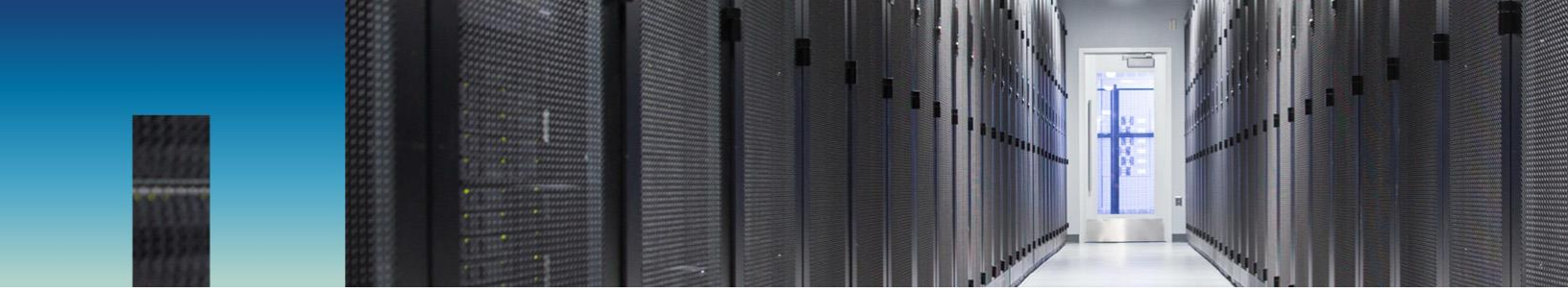


Benchmarking Technical Report

Internet of DNA: Cloud Enabled

By Chad Morgenstern, NetApp Senior Cloud Solutions Architect;
Jeff Whitaker, NetApp Cloud Product and Solutions;

March 2019 | TR-4761

In partnership with

WuXiNextCODE

Dr. Hákon Gudbjartsson, PhD, CIO of WuXi NextCODE

Abstract

Learn how Cloud Volumes Service handles the extreme demands of the genomic analytics workloads required by the WuXi NextCODE platform

NetApp®

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Dr. Hákon Gudbjartsson, PhD, the CIO of WuXi NextCODE, knows what it takes to be a data visionary. Using sequence data, he and his team are enabling organizations to unlock the power of the genome to enhance health and wellness.

With offices in Shanghai, Reykjavik, and Cambridge, Massachusetts, WuXi NextCODE is a genomic information company and global platform for genomic big data. Over the years they have amassed one of the world's largest databases of human genome sequences.

"At WuXi NextCODE, we like to call ourselves the 'internet of DNA'," Dr. Gudbjartsson said.

If you're thinking that genomics isn't a big data challenge, think again.

Although all humans share a similar DNA sequence, it is not 100% unique to the individual. If you and a friend were to compare your DNA, you would find that in the roughly 3 billion letters of the DNA, you differ in about 5 million locations. "The challenge is to take a dataset of 5 million and figure out the differences or mutations that are important—which ones are the causes of rare diseases, which ones are the causes of cancer, and how to treat patients," Dr. Gudbjartsson explained.

At the heart of the WuXi NextCODE platform is the genomic relational database, the only relational database designed and optimized to query and analyze massive genomic data. By taking advantage of NetApp® Cloud Volumes Service, the genome platform makes it possible to integrate data dynamically to deliver unprecedented computational efficiency. "A benchmark analysis for analyzing genomic data is generating the allele frequency of every mutation found in a population of 100,000 individuals," Dr. Gudbjartsson said. "With earlier storage solutions (or self-managed storage), we always had timeouts or file failures. But when we tested this using the NetApp Cloud Volumes Service, it actually finished in less than an hour. That was a great breakthrough for us."

A Demanding Workload

NetApp had the privilege of spending 3 days with Dr. Gudbjartsson and his staff in beautiful Reykjavik—not an overly important fact, but enjoyable nonetheless. The purpose of NetApp's visit was to represent NetApp technical staff during the running of the aforementioned benchmarks.

1.1 The Facts of the Test

The WuXi NextCODE platform gives customers, such as hospitals and pharmaceutical companies, access to WuXi NextCODE's population database, along with their own smaller base of genomes. Sequence read files (FASTQ and BAM) are stored in object storage; however, the important informative sequence variation data, analyzed through the genomically ordered relational database (GORdb) system, is stored on NFS. NFS is of significant value to the WuXi NextCODE platform because of its POSIX file system cache, its shared nature, and its ability to serve both random and sequential workload profiles equally well.

The benchmark use case calls for a massively scaled-out sequential read workload performed against thousands of these genome files (GOR formatted). The files themselves are spread across one-to-many cloud volumes and many-to-many more Amazon Elastic Compute Cloud (EC2) instances. Because the content is genomically ordered, reads are efficient, resulting in the fastest possible analysis. A second use case calls for an equally scaled-out random read workload that's SQL-like and somewhat comparable to what one might expect from Apache's Hive project.

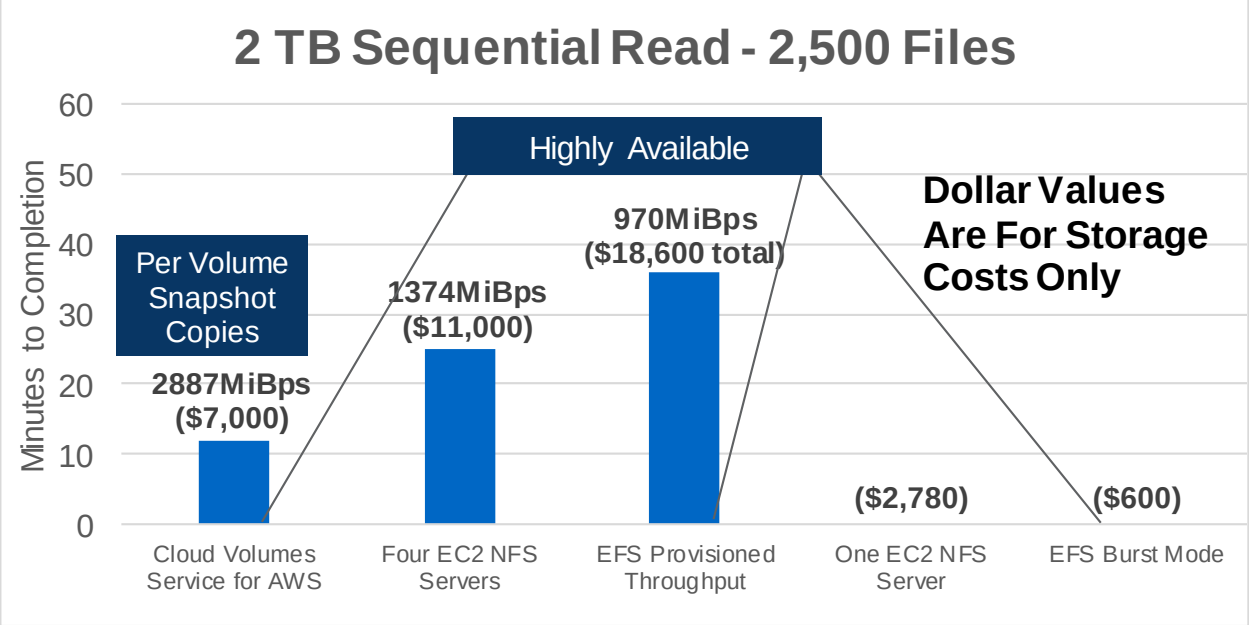
During testing, Cloud Volumes Service was compared against existing cloud-native storage present in Amazon Web Services (AWS), where the tests were run. The sequential read benchmark had a single hour to complete—with a stated goal of 10TIB per hour. The test itself comprised 2,500 GOR files representing 2000TIB of content, and was intended to evaluate several attributes, including:

- Four EC2 NFS servers atop Provisioned IOPS SSD (io1) Amazon Elastic Block Store (EBS) volumes
- A single EC2 NFS server configured similarly to the four servers
- Four EFS volumes configured for a total of 4096MIBps of bandwidth using Provisioned Throughput mode
- One EFS volume configured using Bursting Throughput mode
- One Cloud Volumes Service volume provisioned with the maximum bandwidth supported by Cloud Volumes Service

1.2 The Benchmark Results

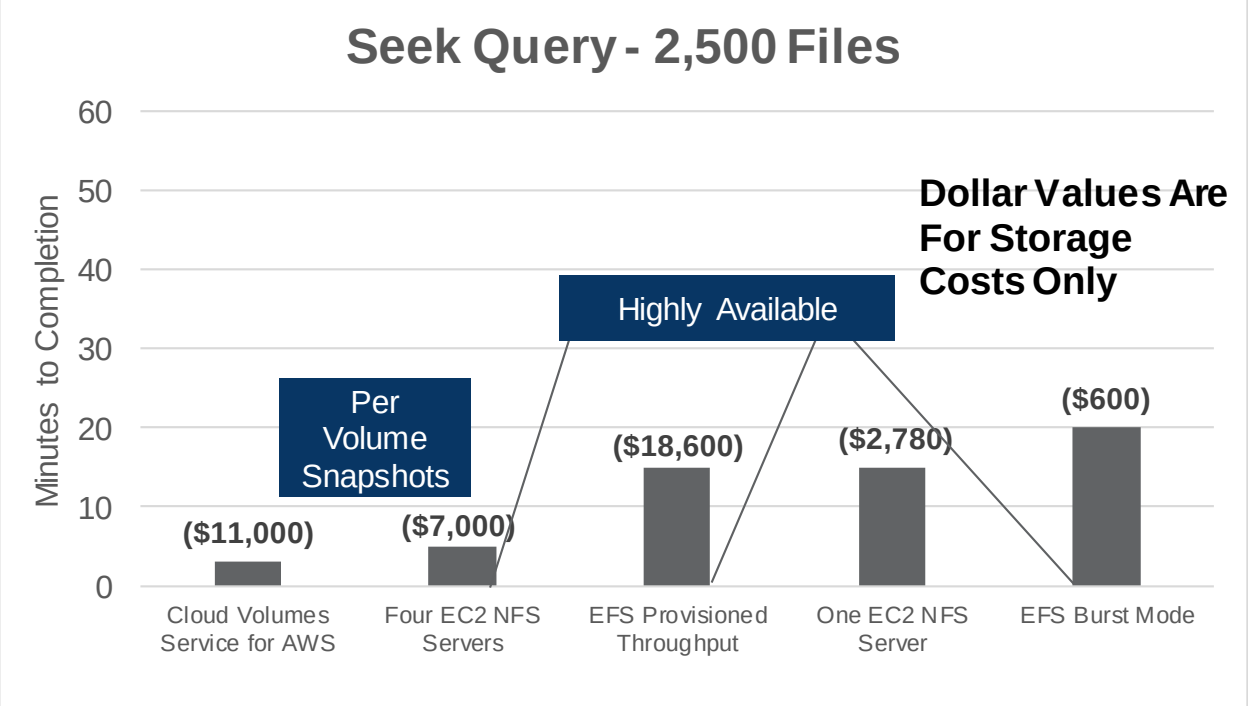
The results of the tests are shown in the following charts. Time to completion is represented on the y-axis of the graph; the lower the value (expressed in MIBps), the better. With a little bit of math, you can see that the throughput achieved using the Cloud Volumes Service volume is equal to 9.91TIB per hour, which is 2.1 times the rate of the four self-managed NFS servers and 3.0 times that of the Provisioned Throughput configured EFS volumes. The data from the tests themselves shows that a handful of workers took longer than the rest of the 2,500 workers, and that the throughput achieved throughout most of the Concurrent Versions System (CVS) test equaled roughly 3,500MIBps.

Figure 1) Benchmark Results.



As additional data points, the following graph shows the results of the second use case: that of a SQL-type query of the 2,500 genomic files. A lower time to completion indicates stronger performance.

Figure 2) Random Query Results.



1.3 Added Value

Cloud Volumes Service was able to access data from 100,000 individuals in less than an hour. This ability is of great value to WuXi NextCODE, considering that the goal of the experiment was to perform 10TIB of sequential reads within that time constraint. But holistically speaking, it's only the beginning. WuXi NextCODE has realized several other benefits since they began to use the NetApp service:

- Since the service is fully managed, site reliability engineering (SRE) resources are freed up to focus on tasks other than managing storage.
- The managed services are highly available and highly durable.
- Cloud Volumes Service boasts per-volume NetApp Snapshot™ copies, enabling rapid restore (through the API) and rapid clones (if they are used).

2 Conclusion: Satisfying the Exteme Demands of Genomic Analytics

Only Cloud Volumes Service is up to the task of satisfying the extreme demands of the genomic analytics workloads required by the WuXi NextCODE platform. Today, this architecture underpins preeminent genomics efforts on four continents and is the emerging global standard for organizing, mining, and sharing large-sequence datasets.

Let the NetApp Cloud Volumes Service enable your scale-out workload, as it did for WuXi NextCODE. Request a demo from a NetApp cloud specialist.

NetApp