White Paper

Building high performance Ceph Object Stores with Cavium ThunderX2® and Micron® NVMe™ SSD Solutions

Private clouds, big data, real-time sensors, self-monitoring and self-reporting devices combined with an ever-changing archival requirement all add up. We are generating and capturing new data at unprecedented rates. Virtualized environments, media streaming, cloud-based infrastructures and a more distributed workforce all need continuous access to that data — at the speed of now. Ceph Storage on all-flash has emerged as an architecture that can address these requirements.

Cavium and Micron are showcasing a pre-engineered, scalable, performance-optimized NVMe hardware solution to better manage rapidly growing storage demands: Built on standard server platforms with Cavium ThunderX2® and Micron 9200 NVMe enterprise SSDs this solution offers an ultra-performance, ultra-dense, all-flash Ceph Storage infrastructure you can rely on.

How this solution delivers?

Internal RADOS Bench based benchmarks run on a 4-node cluster using 4 NVMe drives and a dual socket ThunderX2 processor server showed impressive write and read throughput numbers. Using a 4MB large object size, total write throughput of 10.6GB/s was observed with 4 Micron 9200 NVMe drives connected installed on each machine. On the read throughput test, the solution was able to saturate a 100GbE based network cluster to deliver a maximum throughput of 45GB/s. The graphs provide additional performance details using various numbers of threads and clients for benchmarking. ThunderX2 easily saturates a 100Gb Ethernet link on the read throughput with 4 NVMe drives and provides write throughput that scales linearly as the number of clients increase. This profile suggests that the solution is ideally suited for applications that leverage the Ceph object storage to serve multiple clients uploading and downloading large amounts of data.

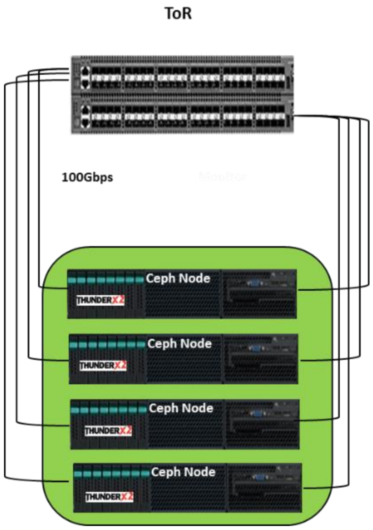
Key Features

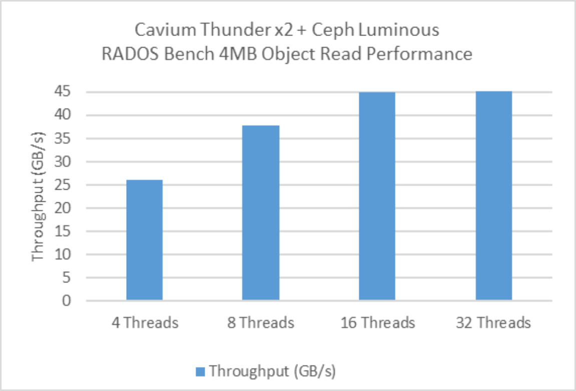
ThunderX2®: Second generation of Cavium's Arm®v8 based server processors

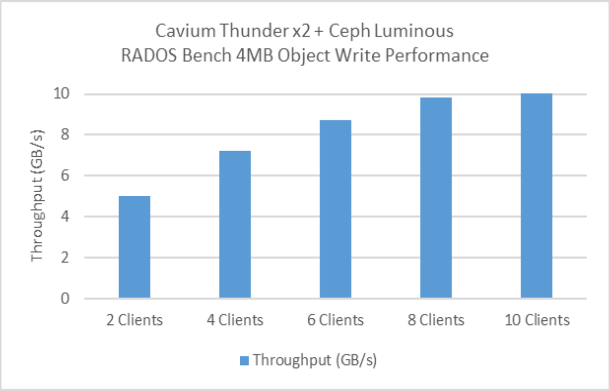
- ✓ Supports dual socket configurations
- ✓ Best in class computational performance
- ✓ outstanding I/O connectivity
- ✓ High memory bandwidth and capacity.
- ✓ fully compliant with Arm's SBSA and SBBR standards
- ✓ Widely supported by industry leading OS, Hypervisor and SW tool and application vendors.

Micron® 9200 NVMe PCIe SSDs:

- ✓ Available in high capacities up to 11TB
- ✓ Transfer speeds up to 3.35 GB/s and read IOPS up to 800K (steady state)
- ✓ Tunable capacity to optimized performance using Micron Flex Capacity™
- ✓ Full enterprise end-to-end data path protection & power-loss protection







Conclusion

As the adoption of cloud and usage of real-time sensors, self-monitoring and self-reporting devices keeps growing, the amount of data generated from these devices is expected to continue growing at an exponential rate [1]. Service providers and operators who need hardware that can handle this data velocity and bandwidth with reliability and predictability should consider the all-flash hardware solution showcased by Cavium and Micron today.

The test environment consists of four Ceph storage nodes, one Ceph monitor node, and 10 load generation servers.

The Cavium Ceph storage nodes are Cavium Thunder x2 servers with 2x Cavium Saber Reference Platform CPUs with 28 cores at 2.2GHz base and 4 threads per core. They have 256GB of DRAM (16 x 16GB Micron DDR4 RDIMM), and 1 dual port Mellanox 100GbE network card. A SATA SSD is used as an OS Drive, while 4 x Micron 9200 NVMe U.2 3.2TB are used for Ceph storage.

The Ceph monitor node is a Supermicro Superserver SYS-1028U-TNRT+ server with 2x Intel 2690v4 Processors, 128GB of DRAM, and a Mellanox ConnectX-4 50GbE network card.The load generation servers are Supermicro Superserver SYS-2028U-TNRT+ servers with 2x Intel 2690v4 processors, 256GB of DRAM (16 x 16GB Micron DDR4 RDIMM), and a Mellanox ConnectX-4 50 GbE NIC.

Networking is handled by 2 Supermicro SSE-3632SR 32-Port 100GbE switches. One switch covers client traffic (Ceph storage nodes, Ceph monitor nodes, and load generation servers), the other covers storage node replication (only Ceph storage nodes).

The software infrastructure is illustrated in the figure above. The software specification is listed below.

- Operating System (Thunder x2 Servers)
  - Ubuntu LTS 16.06 (Arm®64)
- Ceph Version:
  - Ceph Community Edition Latest Stable
- Ceph Luminous 12.2.4 + Bluestore
- Load Generation software
  - RADOS Bench, included in Ceph

Ceph Pool Configuration

All tests use a 2x replicated pool with 8192 placement groups with 2 OSDs per drive

RADOS Bench Workload Overview

RADOS Bench is a tool for measuring object performance built into Ceph. It represents the best-case object performance scenario of data coming directly to Ceph from a RADOS Gateway node.

4MB object writes are measured by running RADOS Bench with a “threads” value of 16 on a load generation server writing directly to a Ceph storage pool. A threads value of 16 simulates a very active consumer of object resources and allows the All-NVMe Ceph cluster to be pushed to its maximum object write performance. The number of load generation servers is scaled up from 2 to 10.

4MB object reads are measured by first writing 15TB of data into the 2x replicated pool using 20 RADOS Bench instances. Once the data load is complete, all 20 RADOS Bench instances are used to run 4MB object reads against the storage pool. It is important to use all 20 RADOS Bench instances for reads so that the entire 15TB dataset is being accessed, otherwise Linux filesystem caching can skew results higher. The thread count of RADOS Bench is scaled up from 4 threads to 32 threads.

Object workload tests are run for 10 minutes, 3 times each. Linux filesystem caches are cleared between each test. The results reported are the averages across all test runs.

CAVIUM™

Follow us:

Corporate Headquarters: Cavium, Inc. | 2315 N. First Street San Jose, CA 95131 | 408-943-7100

This document is provided for informational purposes only and may contain errors. Cavium reserves the right, without notice, to make changes to this document or in product design or specifications. Cavium disclaims any warranty of any kind, expressed or implied, and does not guarantee that any results or performance described in the document will be achieved by you. All statements regarding Cavium’s future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

[1] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html