Technical Report

Performance Characterization of ONTAP Cloud in Amazon Web Services with Application Workloads

NetApp Data Fabric Group, NetApp
July 2017 | TR-4383

Abstract

This technical report examines the performance and fit of application workloads to NetApp® ONTAP® Cloud instances that are running in Amazon Web Services (AWS) EC2.

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

To help NetApp customers select the most appropriate solutions for their IT infrastructure, NetApp provides up-to-date documentation describing its products. This technical report describes the results of performance tests for NetApp ONTAP Cloud data management software running on Amazon Web Services (AWS) EC2 instances. NetApp partners, customers, and employees can use this information to make informed decisions about which workloads are appropriate for ONTAP Cloud.

The test configuration environments described in this report consist of the following components:

- ONTAP Cloud clusters, both single node and high availability (HA) pairs, of the following instance types:
  - c4.8xlarge (Premium license)
  - c4.4xlarge (Premium license)
  - c4.2xlarge (Standard license)
  - m4.xlarge (Explore license)
- A Microsoft Windows Server 2012 R2 host running Iometer. For each test, the instance type used for the Iometer machine matched the ONTAP Cloud system instance type.
- The iSCSI protocol for block workloads.

2 Differences Among Instance Types

The Amazon instance types page describes the capabilities of the different instance types in terms of network, CPU, and memory.

The specific combination of these components is what determines the overall performance of an instance. For ONTAP Cloud, the CPU capability contributes to general performance, whereas memory contributes heavily to read performance. The network capability acts as a throughput limiting factor that is independent of read and write performance levels. All the supported instance types for ONTAP Cloud are Elastic Block Store (EBS)–optimized and thus deliver dedicated bandwidth to Amazon EBS with different maximal throughput, depending on the instance type.

Table 1) Capabilities of supported instance types.

Instance	CPU	RAM	Network	EBS Max Throughput for 128KB Streaming Read Workload (MBps)
m4.xlarge	4	16GB	High	93.75
m4.2xlarge	8	32GB	High	125
m4.4xlarge	16	64GB	High	250
r4.xlarge	4	30.5GB	Up to 10Gb	109
r4.2xlarge	8	61GB	Up to 10Gb	218
c4.2xlarge	8	15GB	High	125
c4.4xlarge	16	30GB	High	250
c4.8xlarge	36	60GB	10Gb	500

3 Test Configurations

All test configurations used the iSCSI protocol for block I/O connectivity. The tests focused on the following:

- Iometer was used to generate I/O workloads. We used different numbers of LUNs and different numbers of outstanding I/O operations to see the effect of concurrency level on IOPS, throughput, and latency.
- Normal writing speed was used for single-node systems. With normal write speed, data is written directly to disk, minimizing the likelihood of data loss if an unplanned system outage occurs. Conversely, when high write speed is used, data is buffered in memory before it is written to disk. This configuration provides faster write performance, but it increases the potential for data loss if an unplanned system outage occurs. High write speed is recommended if fast write performance is required and you can tolerate the risk of data loss due to an unplanned system outage. For example, data loss might be handled by the application.
- We tested single availability zone HA ONTAP Cloud systems and single-node deployments. In the HA tests, a single Iometer client generated load against multiple LUNs that all resided on one of the nodes in the cluster, rather than having LUNs split between the nodes with two clients generating load against them. Read performance can be doubled when LUNs are divided between nodes.
- We tested different instance types. The instance types tested not only differed by their hardware characteristics but also incurred different costs from AWS and NetApp (they belong to different licensing packages). An instance type from each licensing package was picked for the tests.
- All tests were performed with gp2 EBS volumes. ONTAP Cloud supports the following additional EBS volume types: st1, sc1, and io1. Although gp2 and io1 volumes provide the best performance results, gp2 volumes are the recommended configuration because they are much more cost-effective.
- ONTAP Cloud was tested on four different instance types:
  - **m4.xlarge.** Included in the Explore license, which provides up to 2TB of storage.
  - **c4.2xlarge.** Included in the Standard license, which provides up to 10TB of storage.
  - **c4.4xlarge.** Included in the Premium license, which provides up to 368TB of storage.
  - **c4.8xlarge.** Included in the Premium license. This instance type has very strong hardware capabilities. Therefore, its AWS pricing is the highest of all supported instance types.

4 Summary of Test Results

Each tested configuration consisted of a unique workload that is representative of the workloads used in widely deployed POSIX applications:

- **OLTP workload.** 8KB block size, 100% random access I/O, and a mixture of 70% reads. This workload simulates database applications (SAP, Oracle, and SQL) and OLTP servers.
- **Streaming reads.** 64KB block size, 100% reads, and 100% sequential access I/O. This workload simulates applications such as media servers (for example, video on demand) and virtual tape libraries.
- **Streaming writes.** 64KB block size, 100% writes, and 100% sequential access I/O. This workload simulates applications such as media capture, virtual tape libraries, medical imaging, archiving, backup, and video surveillance.
- **Analytics.** 16KB block size, 50% reads, and 100% random access I/O. Analytics workloads are ad hoc in nature. They involve both read and write operations and require high throughput and low latency.

The test workloads differed markedly. Collectively, however, they represent workloads that can be considered for an ONTAP Cloud deployment. Because the workloads were so different, the results of each test are discussed separately in this section.

4.1 OLTP Workload

The 8KB block size workload simulates an OLTP transactional database. Transactional workloads tend to be read-heavy as data about an item is retrieved, but they require a smaller number of writes as transactions are committed. Transactional workloads are highly sensitive to write latency, especially for writes to the transaction log. Typically, it is best for log writes to not exceed 15ms of latency. Lower latency is always better.

The workload tested consisted of an 8KB block size, 70% reads, and 100% random access I/O. The number of outstanding I/O operations was changed to control the number of concurrent I/O operations sent to the storage system. We increased the number of outstanding I/O operations to create the data points necessary to form a knee curve for IOPS and latency.

Figure 1 and Figure 2 show the IOPS and latency of a single-node and HA pair ONTAP Cloud system running on a c4.8xlarge instance type.

Figure 1) OLTP IOPS and latency: ONTAP Cloud with 8KB blocks on a c4.8xlarge single node.

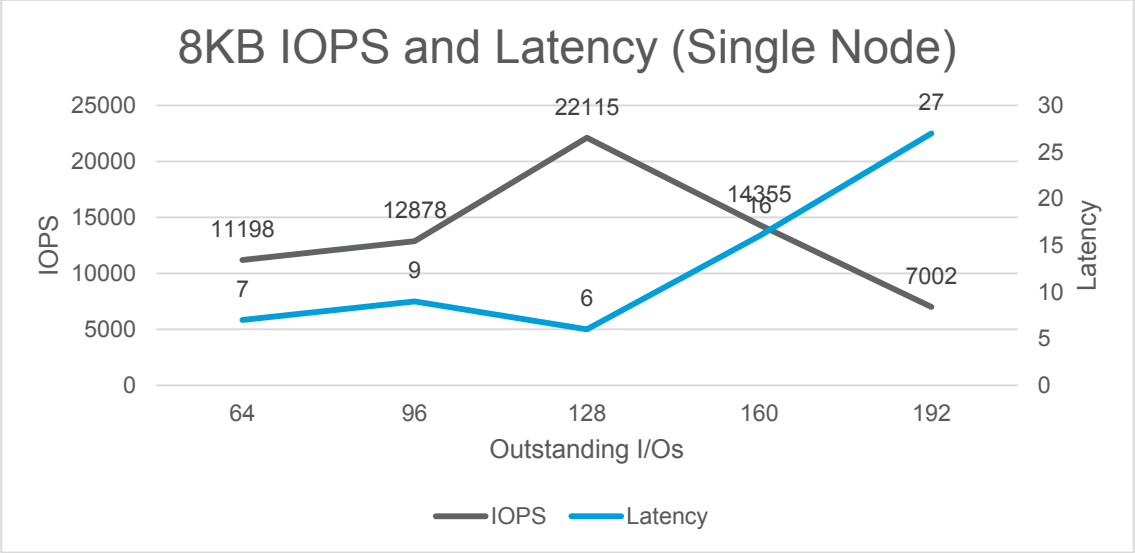

Figure 2) OLTP IOPS and latency: ONTAP Cloud with 8KB blocks on a c4.8xlarge HA pair.

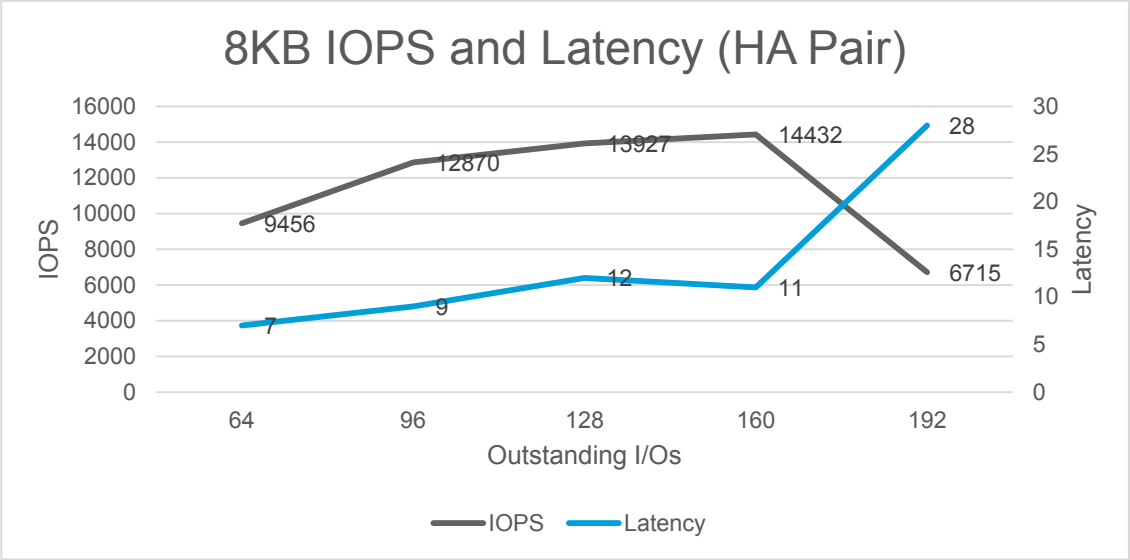

Table 2 shows the results for 8KB blocks, 70% reads, 100% random access I/O for the different instance types.

Table 2) OLTP workload IOPS and latency.

Instance Type	IOPS for Single Node (Ops/s)	Latency for Single Node (msec)	IOPS for HA pair (Ops/s)	Latency for HA pair (msec)	License
c4.8xlarge	22,115	6	14,432	11	Premium
c4.4xlarge	21,942	5	10,690	12	Premium
c4.2xlarge	8,819	8	6,618	19	Standard
m4.xlarge	7,860	10	6,336	11	Explore

4.2 Streaming Read Workload

We tested a workload with a large contiguous read request size that simulates applications such as media servers (for example, video on demand) and virtual tape libraries. The workload consisted of a 64KB block size, 100% reads, and 100% sequential access I/O.

The most significant measurement for this type of workload is throughput, as Table 3 shows.

Table 3) Streaming read workload throughput.

Instance Type	Throughput for Single Node (MBps)	Throughput for HA pair (MBps)	License
c4.8xlarge	204	204	Premium
c4.4xlarge	136	136	Premium
c4.2xlarge	116	121	Standard
m4.xlarge	67	65	Explore

ONTAP Cloud has robust performance for streaming reads. The differences between the instance types are mainly caused by the networking capabilities of the instance and the number of CPUs available. For the HA pair tests, all LUNs were resident on the same node for the tested configuration. Splitting the LUNs between the two nodes can increase the performance even further, with the potential to double it.

4.3 Streaming Write Workload

We tested a workload with a large contiguous write request size that simulates applications such as media capture, virtual tape libraries, medical imaging, archiving, backup, video surveillance, and reference data. The workload consisted of a 64KB block size, 100% writes, and 100% sequential access I/O.

The most significant measurement for this type of workload is throughput, as Table 4 shows.

Table 4) Streaming write workload throughput.

Instance Type	Throughput for Single Node (MBps)	Throughput for HA pair (MBps)	License
c4.8xlarge	194	119	Premium
c4.4xlarge	137	110	Premium
c4.2xlarge	106	102	Standard
m4.xlarge	43	33	Explore

ONTAP Cloud performed well for the streaming write workload. Here too, a single node performed better than an HA pair, because of the propagation of each write to the other node.

4.4 Analytics Workload

We tested a workload with analytics characteristics. Analytics consist of a read and write mixture, for which throughput and latency are the most important measurements. The workload we tested consisted of a 16KB block size, 50% read, and 100% random access I/O.

Table 5) Analytics workload throughput.

Instance Type	Throughput for Single Node (MBps)	Latency for Single Node (msec)	Throughput for HA pair (MBps)	Latency for HA pair (msec)	License
c4.8xlarge	214	6	117	14	Premium
c4.4xlarge	215	6	127	13	Premium
c4.2xlarge	140	12	102	12	Standard
m4.xlarge	46	19	27	16	Explore

The results show that ONTAP Cloud is a good fit for analytics workloads, because it can maintain high throughput under strict latency constraints.

5 Conclusion

ONTAP Cloud was found to be a good fit for OLTP workloads, streaming read workloads, streaming write workloads, and analytics workloads. NetApp has a long history of providing high-performance and feature-rich storage systems. ONTAP Cloud extends this legacy to AWS. With ONTAP Cloud, NetApp continues to develop leading-edge storage solutions that provide the agility and mobility that current

NetApp customers need and that future NetApp customers will want. ONTAP Cloud is part of a family of products that stretch from the private cloud to the hybrid cloud to the public cloud and that run NetApp ONTAP data management software. Understanding the performance characteristics of ONTAP Cloud is critical for setting our customers' expectations and enabling their continued success.

Version History

Version	Date	Document Version History
Version 2.0	June 2017	Updated performance with ONTAP Cloud 9.2 Added performance of high availability
Version 1.2	May 2017	Additional tests
Version 1.1	January 2016	Updated with tests of additional instance types
Version 1.0	February 2015	Initial release

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.