Accelerating Genomic Discovery with Cost-Effective, Scalable Storage

EXTRAORDINARY PERFORMANCE
Accelerated discovery for genomics, 60-90% reduction of data transfer time; 10x faster than traditional NAS

EASY MANAGEMENT
Lossless compression transparent to workflows, single namespace for local and cloud object storage, global sharing of data

BREAKTHROUGH ECONOMICS
Up to 90% reduction of genomic file sizes, integrated storage tiering with object storage, up to 60% reduction in storage costs

UNMATCHED SCALABILITY
Support for trillions of files and file sizes up to 4 PB, scalable performance of world's fastest file system, and capacity of object storage

THE PROMISE OF GENOME SEQUENCING

Genome sequencing offers insights on the genetic origins of certain diseases and holds the promise of potential discovery of new treatments to enhance human health. Genomic profiling is empowering researchers to explore the possibilities of precision, targeted treatments for cancer. Genomic technologies can be leveraged to potentially diagnose, treat, characterize, and provide understanding into foundations of human disease, including very rare ones.

THE TECHNOLOGY CHALLENGES

Since the sequencing of the first human genome in 2001, the massive growth of genomic data has made storage costs prohibitively expensive. Genomic data is predicted to reach 2-40 exabytes/year by 2025[1]. Current analytics platforms struggle to process these massive amounts of data in a timely manner, and storage costs dominate the budgets of large genomics applications. A single human genome file occupies approximately 100 GB of storage, and a large data set is in the multi-petabyte range. As storage costs escalate, the pace of discovery slows.

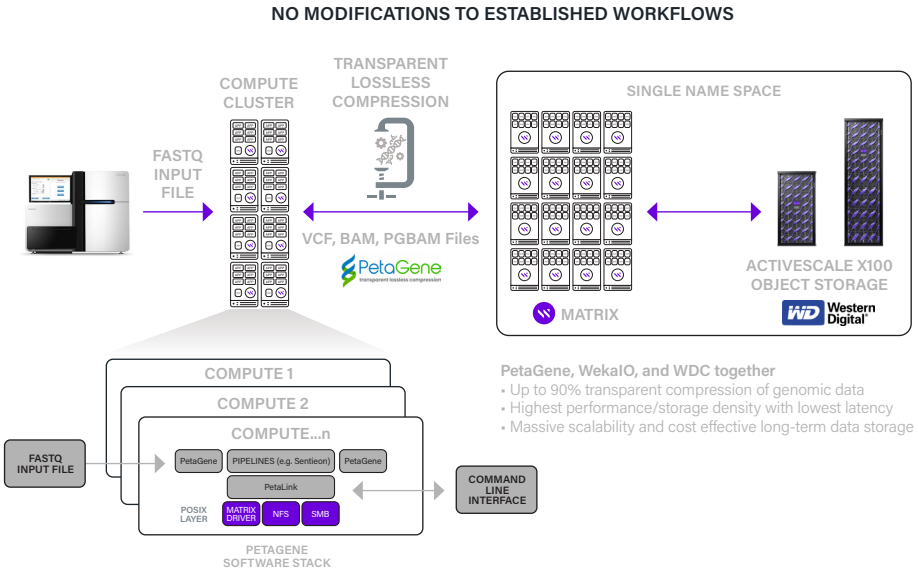

Figure 1 – Sample Workflow Leveraging PetaGene, WekaIO, and Western Digital to Accelerate Genomic Discovery

STORAGE SOLUTION REQUIREMENTS

A comprehensive storage solution must reduce the amount of storage that genomic data requires without losing any information. It must also maintain compatibility and transparency with existing workflows. Finally, it must cost-effectively scale in both performance and capacity.

DATA AND COST REDUCTION

PetaGene lossless data compression dramatically reduces the storage capacity needed for genomic data sets. Together with WekaIO, which leverages NVMe-based flash and a single namespace that spans local storage and cloud object storage, the most cost-effective benefit is derived from both high-performance and long-term storage investments. Discovery is accelerated with NVMe flash, while Western Digital's ActiveScale™ cloud object storage is used for long-term storage and archival. Whether the PetaGene-compressed files are stored locally or in the cloud, PetaGene's PetaLink and WekaIO's Matrix™ file system technologies provide transparent and secure access to this data to all applications, tools, and pipelines without modifications to established workflows.

STORAGE PERFORMANCE AND CAPACITY SCALING

PetaGene data compression accelerates data transfer of genomic files. WekaIO further reduces time to discovery by providing low-latency data access and fast delivery of data to compute servers. WekaIO eliminates the I/O bottleneck and the CPU starvation problems common to genomic and cancer research workloads. These workloads require many runs to be processed concurrently. Matrix is the world's fastest file system, and it ensures that the sequencing workloads are kept CPU-bound. Matrix™ delivers simplified management and data protection, and is 3x faster than local file systems and 10x faster than traditional NAS. With PetaGene compression and integrated tiering and remote backup to the cloud on ActiveScale, WekaIO provides unprecedented storage performance and capacity scaling for genome sequencing workloads.

SCALABLE AND COST-EFFECTIVE STORAGE

With genomic data continuing to grow exponentially, storing, managing, and protecting this critically important information within a tight budget is a challenge. WekaIO and the ActiveScale highly-scalable, high-performance cloud object storage system address this issue with breakthrough economics. WekaIO supports trillions of files and file sizes up to 4 PB, all in a single namespace. ActiveScale provides low-TCO, long-term archival of data that frees up expensive capacity on primary storage and supports multiple workloads often served by traditional NAS solutions. WekaIO and ActiveScale empower research partners around the world to collaborate by allowing sharing of genomic data regardless of device platform. ActiveScale reduces storage capacity requirements by distributing a single copy of data across multiple data centers, reducing storage costs by up to 60%.

STORAGE FOR ACCELERATING GENOMIC DISCOVERY

PetaGene delivers dramatic, lossless compression of genomic data while preserving transparency to established workflows. WekaIO's Matrix file system delivers unprecedented performance and scalability, unified support for local and cloud object storage, simplified management, and reduced time to discovery. ActiveScale delivers an on-premises and hybrid cloud system that extends primary storage capacity with massive scalability, outstanding levels of durability, and affordability. PetaGene, WekaIO, and Western Digital provide storage for genomics that is easy to manage and combines industry-leading density and performance with breakthrough scale, economics, and world-class support.

Find out more about PetaGene, Western Digital Corporation, and WekaIO offerings, and explore the possibilities of streamlining and accelerating your genomic processing while dramatically lowering your storage costs.

[1] "Big Data: Astronomical or Genomical?" by Stephens, Lee, et al., Public Library of Science (PLOS) Journal, July 7, 2015; available online at https://bit.ly/2Vx2lEw

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io

W16R1SB201904