WEKA.IO

Meeting the Storage Challenge of Exponential Data Growth in Genomics

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

UNMATCHED SIMPLICITY
Plug-and-play simplicity, a jointly tested and certified solution that is easy to deploy

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

TRANSFORMATIONAL AGILITY
Independent capacity and performance scaling

The Challenge

The life sciences and healthcare industries are in the midst of a dramatic transformation that will make personalized medicine common-place. 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 explosion in genomic data, driving organizations to find more affordable ways to manage and share the data. But today's complex scientific workflows require high throughput and IOPs at low latencies so that researchers can get to discovery faster. Research is limited by legacy storage systems that don't scale, creating data silos that are difficult and costly to manage. Supporting this workflow is compounded by the large data sets and long retention periods of life science data. What is needed is a solution that protects the valuable digital assets for long periods of time, while not driving up the cost of managing the infrastructure.

MEETING THE STORAGE CHALLENGE OF EXPONENTIAL DATA GROWTH IN GENOMICS

Current generation file systems were architected to support hard disk drives which provide good large-file and sequential access performance, whereas genomic analysis requires small-file and random access performance. Similarly, backup and archive systems are comprised with 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.

WekaIO MATRIX + ACTIVESCALE = AN IDEAL SOLUTION

Life science involves some of the most complex analyses found in scientific research. It's not surprising that researchers have very unique needs in terms of compute 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 Matrix™ and HGST's ActiveScale™ object storage system make 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. MatrixFS 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.

DATA DURABILITY AND INTEGRITY AT SCALE TO SAFEGUARD DATA

ActiveScale system software ensures valuable data is protected and always available with up to 17 nines durability, including site-level fault tolerance in a multi-site configuration. Robust data integrity checks occur automatically and transparently protecting data from silent data corruption known as bit-rot. Each object can tolerate 1000 bit-errors without the loss of data, which can eliminate the risks, costs, and media management activities associated with tape-based archives.
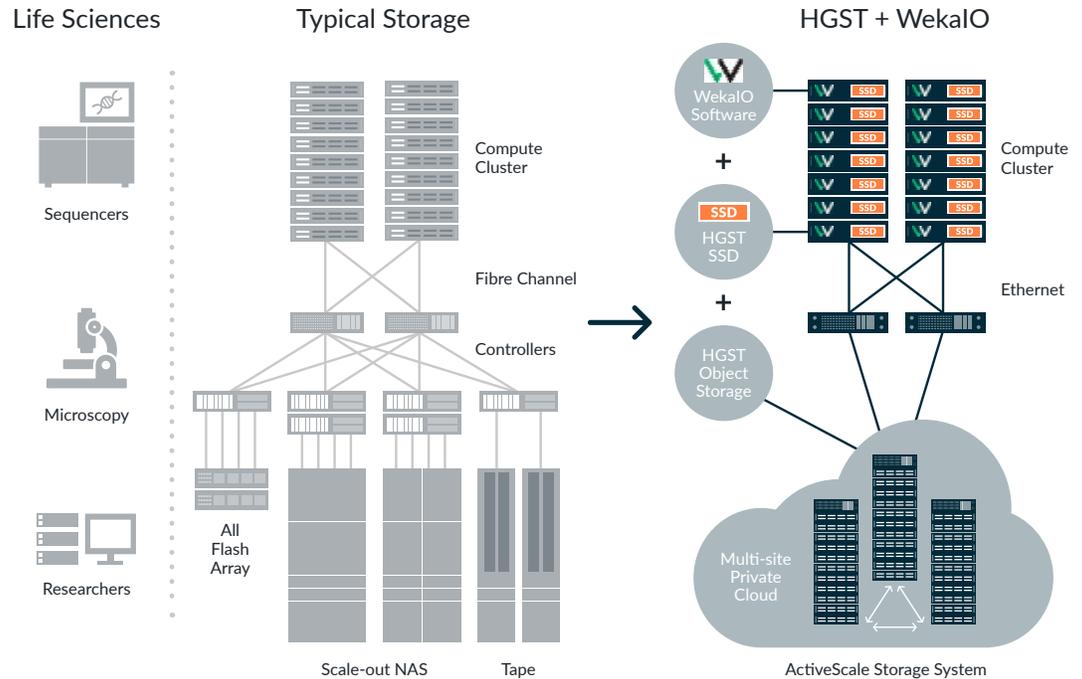
Figure 1 - HGST and WekaIO Life Science storage architecture

EASY DEPLOYMENT AND DATA MANAGEMENT FOR LOWER TCO

The transformative simplicity of the combined solution compared to traditional storage architectures is shown in Figure 1. Multiple tiers of storage are consolidated into a performance tier for hot data and a capacity tier for cold data. The entire solution can be up and running in minutes. Simply run the automated MatrixFS installation procedure on the selected servers, roll the ActiveScale system in place, connect the power, configure the network connections, and it presents an Amazon® S3™-compliant object interface and global namespace to MatrixFS. Data is automatically migrated to the to the ActiveScale system for long-term retention either on-demand or based on policies.

The Solution

Designed for the most demanding research workloads, a comprehensive storage solution for life sciences includes the WekaIO Matrix software platform and HGST's ActiveScale object storage system that when combined can dramatically change the economics of high performance storage. Matrix software converts your existing server cluster into a massively parallel scale-out storage system with linear performance, scalability, and sub-millisecond latency. While ActiveScale object storage system is a complete, easy-to-install, easy-to-manage active archive solution delivering cloud economics and superior resiliency. Leveraging off-the-shelf servers, SSDs, and object storage, the solution delivers the performance of an all-flash-array with the durability and efficiency of on-premises cloud object storage.

To learn more visit http://www.weka.io/ and http://www.hgst.com/products/systems

WEKA.IO

2001 Gateway Place, Suite 400W, San Jose, CA 95110 USA   T 408.335.0085   E info@weka.io   www.weka.io