



Genomics England Improved Scale and Performance for On-Premises Cluster

CHALLENGES

- Scale capacity in a single namespace to store 140 Petabytes of data
- Improve storage performance to accelerate innovation
- No strategy to protect valuable national data from a major disaster
- 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
- Integrated 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

WekaIO allows Genomics England to scale to extreme performance and capacity in their mandate to sequence 5 million genomes by 2023. Genomics England, headquartered in London, England, is owned by the UK Department of Health and Social Care to run the 5 million genomes project, announced in 2018, which 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 allows 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, Genomics England (GEL) has already acquired 21 Petabytes of genome data and it is projected to grow to 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.

THE CHALLENGE: POOR PERFORMANCE AT LARGE CAPACITY SCALE

In 2018, Genomics England approved a new storage platform to support the 5 million genome projected growth through to 2023. 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 is infeasible. This left key national data in a vulnerable position should a major disaster occur.

The GEL infrastructure team determined that they 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 demanding performance requirements of the bioinformatics pipeline and core research
- Provide a disaster recovery solution
- Be as easy to manage while delivering the require enterprise features
- Fit within the allotted budget

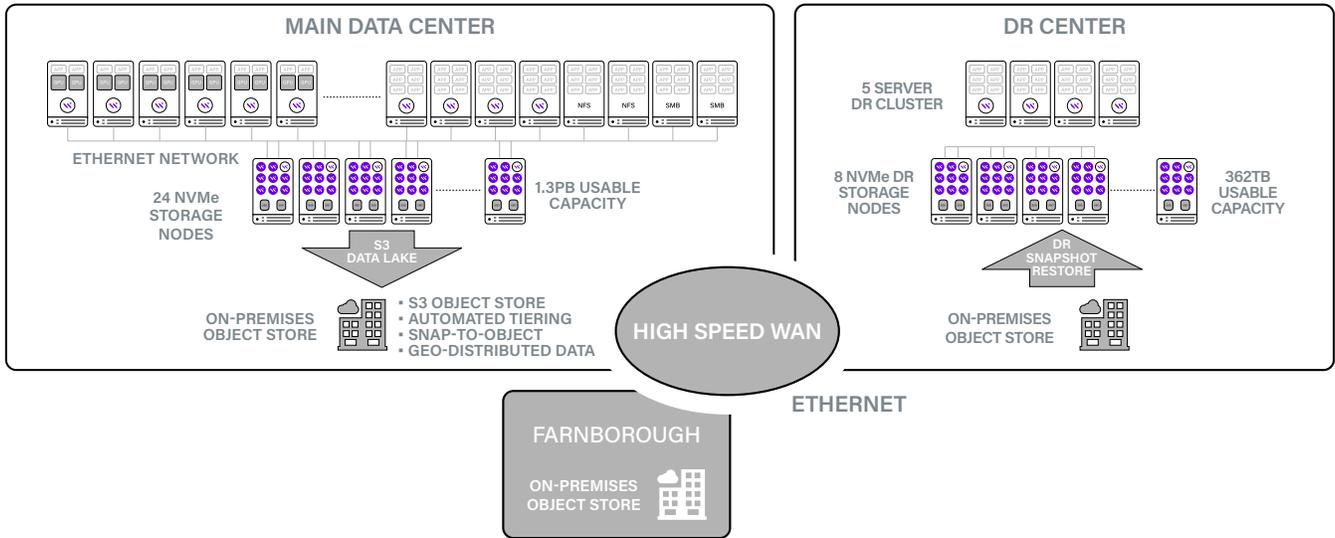
THE SOLUTION: WEKAFS™ SOFTWARE ON INDUSTRY STANDARD SERVER INFRASTRUCTURE

Genomics England evaluated all the major contending solutions through a rigorous RFP process. They rejected parallel file systems due to the 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 who 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 high performance flash technology and cost-effective disk storage. Both data and metadata are 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 allowing for best use of storage media and 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 which 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 there is a major disaster 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 significantly higher data protection, as full back-up is impractical at this scale.



Genomics England Infrastructure with Disaster Recovery

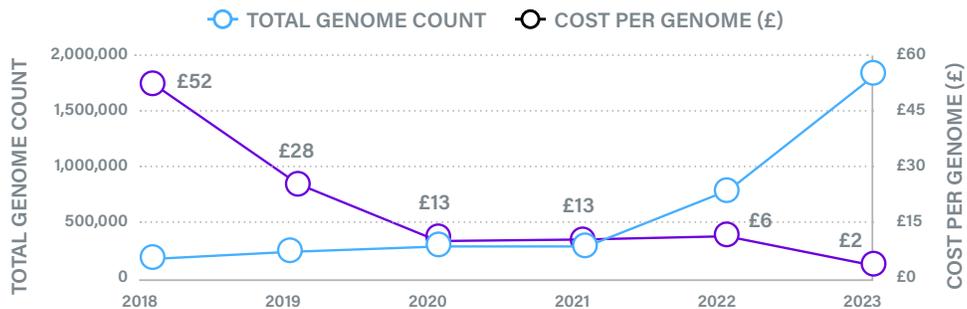
BENEFITS AND RETURN ON INVESTMENT

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

- GEL enjoyed a 10X+ increase in performance compared to their legacy NFS-based NAS. The current system is capable of delivering over 135GBytes/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% drop in storage cost, and is anticipated to reach £2 by 2023, a 96% cost reduction. This has been accomplished due to 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. 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



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

©2020 All rights reserved. WekaIO, WekaFS, WIN, Weka Innovation Network, the Weka brand mark, the Weka logo, and Radically Simple Storage are trademarks of WekaIO, Inc. and its affiliates in the United States and/or other countries. Other trademarks are the property of their respective companies. References in this publication to WekaIO's products, programs, or services do not imply that WekaIO intends to make these available in all countries in which it operates. Product specifications provided are sample specifications and do not constitute a warranty. Information is true as of the date of publication and is subject to change. Actual specifications for unique part numbers may vary.