Genomics England Improved Scale and Performance for On-Premises Cluster

WekaIO (Weka) allows Genomics England to scale to extreme performance and capacity in their mandate to sequence 5 million genomes by 2023

Genomics England (GEL), headquartered in London, England, is owned by the UK Department of Health and Social Care and tasked to run the 5 million genomes project, which was announced in 2018 and aims to sequence 5 million genomes from National Health Service (NHS) patients with rare diseases.

The organization acquires sequenced DNA samples from National Health Services patients and provides access to a team of over 3,000 researchers to use this data for medical research. Having completed the sequencing of the first 100,000 genome project, GEL has already acquired 21 petabytes of genome data and is projected to amass over 140 petabytes by 2023. The research conducted requires access to the entire data set and must allow researchers to query the data in a highly randomized fashion. Therefore, all data has to be stored in a single storage system.

CHALLENGES

- Scale capacity in a single namespace to store 140 petabytes of data
- Improve storage performance to accelerate innovation
- Devise a strategy to protect valuable national data from a major disaster
- Comply with tight budget constraints

BENEFITS

- No limit on capacity scaling
- 10x+ improvement in performance
- 75% reduction in storage cost per genome
- Full disaster recovery strategy in place
- Integration with public cloud for compute elasticity

"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 and wasn't performing as well in these magnitudes — that's what drove us to Weka."

David Ardley, Director of Infrastructure Transformation, Genomics England

THE CHALLENGE: POOR PERFORMANCE AT LARGE CAPACITY SCALE

In 2018, Genomics England approved a new storage platform to support the projected growth to 5 million genomes by 2023. Previously, GEL had implemented a scale-out NAS solution from a leading vendor to support the 100,000 genome project; however, it had already hit its limit on storage node scaling, and performance suffered when the system was near capacity. In addition, the existing solution had no viable disaster recovery strategy, as backing up all 21 petabytes of storage was infeasible. Key national data would be in a vulnerable position if a major disaster were to occur. The GEL infrastructure team determined that it needed a new storage strategy to support the anticipated growth through 2023.

The new solution had to meet several key criteria:

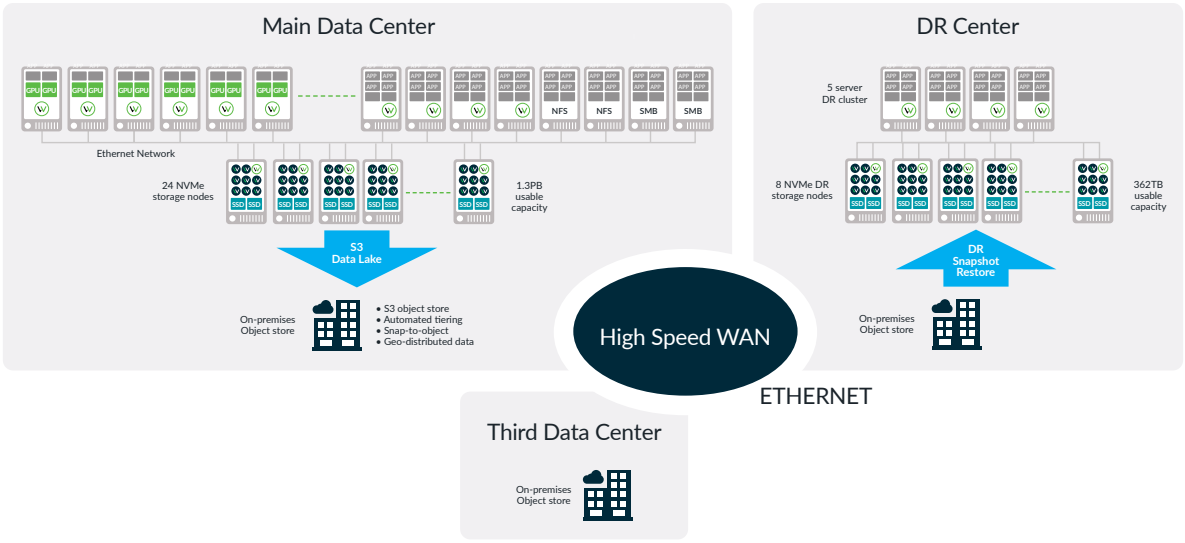
- Scale in line with the anticipated growth to 140 petabytes in a single storage system
- Meet the demanding performance requirements of the bioinformatics pipeline and core research
- Provide a disaster recovery solution
- Be easy to manage while delivering the required enterprise features
- Fit within the allotted budget

THE SOLUTION: WekaFS™ SOFTWARE ON INDUSTRY STANDARD SERVER INFRASTRUCTURE

Genomics England evaluated all the major contending solutions through a rigorous RFP process. They rejected parallel file systems because of their complexity and lack of enterprise features. They rejected all-flash scale-out NAS because the economics were not viable for a project at the scale of GEL's requirements. Ultimately, GEL chose WekaIO because it was the only vendor that could deliver a solution that met all the requirements in a single architecture.

WekaFS is a fully parallel and distributed file system that has been designed from scratch to leverage both high-performance flash technology and cost-effective disk storage. Data and metadata are both distributed across the entire storage infrastructure to ensure massively parallel access to NVMe drives. Data is seamlessly tiered from flash to disk with Weka's internal tiering mechanism, achieving the optimum use of storage media for the best economics.

WekaFS 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 do so 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, 50 miles apart. If a major disaster occurs in the primary location, Weka's unique Snap-to-Object feature allows the system to be re-started in a second location, ensuring continued access to the data. The geo-distribution and object store resiliency provide highly effective and feasible data protection, as full back-up is impractical at this scale.



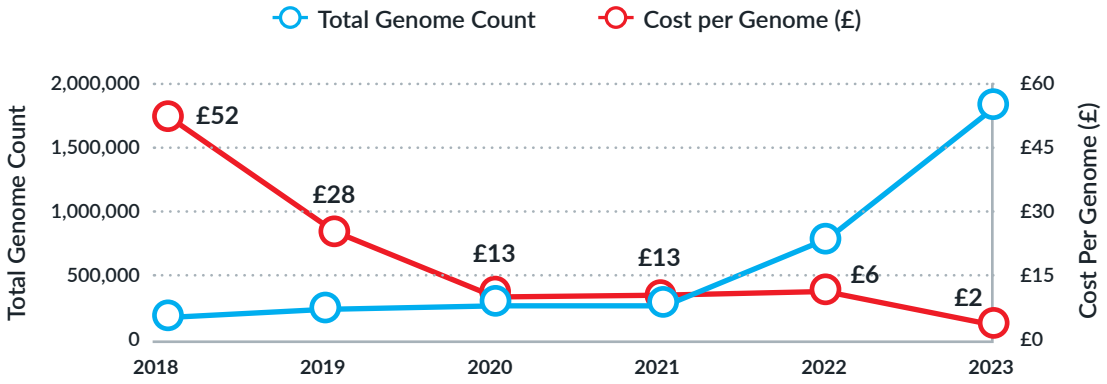
Genomics England Infrastructure with Disaster Recovery

BENEFITS AND RETURN ON INVESTMENT

Genomics England was able to realize several benefits and tremendous return on investment by choosing WekaFS:

- GEL enjoyed a 10x+ increase in performance over its legacy NFS-based NAS. The current system is capable of delivering over 135 GBytes/second from the NVMe tier and performance will continue to scale as the cluster grows.
- The storage cost per genome has dropped from £52 to £13, a 75% reduction in storage cost, and is anticipated to plunge to £2 by 2023, achieving a 96% reduction in cost. This has been accomplished because of the ability to integrate low-cost disk-based object storage into the solution.
- GEL can survive a major disaster at the primary site and still maintain access to the data. The object tier is geo-distributed across three sites, each 50 miles apart. Should the primary site fail, a small disaster recovery cluster is available on a second site and can re-hydrate the file system from the most recent snapshot.
- Critical data sets are fully encrypted from the high-performance compute cluster all the way to the permanent data store, with integration to a key management system. Performance measurements showed no discernable degradation in application performance with encryption enabled. In addition, the system is protected from rogue security threats through a robust authentication mechanism.

Finally, new opportunities have opened up for Genomics England with Weka. As more research comes on-line, GEL would like to integrate public cloud for compute elasticity. The Weka solution will allow GEL to burst to the cloud and leverage on-demand compute resources.



Huge Reduction in Storage Cost per Genome

WEKA.IO
910 E HAMILTON AVENUE, SUITE 430, CAMPBELL, CA 95008 USA T 408.335.0085 E info@weka.io www.weka.io