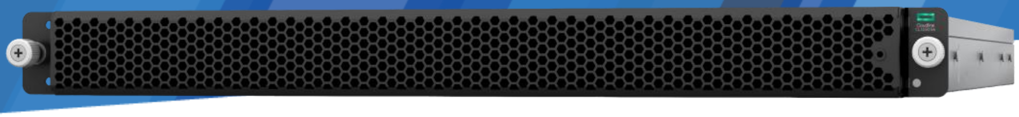
February 2018

HPE Cloudline CL3150 Gen10 Servers Power High-Performance Cloud Solutions

Powerful 100Gb Ethernet networking combined with Excelero NVMesh management provides performance and application availability for cloud environments.



Executive Summary

Hewlett Packard Enterprise believes that customers need the flexibility to select the IT solutions best matched to their operating environment, applications and business needs. HPE Cloudline servers offer solutions that combine open infrastructure efficiencies and economics with the confidence of a world-class product and customer experience, backed by HPE engineering expertise, global manufacturing capabilities, and award-winning support.

The HPE Cloudline CL3150 Gen10 server is a 1U open standards-based ultra-dense storage server, ideal for performance demanding cloud applications. This server supports up to 24 NVMe solid state drives (SSDs) providing very high-performance and plenty of capacity, to meet ever-growing needs. This server is also designed for the AMD EPYC™ 7000 series processor, offering strong performance at a competitive price.

These servers are powered by AMD EPYC processors that provide up to 32 cores, 64 threads and 8 memory channels per socket. These processors also provide 128 PCIe 3.0 lanes, providing cloud environments with direct support for many high performance NVMe storage devices, eliminating performance bottlenecks found in other architectures.

HPE's partnership with Mellanox brings network flexibility by supporting 10GbE, 25GbE and 50GbE open-compute platform (OCP) mezzanine cards and 100GbE connectivity via standard PCIe adapters.

Excelero's Software-Defined Storage NVMesh enables customers to design server infrastructure for cloud-scale applications with the advantage of logically disaggregating storage from the compute platform. By taking advantage of the high-speed data path (100GbE in this case), very high-performance NVMe storage is made available in a seamless fashion to cloud applications. The application or compute node can fail without loss of access to the data, and a new compute node can be added to the cluster for immediate resumption of service.

Key Findings

> Single-socket AMD EPYC CPU utilization was less than 50% for 500,000,000-record database for YCSB operations on compute node.

> Minimum write latencies for all YCSB runs were below 100 microseconds (μs) and minimum read latencies were below 50 microseconds.

> For the database of 700,000 records, the workloads completed in less than one minute.

> The average latencies for workloads A & B for a database size of 500,000,000 records was less than 250 microseconds (μs).

> The 99th percentile latency for workloads A & B for all database sizes was less than 500 microseconds (μs).

> The 99th percentile latency for workload F (read-modify-write) for all database sizes was less than 750 microseconds (μs).

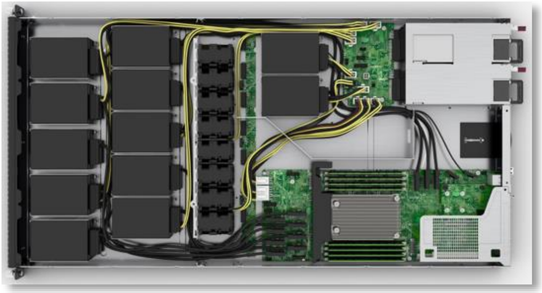
HPE Cloudline CL3150 Gen10 Servers Power High-Performance Cloud Solutions

Demartek
February 2018

HPE Cloudline Servers

HPE Cloudline servers are designed for cloud data centers. They use industry-standard components, common firmware and BIOS, to help reduce costs.

This evaluation deployed three HPE Cloudline CL3150 Gen10 servers. Each server was provisioned with the AMD EPYC 7401 processors (24c/48t), 128 GB RAM and 12 NVMe SSDs.



AMD EPYC Processor

The AMD EPYC processor, designed for cloud deployments, provides up to 32 cores, 64 threads and 8 memory channels with up to 2 TB of memory per socket for new levels of data center performance. It also provides 128 PCIe 3.0 lanes making it the perfect processor for servers with NVMe storage combined with high-speed network adapters.

Excelero NVMesh

Excelero's Software-Defined Storage platform NVMesh enables customers to design Server infrastructures for the most demanding enterprise and cloud-scale applications, leveraging standard servers and multiple tiers of flash.

NVMesh is a Software-Defined Block Storage solution that features Elastic NVMe, a distributed block layer that allows unmodified applications to utilize pooled NVMe storage devices across a network at local speeds and latencies. Distributed NVMe storage resources are pooled with the ability to create arbitrary, dynamic block volumes that can be utilized by any host running the NVMesh block client. These virtual volumes can be striped, mirrored or both while enjoying centralized management, monitoring and administration.

100GbE Switch and Adapters

An HPE StoreFabric M-Series SN2100M 16-port 100GbE switch was used to connect each of the servers. This switch supports 10/25/40/100Gb speeds and has breakout cable options for 10Gb and 25Gb connections. This switch provides:

- Ultra-low Latency: < 300 ns port-to-port
- Zero packet loss at all frame sizes
- 100% forwarding capacity wire rate at all ports concurrently at 100GbE speeds
- Cut-through performance with DCBX, PFC and ECN support

A Mellanox ConnectX-5 100GbE network adapter was used in each server. Features include:

- Low latency, low CPU utilization, high message rate
- Storage capabilities include NVMe over Fabric offloads
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization

Yahoo Cloud Serving Benchmark

The Yahoo Cloud Serving Benchmark (YCSB) is a framework that provides a common set of workloads typically found in cloud data centers.

YCSB includes six workloads that cover many of the workloads found in cloud data centers. These are:

- **Workload A**: Update heavy (50% read, 50% write)
- **Workload B**: Read mostly (95% read, 5% write)
- **Workload C**: Read only (100% read)
- **Workload D**: Read latest (new records inserted and then read)
- **Workload E**: Short ranges (ranges of reads, such as email threads)
- **Workload F**: Read-modify-write

The Evaluation Configuration

The evaluation environment consisted of three HPE Cloudline CL3150 Gen10 servers identically configured with:

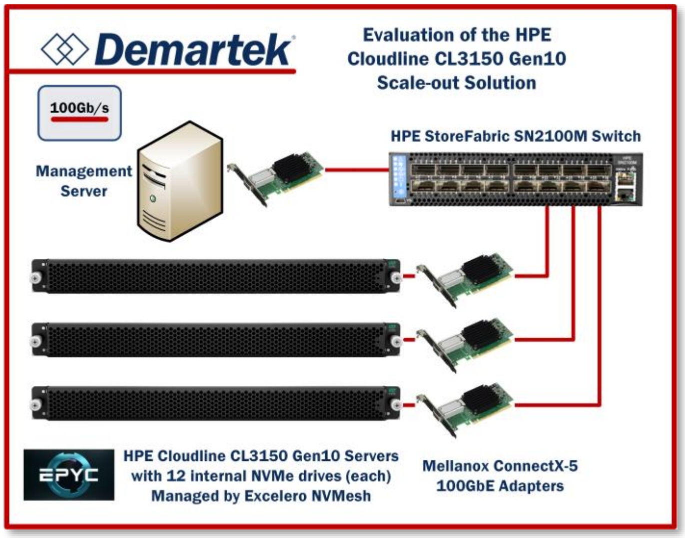
- AMD EPYC 7401, 2.0 GHz, 24c/48t
- 128 GB RAM (4x 32GB 2Rx4 PC4-2666)
- 12x Samsung PM963 1.92TB NVMe SSD
- Mellanox MCX515A-CCAT 100 Gigabit Ethernet PCIe Adapter (ConnectX-5)

Each of these servers, along with a separate management server, were connected to an HPE StoreFabric M-Series SN2100M Ethernet Switch.

Division of Work in the Cluster

For these tests, the YCSB benchmark was run on node 1 and the data it accessed was on nodes 2 and 3, separating the compute node from the storage nodes. In the event of a compute node failure with this type of configuration, the work can be started on a different node without having to move databases.

All the storage traffic from the twelve NVMe drives in each of the storage nodes flowed over our 100GbE network. Based on our measurements, we believe that our results would have been higher with additional network bandwidth available in each server.


Demartek
Evaluation of the HPE Cloudline CL3150 Gen10 Scale-out Solution
100Gb/s
Management Server
HPE StoreFabric SN2100M Switch
EPYC
HPE Cloudline CL3150 Gen10 Servers with 12 internal NVMe drives (each) Managed by Excelero NVMesh
Mellanox ConnectX-5 100GbE Adapters

Results and Analysis

We ran three of the YCSB workloads: **A**, **B** and **F**.

- **Workload A**: Update heavy (50% read, 50% write)
- **Workload B**: Read mostly (95% read, 5% write)
- **Workload F**: Read-modify-write

For these three workloads and given the same database record count, workload B takes the shortest amount of time to run, and, not surprisingly, workload F takes the longest amount of time to run.

Three sets of tests were run with our 3-node cluster, showing the scale-up trajectory of this type of configuration.

- Total Database Records: 700,000 (700K)
- Total Database Records: 200,000,000 (200M)
- Total Database Records: 500,000,000 (500M)

For all runs the YCSB threads = 24.

The YCSB benchmark provides several statistics for each run. The raw output is provided in the tables. The latencies are provided in the graphs.

Latencies are expressed in ***microseconds (μs)***.
(1000 microseconds = 1 millisecond)

Operating System
CentOS: 7.4.1708
Kernel: 3.10.0-693.5.2.el7.x86_x64

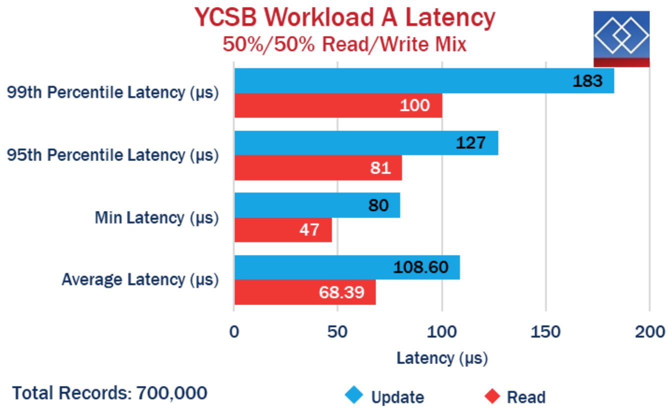
YCSB
YCSB 0.12.0

Database
We used MongoDB version: 3.4.10. No attempt was made to optimize any database settings, engines, etc.

Results for Total Records: 700,000

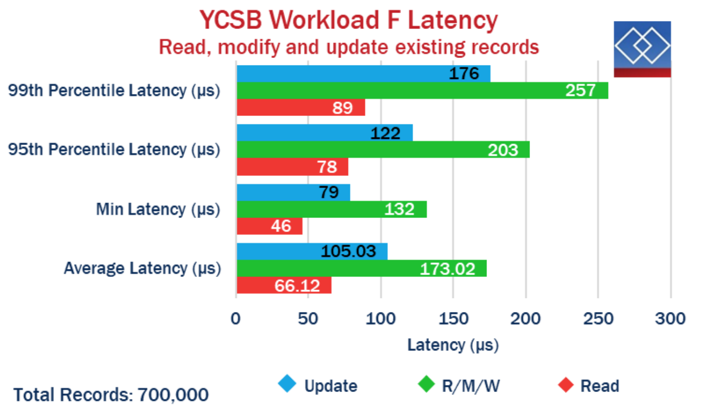
Workload A: 50% Read / 50% Write

Runtime (ms)*	31952
Throughput (ops/sec)	21907.86
Read Operations	349,749
Update Operations	350,251



Workload F: Read / Modify / Write

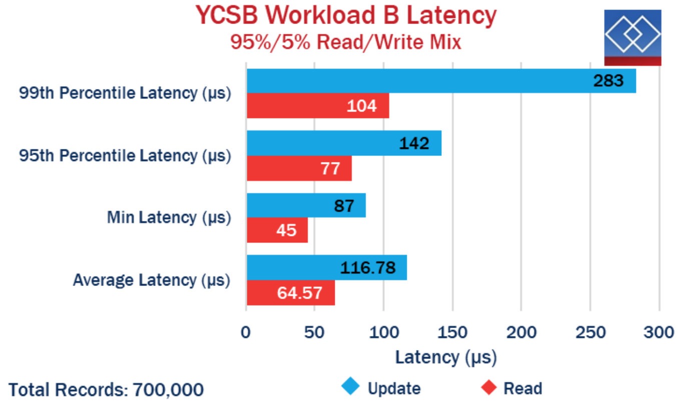
Runtime (ms)*	43120
Throughput (ops/sec)	16233.77
Read Operations	700,000
Read-Modify-Write Operations	349,824
Update Operations	349,824



** Note that these runs with 700,000 records took less than one minute to run.*

Workload B: 95% Read / 5% Write

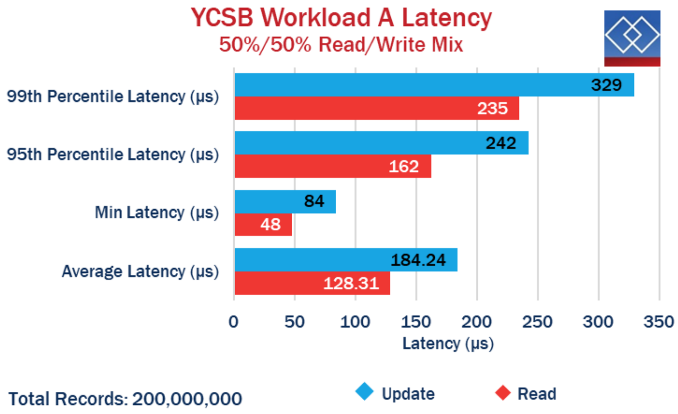
Runtime (ms)*	29435
Throughput (ops/sec)	23781.21
Read Operations	665,020
Update Operations	34,980



Results for Total Records: 200,000,000

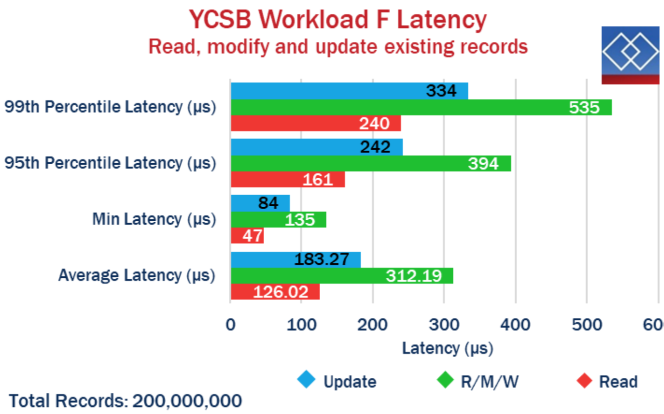
Workload A: 50% Read / 50% Write

Runtime (ms)	1996348
Throughput (ops/sec)	100182.934
Read Operations	99,994,107
Update Operations	100,005,893



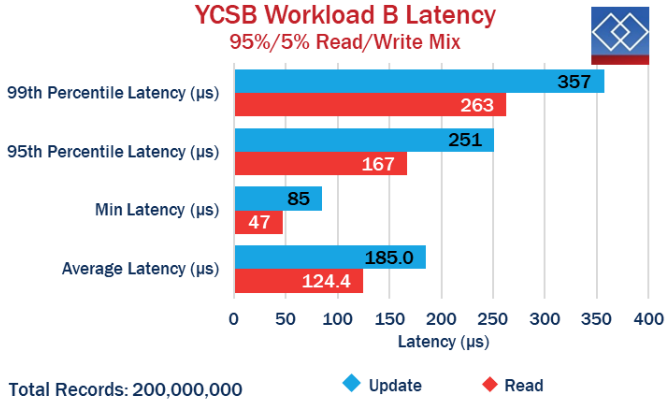
Workload F: Read / Modify / Write

Runtime (ms)	2777729
Throughput (ops/sec)	72001.26434
Read Operations	200,000,000
Read-Modify-Write Operations	100,002,158
Update Operations	100,002,158



Workload B: 95% Read / 5% Write

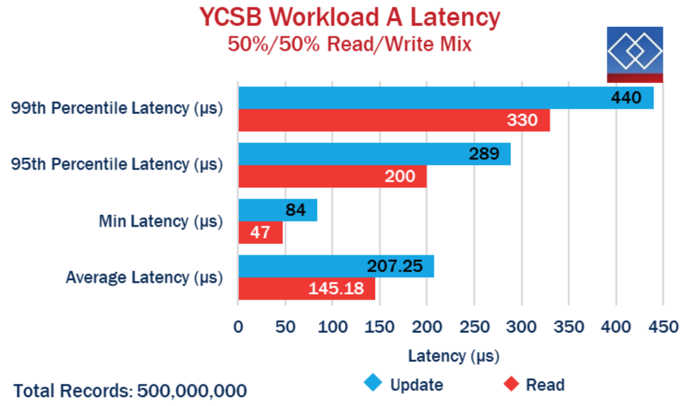
Runtime (ms)	1626376
Throughput (ops/sec)	122972.7935
Read Operations	190,006,021
Update Operations	9,993,979



Results for Total Records: 500,000,000

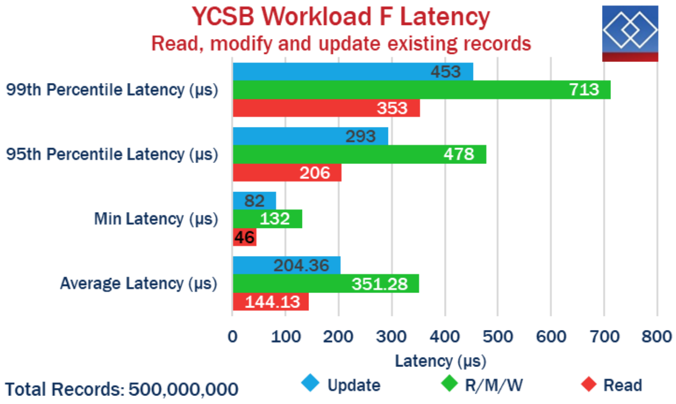
Workload A: 50% Read / 50% Write

Runtime (ms)	3739925
Throughput (ops/sec)	133692.521
Read Operations	250,000,000
Update Operations	250,000,000



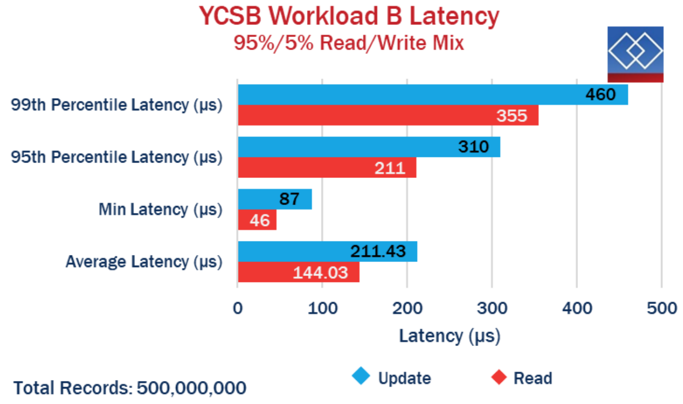
Workload F: Read / Modify / Write

Runtime (ms)	5230121
Throughput (ops/sec)	95600.083
Read Operations	500,000,000
Read-Modify-Write Operations	250,000,000
Update Operations	250,000,000



Workload B: 95% Read / 5% Write

Runtime (ms)	3131714
Throughput (ops/sec)	159656.98
Read Operations	475,000,000
Update Operations	25,000,000



CPU Utilization

CPU utilization is an important metric that can help determine how much load a system can take. Each application affects the host CPU differently because different workloads need differing amounts of CPU, memory, networking and storage resources.

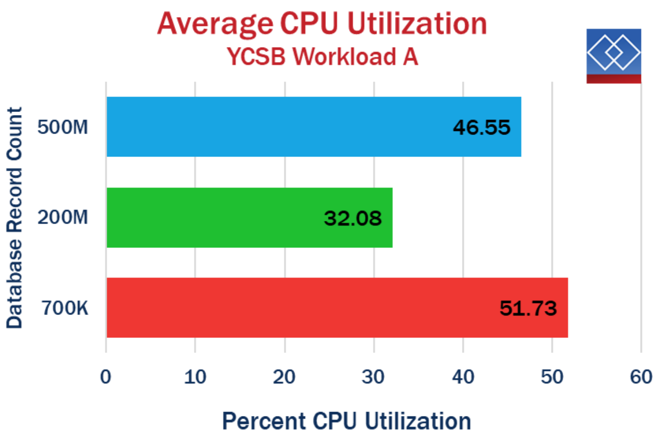
We measured the CPU utilization of node 1 (the compute node) while it was running the workload. Because the storage nodes were separate from the compute node, all of the benchmark computations were performed on a single node.

It is important to note the run times for the different database record counts. Like many applications, YCSB performs some initialization and cleanup at the beginning and end of the tests. The 700K record count runs completed in less than one minute, which means that a relatively large percentage of the work was focused on initialization and cleanup, which skewed the results higher than for the larger database record counts.

Workload Runtimes & CPU Utilization Graphs

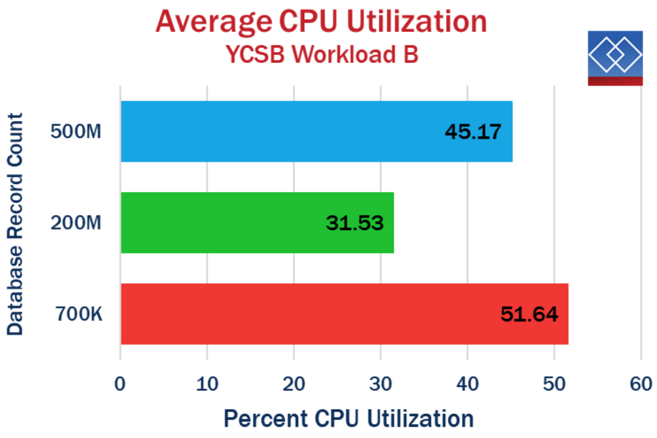
Workload A

	Milliseconds	Seconds	Minutes
700K records	31952	32	0.5
200M records	1996348	1996	32.3
500M records	3739925	3740	62.3



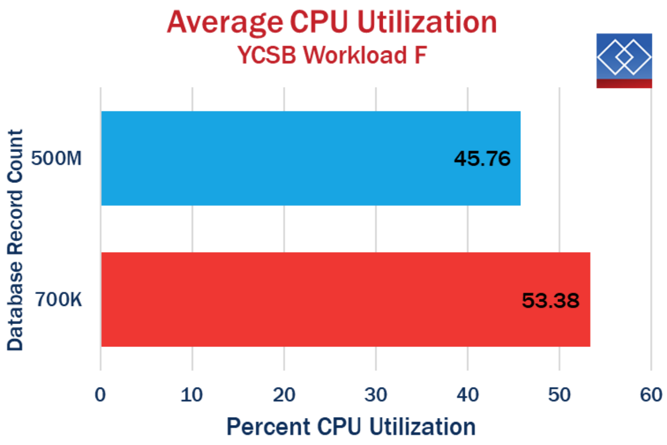
Workload B

	Milliseconds	Seconds	Minutes
700K records	29435	29	0.5
200M records	1626376	1626	27.1
500M records	3131714	3132	52.2



Workload F

	Milliseconds	Seconds	Minutes
700K records	43120	43	0.7
200M records	2777729	2778	46.3
500M records	5230121	5230	87.2



CPU utilization statistics were not captured for the 200M database for workload F.

Comparison to Other Published Results

In order to provide some context for our YCSB results, we reviewed published YCSB results on the Internet. We found a wide variety of published results, but none of the results we found provided enough data to make an "apples-to-apples" comparison with our results.

Several of the published results were run in some of the well-known public clouds. Some ran with three nodes, but many ran with higher numbers of nodes.

Here are a few differences we noted.

- None used AMD EPYC processors.
- Most did not clearly show minimum latency, average latency, 95th percentile latency and 99th percentile latency results for their runs.
- None of the other published benchmarks achieved minimum latencies less than 100 microseconds (μs).
- The average latencies published exceeded 1000 microseconds, or 1 millisecond.
- Some indicated the thread count, others did not.
- Almost all of these benchmarks were run in a 10GbE networking environment, which runs at 1/10th the speed of our 100GbE network
- Most used fewer drives for storage nodes than our configuration.
- None of the published results used NVMe drives for storage.
- As far as we could tell, none were configured so that the database storage was accessed 100% over the network from a separate compute node.
- Some of the results exceeded our throughput results for a three-node cluster, but they configured considerably higher numbers of threads per server node and their latencies were much higher.

Scalability of the Solution

We deployed a three-node cluster of the HPE Cloudline CL3150 Gen10 servers. For most cloud deployments, more nodes would be deployed.

Based on our observations from these tests and the architecture of this solution, we believe that this type of solution would scale well, and probably linearly, with higher numbers of nodes.

Summary and Conclusion

This cloud data center solution provided by a combination of hardware and software from HPE, Mellanox and Excelero is an easy-to-manage, scalable solution designed to provide high performance at a competitive price point. Using the building block approach with an open design specification and off the shelf standard components, customers can integrate the latest technology and features into their systems to maximize productivity.

Even though the compute nodes may have high reliability and high life expectancy, when hundreds or thousands of nodes are operating in a cloud data center, failures are bound to happen. With the combined solution, it is possible to physically disaggregate NVMe flash, make it redundant, and still have it perform like locally installed NVMe. Separating the compute nodes from the storage nodes in a clustered cloud application, such as the Yahoo Cloud Serving Benchmark, assures users and administrators that in the event of a compute node failure, the workload can easily be moved to another compute node without having to move the databases from their storage locations.

The solution we tested using the HPE Cloudline CL3150 powered by AMD EPYC processors provided very high performance and very low latency, better than we have seen with other implementations of the Yahoo Cloud Serving Benchmark. We believe that this solution would work well in many other cloud data center applications that require clustered server nodes. In addition, by separating the compute nodes from the storage nodes while providing very high-speed networking insures minimal disruption in the event of node failures.

The AMD EPYC processor in general and the HPE Cloudline CL3150 specifically seem particularly suited for building high performance and resilient Cloud solutions.

The most current version of this report is available at http://www.demartek.com/Demartek_HPE_CL3150_Gen10_Evaluation_2018-02.html on the Demartek website.