Western Digital.

SOLUTION BRIEF

Remove the Storage Bottlenecks in Genomics Analysis

Highlights

- **Higher storage performance** for diverse genomics analysis pipeline using a parallel distributed file system
- **Increased data accessibility and flexibility** by aggregating disparate storage into a single global namespace
- **Easily integrates** into existing environments with standard file and block support
- **Cost effective** options that use off-the-shelf-hardware
- **Storage optimization** by moving infrequently accessed data to object storage
- **Extreme data durability and integrity** at petabyte scale helps to ensure valuable data is protected long-term
- **Improve IT agility** by quickly provisioning and redeploying storage resources to support new business needs

Challenge

- **Storage bottlenecks slow analysis pipeline and results**
- **Costly to scale performance storage systems** to keep up with data growth and increasing retention times
- **Need higher performance** for faster turnaround time for analysis
- **High management overhead** to enable collaboration and sharing of data among global, distributed teams
- **Cost effectively keep data protected and accessible** over increasingly longer retention periods

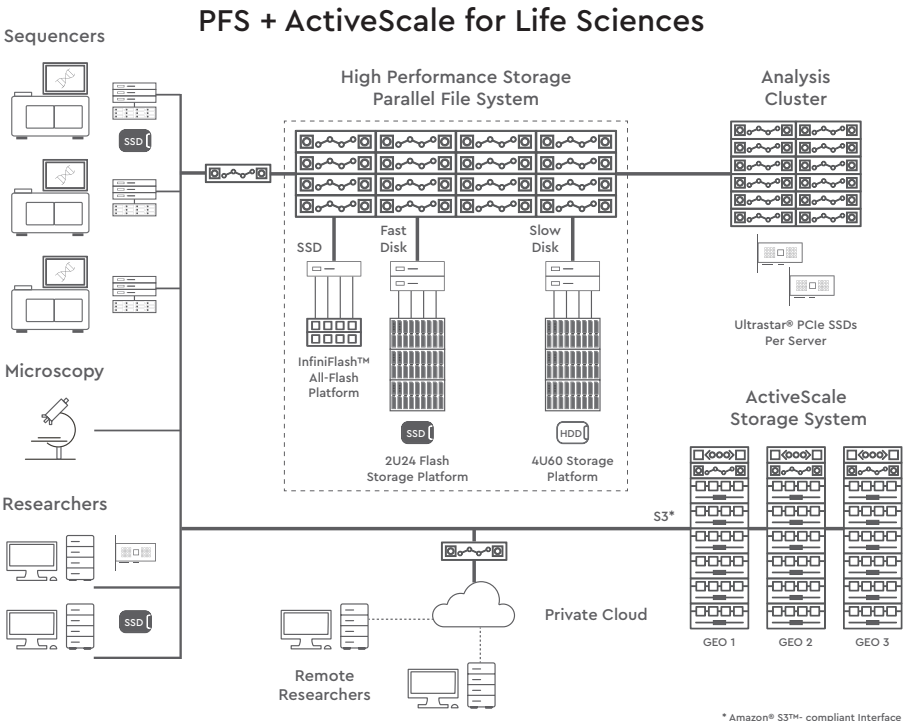
Solution

Scalable high performance storage using a parallel distributed file system and ActiveScale™ object storage system. The combined solution provides extreme performance, durability and scalability to meet the diverse storage workload requirements of genomic analysis.

Rapid advances in life sciences R&D are making it possible to deliver medical diagnoses and treatment based on a person's genetic makeup. Delivering better, faster, and more cost-effective healthcare means life sciences organizations must accelerate time to results. Their ability to store, process, and analyze high volumes of data with greater efficiency is essential.

Next-generation genome sequencers are producing more data than ever. Each run can can produce terabytes of raw data that is typically retained for long periods of time. The raw data is processed through a deep analysis pipeline that can lead to personalized treatment.

New analysis techniques with a greater emphasis on collaboration, along with explosive data growth, are causing organizations to consider new approaches for addressing their compute and storage infrastructure.



Accelerating the Genomics Analysis Pipeline

Storage and data management bottlenecks are perhaps the most challenging issues in life sciences (LS) to enable precision medicine. Reducing analysis and turnaround times are essential to deliver timely personalized treatment.

Deep genomics analysis requires a full pipeline that integrates many data processing and analysis tools. This pipeline typically includes many steps, from initial alignment to data cleaning, to variant calling and geno- typing, and further. Filling the pipeline is an enormous volume and variety of data generated by lab instruments. The sheer breadth of analysis applications required have highly varied compute, memory and I/O requirements. Data access patterns can range many jobs executing concurrently that need simultaneous access to the same small data files, to applications used to process large, 4D super high-resolution bioimaging files.

The diversity of storage workloads in this pipeline bottlenecks traditional storage file systems. High performance computing (HPC) environments have been dealing with similar challenges for years. and have had success deploying storage using parallel distributed file systems. Higher performance is achieved by spreading blocks of data from individual files across many file system storage nodes and reading/writing them in parallel. There are several commercial and open source parallel file systems to choose from. Application performance can be further enhanced using Western Digital PCIe flash devices in the analysis cluster and workstations as shown in the diagram above. Use includes overflow for data sets larger than physical memory, a bigger scratch space, and check pointing.

Cost effective Preservation of Genomic Data

An often-overlooked strategy is storage optimization, with the objective of reducing cost and improving utilization of the most important and expensive storage tier. Moving data that does not need high performance onto a lower cost, capacity-optimized tier like the ActiveScale object storage systems, helps improve primary storage performance and reduce costs. ActiveScale can also serve as a private cloud for sharing data internally and with other organizations around the world. Finding data is easy within a single global namespace vs. the hierarchical limitation of traditional storage architectures.

High performance storage using a parallel distributed file system with ActiveScale object storage, allows life science organizations to gain control of data growth and analysis pipeline with:

Unified storage that supports a diverse set of life sciences applications and workloads where performance, reliability, and availability of data are essential to the business. Native protocol support for NFS, SMB, Object enables seamless integration into existing environments.

Improve storage efficiency by pooling redundant isolated storage resources under a single global namespace. Free up performance storage tiers by transparently moving infrequently accessed data to ActiveScale using automated lifecycle policies.

Accelerate workflow and collaboration to translate genomic research into insights that contribute to better patient care. ActiveScale offers fast access vs. tape and with lower management complexity making it easier for teams to collaborate boosting overall productivity.

Extreme data durability and integrity at petabyte scale helps ensure valuable data is protected and always available. ActiveScale delivers up to 19 nines durability and site-level fault tolerance in a multi-site configuration. Robust data integrity checks occur automatically and transparently protecting long-term archives; each object can tolerate up to 1000 bit-errors without data loss.

Easy to Install, and manage by virtue of being software defined storage. ActiveScale systems are easy to deploy – simply add power and network connections and it is ready to go. The system self-protects and heals requiring significantly less IT intervention compared to traditional storage systems.

Conclusion

The scale and diversity of the data management and storage workload challenges faced by life science organizations are daunting. Sequencers and super high resolution bio-imagery create extreme amounts of data that needs to be quickly stored, processed, analyzed and kept for long periods of time. Storage solutions using high performance parallel file systems combined with ActiveScale deliver the necessary performance, scale, and efficiency to address these challenges vs. traditional approaches.

To learn more about ActiveScale visit:
www.wdc.com/dc-systems

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
US (Toll-Free): 800.801.4618
International: 408.717.6000

www.westerndigital.com