NetApp and PetaGene

Accelerating the Genomics Revolution

Abstract

This document explains how technologies from NetApp® and PetaGene help to accelerate genomic data analysis and biomedical discoveries. It explores how NetApp's cloud and on-premises based solutions can simplify genomic research collaboration, reduce research IT cost, and shorten genomic analysis time.

The goal of the solutions is to achieve true precision medicine so that doctors and other care-givers can provide customised treatments and medications based on an individual's unique genetic profile. Although the old "one-size-fits-all" healthcare treatment model is still practiced, many clinicians believe that personalised care will soon become the standard around the world.

Summary of Benefits

The following brief list desribes the benefits that the NetApp and PetaGene collaboration brings to genomic researchers and biomedical clinicians.

- **Smaller files = faster genomic data transfers**. NetApp recently confirmed that when PetaGene's PetaSuite software shrinks the large genomic files into smaller, more portable files, there are immediate data flow benefits. For example, NetApp verified that the smaller files can be transferred more rapidly across the NetApp® Data Fabric. Faster genomic data benefits all stakeholders.

- **Transparent access to the files in their original format in real time**. PetaGene's software allows researchers and clinicians to continue using FASTQ.GZ and BAM file representations in their existing tools and pipelines, without needing to decompress first. On-the-fly decompression actually speeds up the process.

- **Up to 10x higher data efficiencies.** Another benefit that NetApp verified during the testing is higher levels of data efficiency. PetaGene advertises compression ratios of up to 10x for BAM files and up to 4.3x for FASTQ.GZ files. After running PetaGene's data compression software, NetApp resources reported as much as 4.5x improvement in space efficiency for the FASTQ.GZ files that were tested, as shown in Table 1. These results are consistent with PetaGene figures for those file types.[1]

- **Increase in research collaboration.** Having smaller files to manage makes it easier and faster to share genomic files between labs and research facilities, across campus or around the globe.

- **Cloud enablement (Amazon Web Services, Google, Azure, and more).** NetApp and PetaGene offer multiple options for managing genomic data in cloud environments such as AWS, Google, and others. Genomic files can be seamlessly and securely moved to and from the cloud to support various cloud-based workflows. Also, cold data can be tiered to object storage by using FabricPool, freeing performance tiers for new sequencing projects.

- **Reduced operational costs using Data Fabric**. PetaGene's software helps to decrease research operations costs because it shrinks the amount of data that needs to be managed at different stages of its lifecycle. Having smaller genomic files enables more efficient data transfers and accelerates data tiering. For example, some research data might need to be archived for short-term (or long-term) access by genomic researchers. Decreasing the size of the files by using PetaSuite compression software can save costs and time, regardless of whether the data is in a cloud or on premises. The NetApp Data Fabric allows researchers to seamlessly migrate genomic data where it's needed and when it's needed, locally or globally.

1 http://petagene.com/products

NetApp®

PetaGene transparent lossless compression

Technology Partner

Introduction

Scientists and bioinformaticians have long sought ways to reduce the size of their large genomic datasets by using a combination of data compression and reduction techniques. In the past, the raw sequencer output was often stored for extended periods while bioinformaticians carried out the complex tasks in assembling and aligning the sequencing data. After these steps were complete, data could be used in variant calling and interpretation, which are the vital steps in understanding gene expression and disease.

Today, this process is highly automated and has been greatly accelerated through a combination of parallel processing and the availability of reference genomes. There are now several compressed genomic file formats that reduce the size of an individually stored genome down to a few tens of gigabytes. Efforts that previously took months or years can now be turned around in little more than a day. This has greatly improved the ability of bioinformaticians to work with and transfer data to clinicians quickly and efficently.

We are entering a world of personalised or precision medicine. Faster sequencing and more compact datasets have increased the number of individual sequences that can be performed. Individual patient genomes and even their individual diseases, typically cancers, can be sequenced. (Cancerous tumors have their own genetic makeup that can be differentiated from that of a healthy individual.) These developments carry great hope and opportunity for new insights, while increasing the pressure on data capacity.

Amplify Storage Efficiency

PetaGene's software addresses the challenges resulting from growing volumes of genomics data. PetaSuite is a set of scalable complementary software tools that significantly reduces the size and cost of NGS data for storage and transfer. NetApp's testing showed a reduction of up to 4.5x in both storage costs and data transfer times compared to gzipped FASTQ.GZ files. This result is consistent with PetaGene's advertised figures for those files types. PetaSuite also reduces the size of BAM files by up to 10x. PetaSuite transparently integrates with the existing storage infrastructure and bioinformatics pipelines, while PetaLink provides a powerful virtual file access system. PetaLink produces a high-performance virtual file view of the compressed file. This virtual file can be used transparently just like the original file by Linux toolchains, pipelines, and genome browsers, with a speed-up rather than a slow-down in performance.

Maximum Compression Ratio	Minimum Compression Ratio	Average Compression Ratio
4.50	2.03	3.24

Table 1) Observed Compression Ratios

Increase Data Mobility

NetApp SnapMirror® and NetApp Cloud Sync make it easy to move genomic data across various cloud environments, so that data is readily available where it is needed. This capability is enhanced by PetaGene's unique data compression technique. The following sections describe three reference designs which can be leveraged to efficiently store and access genomic data, improving performance and reducing cost.

Here is a high-level overview of the workflow of the reference designs:

- Raw genetic data is read by the sequencer, and the results are written out in BCL format to a NetApp ONTAP® cluster.
- Because many companies prefer to split the conversion of BCL to FASTQ.GZ and BAM formats across separate environments, the BCL files are transferred using SnapMirror to a second ONTAP cluster. Although not strictly necessary, this convention was followed in the reference design.
- At the secondary site, the BCL files are processed and converted into FASTQ.GZ or BAM format.
- PetaSuite compresses the FASTQ.GZ and BAM files and converts them from BAM to PGBAM, and from FASTQ.GZ to FasterQ formats respectively, while keeping them transparently accessible to the user in their native FASTQ.GZ and BAM formats. The space savings discussed previously are realised at this stage.
- The compressed files are transferred from the secondary site to either volumes within the NetApp Cloud Volumes Service, a NetApp Cloud Volumes ONTAP cluster, or a NetApp Private Storage (NPS) cluster. The transfer mechanism depends on the target—Cloud Sync in the case of NetApp Cloud Volumes Service, or SnapMirror if NetApp Cloud Volumes ONTAP or NPS is used.
- All postprocessing activity is performed in one of the various hyperscaler environments and all file operations occur in the NetApp storage target.
- The resulting datasets can be replicated back to the origin data center if necessary, by using either Cloud Sync or SnapMirror, depending on the cloud solution.
- Aged files can be archived to object storage either in the public cloud (AWS S3, Azure Blob, or a NetApp StorageGRID® Webscale target) or on premises using StorageGRID Webscale. NetApp recommmends using NetApp Cloud Sync or FabricPool to handle the archival process.

Control, Protection, and Efficiency with Cloud Volumes ONTAP

NetApp Cloud Volumes ONTAP delivers enterprise control, protection, and efficiency of your data with the flexibility of the cloud. A software-defined data management service built on ONTAP 9 software, Cloud Volumes ONTAP provides a superior universal storage platform that addresses most cloud data needs. Having the same storage software in the cloud and on premises delivers the value of a Data Fabric, without needing to train the IT administrators in new data-management methods. The SnapMirror feature of ONTAP offers a bandwidth-efficient data replication and transfer mechanism between the clouds and to or from a data center.

Cloud Volumes ONTAP provides a data storage solution that fits many different customer requirements, including disaster recovery, development, and test environments. Cloud Volumes also supports critical applications that require highly available nondisruptive operation such as production business applications and file services using NFS, SMB, and iSCSI. Setup and management of the Cloud Volumes ONTAP environment is simple and intuitive with the NetApp OnCommand® Cloud Manager web interface.

Speed, Scale, and Simplicity with Cloud Volumes Service

NetApp Cloud Volumes Service is a cloud-native file storage service based on proven NetApp technology. This offering combines enterprise-class storage with the simplicity and flexibility of the cloud, resulting in the ability to take your operation from 0 TB to 100 TB in less than 10 seconds. NetApp Cloud Volumes Service supports the NFS v3 and NFS v4 protocols along with SMB.

NetApp Cloud Sync is an intuitive replication and synchronisation service that enables simplified replication into and out of NetApp Cloud Volumes. This software-as-a-service (SaaS) offering enables customers to transfer and synchronise data between source and destination of any type or formats, in the cloud or on premises. Cloud Sync supports NAS data (NFS and SMB), EFS, Amazon S3, and NetApp StorageGRID Webscale appliance.

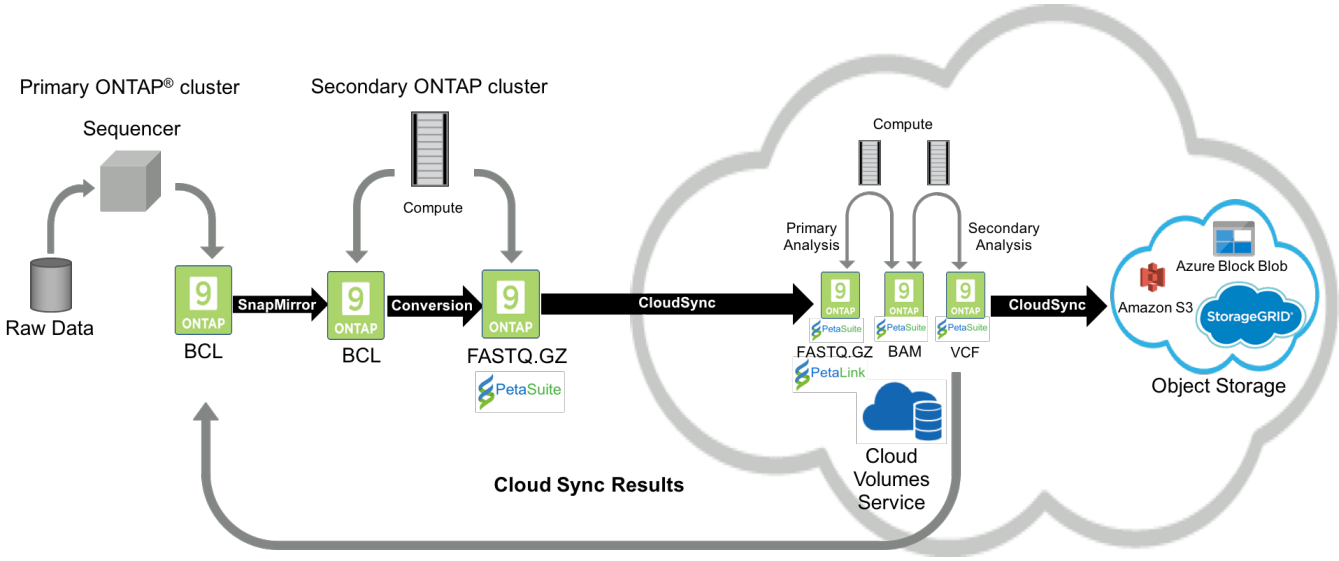

Figure 1: Cloud Volumes ONTAP reference design

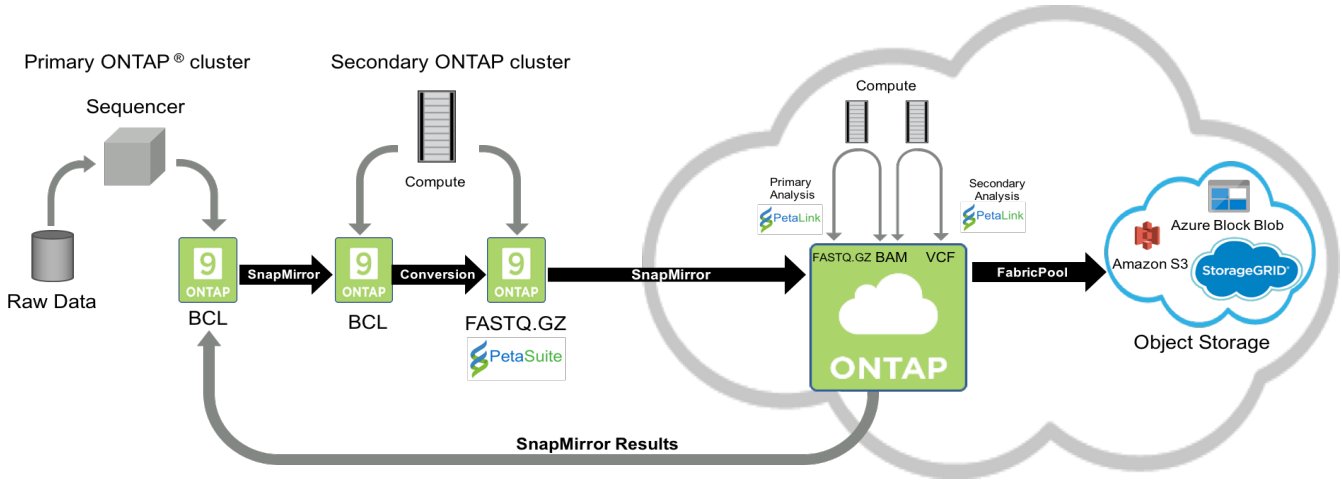

Figure 2: Cloud Volumes Service reference design

Freedom and Flexibility with NetApp Private Storage

NetApp Private Storage (NPS) is a cloud-connected storage solution that puts data near the cloud, providing the freedom and flexibility to run your application or workload on cloud compute while maintaining complete control of your data. NPS connectivity options allow you to choose from an expanding global network of cloud service providers, including AWS, Google Cloud Platform, IBM Cloud, and Microsoft Azure. With NPS, you can easily ensure compliance with HIPAA, GDPR, or any other regulatory requirement.

With NPS, your NetApp storage is housed in colocated cloud-connected data centers, next to major networks and close to all major clouds. Establishing secure, dedicated, high-speed connections to all those clouds is quick and easy, with the added advantage of enhanced performance and reduced cost by bypassing the internet. NetApp makes it easy to move data between clouds and any NetApp data management infrastructure, including public, private, and hybrid clouds. SnapMirror technology enables applications to failover to a secondary system and continue operating with the capability to fail back to the primary location later.

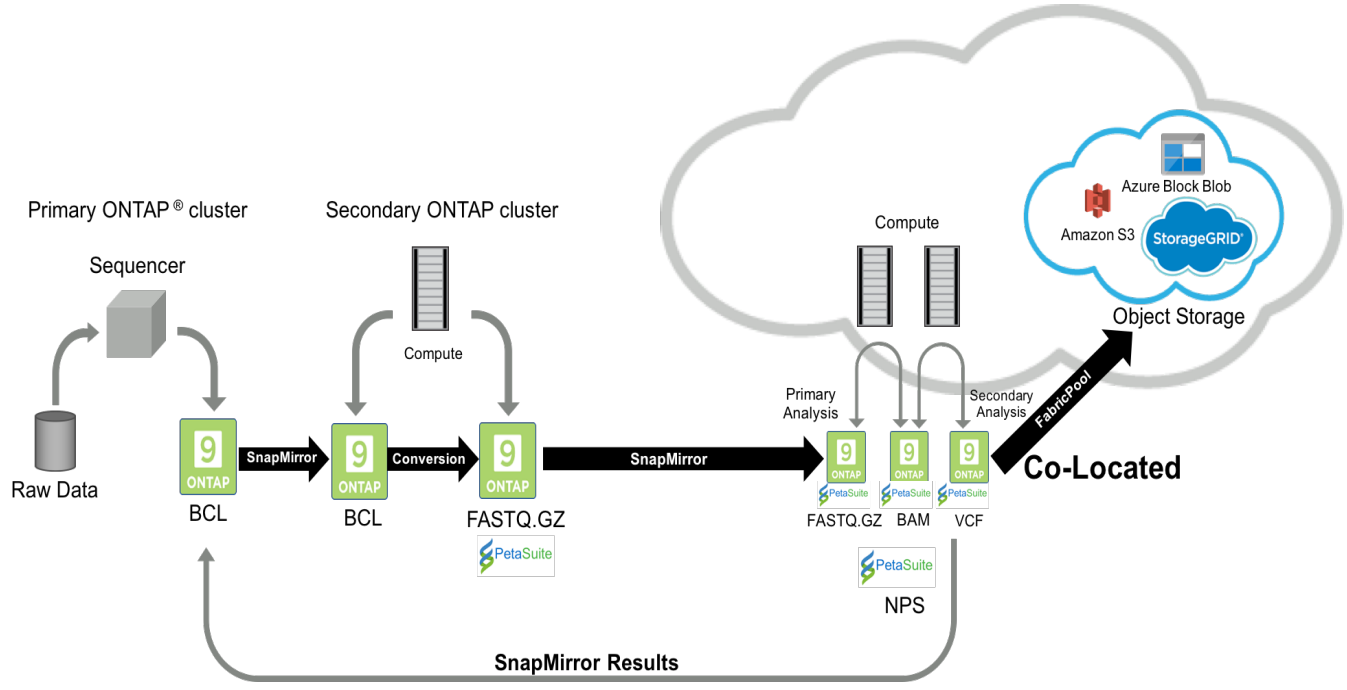

Figure 3: NetApp Private Storage design

	NetApp Private Storage	Cloud Volumes ONTAP	Cloud Volumes Services
Legal restrictions	X		
Data Mobility	X	X	X
Multicloud	X		
Cloud deployment		X	X
Cost model	Opex & capex Models available	Opex	Opex
Simplicity			X
Feature currency	X		
Storage tiering	X	X	X

Table 2) NetApp Cloud Data Services
Comparison matrix for the NetApp Cloud Technologies

About PetaGene

PetaGene was founded in Cambridge, the birthplace of genomics, to address the rapidly growing data management problems of the genomics industry. PetaGene's software enables compression of huge amounts of genomic data without compromising on access or data quality. The company's products go beyond regular data reduction techniques and have twice been recognised by Bio-IT World's Best of Show Award for their industry-leading performance and usability. For more information visit www.petagene.com or email sales@petagene.com.

About NetApp

NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of data across cloud and on-premises environments to accelerate digital transformation. We empower global organisations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation and optimise operations. For more information email genomics@netapp.com. #DataDriven.