SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2016

Hardware Based Compression in Ceph OSD with BTRFS

Weigang Li (weigang.li@intel.com)
Tushar Gohad (tushar.gohad@intel.com)

Data Center Group
Intel Corporation

Credits

This work wouldn't have been possible without contributions from –

- Reddy Chagam (anjaneya.chagam@intel.com)
- Brian Will (brian.will@intel.com)
- Praveen Mosur (praveen.mosur@intel.com)
- Edward Pullin (edward.j.pullin@intel.com)

Agenda

- Ceph
  - A Quick Primer
  - Storage Efficiency and Security Features
- Storage Workload Acceleration
  - Software and Hardware Approaches
- Compression in Ceph OSD with BTRFS
  - Compression in BTRFS and Ceph
  - Hardware Acceleration with QAT
  - PoC implementation
  - Performance Results
- Key Takeaways

Ceph

- Open-source, object-based scale-out storage system
- Software-defined, hardware-agnostic – runs on commodity hardware
- Object, Block and File support in a unified storage cluster
- Highly durable, available – replication, erasure coding
- Replicates and re-balances dynamically

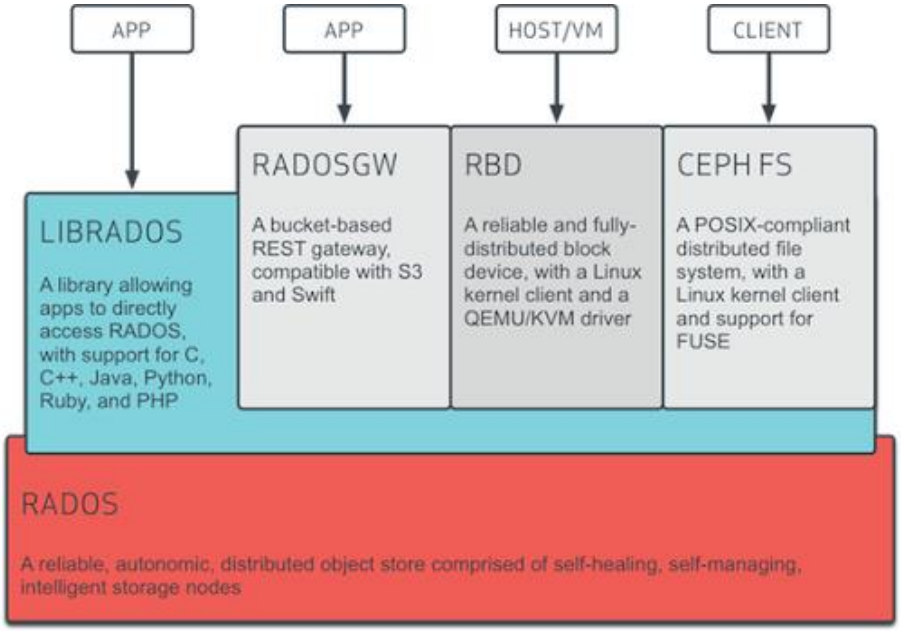
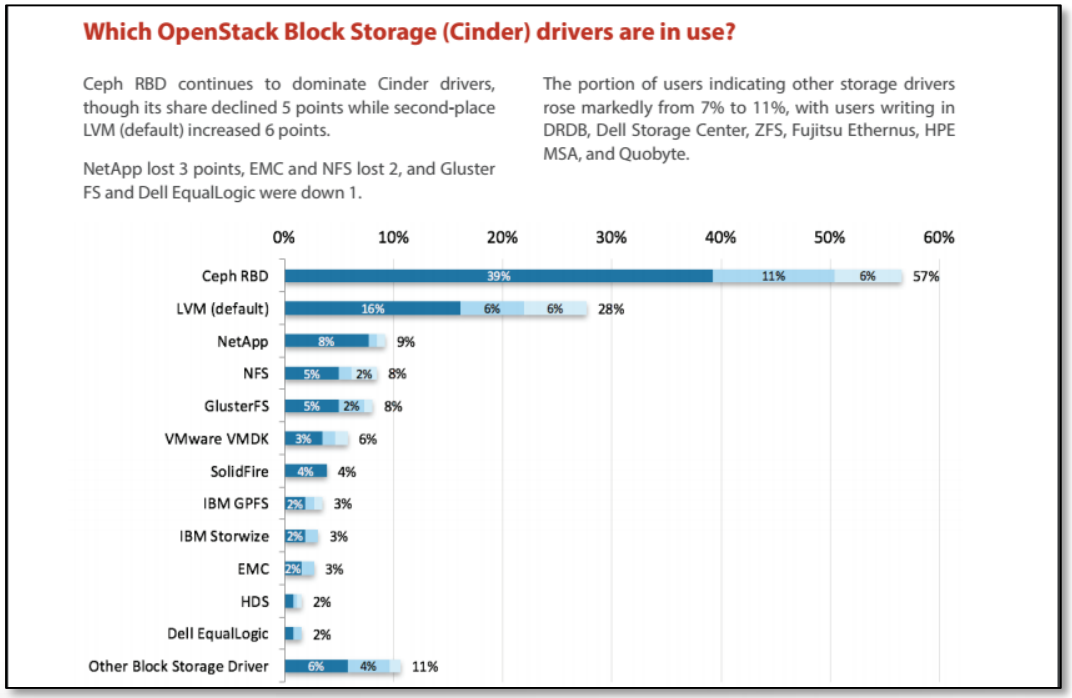


Image source: http://ceph.com/ceph-storage

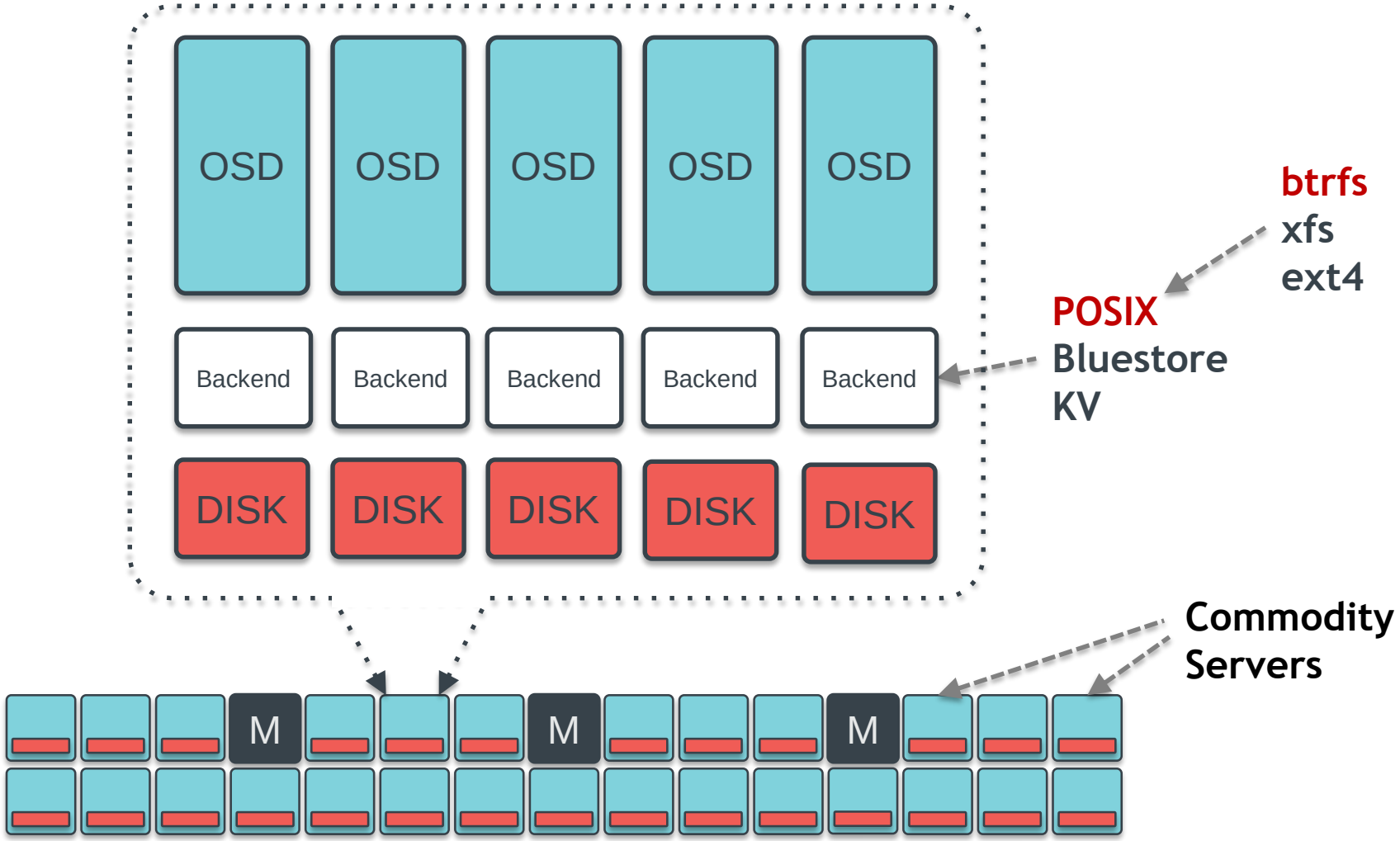
Ceph

- Scalability – CRUSH data placement, no single POF
- Enterprise features – snapshots, cloning, mirroring
- Most popular block storage for Openstack use cases
- 10 years of hardening, vibrant community



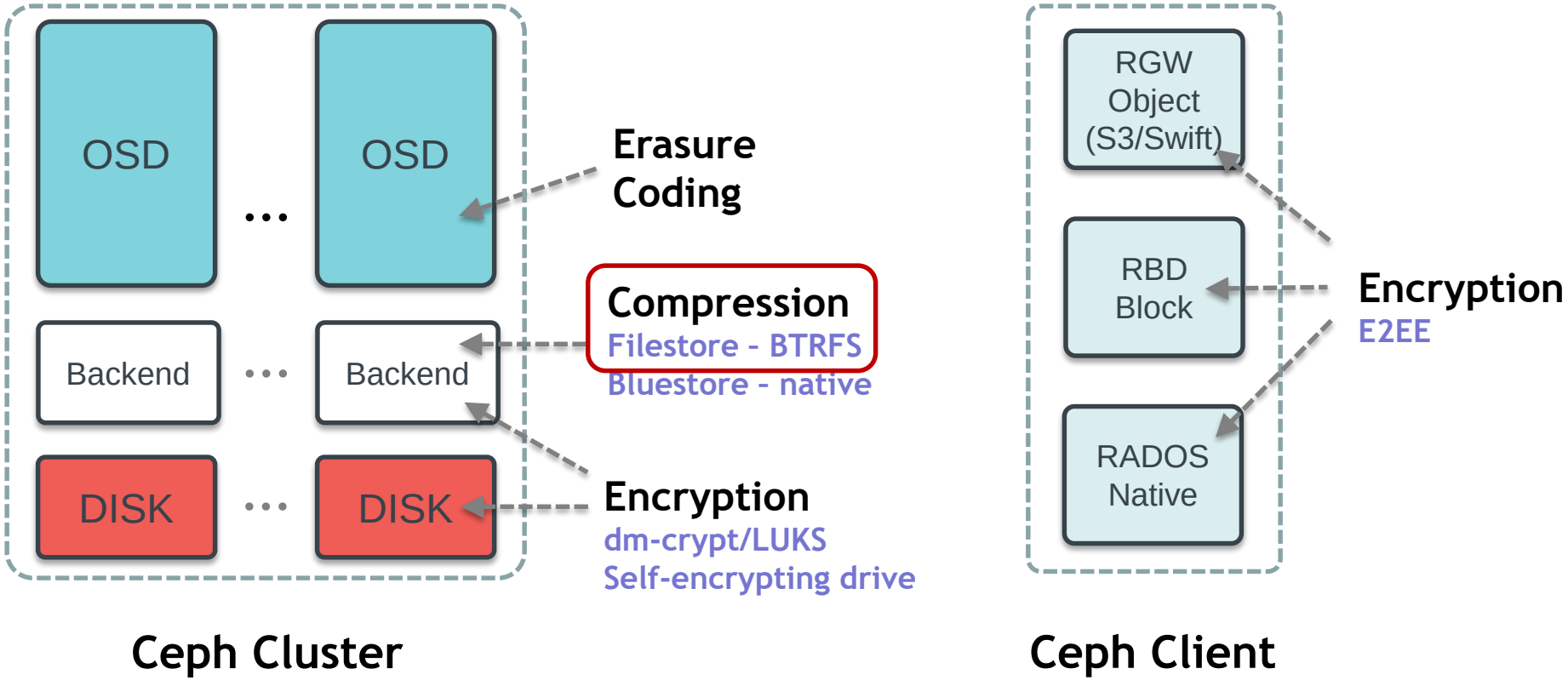
Source: http://www.openstack.org/assets/survey/April-2016-User-Survey-Report.pdf

Ceph: Architecture



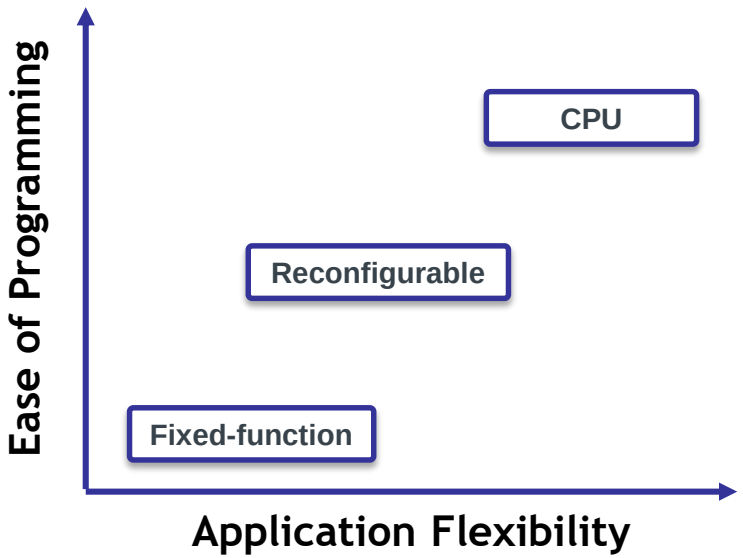
Ceph: Storage Efficiency, Security

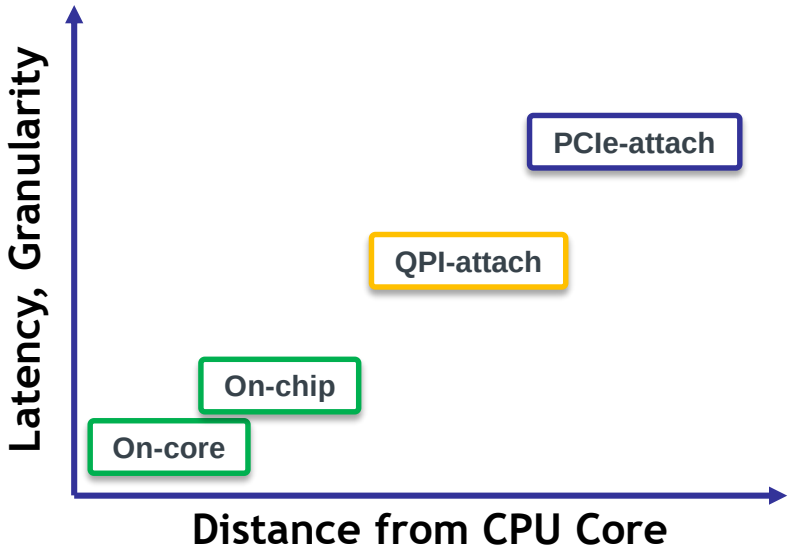
Erasure Coding, Compression, Encryption



Storage Workload Acceleration

Software and Hardware-based Approaches





ISA-L (SIMD)

Software-based Approaches

ASIC (QAT) | FPGA | GPGPU

Fixed-function, Reconfigurable Approaches

Software-based Acceleration with ISA-L

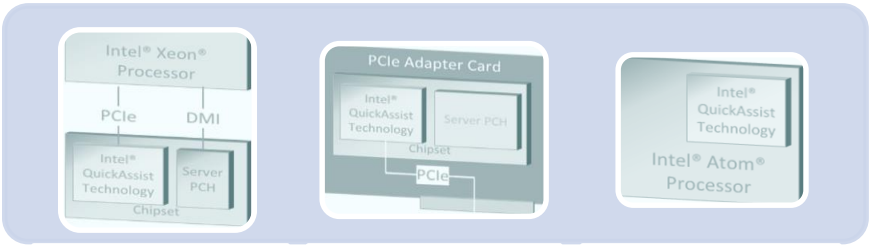
Intel® Intelligent Storage Acceleration Library

- Collection of optimized low-level storage functions
- Uses SIMD primitives to accelerate storage functions in software
- Primitives supported
  - Compression – gzip
  - Encryption – AES block ciphers (XTS, CBC, GCM)
  - Erasure Coding – Reed-Soloman
  - CRC (iSCSI, IEEE, T10DIF), RAID5/6
  - Multi-buffer hashes (SHA1, SHA256, SHA512, MD5), Multi-hash
- OS agnostic: Linux, FreeBSD etc
- Commercially-compatible, Free, Open Source under BSD license
  - https://github.com/01org/isa-l_crypto
  - https://github.com/01org/isa-l
  - https://software.intel.com/en-us/storage/ISA-L

Hardware-based Acceleration

Intel® QuickAssist Technology

Bulk Cryptography	• Security (symmetric encryption and authentication) for data in flight and at rest
Public Key Cryptography	• Secure Key Establishment (asymmetric encryption, digital signatures, key exchange)
Compression	• Lossless data compression for data in flight and at rest



Chipset

Connects to CPU via on-board PCI Express* lanes

PCI Express* Plugin Card

SoC

Connects to CPU via on-chip interconnect

Open-source Software Support

Cryptography	OpenSSL libcrypto, Linux Kernel Crypto Framework
Data Compression	zlib (user API), BTRFS/ZFS (kernel), Hadoop, Databases

Hardware acceleration of compute intensive workloads (cryptography and compression)

Support for Offloads in Upstream Ceph

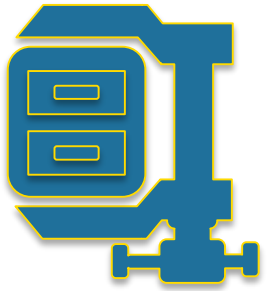
Intel® ISA-L and QAT

- Erasure Coding
  - ISA-L offload support for Reed-Soloman codes
  - Supported since Hammer
- Compression
  - Filestore
    - QAT offload for BTRFS compression (kernel patch submission in progress)
  - Bluestore
    - ISA-L offload for zlib compression supported in upstream master
    - QAT offload for zlib compression (work-in-progress)
- Encryption
  - Client-side (E2EE)
    - RADOS GW encryption with ISA-L and QAT offloads (work-in-progress)
    - Qemu-RBD encryption with ISA-L and QAT offloads (under investigation)

Today's Focus: Compression

More Data

10x Data Growth from 2013 to 2020[1]

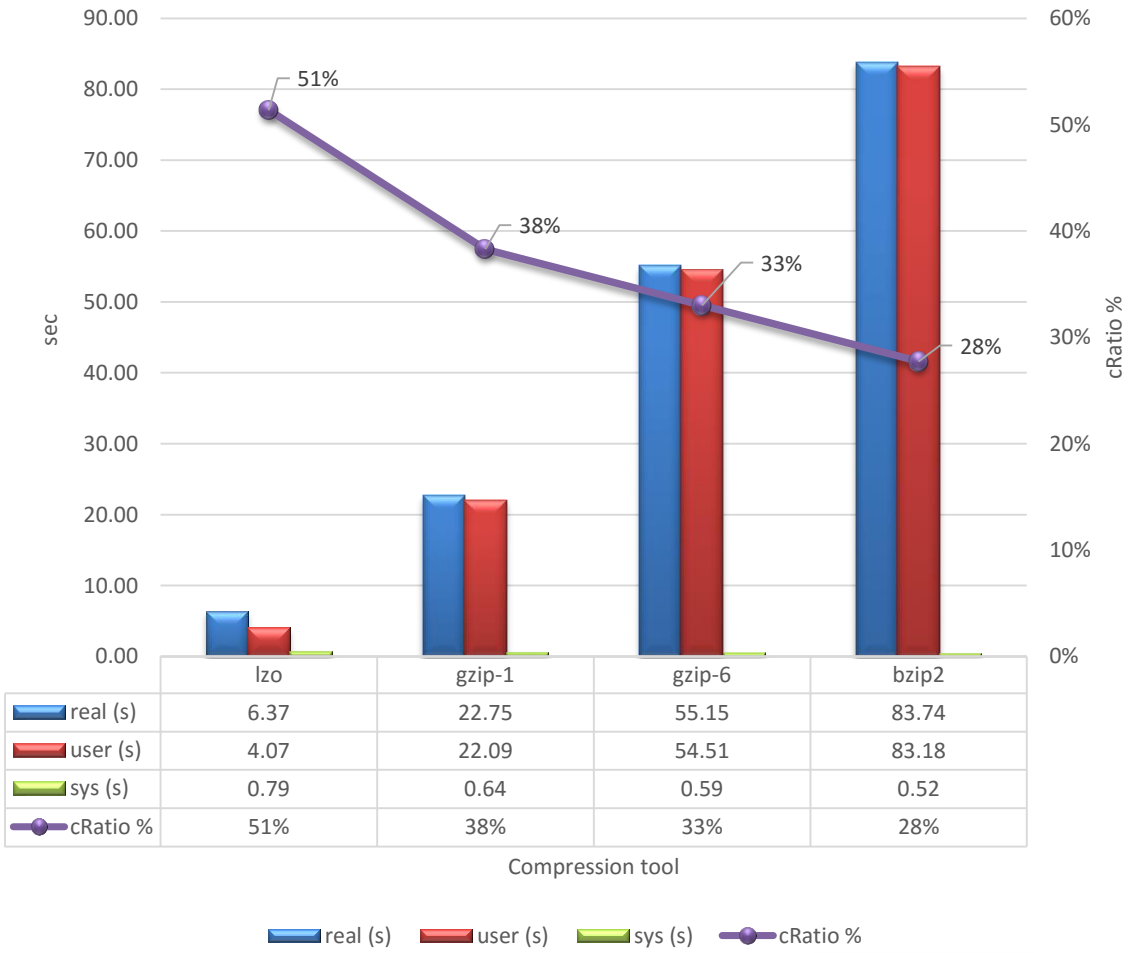


- Digital universe doubling every two years
- Data from the Internet of Things, will grow 5x
- % of data that is analyzed grows from 22% to 37%

Compression can save storage capacity

[1] Source: April 2014, EMC* Digital Universe with Research & Analysis by IDC*

Compression: Cost



	lzo	gzip-1	gzip-6	bzip2
real (s)	6.37	22.75	55.15	83.74
user (s)	4.07	22.09	54.51	83.18
sys (s)	0.79	0.64	0.59	0.52
cRatio %	51%	38%	33%	28%

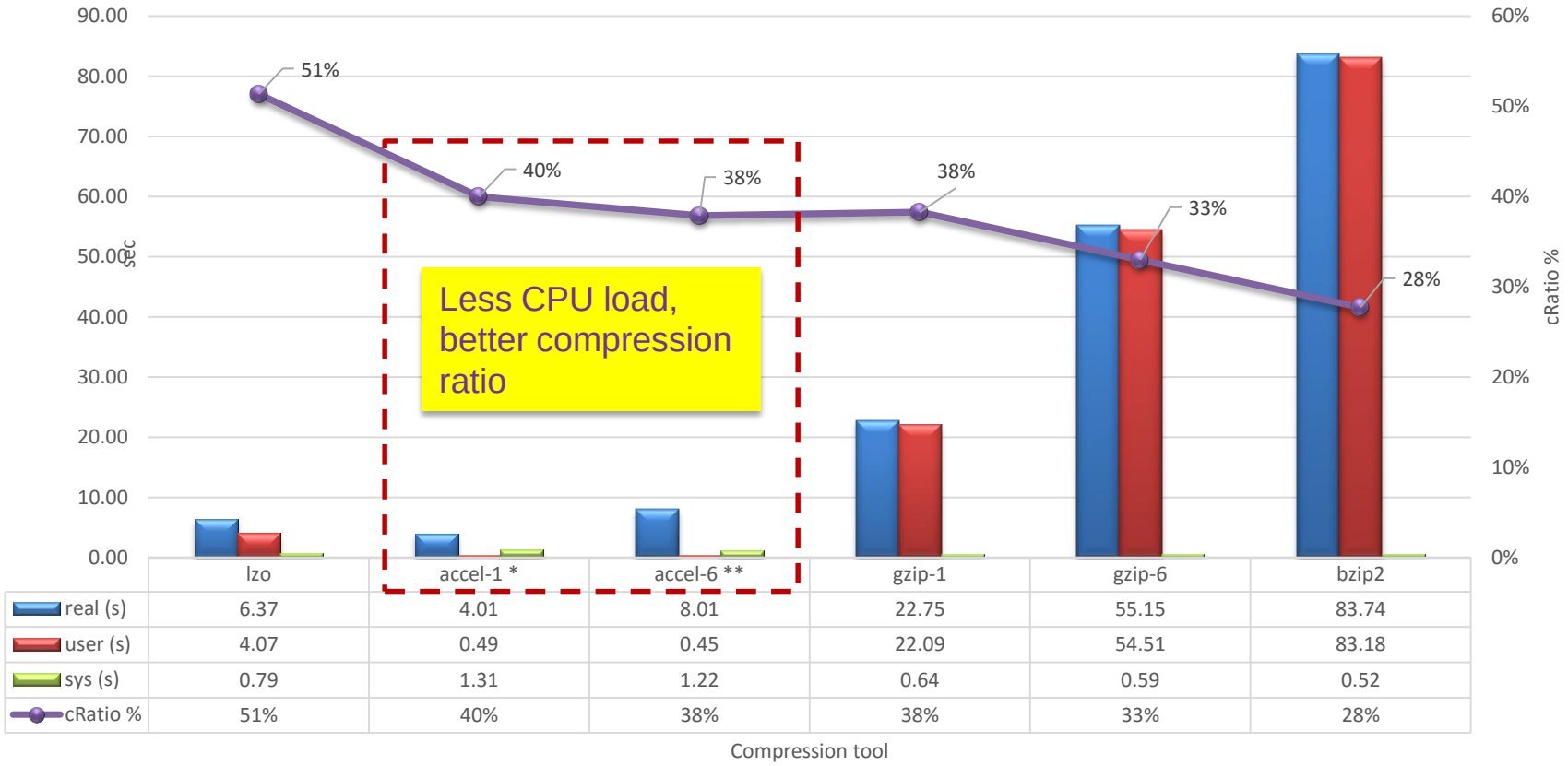
- Compress 1GB Calgary Corpus* file on one CPU core (HT).
- Compression ratio: less is better

  cRatio = compressed size / original size
- CPU intensive, better compression ratio requires more CPU time.

Source as of August 2016: Intel internal **measurements** with dual E5-2699 v3 (18C, 2.3GHz, 145W), HT & Turbo Enabled, Fedora 22 64 bit, DDR4-128GB
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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. Any difference in system hardware or software design or configuration may affect actual performance. 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. **For more information go to** http://www.intel.com/performance

*The *Calgary corpus* is a collection of text and binary data files, commonly used for comparing data compression algorithms.

Benefit of Hardware Acceleration



	lzo	accel-1 *	accel-6 **	gzip-1	gzip-6	bzip2
real (s)	6.37	4.01	8.01	22.75	55.15	83.74
user (s)	4.07	0.49	0.45	22.09	54.51	83.18
sys (s)	0.79	1.31	1.22	0.64	0.59	0.52
cRatio %	51%	40%	38%	38%	33%	28%

* Intel® QuickAssist Technology DH8955 level-1

** Intel® QuickAssist Technology DH8955 level-6

Compress 1GB Calgary Corpus File

Source as of August 2016: Intel internal **measurements** with dual E5-2699 v3 (18C, 2.3GHz, 145W), HT & Turbo Enabled, Fedora 22 64 bit, 1 x DH8955 adaptor, DDR4-128GB
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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. Any difference in system hardware or software design or configuration may affect actual performance. 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. **For more information go to** http://www.intel.com/performance

Transparent Compression in Ceph: BTRFS

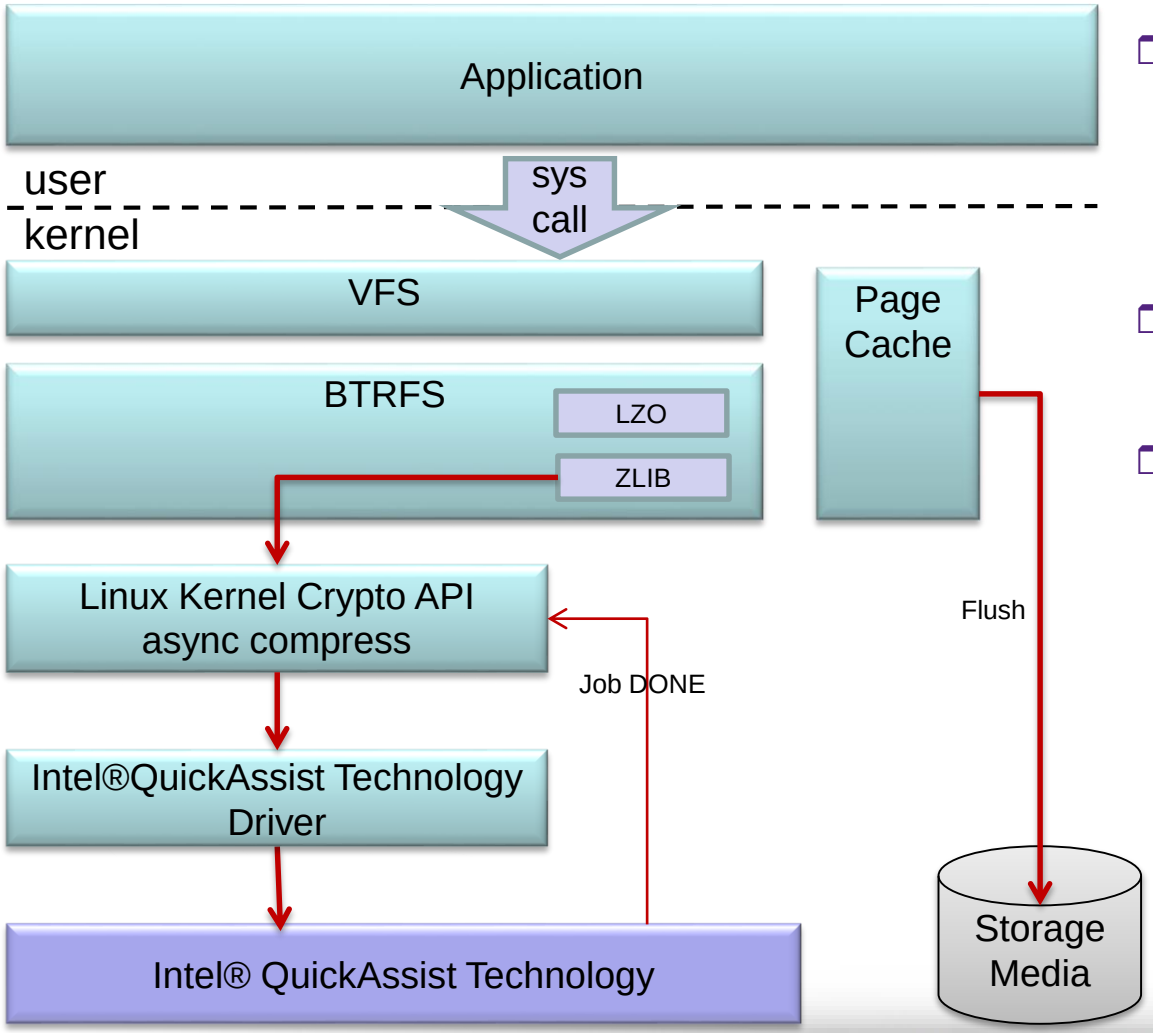
- Copy on Write (CoW) filesystem for Linux.
- *"Has the correct feature set and roadmap to serve Ceph in the long-term, and is recommended for testing, development, and any non-critical deployments… This compelling list of features makes btrfs the ideal choice for Ceph clusters"**
- Native compression support.
  - Mount with "compress" or "compress-force".
  - ZLIB / LZO supported.
  - Compress up to 128KB each time.

* http://docs.ceph.com/docs/hammer/rados/configuration/filesystem-recommendations/

BTRFS Compression Accelerated with Intel® QuickAssist Technology

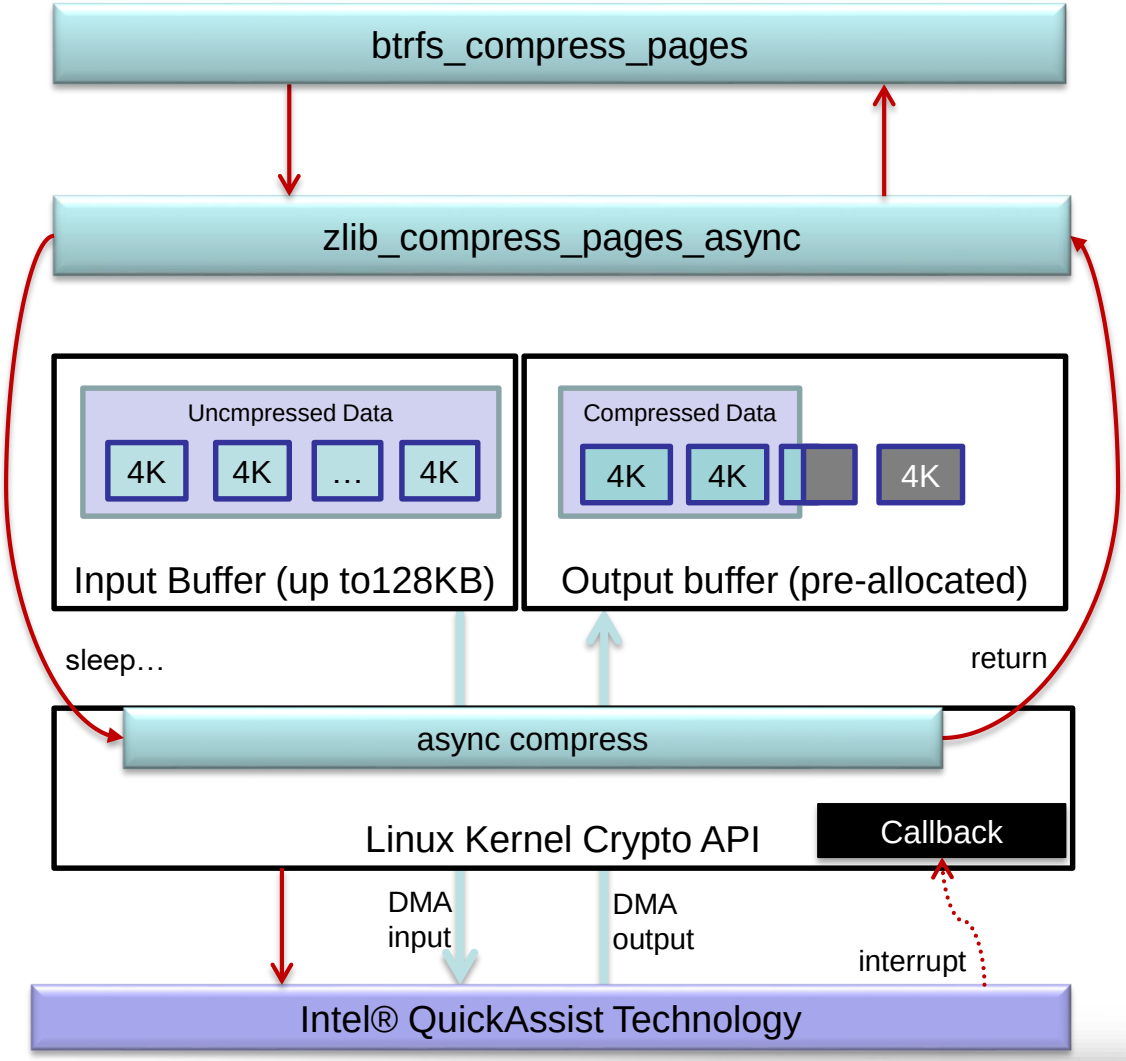
- BTRFS currently supports LZO and ZLIB:
  - The Lempel-Ziv-Oberhumer (LZO) compression is a portable and lossless compression library that focuses on compression speed rather than data compression ratio.
  - ZLIB provides lossless data compression based on the DEFLATE compression algorithm.
    - LZ77 + Huffman coding
    - Good compression ratio, slow
- Intel® QuickAssist Technology supports:
  - DEFLATE: LZ77 compression followed by Huffman coding with GZIP or ZLIB header.

Hardware Compression in BTRFS



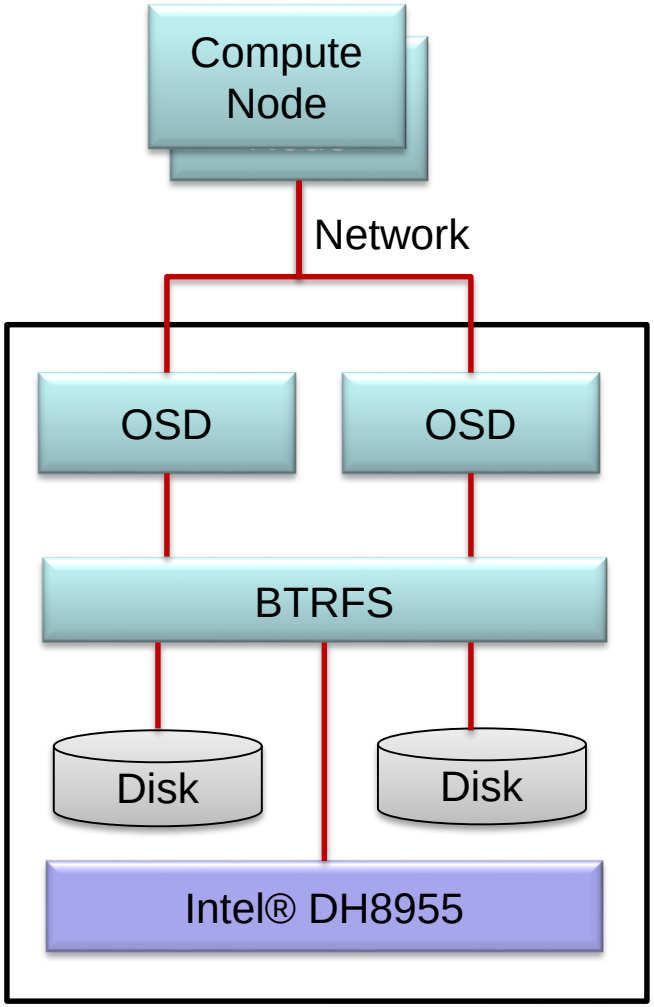
- BTRFS compress the page buffers before writing to the storage media.
- LKCF select hardware engine for compression.
- Data compressed by hardware can be de-compressed by software library, and vise versa.

Hardware Compression in BTRFS (Cont.)



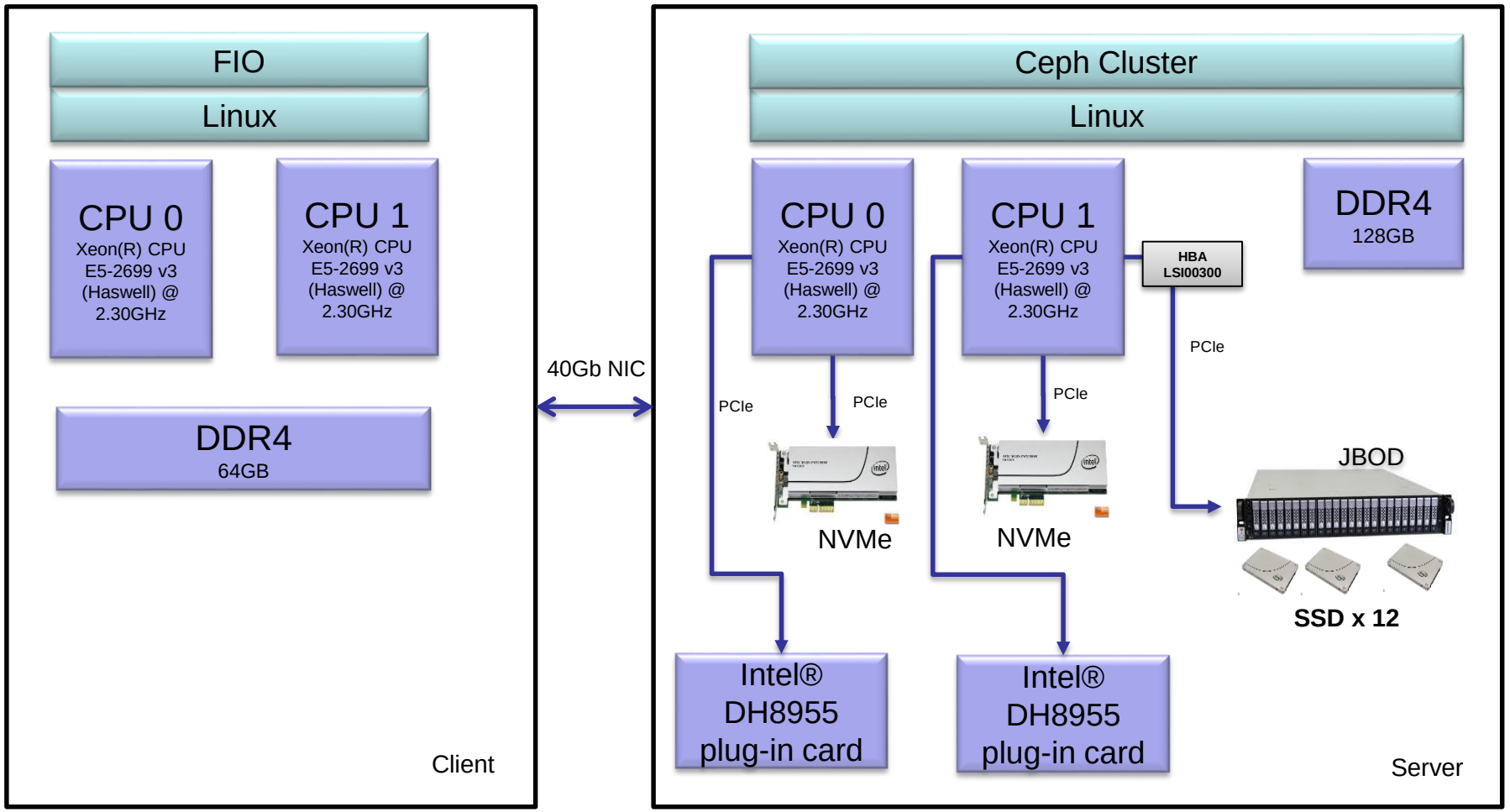
- BTRFS submit “async” compression job with sg-list containing up to 32 x 4K pages.
- BTRFS compression thread is put to sleep when the “async” compression API is called.
- BTRFS compression thread is woken up when hardware complete the compression job.
- Hardware can be fully utilized when multiple BTRFS compression threads run in-parallel.

Compression in Ceph OSD

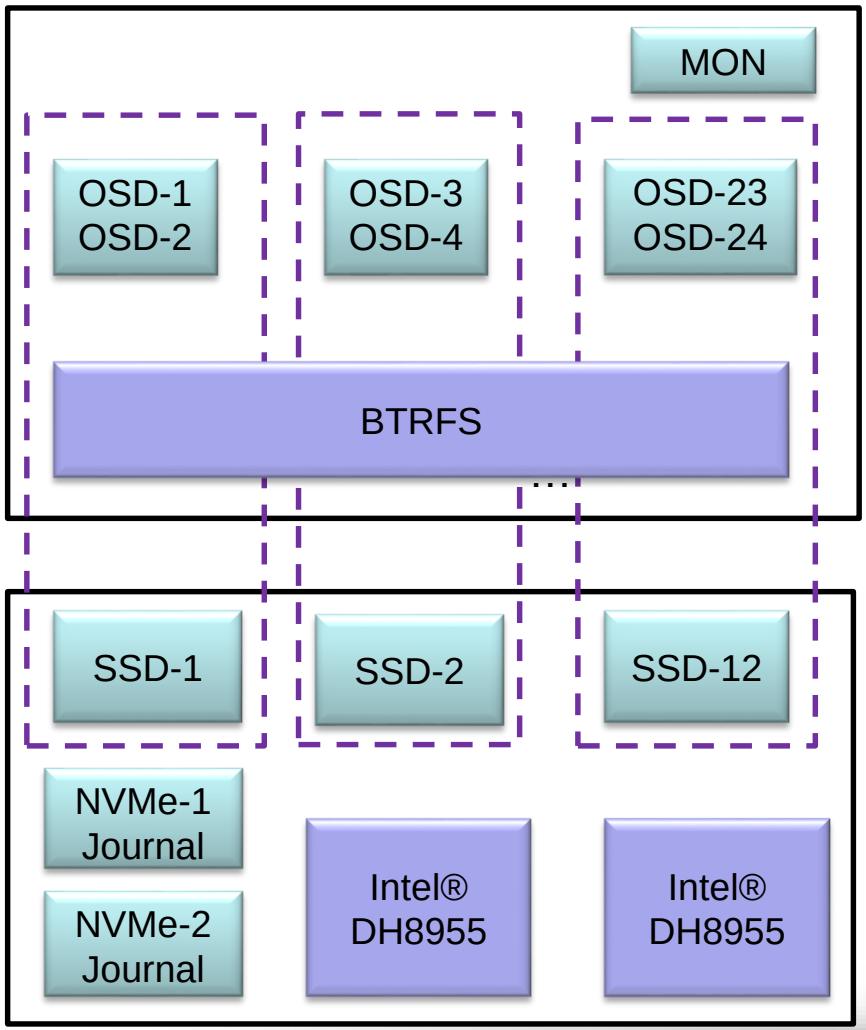


- Ceph OSD with BTRFS can support build-in compression:
  - Transparent, real-time compression in the filesystem level.
  - Reduce the amount of data written to local disk, and reduce disk I/O.
  - Hardware accelerator can be plugged in to free up OSDs' CPU.

Benchmark - Hardware Setup

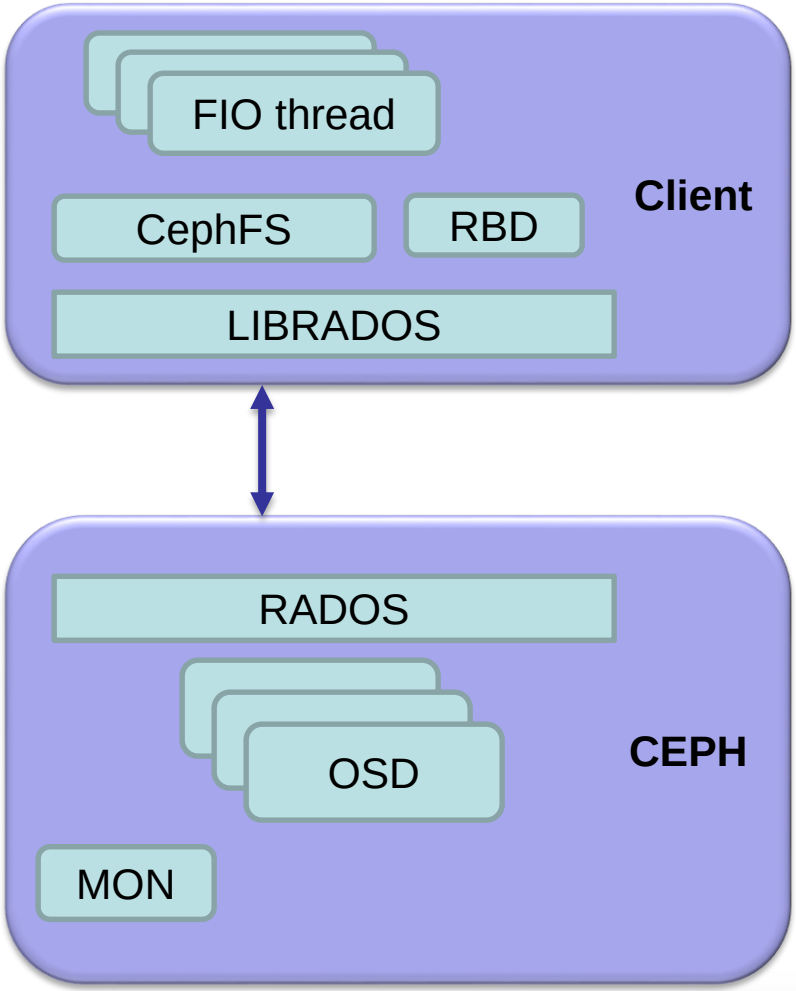


Benchmark - Ceph Configuration



- Deploy Ceph OSD on top of BTRFS as backend filesystem.
- Deploy 2 OSDs on 1 SSD → 24x OSDs in total.
- 2x NVMe for journal.
- Data written to Ceph OSD is compressed by Intel® QuickAssist Technology (Intel® DH8955 plug-in card).

Test Methodology

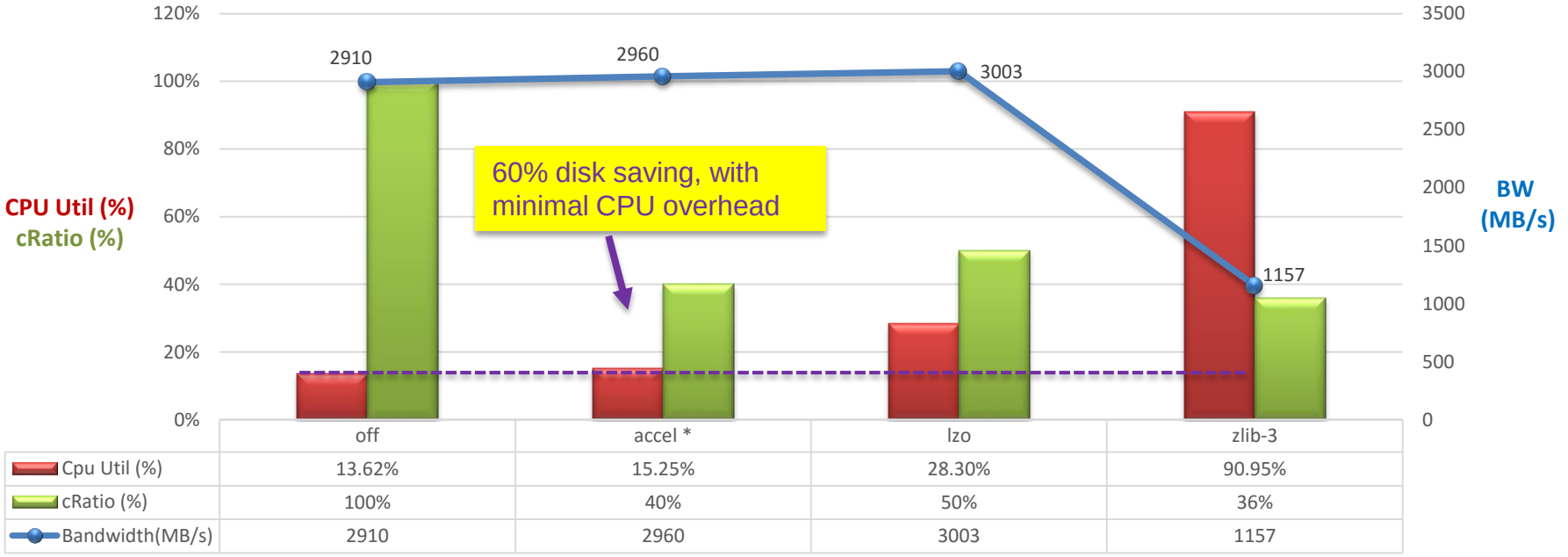


- Start 64 FIO threads in client, each write / read 2GB file to / from Ceph cluster through network.
- Drop caches before tests. For write tests, all files are synchronized to OSDs' disk before tests complete.
- The average CPU load, disk utilization in Ceph OSDs and FIO throughput are measured.

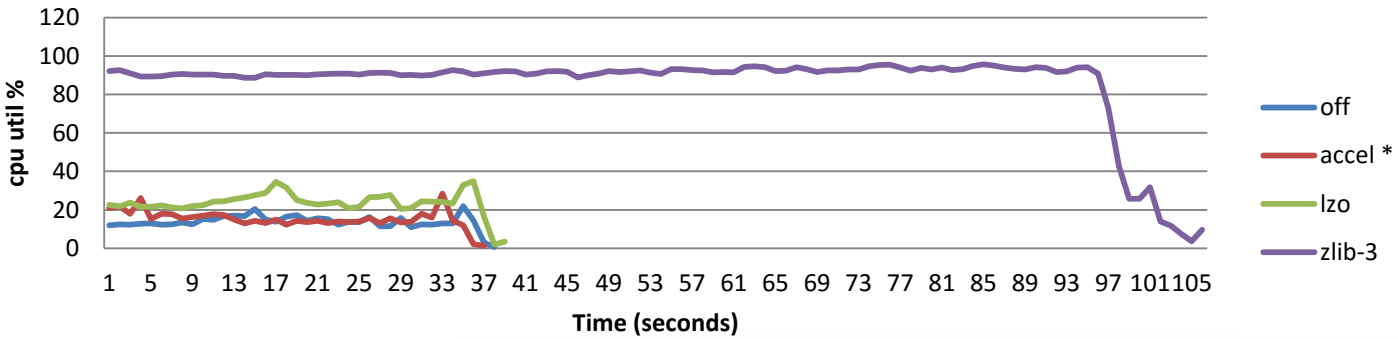
Benchmark Configuration Details

Client	
CPU	2 x Intel® Xeon CPU E5-2699 v3 (Haswell) @ 2.30GHz (36-core 72-threads)
Memory	64GB
Network	40GbE, jumbo frame: MTU=8000
Test Tool	FIO 2.1.2, engine=libaio, bs=64KB, 64 threads
Ceph Cluster	
CPU	2 x Intel (R) Xeon CPU E5-2699 v3 (Haswell) @ 2.30GHz (36-core 72-threads)
Memory	128GB
Network	40GbE, jumbo frame: MTU=8000
HBA	HBA LSI00300
OS	Fedora 22 (Kernel 4.1.3)
OSD	24 x OSD, 2 on one SSD (S3700), no-replica 2 x NVMe (P3700) for journal 2400 pgs
Accelerator	Intel® QuickAssist Technology, 2 x Intel® QuickAssist Adapters 8955 Dynamic compression Level-1
BTRFS ZLIB S/W	ZLIB Level-3

Sequential Write



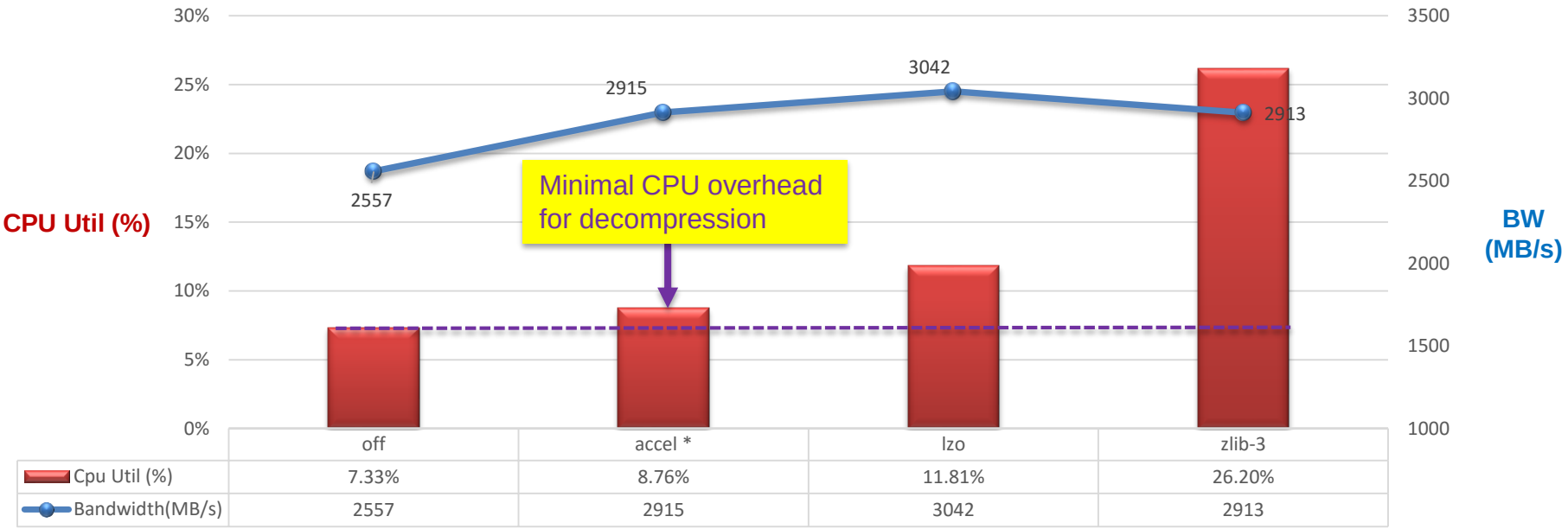
	off	accel *	lzo	zlib-3
Cpu Util (%)	13.62%	15.25%	28.30%	90.95%
cRatio (%)	100%	40%	50%	36%
Bandwidth(MB/s)	2910	2960	3003	1157

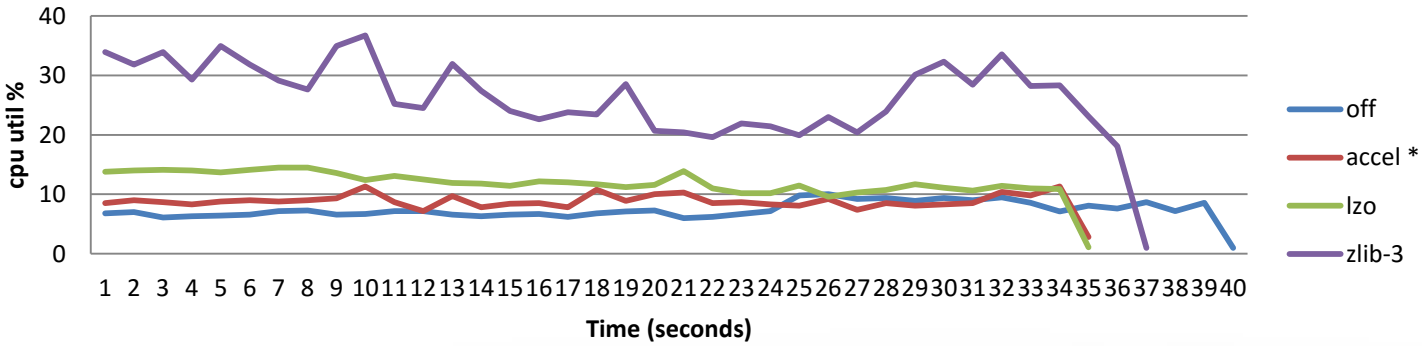


* Intel® QuickAssist Technology DH8955 level-1
** Dataset is random data generated by FIO

Source as of August 2016: Intel internal **measurements** with dual E5-2699 v3 (18C, 2.3GHz, 145W), HT & Turbo Enabled, Fedora 22 64 bit, kernel 4.1.3, 2 x DH8955 adaptor, DDR4-128GB Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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. Any difference in system hardware or software design or configuration may affect actual performance. 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. **For more information go to** http://www.intel.com/performance

Sequential Read





* Intel® QuickAssist Technology DH8955 level-1

Source as of August 2016: Intel internal **measurements** with dual E5-2699 v3 (18C, 2.3GHz, 145W), HT & Turbo Enabled, Fedora 22 64 bit, kernel 4.1.3, 2 x DH8955 adaptor, DDR4-128GB Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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. Any difference in system hardware or software design or configuration may affect actual performance. 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. **For more information go to** http://www.intel.com/performance

Key Takeaways

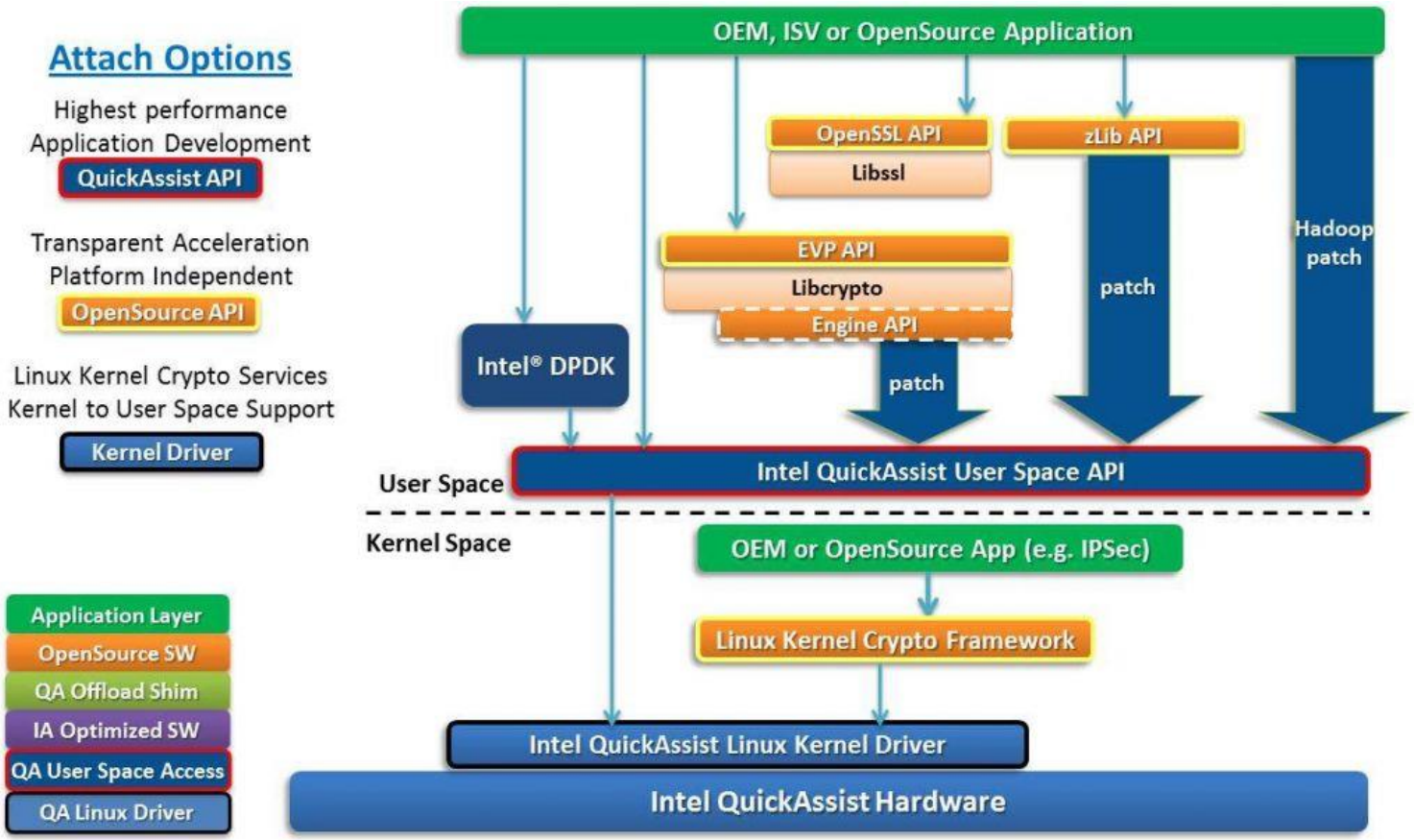
- Data compression offers improved storage efficiency
- Filesystem level compression in OSD is transparent to the Ceph software stack
- Data compression is CPU intensive, getting better compression ratio requires more CPU cost
- Software and hardware offloads methods available for accelerating compression, including ISA-L, Intel® QuickAssist Technology and FPGA
- Hardware offloading method can greatly reduce CPU cost, optimize disk utilization & IO in Storage infrastructure

Additional Sources of Information

- For more information on Intel® QuickAssist Technology & Intel® QuickAssist Software Solutions can be found here:
  - Software Package and engine are available at 01.org: Intel QuickAssist Technology | 01.org
  - For more details on Intel® QuickAssist Technology visit: http://www.intel.com/quickassist
  - Intel Network Builders: https://networkbuilders.intel.com/ecosystem
- Intel®QuickAssist Technology Storage Testimonials
  - IBM v7000Z w/QuickAssist:http://www-03.ibm.com/systems/storage/disk/storwize_v7000/overview.html
    - https://builders.intel.com/docs/networkbuilders/Accelerating-data-economics-IBM-flashSystem-and-Intel-quick-assist-technology.pdf
- Intel's QuickAssist Adapter for Servers: http://ark.intel.com/products/79483/Intel-QuickAssist-Adapter-8950
- DEFLATE Compressed Data Format Specification version 1.3 http://tools.ietf.org/html/rfc1951
- BTRFS: https://btrfs.wiki.kernel.org
- Ceph: http://ceph.com/

Additional Information

QAT Attach Options



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.

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%20 Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations

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

Intel, Intel logo, Intel Core, Intel Inside, Intel Inside logo, Intel Ethernet, Intel QuickAssist Technology, Intel Flow Director, Intel Solid State Drives, Intel Intelligent Storage Acceleration Library, Itanium,, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

64-bit computing on Intel architecture requires a computer system with a processor, chipset, BIOS, operating system, device drivers and applications enabled for Intel® 64 architecture. Performance will vary depending on your hardware and software configurations. Consult with your system vendor for more information.

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology is a security technology under development by Intel and requires for operation a computer system with Intel® Virtualization Technology, an Intel Trusted Execution Technology-enabled processor, chipset, BIOS, Authenticated Code Modules, and an Intel or other compatible measured virtual machine monitor. In addition, Intel Trusted Execution Technology requires the system to contain a TPMv1.2 as defined by the Trusted Computing Group and specific software for some uses. See http://www.intel.com/technology/security/ for more information.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

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

Other vendors are listed by Intel as a convenience to Intel's general customer base, but Intel does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices. This list and/or these devices may be subject to change without notice.

-