Intel® Cache Acceleration Software (Intel® CAS) for Linux*

Non-volatile Memory Solution Group

Dave Leone
December 2015

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.

This document contains information on products in the design phase of development.

Material in this presentation is intended as product positioning and *not* approved end user messaging.

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Results have been simulated and are provided for informational purposes only. Results were derived using simulations run on an architecture simulator or model. Any difference in system hardware or software design or configuration may affect actual performance.

Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

Code names are internal project names used solely as identifiers and are not intended to be used as trademarks or publicly disseminated.

Intel and the Intel logo 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.

Copyright © 2014 Intel Corporation. All rights reserved.

Agenda

- Is Intel® CAS Right for You?
- Intel® CAS for Linux* Features
- How It Works?
- Get Started (install, configure, manage)
- Benchmark BKMs
- Use Cases
- FAQ
- Future Capabilities

Is Intel® CAS Right for You?

- **Is I/O your performance bottlenecked or sub-optimal?**
  If you don't have I/O problem, Intel® CAS for Linux* can't help you.

- **Is your system OS and virtualization configuration supported by Intel® CAS for Linux*?**
  We validate against the most widely used Linux distributions. Please check Admin Guide for supported OS.

- **What if I don't know if or how big my IO problem is?**
  Use iostat and top to look at IO and CPU utilization. Evaluate Intel® CAS for Linux* performance improvement using Intel® CAS for Linux* Trial Software.

Identify I/O Problem

- If CPU utilization is low, could be an I/O problem.
- If the disk queue is greater than 1, could be an I/O problem.
- If I/O latency is high, likely I/O problem.
- Use top and iostat to check CPU utilization, queue depth, and latency.

Platform Connected:
Enabling New Usage Models

Intel® Data Center SSD

+

Intel® Cache Acceleration Software (CAS)

Up to…
1400X IOPS[1]
57X OLAP[2]
3X OLTP[2]

Accelerate Your Data Center… Without Application, SAN, or NAS Changes

Intel technologies may require enabled hardware, specific software, or services activation. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as IOMeter, 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.

1. Configuration used: Intel® Server model 2600GZ (Grizzly Pass); Dual Intel® E5-2680 processor (2.7GHz), 32GB memory; Seagate ST1000NC000 SATA HDD Microsoft* Windows 2012R2 SP1**, Intel® SSD DC P3700 -800GB, Intel® CAS 2.6 release, L2 cache on** ; IOmeter 10.22.2009 ; 4K Random Read test; 8-queue depth x 8 workers
2. Configuration used: Intel® Server model 2600GZ (Grizzly Pass); Dual Intel Xeon E5-2680 processor (2.7GHz), 96GB DDR3, VMware* 5.5, **Intel® SSD DC P3700 -800GB, Intel® CAS 2.6,, L2 cache off,** 8xSeagate 146GB SAS in RAID5,VMs: Microsoft Server 2008R2, 8GB, 2 Cores, IOMeter workloads: Media Player ,Exchange Server, Web Server, 4K OLTP using QD4.1 Worker

Intel® CAS for Linux* Features

Supported Linux Distributions:

- Intel® CAS for Linux* v3.0 will compile from source upon install on *any* distro and is fully validated on the following distros:
  - RHEL 6.6
  - RHEL 7.0-7.1
  - CentOS 6.6
  - CentOS 7.0-7.1
  - SLES 11 SP3

Supported Hypervisors:

Hypervisor	Supported Configuration	Notes
Xen*	Supported in hypervisor or guest.	Paravirtualized drivers are not supported.
KVM*	Supported in hypervisor or guest.	
VMWare*	Supported in guest.	

Intel® CAS for Linux* Features

Supported file systems:

- ext3 (limited to 16TB volume size), ext4, xfs

Caching Modes

- Write-Through
- Write-Back
- Write-Around

Multi-Tiered Caching

- Intel® CAS for Linux* supports multi-level caching.
- E.g., the user can cache HDD to SSD in write-back mode, and cache SSD to RAMDisk in write-through mode.

I/O Classification

- Ability to selectively cache I/O and prioritize eviction.

Intel® CAS for Linux* Features

Include Files

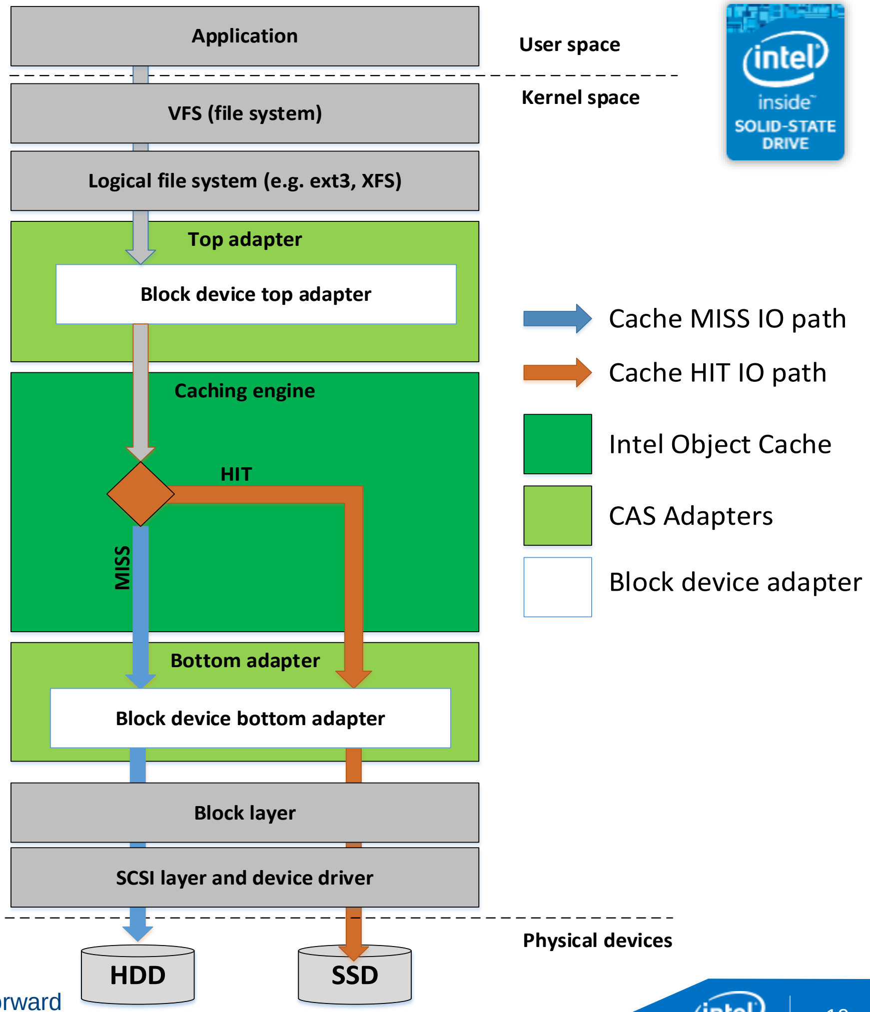
- By default, Intel® CAS for Linux* caches everything.
- User can provide an "include files" list, which is translated into a block list. Only those blocks will be cached, avoids cache pollution.
  *(**NOTE:** If the included files grow, shrink, or move, the "include files" command must be re-issued to cache the new blocks.)*

Performance Monitoring

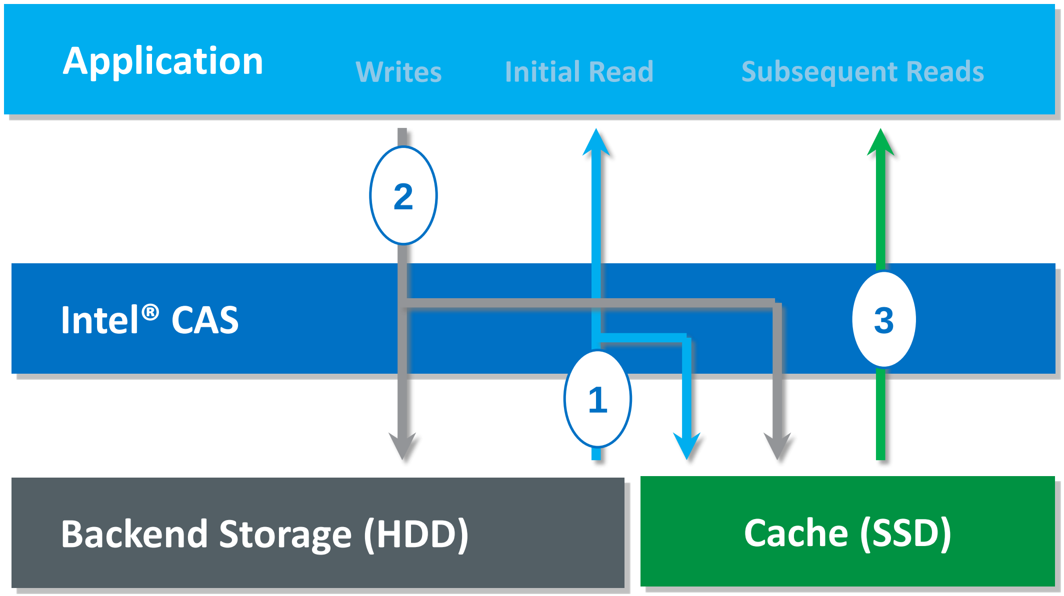
- Intel® CAS for Linux* provides detailed performance statistics (*casadm --stats –i 1*).
- Statistic include:
  - # of reads and writes
  - # of cache hits and misses
- This data can help to determine whether caching can improve your workload performance, if you need a bigger cache, etc.

How it works?

- Intel® CAS for Linux* is installed as a loadable kernel module and user-space administration tool
- Intel® CAS for Linux* is deployed between the logical file system and the underlying block device
- Cache pairing exposes a new mountable device
- Intel® CAS for Linux* v2.9+ will force recompile from source via DKMS on kernel update



How it works? - Write-Through

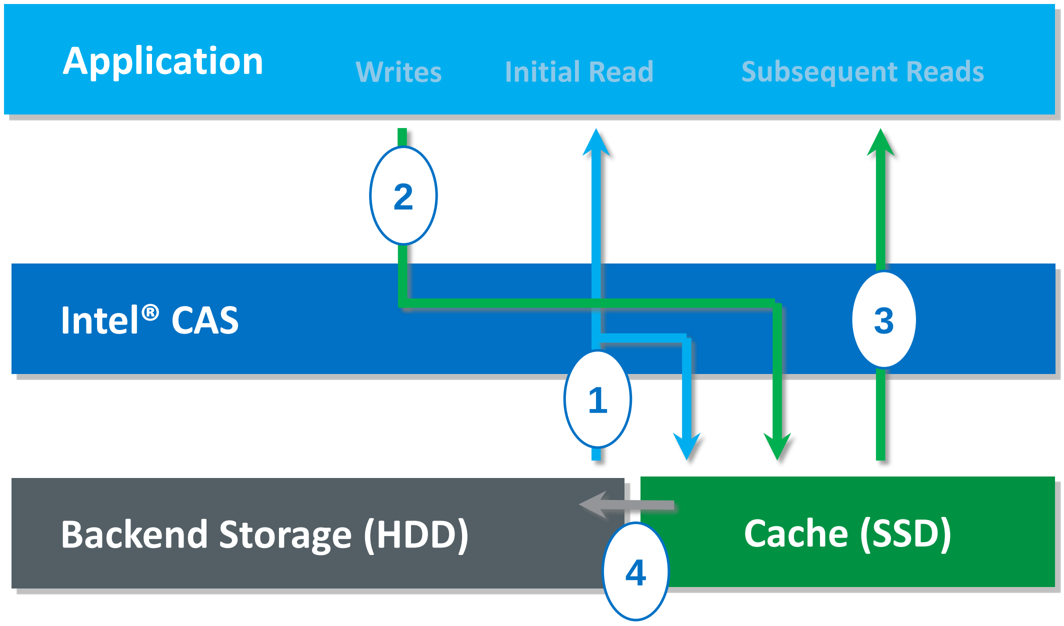


1. Data is read from backend storage and copied to the cache on SSD
2. All data is written synchronously to backend storage and cache
3. Subsequent reads of cached data are returned at high-performance SSD speed

- Benefits random & repeated reads

"Accelerates re-reads of data that was read or written"

How it works? - Write-Back

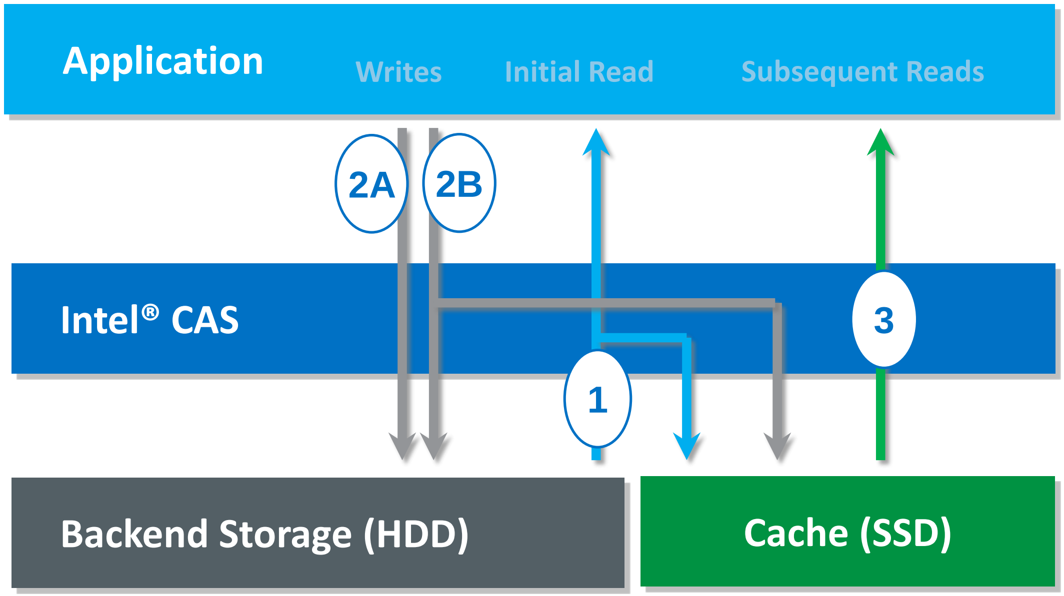


"Accelerates both writes and re-reads"

1. Data is read from backend storage and copied to the cache on SSD
2. All writes go to the cache.
3. Subsequent reads of cached data are returned at high-performance SSD speed
4. Dirty data is written opportunistically to backend storage.

- Benefits reads & writes

How it works? - Write-Around



“Enhanced write-through to avoid cache pollution when data is written but not often re-read”

1. Step 1 same as previous policies.
2. Writes:
   A. If block has never been read before, data is written directly to backend storage
   B. If block *has* been read before, data is written synchronously to backend storage and cache
3. Subsequent reads from SSD

- Enhanced write-through mode to avoid cache pollution

How it works? - I/O Classification

I/O Classification

- Classifies I/O requests in software
- Assigns policies to I/O classes (priority, allocation)
- Enforces policies in the storage system
- Evict based on priority

Similar to DiffServ in networking

- IP packet classification for network QoS

Very useful for software defined storage systems such as Ceph, Swift, and Lustre.

- Where filesystem metadata is accessed much more often than file data.
- Metadata becomes the factor limiting throughput and causing high latency.
- Enables caching of *just* filesystem metadata on storage nodes
- Results in increased throughput and decreased latency.

How it works? - I/O Classification

I/O Classification Schema

- Intel® CAS operates below the software stack at the Local filesystem block layer
  - No modification to the Ceph*/Swift*/Lustre* stack required
- Ability to selectively cache & evict based on block type & priority
- Enables a new approach of using a very small cache for the best price-performance trade-off for a given workload/usage

CAS DSS IO Classes
Unclassified
Metadata (Superblock,GroupDesc,BlockBitmap,InodeBitmap,Inode,IndirectBlk,Directory,Journal,Extent,Xattr)
<=4KiB
<=16KiB
<=64KiB
<=256KiB
<=1MiB
<=4MiB
<=16MiB
<=64MiB
<=256MiB
<=1GiB
>1GiB
O_DIRECT
Misc

Install and Configuration

Short videos on how to install, configure, and test (click link to watch):

- INSTALL - Intel® Cache Acceleration Software for Linux (English, Chinese)
- CONFIGURE - Intel® Cache Acceleration Software for Linux (English, Chinese)
- TEST - Intel® Cache Acceleration Software for Linux (English, Chinese)

Quick Start Guide

Administrator Guide has:

- System Requirements
- Supported Distributions
- Installation and configuration details
- Detailed instructions of command-line based configuration

Benchmarking BKMs

#1 BKM: Pre-condition the SSD prior to cache creation.
SSD Firmware has optimizations for blocks that have never been used in which requests from those blocks will be served from firmware, not from NAND. This results in unrealistically high throughput results.

Recommend secure-erase of SSD an dd zeroes to the whole SSD prior to executing the casadm --start-cache command (and prior to executing the fio test script)

#2 BKM: Make sure cache is started correctly, and correct device is passed into the script.

Ex.: casadm --start-cache --cache-device */dev/nvme0n1* --cache-mode wb
casadm --add-core --cache-id 1 --core-device /dev/sdc

The above will result in creation of /dev/intelcas1-1 caching device. This is the device that should be provided to the script for fio operations.

Benchmarking BKMs

#3 BKM: Use latest fio (currently v2.6).
Older fio releases had bug affecting randomness that resulted in poor caching results.

#4 BKM: Use zipfian distribution (--random_distribution=zipf:1.2).
By default, fio will use pure random distribution. A pure random distribution has no "data hot spots" and is not good for caching. Many studies have found that a Zipfian distribution with 1.2 theta is representative of typical real-world workloads including web traffic(1&2), blog traffic(3), video-on-demand(4) and live streaming media(5) traffic, big data map-reduce workloads(6):

1. *"Glottometrics" (see page 143, "Zipf's law and the internet")* - http://www.arteuna.com/talleres/lab/ediciones/libreria/Glottometrics-zipf.pdf#page=148
2. *"Zipf Curves and Website Popularity"* - http://www.nngroup.com/articles/zipf-curves-and-website-popularity/
3. *"Web Caching and Zipf-like Distributions: Evidence and Implications"* - http://others.kelehers.me/zipfWeb.pdf
4. *"Understanding User Behavior in Large-Scale Video-on-Demand Systems"* - https://www.cs.ucsb.edu/~ravenben/publications/pdf/vod-eurosys06.pdf
5. *"A Hierarchical Characterization of a Live Streaming Media Workload"* - http://www.cs.bu.edu/faculty/best/res/papers/imw02.pdf
6. *"Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads"* - http://www.eecs.berkeley.edu/~alspaugh/papers/mapred_workloads_vldb_2012.pdf

Recommend using a Zipfian (zipf) distribution with theta value of 1.2.

Benchmarking BKMs

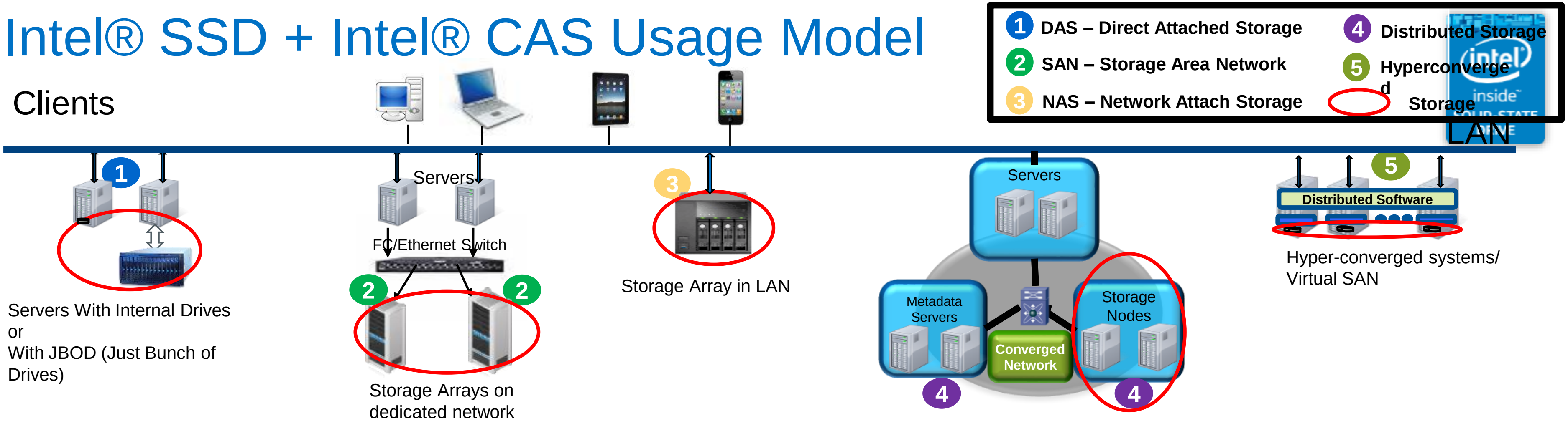
#5 BKM: Warm cache prior to collecting results
Need to run the workload for a period of time to achieve “steady state” for caching. This represents the long-term realistic performance of your caching.

Recommend running for 2 hours of 128K sequential writes to the /dev/intelcas1-1 device prior to beginning to gather benchmark data.

#6 BKM: Run the test sequence multiple times.
Need to run the test at least 4 times to see the trend and know that the results are realistic and consistent (low variance). Intel runs the test 5 times, then we throw out first result and use the average of the remaining 4 results.

Recommend repeating entire test sequence 5 times.

Intel® SSD + Intel® CAS Usage Model



Intel® SSD DC + Intel® CAS v3.0 for Windows* Available Q4'15

Files copied from local or remote storage on to a fast SSD -> accelerates performance without modifications to the Application or Storage system

- ***NEW:* Write-back** caching for optimal caching performance
- 100% SSD for maximum system performance

- ***NEW:* Write-back** caching for optimal caching performance to address network I/O bottlenecks

- Client caching for NAS in Windows* is dependent on specific usage model requirements to address network I/O bottlenecks

Under Exploration
Refer to NSG Roadmap

Under Exploration
Refer to NSG Roadmap

Intel® SSD DC + Intel® CAS v3.0 for Linux* Available Q4'15

Blocks copied from local or remote storage to a fast SSD -> accelerates performance without modification to the Application or Storage system

- Write-back caching for optimal caching performance
- 100% SSD for maximum system performance

- Write-back caching for optimal caching performance to address network I/O bottlenecks

- Client caching for NAS in Linux* (NFS Client) planned for 2H'16 to address network I/O bottlenecks

- ***NEW:* Selective hint based caching for Ceph/SWIFT/Lustre** optimal for small random files to improve overall cluster performance ***(2X throughput & ½ the latency)***[1]
- 100% SSD for maximum system performance

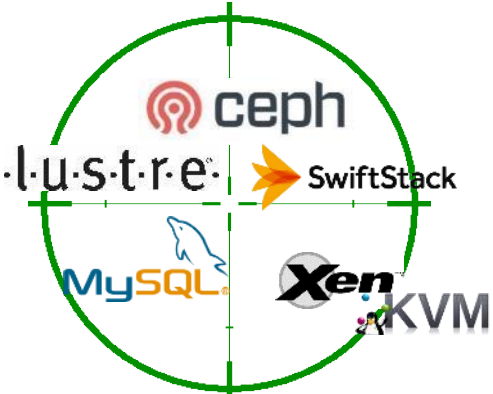
Under Exploration
Refer to NSG Roadmap

http://intelstudios.edgesuite.net//idf/2015/sf/aep/SSDS002/SSDS002.html

Intel® Cache Acceleration Software Product Line

Intel® CAS for Linux v3.0

- ***NEW*** Selective caching based on hinting
- **Ceph Cluster *2x /½ Latency* Gains, QoS, SLA**
  - Save on overprovisioning costs in > $250Ks / cluster
- **Lustre Cluster *2X* Performance Gains**
  - Save on overprovisioning costs in ~ $50Ks / cluster

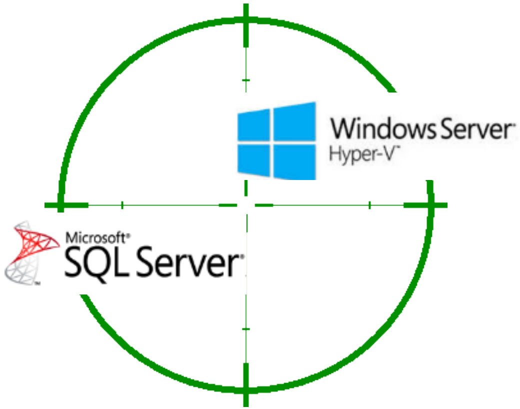


Storage Clusters (Ceph/Luster)

Database / Big Data Analytics

Intel® CAS for Windows* Enterprise v3.0

- ***NEW* Write-back caching performance gains**
- ***~24X* Faster vs CAS 2.71 in benchmarks**
- **Improves existing apps & existing environment**
- **MSFT SQL Server Enhancements**



Microsoft Applications / Database

Software Defined Storage (SDS):

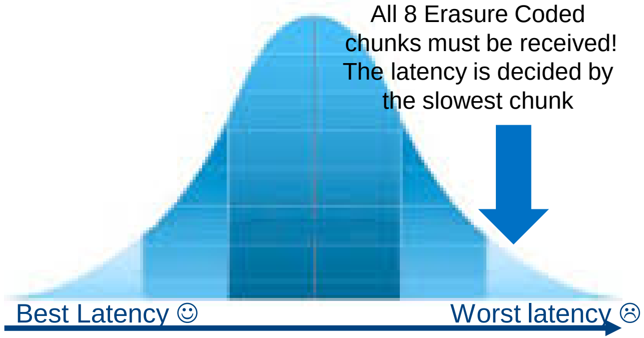
Ceph, Yahoo!* Case Study*[1,2]

Environment:

- Ceph* chosen due to inherent architectural support for Object, Block, File & self recovery/flexibility
- Erasure coding for best cost profile and disk utilization
- Flickr* initially for a multi-Petabyte deployment
- Scaling out for Tumblr*, Yahoo! Mail and more as improve latency and performance
- Read more here[1,2]

Problem:

- Yahoo Mail, Flikr, and Tumblr data comprised of small files
- 8+3 Erasure coding algorithm further breaks small files down into 11 smaller erasure coded slices
  - eg,. 1M photo becomes 11 x 128K files
  - End result are hundreds of millions of small files on the underlying filesystem on each disk
- Rehydrating an object requires gathering at least 8 of the 11 slices
- Getting each slice requires walking through 4-6 inodes on the disk to find the slice in the filesystem
- Overall object rehydration latency dependent on the worst latency (tail latency)



[1]IDF 2015 Class SDS002 http://intelstudios.edgesuite.net//idf/2015/sf/aep/SSDS002/SSDS002.html
[2]Yahoo http://yahooeng.tumblr.com/post/116391291701/yahoo-cloud-object-store-object-storage-at

*Other names and brands are the property of their respective owners

SDS: Intel® NVMe SSD + Intel® Cache Acceleration Software (Intel® CAS) for Linux*

Solution:

- Intel® NVMe SSD – consistently amazing
- Intel® CAS v3.0 featuring new hint-based selective caching
- Ceph optimization: cache Linux metadata (ie Inodes) only with a ~5% SSD cache size of total backend data store on each node
- Operates below the Ceph* S/W stack in the Linux* Local FS

Unique[1] Host Side Caching Solution Improves Entire Cluster Performance up to:[2,3]

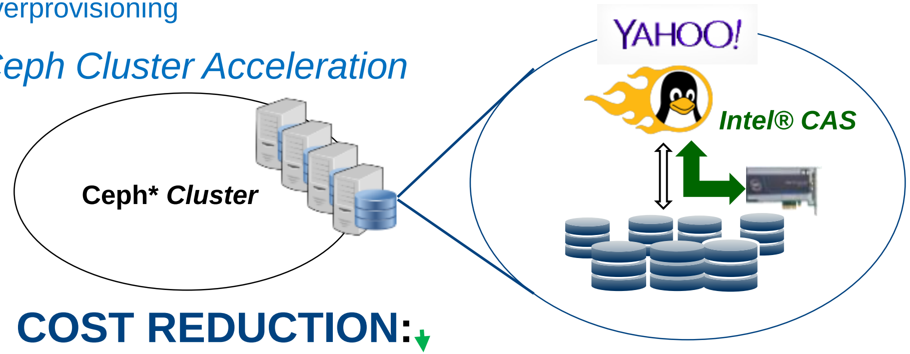
2X Throughput

1/2 LATENCY

Technology Benefits:

- Delivers data at consistent SSD latencies
- Greatly reduces the worst case tail latency
- Reducing latency variability improves deployment & support predictability
- Increasing throughput with average lower read/write latency reduces overprovisioning

Ceph Cluster Acceleration



COST REDUCTION:

- ***CapEx savings*** *(over-provision)*
  OpEx savings *(Power, Space, Cooling…)*
- ***Improved scalability planning*** *(Performance and Predictability)*

[1] Intel® CAS v3.0 is the first commercial software caching solution (to our knowledge) to utilize Linux metadata types for selective priority-based caching based on Intel® Labs DSS technology innovation
[2] Common to each server used in the Ceph Cluster (2ea Xeon® X5620 x2, Intel® 5520 chipset, 48GB DDR3 1333Mhz RAM, 10ea Seagate* 6T 7.2K SATA HDD, 2ea HP NC362i GbE Public Network connection, 2ea Intel® 82599EB 10GbE private LAN connection, Linux RHEL 6.5, kernel 3.10.0-123.4.4.el7
Ceph OSD Storage Node (8 servers) unique configuration: 1per server Intel® SSD Series P3600 1.6TB, Intel® CAS-L v3.0, caching configuration to cache all local file system metadata on the OSD server node
Ceph Admin/Gateway Nodes (3 servers): One server configured with rest-bench for Ceph version 0.87.1 for latency/throughput testing – Throughput & Latency data collection comparison between intel® CAS v3.0 on/off
[3] IDF 2015 Class SDS002 http://intelstudios.edgesuite.net//idf/2015/sf/aep/SSDS002/SSDS002.html

*Other names and brands are the property of their respective owners

BEFORE/BASELINE

CEPH CHALLENGE:

- Erasure coding turns each user IO into multiple disk IOs. Performance is bottlenecked by the slowest disk IO.
- Huge amount of files requires several file system metadata accesses to determine where the file is.
- Deployments over provision Ceph clusters to meet SLA requirements (example **4:1**)

CEPH OBJECT STORE SERVER (0SS)
8 ea per Cluster
Supermicro OSD SSG6028R0sD072[1]: 72TB/No SSD $7,859

Ceph Cluster

CEPH MONITOR/GATEWAY
2 ea per Cluster
Supermicro SYS-6017R-MON1[2]: $5,999

CLUSTER COST = $74.9K
(2xMON, 8xOSS)

CLUSTER PERFORMANCE = 1.0

AFTER/BENEFIT

__NOTE:__ Intel Estimates Only Based on __Publicly Available__ Pricing of Comparable Solutions[1,2,3]

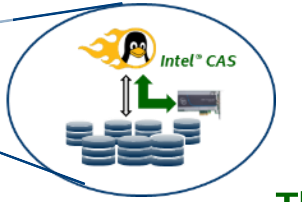
INTEL CEPH NVME™ SOLUTION:

- Selective File Metadata identification, classification, and caching reduces disk accesses, improving overall cluster performance
- Reduce overprovision deployments to meet SLA requirements in half (example **2 :1**)

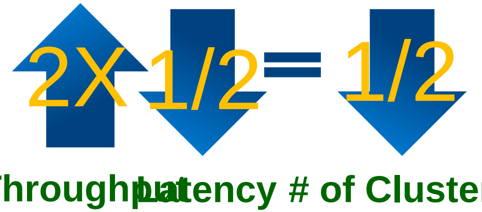
CEPH OBJECT STORE SERVER (0SS)

Ceph Cluster

CEPH MONITOR/GATEWAY



INTEL CEPH NVME™ SOLUTION per OSS
P3700 1.6TB Solution Price Estimate[3] ~$3,400
Cluster Cost Adder (8*$3.4K) = $27.2K



	Before	After[4]
Cluster Throughput	1.0	2X
Cluster Latency	1.0	½
Per Cluster Cost	$74.9K	$102.1K
# Clusters Required	*4X*	*2X*
CapEx	$299.6K	$204.2K **LESS $95K in CapEx**
OpEx	Baseline	**LESS $$:** Power, Cooling, Space, Networking
DevEx	Baseline	**LESS $$:** Cluster Predictability & Scalability

CLUSTER COST = $102.1K
(2xGTWY, 8xOSS, SAME) $74.9K
Add per node caching solution +$27.2K

CLUSTER PERFORMANCE > 2X

Invest ~$27K to Save >$100K

[1]http://www.compsource.com/pn/SSG6028ROSD072/Supermicro-428// as of 9.22.15
[2]http://www.atacom.com/program/print_spec.cgi?Item_code=SYI1_SUPE_60_65 as of 9.22.15
[3]http://softwarestore.ispfulfillment.com/Store/Product.aspx?skupart=I24S10 as of 9.22.15
[4]IDF 2015 Class SDS002 http://intelstudios.edgesuite.net//idf/2015/sf/aep/SSDS002/SSDS002.html
Ordering information: MMID 942112, P/N SSDPEDMD016T4U1, Desc: Intel® SSD DC P3700 Series (1.6TB, 1/2 Height PCIe 3.0, 20nm, MLC) Bundle with Intel® CAS for Linux*

FAQs

Q: How do I contact technical support?

A: Contact technical support by phone at 800-538-3373 or at the following URL: *http://www.intel.com/support/ssdc/cache/cas*.

Q: What are the Supported Distros and System Requirements?

A: Please read Admin Guide Chapter 2, “Product Specifications and System Requirements”

Q: Why does Intel® CAS for Linux* use some DRAM space?

A: Intel® CAS for Linux* uses some memory for metadata, which tells us which data is in SSD, which is in HDD. The amount of memory we need is proportional to size of caching. This is true for any caching software solution. You could add more memory or shrink the size of caching.

Q: Does Intel® CAS for Linux* work with non-Intel SSDs?

A: Yes, Intel® CAS for Linux* will work with *any* SSD but we validate only on Intel SSDs. Additionally, Intel® Cache Acceleration Software is favorably priced when purchased with Intel SSDs.

FAQs

Q: How do I test performance?

A: In addition to the statistics provided (see Admin Guide Chapter 7, *"Monitoring Intel® CAS"* for details), third-party tools are available that can help you test I/O performance on your applications and system, including:

- FIO (*http://freecode.com/projects/fio*)
- dt (*http://www.scsifaq.org/RMiller_Tools/dt.html*) for disk access simulations

Q: Where are the cached files located?

A: Intel® CAS for Linux* does not store files on disk; it uses a pattern of blocks on the SSD as its cache. As such, there is no way to look at the files it has cached.

Q: How do I delete all the Intel® CAS for Linux* installation files?

A: Stop the Intel® CAS software as described in Admin Guide Chapter 5.3, *"Stopping Intel® CAS"*, then uninstall the software as described in Chapter 3.4, *"Uninstalling the software"*.

Q: Does Intel® CAS for Linux* support write-back caching?

A: Yes, Intel® CAS for Linux* v2.6 and newer supports write-back caching. See Admin Guide Chapter 4.3, *"Configuration for write-back mode"* for details.

FAQs

Q: Must I stop caching before adding a new pair of cache/core devices?

A: No, you can create new cache instances while other instances are running.

Q: Can I assign more than one core device to a single cache?

A: Yes. With Intel® CAS for Linux* v2.5 and newer, many core devices (up to 32 have been validated) may be associated with a single cache drive or instance. You can add them using the `casadm -A` command.

Q: Can I add more than one cache to a single core device?

A: No, if you want to map multiple cache devices to a single core device, the cache devices must appear as a single block device through the use of a system such as RAID-0.

Q: Why do tools occasionally report data corruption with Intel® CAS?

A: Some applications, especially microbenchmarks like *dt* and *FIO*, may use a device to perform direct or raw accesses. Some of these applications may also allow you to configure values like a device's alignment and block size restrictions explicitly, for instance via user parameters (rather than simply requesting these values from the device). In order for these programs to work, the block size and alignment for the cache device must match the block size and alignment selected in the tool.

FAQs

Q: Do I need to partition the cache device?

A: No. If you do not specify a partition, Intel® CAS uses the entire device as the cache device.

Q: Can I use a partition on a SSD as a cache device?

A: Yes, however, using the entire SSD device as the cache is highly recommended for best performance.

Q: Do I need to format the partition or the device configured as the cache device?

A: No, the cache device has no format requirement. If any formatting is used, it is transparent to the caching software.

Q: What is the default/optimal block size for the core device?

A: With Intel® CAS for Linux* v2.8 (GA) and later 512 Byte blocks or larger (4 KiB or 4096 byte blocks are also supported as before).
*(**NOTE:** There may be a performance impact if most transactions are 512 bytes and there are a significant number of cache misses.)*

Q: Where is the log file located?

A: All events are logged in the standard Linux* system logs. Use the *dmesg* command or inspect the */var/log/messages* file. To log all messages during testing or kernel debugging, use the command `echo 8 > /proc/sys/kernel/printk`.

Future Capabilities

Intel® CAS for Linux v3.1 (June 2016):

- NEW! In-Flight Upgrade Capability
- Selectable cache line size (4k, 8k, 16k, 32k, 64k)
- Automatic Partition Mapping
- Continuous I/O during flushing
- Improved device I/O error handling

Intel® CAS for Linux v3.5 (2H 2016):

- NFS caching
- …stay tuned!...

Thank You

Backup

- Case Study
- Intel S3700, P3700
- Competition

Trial Questionnaire

1. Do you have an SSD (Solid State Drive) to test our software (60GB or larger)? Yes/No

2. What OS's will be Used?

RHEL/CentOS 6.x __ kernel-_________________________
RHEL/CentOS 7.x __ kernel-_________________________
SLES 11 SP3 kernel-_________________________
Other - ____________________

3. Are you deploying in a VM (Virtual Machine)? No, VMware (ESX or ESXi), Hyper-V, KVM, XenServer, Other (put in blanks for an answer)

4. What applications are you accelerating? Microsoft SQL Server, Microsoft Exchange, Microsoft Sharepoint Server, Microsoft Dynamics NAV, Oracle, MySQL, SAP, BI Application, CRM Application, Other _________________

5. What is the size of the Backend Data? Less than 100GB, 100 to 500GB, 500GB to 2TB, 2 to 10TB, 10TB to 1PB, I do not know

6. What is the I/O Ratio (Read/Write mix)? Mostly Reads, Mostly Writes, Approximately Equal Reads and Writes, Don't Know

7. Is the Application Clustered? No, Yes, in an Active/Passive mode, Yes, Application is Load-Balance, Other ___________________________

8. Do you have any other information to provide?

Troubleshooting

For all issues, send us the following information:

- Platform information:
  - OS and kernel version
  - CPU quantity & model
  - RAM size
  - Cache and core device details – disk model, storage controller, and device capacity
  - Intel® CAS for Linux* version
- Workload description:
  - Type of workload – eg, sysbench, swingbench, fio, live Oracle DB, etc.
  - read vs. write % mix
  - data localization (“hot spot”) parameters
- Steps to reproduce
- Expected results
- Actual results
- *dmesg* output

Troubleshooting

Prior to reproducing the issue:

- Clear cache statistics (eg., *casadm --reset-counters --cache-id 1 --core-id 1*)
- Ensure appropriate logging level (*dmesg -n 8*)
- Ensure that kernel hung task detector is enabled (*echo 60 > /proc/sys/kernel/hung_task_timeout_secs* – this will print the stack trace of the hung task after 60s)

Additional information to provide for functional issues, depending on the defect type:

- System crash:
  - */var/log/messages* output
  - serial debug logs
  - */proc/slabinfo* dump
- Performance issue:
  - Extended cache statistics before run (eg., *casadm --stats --cache-id 1*)
  - During run:
    - IO information (*iostat -xmt 1*)
    - Examine CPU utilization (*top*). If CAS processes are contributing to high CPU utilization, note the PIDs and capture stack dump for each (*cat /proc/<cas_pid>/stack*)
    - *memstat* output
  - Extended cache statistics after run (eg., *casadm --stats --cache-id 1*)

Case Study – Caching from DAS

Challenge

Poor/inconsistent application performance caused by I/O bottleneck

HDD Storage can't keep pace with CPU/Server performance

Solution

Add Intel® CAS and Intel® SSD 910 as a caching layer on the application Server

- Lowest latency with direct attached PCIe SSD card
- Most active data automatically placed on the SSD

Up to 50X IOPS Increase

1. Intel® Server model 2600CO (Copper Pass); Intel® E5-2680 processor (2.7GHz); 32GB DDR2/1333 memory; Microsoft* Windows 2008R2 SP1, Intel® CAS 2.0 release candidate 1; IOMeter 10.22.2009; 4K Random Read test; 32-queue depth; 800Gig Intel SSD 910 series or Intel SSD DC 3700 Series, Intel RAID RS25AB080 with MR54p1 firmware; 8 x 10K SAS HDD in a RAID0 array. *Other names and brands may be claimed as the property of others.

Case Study – Caching from SAN/NAS

Caching from SAN/NAS

Challenge

Existing SAN or NAS with HDD Storage can’t keep pace with CPU/Server performance

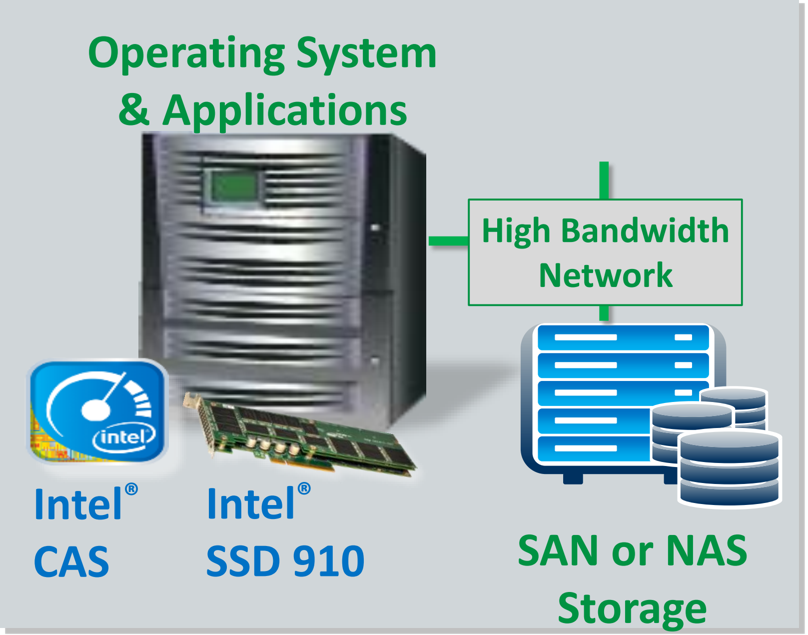
Latency of off-server provisioned storage

Solution

Add Intel® CAS and Intel® SSD 910 as a caching layer to the application Server

- No changes to SAN or NAS storage
- Data read from the storage node and cached locally on the application server
- No application changes, data migration, or manual tiering

Near SSD speed, no migration



1. Intel® Server model 2600CO (Copper Pass), Intel® E5-2680 processor (2.7GHz), 32GB DDR2/1333 memory; Microsoft* Windows 2008R2 SP1, Intel® CAS 2.0 release candidate 1; IOMeter 10.22.2009 ; 4K Random Read test; 32-queue depth; 800Gig Intel SSD 910 series or Intel SSD DC 3700 Series, Intel RAID RS25AB080 with MR54p1 firmware; 8 x 10K SAS HDD in a RAID0 array *Other names and brands may be claimed as the property of others

Case Study – Caching in VM environment

Challenge

Need consistent virtualized application performance to meet customer SLAs

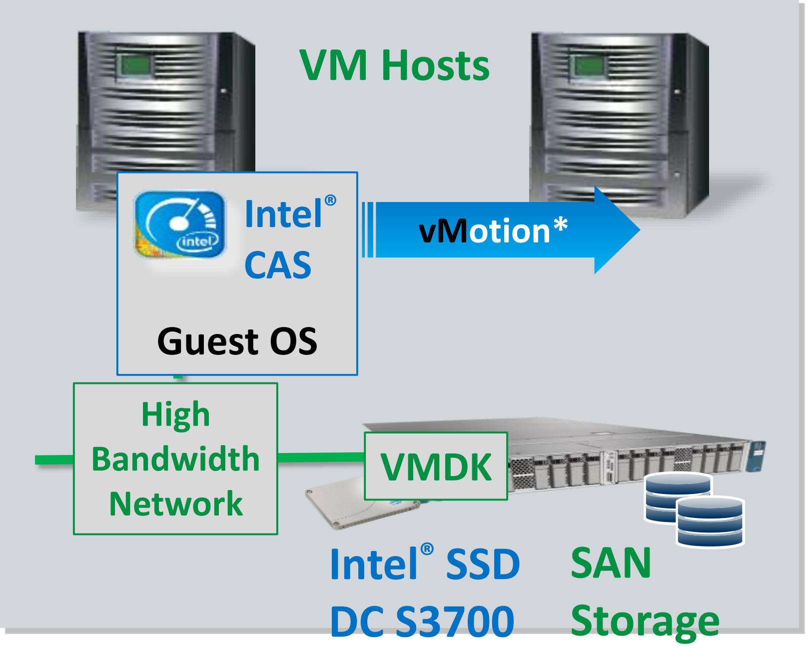
VM and application performance restricted by I/O blender effect

Solution

Intel® CAS installed in Guest OS, caching to Intel® SSD DC S3700 on shared storage

- Live VMware* vMotion* while caching
- VMs move automatically for Load Balancing and Quality of Service
- VM arrives on target host with hot cache intact

Live migration with hot cache



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

Intel® Solid-State Drive Data Center P3700 Series

Consistently Amazing

Consistent, Native PCIe* Performance

Up to 6X throughput, up to 2X latency reduction vs. SATA[1]

- 2800/1900 MB/s (Read/Write)
- 460/180K 4K Random Read IOPS
- 250K 4K 70/30 Read/Write Mixed Workload IOPS

Stress-Free Protection

Data storage with multiple checkpoints helps protect data

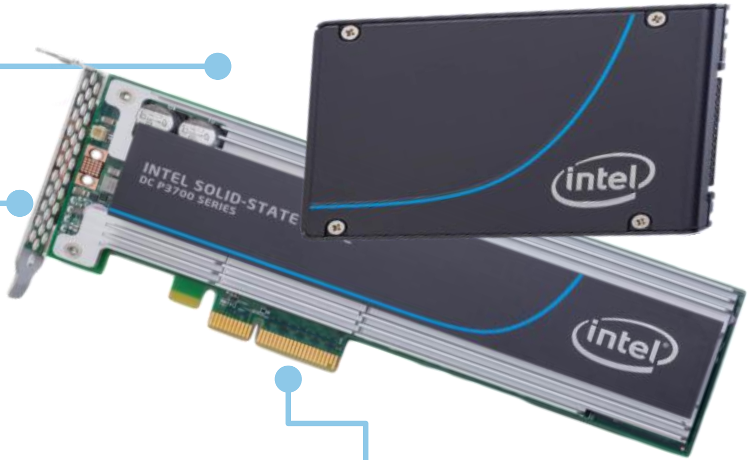
- End-to-end data path protection
- Power loss protection with built in self-test
- 2.5" hot-plug capability
- Uncorrectable Bit Error Rate (UBER): 1 sector per 10^{17} bits read

230 yrs

High Endurance

Leading technology provides optimal price performance

- Intel® 20nm MLC NAND with High Endurance Technology
- 10 full capacity drive writes per day over 5 years*



High Capacity

Broad range of capacities in a single volume

- 400/800 GB 1.6/2.0 TB
- 2.5" x 15mm, SFF-8639
- Add-In-Card (AIC) half-height, half-length card

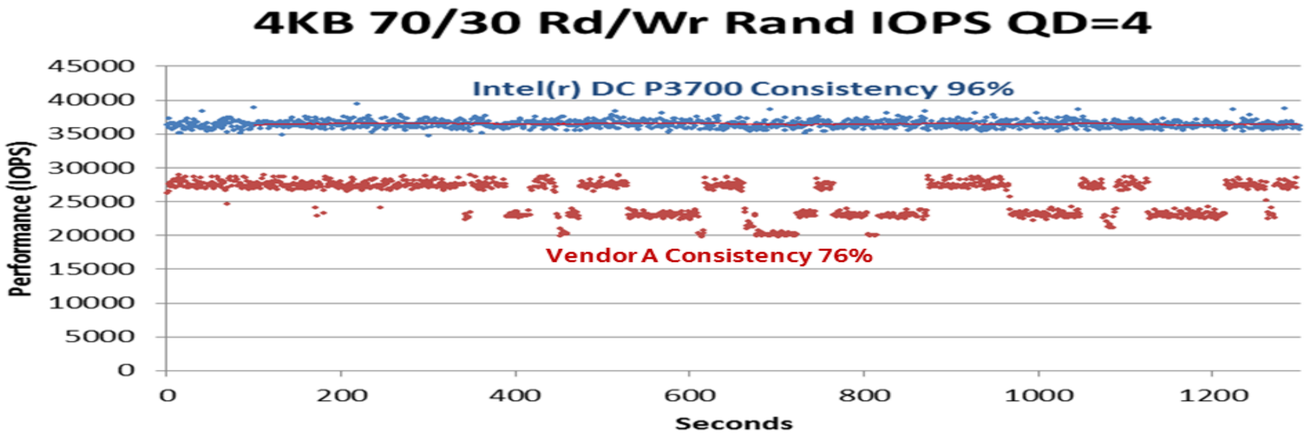
Modernize Your Data Center Storage with Breakthrough Performance

Non-Vo

* Other names and brands are property of their respective owners

[1]Data source: Intel® SSD DC P3700 Series datasheet based upon internal Intel testing *JESD 219 workload

Intel® SSD DC P3700 Series Consistently Amazing



Source: Intel measured on Dell R920, Microsoft Server 2012, DC P3700 1.6TB vs Vendor A NVMe* 1.6TB, 70% Rd/30% Wr, 4KB transfer, queue depth = 4

Designed for Real Data Center Applications

- ✓ High consistency enables scalable performance across RAID sets
- ✓ Right balance of read/write performance optimizes mixed workloads
- ✓ Low latency at low queue depths delivers high performance

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.

* Other names and brands are property of their respective owners

High-Endurance Technology (HET[1])

SLC-Like Endurance

Deploy with confidence
Intel® SSD Data Center P3700 Series with HET provides SLC-like endurance for data center peace of mind

Intel® HET = firmware + hardware optimizations

Meet the most intensive needs
You can write the entire capacity of an Intel SSD Data Center P3700 Series 10 times a day for five years (36PB[5] total)

10 drive writes per day (DWPD)

How much is 36PB of data?

459 years of HDTV video[2]

Stack of CDs **21 miles** high[2]

More than **12x** the memory capacity of the human brain[3]

Roughly **3452x** the size of the Library of Congress print collection[4]

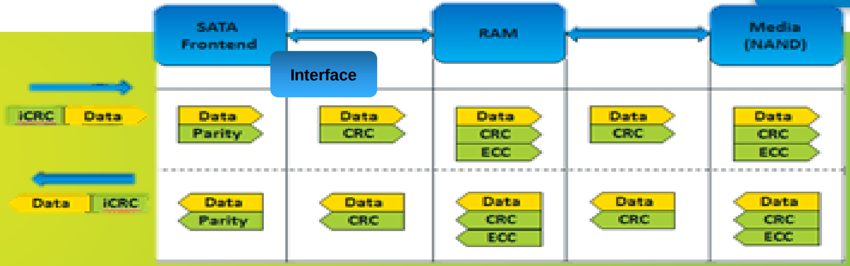
[1] High-Endurance Technology
[2] *Nature International Weekly Journal of Science* (1/19/11)
[3] http://www.geek.com/articles/chips/blue-waters-petaflop-supercomputer-installation-begins-20120130.
[4] http://en.wikipedia.org/wiki/List_of_unusual_units_of_measurement.
[5] 14.6 PB refers to the 2TB capacity point.

Stress-Free Protection

Data Integrity, Reliability

Enable accuracy

- End-to-end data protection feature provides security checkpoints from interface to NAND
- Parity, CRC, ECC, and LBA tag validation[1]



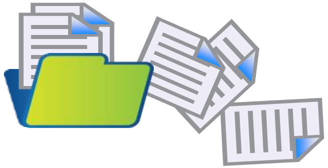
Protect against data loss

- Power Loss Imminent (PLI[1]) with self-test enables in-flight data is written to NAND before shutdown
- NAND XOR automatically recovers data from die, block, or page failure
- Low Density Parity Check ECC NAND (new for 20nm!)
- 2 Million hour MTBF, 230 years
- UBER, 1 sector per 10^{17} bits read

NAND XOR

RAID 5 w Distributed Parity

The Evidence Shows…



Greater risk of data loss without data protection features

Enhanced protection with SSD DC P3700, P3600, and P3500

[1] Abbreviations : CRC – Cyclical Redundancy Check, LBA – Logical block Address, PLI – Power Loss Imminent

Intel® Solid-State Drive DC S3700 Series

sistently Amazing

Fast and Consistent Performance

Deliver data at a breakneck pace, with consistently low latencies and tight IOPS distribution.

- 75K Random Read IOPS[1]

Stress-Free Protection

Protect your data center applications with multiple secure checkpoints that provide protection against data loss and corruption.

- Full data path and non-data path protection
- Power safe write cache with built in self-test

High-Endurance Technology

Meet your most demanding needs with marathon-like write endurance of 10 full drive writes per day over five years

[1] 4K Random Reads [2] As measured by Intel:100GB 4K Random Writes QD=1 at 99.9 %of the time across 100% span of the drive
Configuration: Intel DH67CFB3; CPU i5 Sandy Bridge i5-2400S LGA1155 2.5GHz 6MB 65W 4 cores CM8062300835404; Heatsink: HS - DHA-B LGA1156 73W Intel E41997-002 and E97379-001; Memory: 2GB 1333 Unbuf non-ECC DDR3 ; 250GB HDD 2.5in SATA 7200RPM Seagate ST9250410AS Momentus 3Gb/s; Mini-ITX Slim Flex w/PS Black Sentey 2421; Ulink Power Hub; SATA Data and Power Combo 24 in. Orange EndPCNoise Sata fp7lp4

platform connected, customer inspired, technology driven