Accelerating Ceph Distributed Database Using ThunderX and CloudSpeed Ultra

Whitepaper Sponsored by Cavium and Western Digital
June 20, 2017

Overview

Cavium and Western Digital set out to prove the value of a new cloud service provider (CSP) distributed storage platform that is price per performance competitive with current distributed storage platforms. Current storage platforms typically leverage high-priced Intel Xeon processors with legacy hard disk drives (HDDs). The proposed platform combines the performance and value of Cavium's ThunderX ARMv8-based system-on-chip (SoC) and Western Digital's 3D NAND flash-based SanDisk CloudSpeed Ultra Gen II SATA second-generation solid-state drives (SSDs). Target use cases for this configuration are workloads that are highly variable in nature, but with requirements for low latency I/O access.

Cavium's ThunderX implementation of the 64-bit ARMv8 instruction set is competitive with mainstream x86 servers used in cloud server deployments. But, with its lower pricing, ThunderX has a much better price to performance ratio than Intel's Xeon processors.[1] Cavium has been shipping ThunderX for two quarters and it meets Microsoft's OCP Project Olympus requirements for cloud data centers.[2]

SSDs now have orders of magnitude higher reliability than hard disk drives (HDD). In addition, falling SSD pricing plus lower replication requirements (based on better reliability) and cloud-targeted features sets like those in SanDisk CloudSpeed SSDs are enabling SSDs to replace performance-grade rotating drives (10,000 and 15,000 RPM) as the primary data storage volumes in distributed storage systems. Western Digital has been shipping SanDisk CloudSpeed Ultra SSDs for two years and it has been qualified at the largest cloud providers in North America, China, and in Europe, Middle East, and Africa (EMEA). Western Digital has been instrumental in upstreaming Linux kernel-level optimizations and Ceph performance enhancements for the past few years.[3][4]

Ceph is a good proxy for a broad class of distributed storage environments, because it has become a de facto open-source choice for public and private cloud deployments.

This paper describes the first performance assessment of a Ceph distributed block storage system using solely flash-based SSD coupled with modern 64-bit ARM processors. No rotating drives were used in the systems under test. It is a first look at what is possible today using ARMv8 processors and SSDs in branded or whitebox at-scale distributed storage systems.

---

[1] http://www.tiriasresearch.com/downloads/nginx-cdn-using-thunderx/
[2] https://www.forbes.com/sites/tiriasresearch/2017/03/10/microsoft-taps-arm-for-datacenter/
[3] https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160810_K21_Samuels.pdf
[4] https://www.sandisk.com/about/media-center/press-releases/2015/innovations-and-contributions-to-the-open-source-community

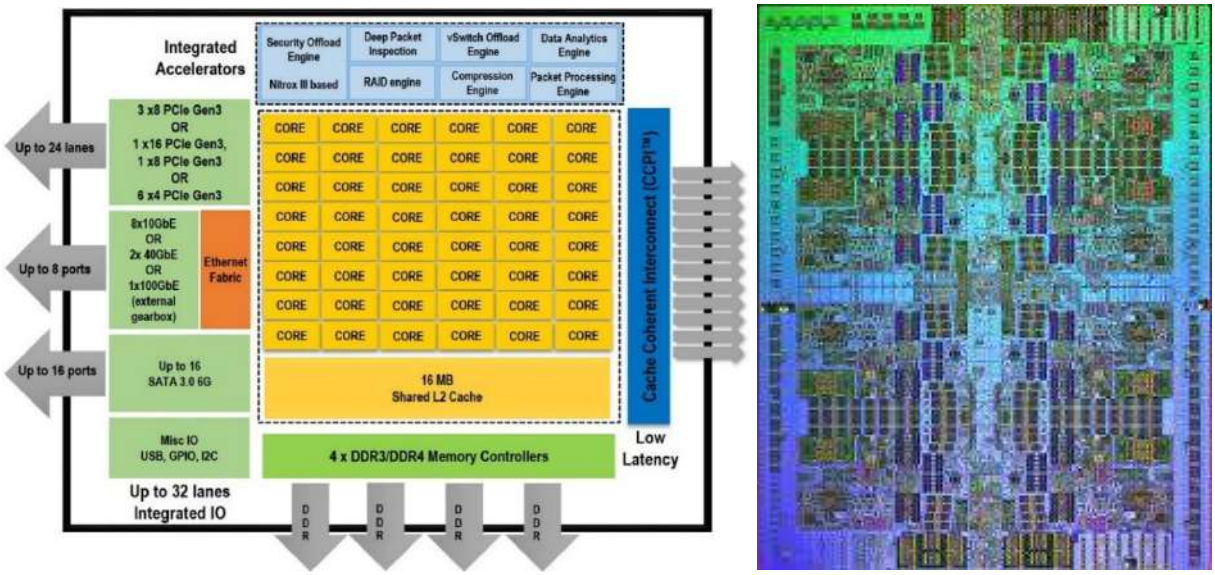
Cavium ThunderX

ThunderX is a Cavium designed custom processor architecture that implements an ARMv8 compliant 64-bit instruction set architecture (ISA). Cavium optimized the ThunderX core and SoC architecture for cloud workloads. The ThunderX SoC implements various configurations of up to 48 single-threaded processor cores on one die. ThunderX is fully compliant with ARM's Server Base System Architecture (SBSA) specification.

Cavium offers several models of the ThunderX SoC[5], each tuned for different workloads via different integrated accelerator, networking, and I/O configurations:

- ThunderX_CP: Compute – public and private cloud servers
- ThunderX_ST: Storage – distributed storage servers
- ThunderX_SC: Security – appliances and cloud radio access networks (C-RAN)
- ThunderX_NT: Network – media servers, scale-out embedded applications, and network function virtualization (NFV)

Figure 1: Cavium ThunderX_AAP All Accelerator Processor (AAP) Block Diagram and Die Photo



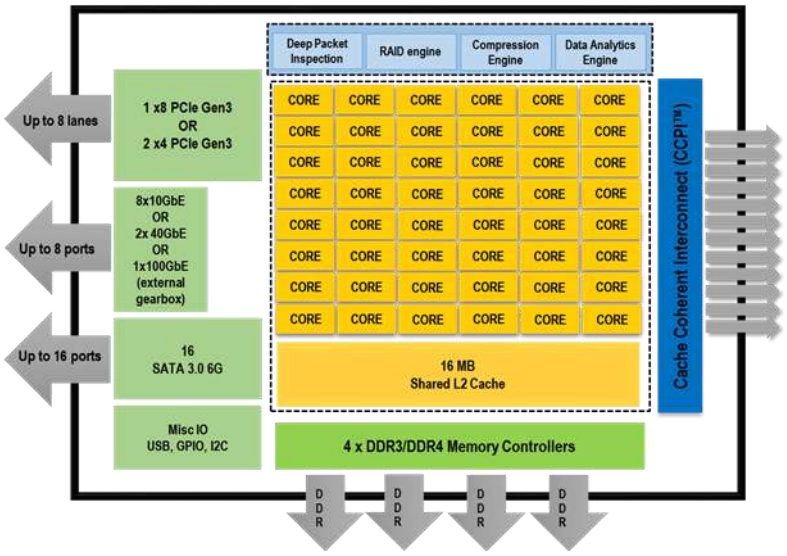
[Source: Cavium]

Cavium also designed the ThunderX as a bootable SoC with all necessary I/O interfaces, as opposed to Intel's Xeon processors which require an additional discrete southbridge chip.

[5] http://www.cavium.com/ThunderX_ARM_Processors.html

Cavium first sampled its ThunderX architecture in 1Q2015, using an "all inclusive" version of the SoC design called "all accelerator processor" or ThunderX_AAP (see Figure 1). This part enables potential customers to evaluate the overall ThunderX SoC architecture before committing to a specific market optimized product model. However, Cavium will not sell the AAP part in production quantities.

Figure 2: Cavium ThunderX_ST Block Diagram



[Source: Cavium]

Cavium started production of the ThunderX_ST SoC in 1Q2016. ThunderX_ST supports dual-socket SoC configurations by including Cavium Coherent Processor Interconnect (CCPI) bus. The CCPI bus directly links two ThunderX processor sockets so that they can share memory coherently without using a southbridge architecture – an industry first for the ARM server market. Each ThunderX_ST SoC also integrates four dual-channel memory controllers and is configurable for up to eight 10Gbps or two 40Gbps Ethernet ports without requiring an add-in NIC (see Figure 2).

Figure 3: AEWIN SCS-4201 ThunderX Processor CN88XX-based Storage Server

[Source: TIRIAS Research]

Cavium optimized its ST family for high network and storage I/O traffic. ThunderX_ST integrates hardware accelerators for compression and for data integrity and protection, plus support for NVMe attached SSDs. Each ThunderX_ST SoC can be configured to directly attach up to 16 SATA 6 Gbps ports, with dual-socket configurations directly attaching up to 32 SATA ports, reducing the need for add-in PCIe-based host baseband adaptors (HBA) to attach more drives.

SanDisk CloudSpeed SSD

Western Digital sells a diverse set of storage products into the data center market under the brands SanDisk and HGST. The Western Digital product portfolio spans a wide range of price and performance using SATA, SAS, and PCIe NVMe interfaces for SSD form factors, and an offering of data center HDDs.

Western Digital has been investing in open-source software for several years, with a focus on improving Ceph's performance using SSDs[6][7] as well as more general cloud database-as-a-service (DBaaS).[8] In 2013 Western Digital first started working on SSD improvements to Ceph's physical storage back end, which became BlueStore[9] in 2016. Western Digital continues to upstream improvements to BlueStore. While BlueStore was not used in this performance analysis, Western Digital's in-depth knowledge of Ceph operations and settings contributed to both hardware choice and software configurations.

Cavium and Western Digital collaborated to select SSDs that represent the best price per performance trade-off available today. Cavium's ThunderX_ST has a high value position for compute and networking capabilities,[10] and so, based on Western Digital's experience with Ceph, ThunderX was paired with SanDisk CloudSpeed Ultra Gen II (second generation) SATA drives for this evaluation.

Western Digital offers two variants of their SanDisk CloudSpeed SSDs – Eco and Ultra. CloudSpeed Eco is intended to lower read latencies and maximum sustained data throughput for read-intensive workloads, such as media streaming and content repositories.

CloudSpeed Ultra is a better choice for Ceph, because it is intended for a variety of mixed-use cloud service provider workloads that demand low-latency response times and predictable random-

[6] https://www.sandisk.com/business/datacenter/products/flash-software/open-source
[7] https://events.linuxfoundation.org/sites/events/files/slides/optimizing_ceph_flash.pdf and in more detail at https://www.terena.org/activities/tf-storage/ws19/slides/20151014-flash.pdf
[8] https://www.sandisk.com/content/dam/sandisk-main/en_us/assets/resources/enterprise/white-papers/how-sandisk-removes-dbaas-adoption-obstacles.pdf
[9] http://7xweck.com1.z0.glb.clouddn.com/cephdaybeijing201608/10-Ceph全闪存存储-周皓.pdf
[10] http://www.tiriasresearch.com/downloads/nginx-cdn-using-thunderx/

write performance. Western Digital optimized CloudSpeed Ultra firmware for a 70% to 30% balance of reads to writes at small file sizes (512 bytes to 4KB) for typical scale-out deployments based on CSP feedback over the past several years.

Figure 4: SanDisk CloudSpeed Ultra Gen II SATA 800GB SSD

[Source: Cavium]

Access density, defined as Input/output Operations per Second (IOPS) per drive capacity, has been decreasing over the past decade. This is happening because drive capacity is increasing faster than drive utilization. Some of this is the result of drive capacities increasing faster than the drive interface speeds, and some of it is due to storage workload demands not keeping pace with capacity. The net effect is low drive utilization that increases both capital expenses (buying more than the optimal number of drives) and operational expenses (powering and supporting those extra drives). A midrange SSD capacity of 800GB was used for this analysis as a means of optimizing access density.

SSDs have compelling advantages over HDDs in storage server applications:

- Higher performance via shorter read and write latencies compared to HDDs
- Lower power consumption of a comparable HDD, roughly half in many cases
- Higher reliability and lower cost because eliminating the spinning media also eliminates vibration and vibration counter measures. This increases the reliability of the rest of the components and reduces server chassis cost and complexity.
- Higher data integrity with fewer replaced copies, as measured by an Unrecovered Bit Error Rate (UBER)[11] of 1 in 10^{18} bits for SanDisk CloudSpeed SSDs, as opposed to 1 in 10^{14} to 10^{16} bits for HDDs

[11] https://www.jedec.org/standards-documents/dictionary/terms/uncorrectable-bit-error-rate-uber

Orders of magnitude lower UBER means that most distributed storage systems may consider implementing 2x replication (two copies of each block, file, or object), rather than more traditional 3x or higher replication factors. Data center grade SSDs have reached the point of reliability where a leading DBaaS provider uses SanDisk CloudSpeed SSDs at RAID 0 (with availability zone replication) for its NoSQL-as-a-Service, with copies across multiple availability zones.

Ceph Block Storage Architecture

Ceph[12] is a distributed, massively scalable software-defined storage (SDS) platform built from open source software. Ceph was designed to be provisioned on scale-out Linux-based servers using locally attached commodity storage devices, such as conventional HDDs, SSDs, or PCIe attached storage devices. Ceph's architecture turns clusters of standard servers, each with local commodity storage, into redundant, distributed, high-performance pools of storage.

Ceph is best known for object storage because it is a commonly used object storage system for OpenStack users, but it also includes interfaces for file and block storage built on top of its object storage. The object storage system allows users to mount Ceph as a thinly-provisioned block device. Thin provisioning is a storage virtualization technique and is based on the notion that users typically do not consume all the storage they have been allocated. As users consume more of their virtual storage allotments, more physical storage can be added over time.

There are three logical components to a generic Ceph storage cluster:

- **Object Storage Device (OSD)**: stores data, retrieves data, handles replication and balancing, and all other core data storage functions. An OSD manages its own logical storage volume of direct attached storage (DAS). Ceph recommends that one OSD controls one physical hardware storage device. There must be at least two OSDs per cluster, which means two physical storage volumes, because two OSDs defines a baseline 2x replication set. An OSD is an independent application instance and is performance sensitive. Ceph recommends that each OSD run on a dedicated hardware thread – either a dedicated single-threaded core or a dedicated hardware thread on a simultaneous multithreading capable core.
- **Monitor**: tracks the cluster topology and state. Think of it as an executive manager; there must be at least one monitor in each cluster, and it tracks the OSDs. High availability (HA) requires at least three monitors; more specifically, an odd number of monitors with at least one in each availability failure domain.[13] Monitors are not performance sensitive, they are

[12] http://ceph.com/
[13] http://docs.ceph.com/docs/hammer/rados/deployment/ceph-deploy-mon/

not directly involved in the data flow between users and OSDs. Monitors communicate with each other to scale their respective OSDs into a single distributed storage system, spanning hundreds of thousands of storage devices.

- **Meta Data Server (MDS)**: makes it possible for POSIX file system users to run basic file system commands without affecting the performance of the cluster. We mention MDS here for completeness, but it is not used in Ceph block storage clusters.

Unlike network attached storage (NAS) or storage area networks (SAN), Ceph couples and replicates/duplicates the storage control function with the storage media to prevent single points of failure. This is the heart of SDS design philosophy, and enables a small set of simple rules to generate sophisticated and reliable large-scale behavior.

Thus far, Ceph has focused on optimizing its performance for HDD storage devices. Only recently has Ceph supported SSD devices for caching data in-flight and for keeping OSD journals and file system metadata. Because of high historical costs relative to HDDs, SSDs are dismissed as appropriate devices for storage volumes:

"While SSDs are cost prohibitive for object storage…"[14]

The original implementation of Ceph was written before data center grade SSDs became prevalent in the marketplace. Short-stroke (refers to the mechanical disk read head mechanism) 10,000 and 15,000 revolutions per minute (RPM) spinning HDD arrays had been the traditional storage media of choice for block-based performance. SSDs do not require those read wait states and have different write characteristics.

BlueStore is an on-going effort to better align the value of SSD with strong market demand for all-flash Ceph block-based deployments. BlueStore's goal is to align Ceph performance with SSD configurations at lower read and write latency thresholds. [Note that the configuration described in this report did not use BlueStore.]

Therefore, without such modifications to BlueStore, Ceph OSDs spend a lot of time waiting for data when it does not need to. This behavior inhibits SSD adoption, even though the rest of SSD economics and performance are moving in the right directions for performance-oriented adoption.

A further challenge for SSDs in cloud and hyperscale workloads is that, from a storage point of view, those workloads are boring. Compared to traditional on-prem enterprise workloads, cloud workloads tend to demand:

[14] http://docs.ceph.com/docs/master/start/hardware-recommendations/#solid-state-drives

- More random reads and writes
- Higher data transfer rates and IOPS
- Less idle I/O time
- Greater standardization of back-end infrastructure

CSPs are using object storage layered on top of block storage to cost-effectively scale out their storage back-end. This means they can buy large volumes of block I/O configured HDD or SSD products. However, cloud workload application stacks may not fully utilize all of the provided NVMe bandwidth and may not be able to justify NVMe's higher cost structures. While NVMe SSD IOPS benchmarks look impressive, practically speaking CSPs are reluctant to fully standardize on NVMe-based platforms until application stacks are tuned to take advantage of NVMe specifications.

Inexpensive, low latency, small block size SATA SSDs can be used today to build performant scale-out solutions. CSPs are evaluating SSDs for Ceph volume storage. But for SSDs to be widely deployed, the Ceph developer community must continue to improve and optimize BlueStore.[15]

Software Assessment Methodology

This section describes how the system was tested to assess Ceph performance using Cavium ThunderX SoCs and SanDisk CloudSpeed SSDs, and justification for this methodology. The appendices describe the software used for the performance analysis.

First, only OSD performance was tested:

- SSDs were not RAIDed, as Ceph is responsible for replication
- As the MDS tracks file hierarchy and metadata, Ceph block device and RADOS-level benchmarks do not require metadata, this is strictly a block storage test[16]

As this is a straightforward block storage assessment, much like testing any local storage device, RADOS Block Devices (RBD) engine[17], a Ceph wrapper for the Flexible I/O Tester (FIO)[18], was used to assess storage performance under two test conditions:

[15] http://sched.co/7w3A and http://sched.co/81qU
[16] Ceph introduces a new set of acronyms. Rather than spell all of them out please refer to Ceph's glossary at http://docs.ceph.com/docs/master/glossary/
[17] https://github.com/axboe/fio/blob/master/engines/rbd.c
[18] https://github.com/axboe/fio

- Random read IOPS at 4KB and 8KB block sizes
- Random write IOPS at 4KB and 8KB block sizes

The FIO benchmark parameters that have the most impact to this performance assessment are:

Queue Depth (QD): the number of read or write requests that can be issued but outstanding (unserved) at any point in time. At QD = 1 each request must be served before the next request can be issued (in the processor world this is called "in-order" execution). At QD = 32, there can be 32 outstanding requests at any time; as each request is fulfilled, another may be issued (called "out-of-order" execution). One and 32 are considered to be benchmarking bookends for driving access density and were used in this analysis. Higher QD factors present diminishing returns. For example, QD = 16 is not twice as fast as QD = 8.

Replication Factor: the number of copies of each data block maintained by the Ceph cluster. Ceph defaults to three copies, a replication factor of three (3x). As mentioned above, SSDs are reliable enough that most applications can achieve high availability with only 2x replication, but both factors were used for this analysis.

CSPs have settled on a performance ratio of 70% read to 30% write drive optimization as representative of most of their large-scale workloads. They use 90% read to 10% write for read-intensive workloads, but 100% read is close enough to gage read-intensive performance. Because CSPs do not distribute production code samples to component manufacturers, there are no other widely accepted performance analysis ratios or code bases.

See the Appendix for much more detail regarding software used.

Hardware Assessment Methodology

Ceph recommends the following minimum hardware capabilities for OSD storage nodes[19] (see Table 1).

Ceph calls for the server OS to be in a separate partition if it is on the same drive as the storage volume. The test chassis each dedicated one SSD to the OS partition. Each storage volume SSD was mapped to specific ThunderX socket, as they were connected by the ThunderX built-in SATA ports. The performance impact is that each of the three chassis were tested for 23 OSDs paired to 23 SSDs, with the 24th SSD dedicated to the chassis OS and does not contribute to chassis storage capacity or performance.

[19] http://docs.ceph.com/docs/master/start/hardware-recommendations/

Table 1: Ceph OSD Node Hardware Recommendations and System as Implemented

Component	Recommended Cluster Minimum		Implemented	
	Quantity	Type	Quantity	Type
Processor	1	64-bit x64 or 32-bit ARM dual-core	1	64-bit ARMv8 with one core per OSD
Memory	~1	GB RAM per TB storage	13*	GB RAM per TB storage
Volume Storage	1	Storage drive per OSD	1	SSD per OSD
Journal	1	Optional	0	Not Implemented
Network	2	1Gbps Ethernet ports†	2	40Gbps Ethernet ports

* 256GB system memory per server, split evenly between 24 active threads, where each OSD thread manages one 800GB SSD

† Named in Ceph documentation as "GB Ethernet NICs", which is incorrect for both Ethernet rated speed (GB vs. Gbps) and for level of integration (add-in NICs vs. integrated ports)

Source: TIRIAS Research

The monitor was run on a separate server from the three-server chassis being analyzed. As a result, the monitor had no impact on this performance assessment.

The three systems networked into a Ceph storage cluster used Gigabyte's MT60-SC0 motherboard with the following configuration per server chassis (see Table 2).

Table 2: Ceph OSD System Under Test Specifications

Component	Count	Capacity Each	Total Used Per Chassis	Total Used for 3 Chassis
ThunderX CN8890_AAP	2	48 Cores	23 Cores	69 Cores
DDR4 RDIMM 1866 MHz	16	16 GB	256 GB	768 GB
CloudSpeed Ultra Gen. II 2.5-inch SATA SSD, 6Gb	24	800 GB	18.4 TB	55.2 TB
Integrated 40 Gbps NIC	2	40 Gbps	80 Gbps	240 Gbps

Source: TIRIAS Research

Each of the three servers was a dual-socket system with ThunderX processors in both sockets. However, only 23 cores out of 48 total cores were running OSD threads in each of the three actively scheduled ThunderX sockets – one core per each physical storage volume SSD. There were no other threads running on the ThunderX processors.

Within each chassis, the two SoC sockets are connected by CCPI. Instead of using a single-socket with an add-in PCIe-based SATA HBA, twelve SSDs were directly attached to each ThunderX socket via SATA 3.0 (6Gpbs), for a total of 24 SSDs per chassis and 72 SSDs across the three

chassis. An add-in PCIe-based SATA HBA used in a single-socket ThunderX_ST server would have similar performance to the SATA ports integrated on the core-disabled socket.

Cavium ThunderX_AAP was used because Cavium's in-house testing facility is designed to benchmark a wide range of potential customer applications. ThunderX_AAP performance is identical to ThunderX_ST performance when running Ceph, but the full complement of accelerators integrated into ThunderX_AAP might have added a little power consumption over ThunderX_ST. ThunderX_AAP is otherwise identical to ThunderX_ST in terms of design revision and speeds supported.

Cavium's ThunderX SoCs can address up to 512 GB of memory per socket, or a full terabyte (TB) in a dual-socket configuration, but its maximum 2S memory configuration was not evaluated due to time constraints.

Preliminary Results

The first series of performance analysis results are shown in Table 3. They demonstrate the expected effects:

- Small blocks have a higher IOPS than larger blocks (4KB vs 8KB)
- Higher IOPS can be attained against fewer replications (2x vs. 3x)
- Queue depth allows much greater performance (one request at a time vs. a large out-of-order queue, with more work needed to identify the diminishing returns curve)
- Read performance should be the same regardless of the replication factor, however, for small sizes we expected and observed that read performance is CPU limited

Table 3: Ceph IOPS Performance Using ThunderX_ST and SanDisk CloudSpeed Ultra

	Queue Depth	1		32	
	Replication	3X	2X	3X	2X
Block Size	Operation				
4KB	Write 100%	2,481	2,665	27,300	42,200
	Write 30%	**1,256**	**1,381**	**19,600**	**22,600**
	Read 70%	**2,931**	**3,224**	**45,600**	**52,700**
	Read 100%	7,352	7,164	87,200	87,900
8KB	Write 100%	2,162	2,393	21,500	32,900
	Write 30%	**1,109**	**1,214**	**16,700**	**19,800**
	Read 70%	**2,587**	**2,835**	**37,600**	**46,200**
	Read 100%	6,923	7,065	96,600	98,900

Data Source: Cavium

Access densities are shown in Table 4. Because SSD capacities were not changed, either through using different drive models with different storage capacities or by disabling OSDs, the table becomes a simplified version of Table 3.

We observe that Cavium's ThunderX SoC is keeping up with the OSD's workload demand for both 100% reads and 100% writes at QD = 1, there is trivial difference in Access Density from 2x to 3x replication. We also see the same behavior in 100% read performance at QD = 32. However, buffered (but single-threaded) 100% write performance slows by 53% (8KB blocks) and 54% (4KB blocks) from 2x to 3x replication, which reflects the time spent making one more copy (a little over the theoretical minimum increase of 50%).

Table 4: Ceph Access Density (IOPS per GB capacity) Using ThunderX_AAP and SanDisk CloudSpeed Ultra

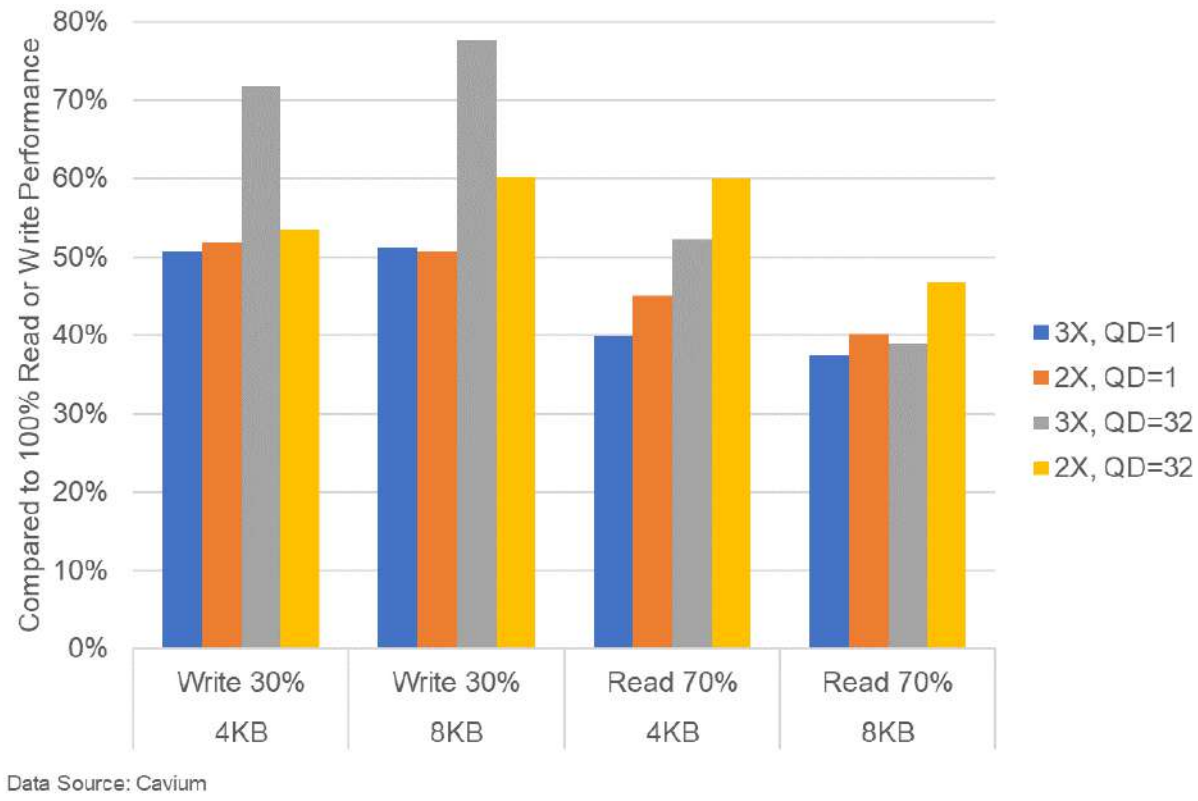
	Queue Depth	1		32	
	Replication	3X	2X	3X	2X
Block Size	Operation				
4KB	Write 100%	0.04	0.05	0.49	0.76
	Write 30%	**0.02**	**0.03**	**0.36**	**0.41**
	Read 70%	**0.05**	**0.06**	**0.83**	**0.95**
	Read 100%	0.13	0.13	1.58	1.59
8KB	Write 100%	0.04	0.04	0.39	0.60
	Write 30%	**0.02**	**0.02**	**0.30**	**0.36**
	Read 70%	**0.05**	**0.05**	**0.68**	**0.84**
	Read 100%	0.13	0.13	1.75	1.79

Data Source: Cavium

Figure 5 highlights that for the 70% read / 30% write environments, write performance is comparatively good, at over half of the access density of 100% writes. However, the read mix comes in at less than half of the access density of 100% reads, except for the smaller block size of 4KB at QD = 32.

The goal for Ceph SSD performance tuning is to optimize workload access density considering OSD compute resources and network architecture. This performance analysis shows that there is still headroom for Ceph performance using both Cavium's ThunderX SoC and SanDisk CloudSpeed Ultra SSDs. These preliminary numbers justify further investigation.

Figure 5: Ceph 70% Read / 30% Write Mix Compared to 100% Read and Write



Data Source: Cavium

As a side note, Micron[20] and Samsung[21] have performed a similar analysis of SSDs as Ceph volume stores, but both used PCIe connected NVMe SSDs instead of SATA connected SSDs. NVMe performance results appear to be roughly in line with respect to SATA results, but the evaluation systems are different enough and Ceph tuning for SSD volume stores is embryonic enough that a direct comparison of the results will not point toward solid, long-term competitive directions.

In addition, PCIe connected NVMe SSDs price per performance suffers due to today's much higher prices compared to SATA SSDs. We previously compared Cavium's ThunderX-based systems to Intel Xeon-based systems – although performance is equivalent, Intel's price per performance ratio suffers because of higher Xeon pricing:

[20] https://www.micron.com/resource-details/0386c85e-d072-455d-bd71-6f0297b99dd3
[21] http://www.samsung.com/semiconductor/support/tools-utilities/All-Flash-Array-Reference-Design/downloads/High-Performance_Red_Hat_Ceph_Storage_Using_Samsung_NVMe_SSDs-WP-20160622.pdf

High Performance MySQL Database Using ThunderX,[22] Nov 14, 2016

MySQL solutions based on Cavium's first generation ThunderX_ST are roughly equivalent in performance and power consumption to Intel's mainstream Xeon E5 family. There is a 44% partial TCO advantage for dual-socket systems based on favorable Cavium SoC pricing compared to Intel's processor pricing.

We consider the power consumption and latency of these two systems to be roughly equal over a wide range of test runs. We called a tie for performance as well, especially for customers who want to operate a database server near its maximum load – ThunderX can serve more client requests while keeping a flat Transactions per Second response rate.

High Performance Memory Caching Using ThunderX,[23] Nov 14, 2016

Memcached solutions based on Cavium's first generation ThunderX_CP are roughly equivalent in performance and power consumption to Intel's mainstream Xeon E5 family. There is a 30% partial TCO advantage based on favorable Cavium SoC pricing compared to Intel's processor pricing, which we do not expect to narrow significantly over the next few quarters.

Given that the price gap between SATA and PCIe connected NVMe SSDs is unlikely to close completely in the next few years, SATA will remain a better value.

Conclusion

SSDs have recently crossed a cost/benefit threshold for the storage server community. As a result, now is the right time for CSP CTOs, VPs of Engineering or Infrastructure, cloud architects, and infrastructure managers to assess Ceph through a different price per performance lens. Cavium's ThunderX_ST is specifically designed to support a large number of SATA drives. SanDisk CloudSpeed Ultra Gen II SATA SSD is specifically optimized (with power protection, data center grade endurance, and tuned firmware) for CSP workloads. Because of their workload focus, ThunderX_ST and CloudSpeed Ultra SSDs have low price per performance ratios.

We have documented as much of the process and configuration as possible as a starting point for follow-on work. ARM's server software development ecosystem is maturing enough that installing and configuring Ceph does not present any challenges. ARMv8 solutions such as Cavium's ThunderX SoC seem like a good match for storage servers, but proving that they are a good match

[22] http://www.tiriasresearch.com/downloads/high-performance-mysql-database-using-thunderx/
[23] http://www.tiriasresearch.com/downloads/high-performance-memory-caching-using-thunderx/

in an evolving storage market will be an ongoing process. Cavium ThunderX SoCs are shipping in OEM branded and ODM white box storage servers today, plus motherboards and chassis are also available to configure do-it-yourself ThunderX-based whitebox storage servers. That means performing SSD volume storage analysis on ARMv8-based storage servers is only dependent on buying enough SSDs and server chassis to start experimenting.

This first analysis is the start of Ceph price per performance assessment and tuning, both on ARMv8 SoCs and SATA SSDs. It is a demonstration that SSD volume storage can be performant in highly scalable ARM-based distributed storage solutions. Future work will further to quantify costs vs. benefits, and will offer more competitive context against both x86 and NVMe solutions.

Cavium and Western Digital invite the rest of ARM SoC community to continue to help improve Ceph's performance tuning for SSD-only deployments. They are both investing to make it happen.

APPENDIX

ThunderX based cluster

- Root file system: Ubuntu 16.04 LTS
- Ceph Version: Jewel
- Compiler: default flags with kernel 4.2, plus Trusty default (gcc 4.8.4) with default flags

System settings:

- sysctl -w kernel.pid_max=131072
- sysctl -w vm.swappiness=1
- disk write cache turned off for OSD drive
- Ceph version: 9.2.0 (Inferalis)

Western Digital provided 72 800GB SanDisk CloudSpeed Ultra Gen II SATA SSDs[24] for this analysis.

Client systems

- 4 client machines running RADOS-bench and FIO and connected via a 10Gbps Ethernet switch
- RADOS bench setup – 4 RADOS-bench threads with four pools and 2x and 3x replication
- FIO setup – block device read write on one pool with 2x and 3x replication

Building Ceph with default compiler flags on Ubuntu

- Compiler Check
  Make sure to use the cpp compiler
  Ceph has ARM CRC support built into the code since release 9.2.0. Make sure to set the right flags in the header files.
  Search for file hwcap.h in asm directory, make sure HWCAP_CRC32 flag defined
  If more than one asm/hwcap.h in the system, make sure the one with this flag is being used
- Download Source Code
  - Ceph Source can be downloaded from GitHub for any particular branch. This example is for release Infarnalis.
  - git clone –b v9.2.0 https://github.com/ceph/ceph.git
- Building Ceph
  - Go into the Ceph directory (root of all source code).
  - Run the following commands in order:
    - ./install-deps.sh
    - ./autogen.sh
    - ./configure
    - dpkg-buildpackage (use –j45 to use 45 cores to build simultaneously and faster on ThunderX)

[24] https://www.sandisk.com/business/datacenter/products/flash-devices/ssds/sata-ssd/cloudspeed-gen2

- Ceph binaries from previous step are in the parent directory of the build process. (.deb files)
- To install from .deb files, copy the files to a folder on any system and:
    - sudo dpkg -i *.deb

  Use this command to install all dependencies as the next step and finish installation:
    - sudo apt-get -f install

Example ceph.conf file

```
[global]
fsid = 53ffdbad-cc8e-4696-985f-70633d081064
public_network = 10.18.240.0/8
cluster_network = 192.168.240.0/8
mon_initial_members = ceph1
mon_host = 10.18.240.101
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
filestore_xattr_use_omap = true
osd_mkfs_options = -f -i size=2048 -n size=64k
osd_mount_options_xfs = inode64,noatime,logbsize=256k
filestore_merge_threshold = 40
filestore_split_multiple = 8
osd_op_threads = 12
osd_pool_default_size = 2
mon_pg_warn_max_object_skew = 100000
mon_pg_warn_min_per_osd = 0
mon_pg_warn_max_per_osd = 32768
```

Files used:

testwrite-32-3x.fio
testwrite-32-2x.fio
testwrite-1-3x.fio
testwrite-1-2x.fio
testread-32-3x.fio
testread-32-2x.fio
testread-1-3x.fio
testread-1-2x.fio
testmix-32-3x.fio
testmix-32-2x.fio
testmix-1-3x.fio
testmix-1-2x.fio