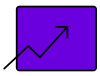
Meeting the Storage Challenge of Exponential Data Growth in Genomics

BREAKTHROUGH ECONOMICS
Consolidate multiple tiers of storage, saving power, cooling, and rack space.

EASY MANAGEMENT
Policy-based tiering optimizes data placement to simplify data management.



INFRASTRUCTURE AGILITY
Independent capacity and performance scaling.

> **We needed something that's much more scalable than existing NAS solutions — an infrastructure that could grow to hundreds of petabytes. Our existing solution couldn't provide that scale — that's what drove us to Weka.**
>
> ***David Ardley,*** *Director of Infrastructure Transformation, Genomics England*

THE CHALLENGE

Completing the sequencing of the first human genome in 2003 was a key scientific breakthrough that took 10 years to complete. Since then, high-throughput sequencing machines have allowed researchers to perform sequencing runs in a matter of hours, greatly increasing the pace of scientific discovery. The result has been an explosive growth of genomic data, driving organizations to find more affordable ways to manage and share the data. Based on current sequencing rates, data storage demands are doubling every seven months and are forecasted to reach over 40 exabytes of capacity by 2025, just for the human genome alone. But today's complex scientific workflows require high throughput and IOPS at low latencies so that researchers can achieve faster discovery. Legacy storage systems that have limited scale are holding back research and creating data silos that are difficult and costly to manage. Supporting genomic sequencing workflows is further complicated by the large data sets and long retention periods of life science data. A solution is needed to protect valuable digital assets for long periods of time without driving up the cost of managing the infrastructure.

MEETING THE STORAGE CHALLENGE OF EXPONENTIAL DATA GROWTH IN GENOMICS

Legacy file systems were architected to support hard disk drives that provide good large-file and sequential access performance, whereas genomic analysis requires small-file and random access performance. Similarly, traditional backup and archive systems were designed for multiple tiers of disk and tape-based systems. This combination adds cost, complexity, and considerable management overhead to ensure that the data is accessible and readable when needed.

WekaFS™ + OBJECT STORAGE = A WINNING SOLUTION

Life science involves some of the most complex analyses found in scientific research. It is not surprising that researchers have very unique needs in terms of computing and storage performance, scalability, and accessibility. File-based storage is ideal for high-performance compute clusters used during the analysis phase of the workflow. For long-term storage and sharing of valuable research data, a cloud-scale active archive is a more effective approach. The combination of WekaIO's WekaFS, the modern file system that is uniquely built to solve big problems, and an object storage data lake provides an ideal two-tier storage solution, offering the performance, scalability, and data resiliency critical to accelerating discovery and protecting valuable research results.

PARALLEL FILE ACCESS TO ACCELERATE YOUR BUSINESS

Performance starts with the file system. WekaFS is a distributed, scale-out, POSIX-compliant file system that runs on your existing compute cluster and uses off-the-shelf SSDs, greatly improving storage system performance. With data on flash-based storage inside the server and part of a global namespace, access is near-instantaneous. Valuable data is protected using patented data protection and distribution algorithms that allow the system to sustain up to four simultaneous node or SSD failures.

Genomics england

DATA DURABILITY AND INTEGRITY AT SCALE TO SAFEGUARD DATA

Industry-leading object storage system software ensures that valuable data is protected and always available with high data durability and site-level fault tolerance in a multi-site configuration. Most object storage systems can tolerate bit-errors without the loss of data, which can eliminate the risks, costs, and media management activities associated with tape-based archives.

The WekaFS hybrid solution presents a performance tier for hot data and a capacity tier for cold data. The entire solution can be up and running in hours. A single global namespace is presented to the applications and data is automatically migrated to the object storage system for long-term retention either on-demand or based on policies. All file metadata remains on the performance tier so any file is easily retrieved at a later date if needed by an application.

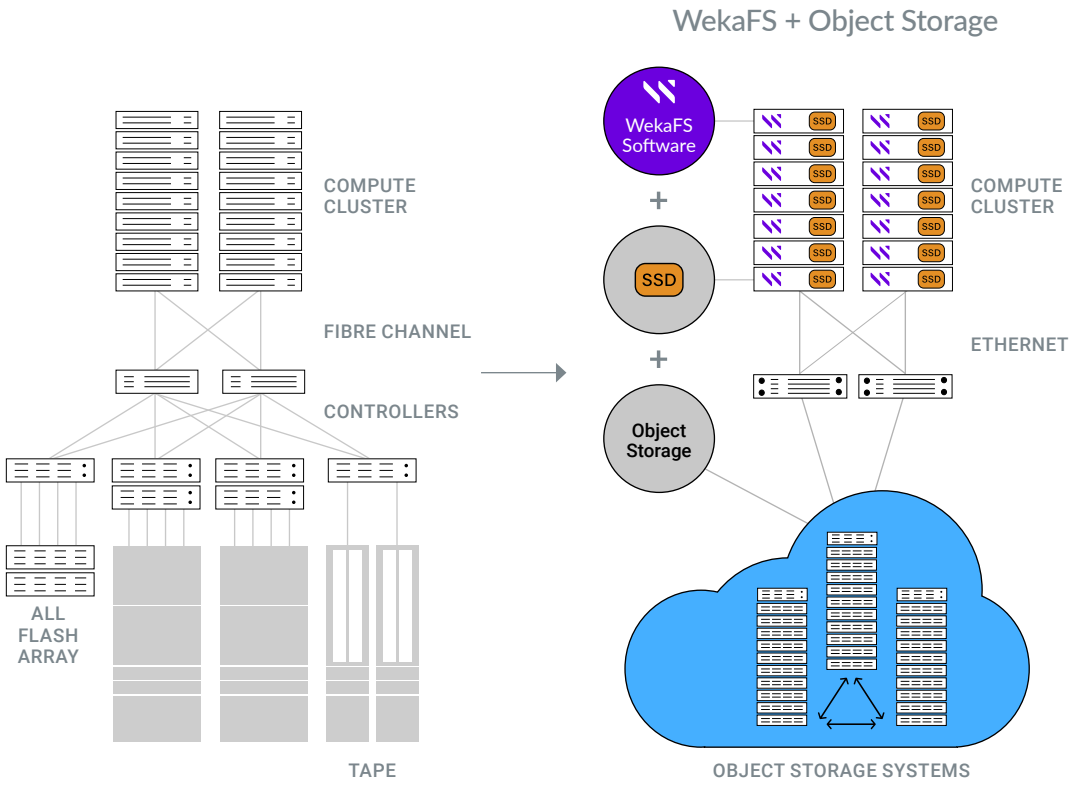


Figure 1 – WekaFS and Object Storage Life Science storage architecture

THE SOLUTION IN PRODUCTION

Genomics England (GEL) needed a solution to support the UK National Health Service 5 Million Genomes Project but could not scale with its existing NAS solution from a leading vendor. Weka delivered a two-tier architecture that takes commodity flash and disk-based technologies and presents it as a single hybrid storage solution. The primary tier consists of 1.3 petabytes of high-performing NVMe-based flash storage that supports the working data sets. The secondary tier consists of 40 petabytes of object storage to provide a long-term data lake and repository. Weka presents the entire 41 petabytes as a single namespace. Each of the tiers can scale independently: should GEL require more performance on the primary tier, it can scale its performance independently of the data lake. The system takes advantage of the geo-distributed capability of the object store, and data is protected across three locations that are 50 miles apart from one another.

WEKA

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

W19SB202003