksum=1 \
    --delete_obsolete_files_period_micros=$((60 * MiB)) \
    --statistics=1 --stats_per_interval=1 --stats_interval=$((1 * M)) \
    --histogram=1 --memtablerep=skip_list --bloom_bits=10 \
    --num_multi_db=1 \
    --open_files=$((20 * KiB)) \
    --max_background_compactions=32 --max_background_flushes=32 \
    --level0_file_num_compaction_trigger=7 --level0_slowdown_writes_trigger=16 --
level0_stop_writes_trigger=24 \
    --benchmarks=fillseq --use_existing_db=0 --num=${key_no} \
    --threads=1
```

To run benchmark:

```
# db_bench --db=/mnt/rocksdb \
    --num_levels=6 --key_size=${KEY_SIZE} --value_size=${VALUE_SIZE} \
    --block_size=4096 --cache_size=$((8 * GiB)) \
    --cache_numshardbits=6 --compression_type=none --compression_ratio=1 \
    --hard_rate_limit=2 --rate_limit_delay_max_milliseconds=1000000 \
    --write_buffer_size=$((1024 * MiB)) --max_write_buffer_number=4 \
    --target_file_size_base=134217728 --max_bytes_for_level_base=1073741824 \
    --sync=0 --verify_checksum=1 \
    --pin_l0_filter_and_index_blocks_in_cache=false \
    --cache_index_and_filter_blocks=false \
    --mmap_read=0 \
    --max_background_compactions=32 --max_background_flushes=32 \
    --disable_auto_compactions=0 --statistics=1 --stats_per_interval=2 \
    --histogram=1 --memtablerep=skip_list --bloom_bits=10 \
    --use_direct_reads=1 --open_files=-1 --level0_file_num_compaction_trigger=8 \
    --level0_slowdown_writes_trigger=16 --level0_stop_writes_trigger=24 \
    --benchmarks=readwhilewriting --use_existing_db=1 --stats_interval=5000000 \
    --num=$((600*M)) --threads=4
```

INTEL VOLUME MANAGEMENT DEVICE (VMD)

Intel® Volume Management Device (Intel® VMD)

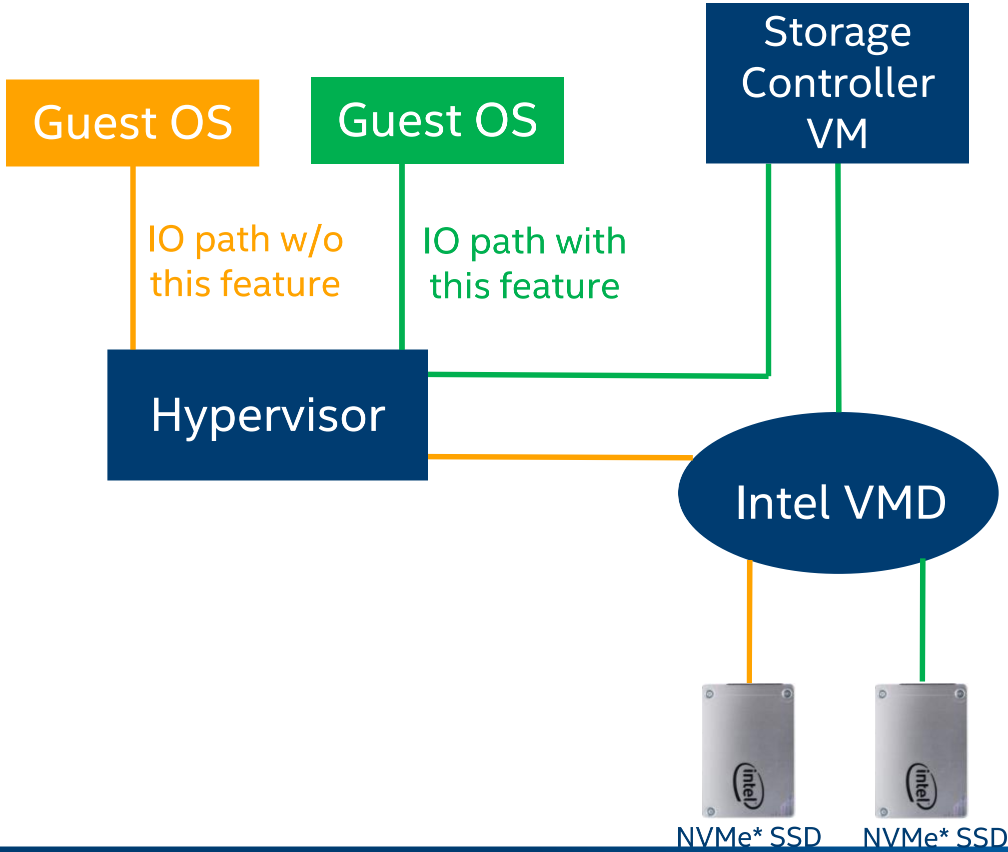


- Intel® VMD: HW logic inside Intel® Xeon® Scalable processors to manage and aggregate NVMe* SSDs

- Benefits are:
  - surprise hot-plug
  - status LED management
  - Bootable RAID
  - Direct assign VMD using VT-d

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

VMD Direct Assignment



- Controller VM is a typical HCI (Hyper-Converged Infrastructure) architecture.
- Today controller VM is implemented by using VT-d technology to direct assign SATA/SAS HBA

Call to Action

- Stop by open CAS in SPDK demo in aisle
- Join in OCF lab in “SPDK hands on lab” session @3:25pm today
- Check out OCF code in SPDK (**spdk/ocf/*, spdk/lib/bdev/ocf/***)

intel

CONNECTED PLATFORM

MEMORY & STORAGE

INTEL® OPTANE™ AND INTEL® QLC 3D NAND

ACCELERATE STORAGE ARCHITECTURE EVOLUTION

DATA CENTER/ CLOUD

WORKING DATA

Fast Access

CAPACITY DATA

Large Capacity

What is Intel® CAS

File system metadata,
File name and director,
IO size, data lifetime ......

Intel® CAS accelerates storage performance by caching specified data classes

Examples of CAS usage

Application Acceleration

Virtualization

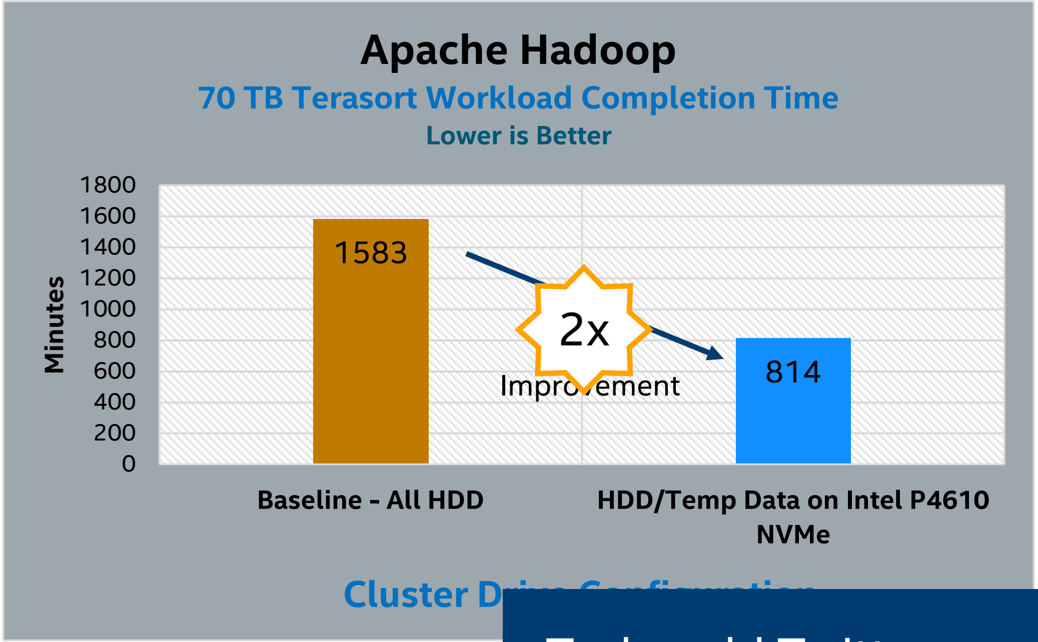
SDS Acceleration

CAS Features and Benefits

- Smarter caching with **I/O classification**
- Incredible performance
- Optimized for Intel® Optane™ SSDs
- Improved application SLA
- Enterprise validated and supported
- Robust roadmap
- **No application changes**

For more information: intel.com/CAS

Intel CAS Unlocks Hadoop* Bottleneck for Twitter[3]



NVMe*-based Intel® SSD w
Software doubles H

Todo: add Twitter paper link

Comparison Benchmark
70TB Terasort Workload

Hadoop* Analytics Baseline
All HDD back-end storage

New Intel® CAS Powered Solution
Direct YARN data to an NVMe-based 6.4TB Intel® SSD DC P4610

Net benefits
- Increase performance by **2x**
- Reduce 30% TCO

Intel's compilers may or may not optimize to the same degree for non-Int ... cessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionalit ... red 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. See appendix for configuration information. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

System Configurations

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 www.intel.com/benchmarks. 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.

1. MySQL (slide 7): Source: Intel. System configuration –Red Hat Enterprise Linux 7.3, Kernal 3.10.0-514.el7.x86_64 #1 SMP Wed Oct 19 11:24:13 EDT 2016, Purley Silver Wolf Pass S2600WFQ, BIOS Version: SE5C620.86B.0X.01.0107.122220170349, BIOS Release Date: 12/22/2017, Skylake H0 (2 Processors)(24 cores each processor, hyper-threading is enabled in BIOS so thread count per processor is 48) Intel® Xeon® Platinum 8160T CPU @ 2.10GHz, Intel(R) Rapid Storage Technology enterprise PreOS Version : 5.3.0.1052, 256GB Physical RAM installed but set to 128GB in the grub2 configuration, Intel 82574L Gigabit Ethernet Adapter, VMD enabled in BIOS and VROC HW key (Premium) installed and activated., Package C-State set to C6(non retention state) and Processor C6 set to enabled in BIOS, P-States set to default in BIOS and SpeedStep and Turbo are enabled, BMC version: 1.43.33e8d6b4 ME version: 4.00.04.309 SDR Package version: 1.43, fio version: fio-3.5-86-gcefd2, (VROC) mdadm - v4.0 - 2017-09-22 Intel build: RSTe_5.3_WW38.5, kmod-md-rste-5.3-514_4.el7_3.x86_64

1. Oil Search (slide 7): Testing was performed by Xenon and additional information can be found here: http://www.xenon.com.au/wp-content/uploads/2016/06/F_Oil-Search-Case-Study_07112016_webv2.pdf

1. Ikea (slide 7): Source: Intel and Ikea. System Configuration: Dell Precision T7910, CPU: Intel Xeon E5-2699 v3 x2 2.3GHz, RAM 192GB, HDD: ATA 1TB ST1000DM003-1ER1, CAS Disk: INTEL SSDSC2BF36333.5G, GPU: NVIDIA Quadro K6000 x3

Source: Intel. Baseline 4-Node Cluster: HDD OSD Drives with Journals on Intel S4600 SSD's: 3x OSD 1x Mon/RGW Nodes: Server Intel S2600GZ (Grizzly Pass), CPUs 2x Intel® Xeon® Ivy Bridge E5-2660v2 @ 2.20GHz, 64GB Mem, SATA Boot SSD 1 x 800GB Intel® SSD DC S3700, OSD HDD 7 x 4TB WD* WDC_WD4003FZEX (excl. Mon/RGW), SATA Journal SSD 1 x 2TB Intel® SSD DC S4600, Network 2 x Intel® X540-AT2 10Gbe NICs; Ceph journal size: 10GB x 7. Value 4-Node Cluster: HDD OSD Drives with Journals on Optane, with/without CAS: Same as Baseline except NVMe Journal and cache 2 x 375GB Intel P4800x Optane; Ceph Journal size: 10GB x 7, Cache Size: 320GB x 2. Software: Ceph Luminous v12.2.3, RHEL 7.4 Updated, COSBench 0.4.2.c4, Intel CAS 3.5.1 (Value)

1. Aliyun (slide 7): Source: http://docs-aliyun.cn-hangzhou.oss.aliyun-inc.com/pdf/ecs-user-guide-intl-en-2017-10-02.pdf?spm=a3c0i.o48226en.a3.8.1f60414jS1ogb&file=ecs-user-guide-intl-en-2017-10-02.pdf. September 2017, Alibaba system, Alyun virtualization Team testing on Intel Broadwell Xeon™ CPU based Broadwell Servers, Ali O/S version 7.2, Intel Cache Acceleration Software for Linux version 3.5, Intel NAND NVMe SSD DC P3700, Intel® Optane® DC P4800x 375GB SSD, Workload FIO 4K Random Reads on warmed cache with primary storage being the Aliyun Basic Cloud Service over the network., Intel CAS configured a write-through

System Configurations (2)

1. Ceph (slide 7): Results based on Independent Redhat* testing of Ceph S3 object performance with Intel CAS + NVMe (64k to 64M object sizes, 130M objects). https://www.redhat.com/cms/managed-files/st-ceph-storage-qct-object-storage-reference-architecture-f7901-201706-v2-en.pdf. System Configuration: 6 Ceph* storage nodes, each server: 2x Intel® Xeon® processor E5-2660 v3, 128GB RAM, thirty-five 6TB Seagate Enterprise* SAS hard drives, and two 800GB Intel® Solid-State Drive (SSD) DC P3700 NVMe* drives with one 40GbE Mellanox ConnectX-3 Network Adapter.

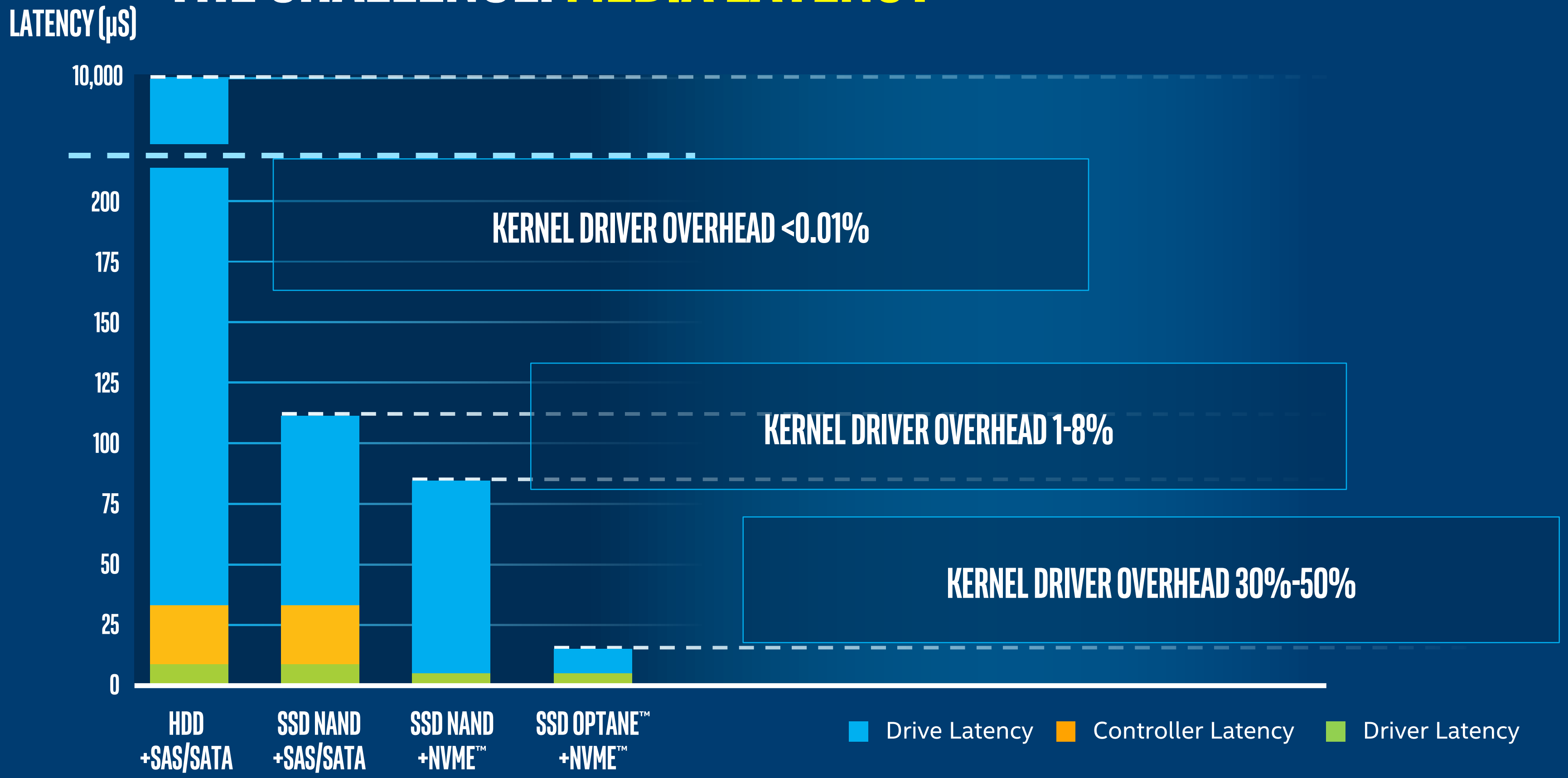
1. Swift (slide 7):Source: Intel. Testing performed by Intel on various Swift cluster sizes (5-15 object storage nodes and several proxy nodes). Cluster configuration: Processors: 2x Intel® Xeon® E5-2699 (45MB cache, 2.3GHz, 18 cores), DRAM 128 and 256GB Memory, HDD 8 x 2TB Seagate* ST2000NX0403, NVMe SSD 1 x 2TB Intel® SSD DC P3520, Intel® RMS3CC080 RAID. Controller: SAS 3.0 Dell Mezzanine* SAS/SATA 8 Port 1GB. Network: 2 x Intel® X540 10Gbe NICs, 2 x Niantic NIC, 2 x Intel® X520-DA2 and 2 x 10G NIC - Intel® X540. Operating system: Ubuntu 14.04.5 kernel revision: 4.4.0-47-generic, Swift Version: 2.9.0.2-4~trusty, SwiftStack Controller version 4.7.0.1. Cache Acceleration Software (CAS) 3.1.1, COSBench version - 0.4.2.c4

2. Hadoop (slide 8): 1x Name Node: CPU 2x Intel® Xeon® E5-2699 v4@2.20GHz (2Socket x 22) Memory: 128GB DDR4-2666 ECC Intel® SSD DC S4600 (Boot drive, 240GB) 2x Intel Corporation Ethernet Controller 10-Gigabit X540-AT2 (rev 01)
9x Data Node: CPU 2x Intel® Xeon® Platinum 8180 Processor @ 2.5GHz (SkyLake 28 cores with 36MB L3 cache) Memory: 128GB DDR4-2666 ECC Intel® SSD DC S4600 (Boot drive, 240GB) 4x Intel Corporation Ethernet Controller X710/X557-AT 10GBASE-T (rev 02) 8x HDD Seagate 4TB 7200RPM SATA ST4000NM0085 1x NVMe Intel P4610 6.4TB SSD
Software Specifications: OS: CentOS 6.9 with custom 4.14 Kernel 2.6.74-t1.el6.x86_64
Application: Apache Hadoop 2.9 Replication Factor 3
Network Interface Bonding: 2x10 Gbps inteface bonding 20Gbps Mode 4 LACP
Intel® CAS v3.9: Yarn directories and metadata cached
Terasort 70TB workload. Sized to overflow the NVMe device and confirm CAS protection against failed jobs due to drive full condition

3. OCF + SPDK Demo (Slide 22): System configuration: Server model: SYS-6029U-TR4T; MB: X11DPU; CPU: Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz, 28C/56T, 38.5 MB L3 Cache, Turbo, HT (205W); Mem: 8x32GB Hynix HMA84GR7AFR4N-VK DIMMs (256GB), DDR4-2666; NICs: 4x Embedded Intel X710/X557 10GbE LAN; BIOS Version: 1.10; Operating System: Red Hat Enterprise Linux Server release 7.4; Kernel Version: 4.20.12-1.el7. TLC config: 3x Intel SSD DC P4510 (8TB) in RAID5; QLC Config: 2x Intel® Optane™ SSD DC P4800X (375GB) in RAID0 for caching, 3x Intel SSD DC P4320 (7.68TB) in RAID5 for backend storage; Software Configuration: SPDK Version 19.04 beta, OCF Version 19.04 beta, FIO Version 3.3. P4800X RAID0 used for write-back caching, cache size is 3% of the 1500GB fio workload (~45GB). Workload: FIO, 3 trials after one single 2 hr ramp time, each trial with: size=1500GiB, block size 16KB, zipf random distribution (theta = 1.1) , random readwrites, 70/30 rw mix, 8 IO depth, 4 jobs. Performance results are based on testing as of April 10, 2019 and may not reflect the publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. 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. For more complete information visit www.intel.com/benchmarks.

BACKUP

THE CHALLENGE: MEDIA LATENCY



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.

STORAGE PERFORMANCE DEVELOPMENT KIT

DPDK
SPDK
DPDK与SPDK开源社区

Scalable and Efficient Software Ingredients

- User space, lockless, polled-mode
- Extreme performance (over 10 million of IOPS/core)
- Minimize average and tail latencies
- Designed for non-volatile media

Storage Reference Software

- Optimized for latest generation CPUs, SSDs & SmartNICs
- Provides Future Proofing
- Extends to Storage Virtualization and Networking

Open Source community

- Open source building blocks (BSD licensed)
- Active Community (~50 contributors each quarter)
- Faster TTM, fewer resources required

UNLEASH ENDLESS POTENTIAL OF DATA VALUE WITH INTEL® OPTANE™ & QLC ® TECHNOLOGIES AND SPDK

- Enjoy better compute efficiency with Intel® Optane™ technologies
- Improve storage scalability and TCO with Intel® QLC® technologies
- Accelerate storage arch innovations with Intel® Optane™ And QLC ® Technologies

intel®

How Does CAS-Linux* Work?

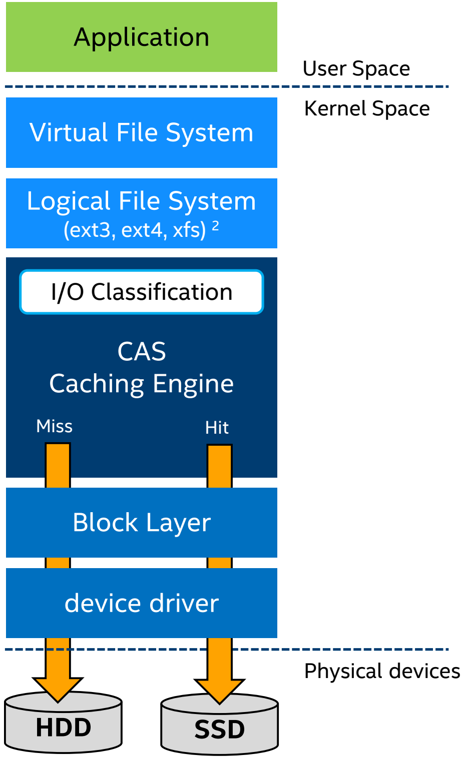
Installed as loadable kernel module

- Configuration via a user-space administration tool
- Deployed at the block layer

Caches "hot data" on a fast SSD

- Validated on common Linux* distros and kernels

Operating modes: Write Thru, Write Back, Write Around[3]



Footnotes:
2. ext3 supports up to 16TB volume sizes.
3. See Admin Guide for modes supported

CAS Unique Differentiation

CAS Generates I/O Classification: CAS identifies I/O by classification and prioritizes caching by class

Flexible Cache Replacement Policy: Multiple LRU (Least Recently Used) caching vs. traditional single LRU caching

Flexible fine tuning:

- Ability to cache just the hottest classes (e.g. metadata)
- Boost performance with a very small cache, keeping cost low.
- Results in improved application and user response time.

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

CAS generates I/O classification to intelligently cache the desired/**hottest** data

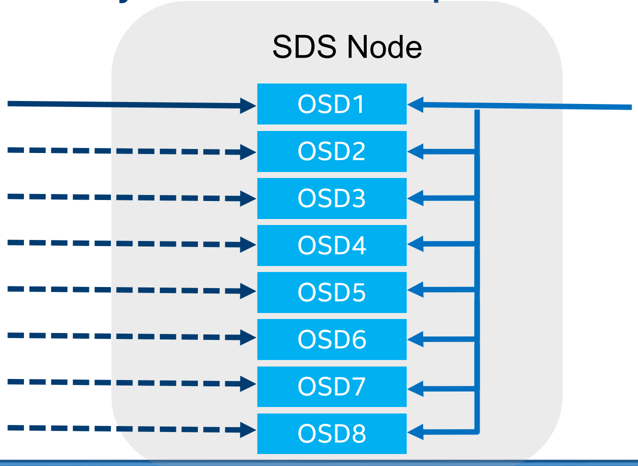
CAS-Linux* Benefits & Capabilities

In-Flight Upgrades

- Enables upgrade to new version of Intel® CAS without stopping I/O
- Upgrade all nodes at the same time
- Improves operational efficiency and reduces performance impacts

Legacy upgrade process (under Ceph*):

1. First OSD is put into Maintenance mode
2. All incoming writes are rebalanced to remaining nodes.
3. Perform software upgrade
4. Return the node to normal operation and Ceph rebalances the cluster again (impacting performance and reliability).
5. Repeat for each OSD

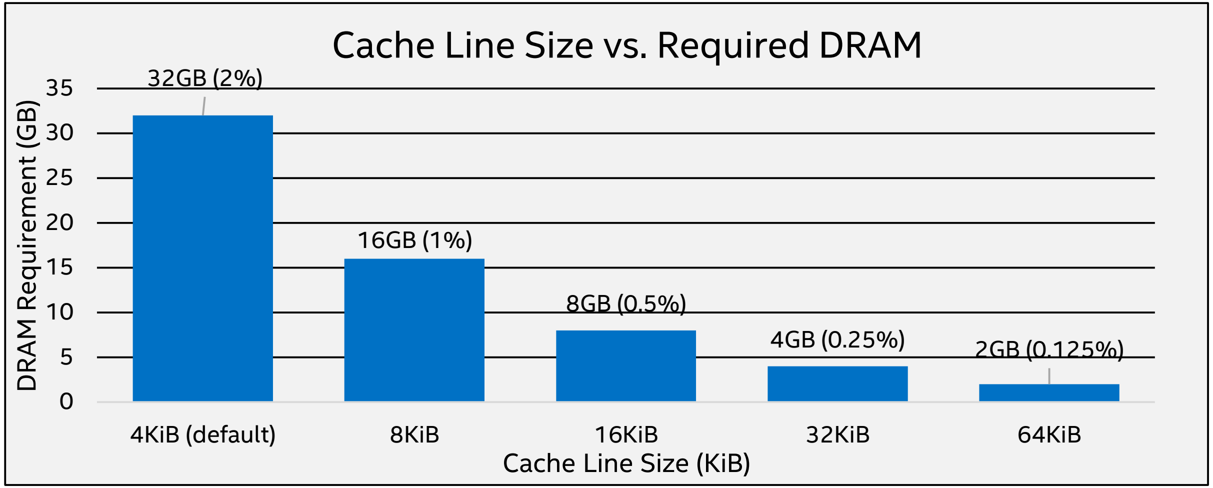


CAS upgrade process:

1. CAS is installed on each OSD simultaneously without stopping the I/O. Therefore, no rebalancing is necessary.

In-flight upgrades increases CAS **USABILITY**

Intel® CAS-Linux* Benefits & Capabilities (cont.)



Calculations based on 1.6TB Cache SSD

Cache line size is user-selectable via administration tool

Doubling Cache Line Size reduces CAS DRAM requirement by half.
(DRAM stores the cache metadata (LBA mapping, valid/invalid bit, etc.)

BOM savings possible by reducing required DRAM

Do You Use QEMU*/KVM* for Virtual Machine (VM) Management?

Would you like to improve your VM performance > 6x?

- Would you like to reduce the network traffic on the server racks?
- Would you like to offer your customers a better SLA?
- Would you like to allocate “chunks” of a big SSD to each VM?
- Would you like to setup a caching policy to allow each VM to optimize the use of its SSD “chunk”?

Intel® Optane™ SSDs running CAS may offer you:

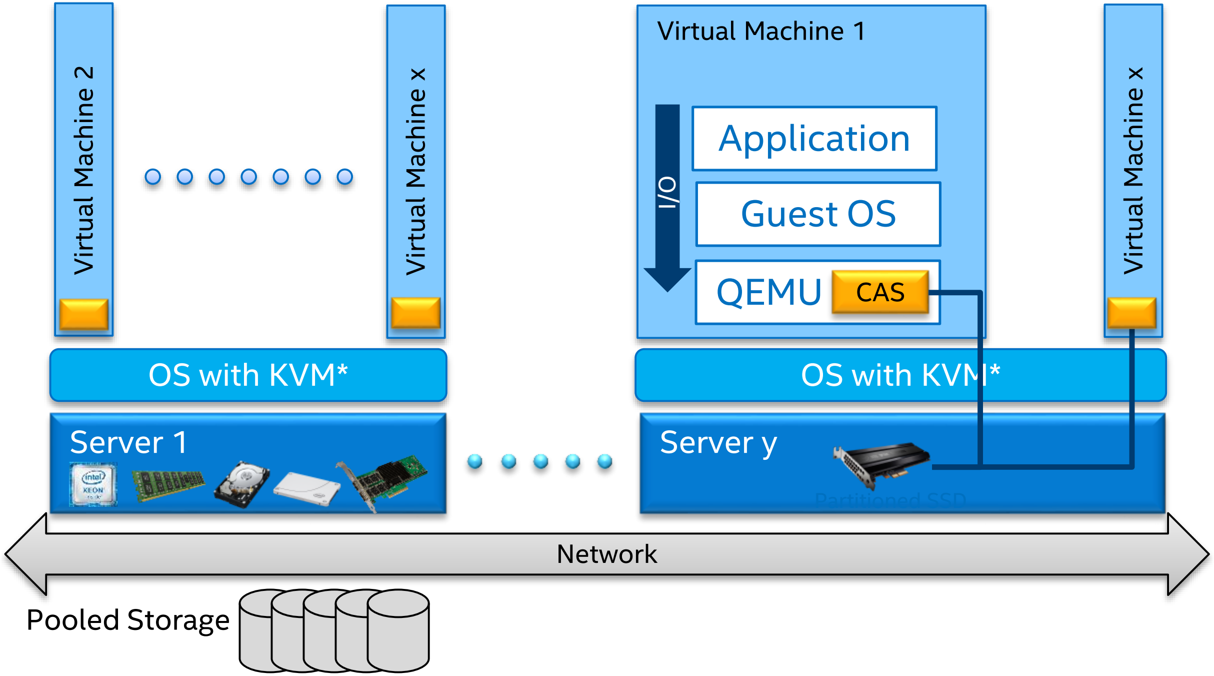
- Outstanding VM performance
- The ability to provide your customers with premium differentiated services

See Footnotes page(s) for system configurations used for performance claims

Virtual Machine Acceleration: QEMU*/KVM* Cloud Stack for Customer VMs

CAS runs in each instance of QEMU*.

- CAS caches the user data from pooled storage in the VM's partition on the SSD
- CAS runs in user space, so latency is very low
- CAS resolves the 'noisy neighbor' problem and is ideal for multi-tenant sharing of the caching SSD by isolating each cache instance.
- CAS offers write-thru caching to ensure data integrity



CAS improves TCO and allows SLAs to be met without overprovisioning

Intel® Cache Acceleration Software Typical Usages[1]

Application Acceleration

- MYSQL* UP TO 5.1x AS FAST W/CAS + INTEL® OPTANE™ SSD
- OIL SEARCH* USING CAS UP TO 5x PRODUCTIVITY INCREASE
- IKEA* UP TO 2x DAILY PRODUCTIVITY GAINS

Virtual Machine Acceleration

- ALIYUN* VM PERFORMANCE 6.1x FASTER W/ CAS & INTEL® OPTANE™ SSD

Cloud Storage Acceleration

- CEPH* READS UP TO 4.9x FASTER WITH CAS + INTEL® OPTANE™ SSD
- CEPH* WRITES UP TO 4.8x FASTER WITH CAS + INTEL® OPTANE™ SSD
- SWIFT * WITH CAS + NVME* SSD 3x THROUGHPUT INCREASE