DELL Technologies

Solution Brief

DELL EMC PowerScale & Genomic Data Compression Solutions

Dell EMC and PetaGene deliver ultra-high capacity genomic data storage and accelerated data transfers.

PetaGene
transparent lossless compression

GENOMIC DATA COMPRESSION

- **Optimized for Genomic Data:** Traditional deduplication and compression strategies struggle with genomic data. PetaGene compression delivers 60-90% reduction.
- **Verifiably Lossless Compression:** each compressed file is validated for match to the original data by an MD5 or SHA256 hash.
- **Transparent Readback:** PetaGene genomic data compression is unique in that the compressed files appear and act as the original data files, only smaller. No new file formats or interfaces to integrate!
- **Faster Data:** PetaGene compressed files are smaller, facilitating faster data transfers and faster processing of genomic data. With the efficient fetch and streaming capabilities of PetaLink, processing PetaGene compressed data is 2-3x faster than native genomic data.
- **High Performance Software:** both compression and readback software are highly optimized for speed and compression with a low memory footprint.
- **Direct Cloud Access** with Cloud Edition, users can stream directly to or from public cloud storage or Dell EMC ECS Object Storage systems.
- **Guaranteed Savings:** Using PetaGene genomic data compression, users save at least 50% on the cost of maintaining genomic data assets.

Life Sciences Organizations are Data First

Hospitals, research facilities, and pharmaceutical companies have long understood the data first mindset: the idea that genomic data is crucial to your operations and the advancements of science and healthcare. Applications of genomic data for research, drug development, and precision medicine are expanding. As a result, the demand for genomic data is growing and with that, the need for data storage solutions to manage the full lifecycle of genomic data. The Dell EMC PowerScale is a scale-out, network attached storage systems that provide Life Sciences organizations with a simple and proven storage solution responsive to dynamic scientific and business requirements while minimizing the need to re-architect existing infrastructure. Combined with PetaGene genomic data compression and encryption, PowerScale solutions provide ultra-high storage capacity and processing performance for all scales of genomic data.

Designed for Next-Generation Sequencing (NGS)

Data generation, analysis, and long-term retention requirements represent the data life-cycle for NGS data. Each phase of the data life-cycle has its own unique set of performance and capacity storage requirements. No matter what the storage requirements are, Dell EMC Unstructured Data Solutions are proven to address the NGS data life-cycle end-to-end.

The Dell EMC PowerScale Storage Portfolio offers scale-out NAS solutions that provide a highly available and reliable, file system, PowerScale OneFS, for NGS workflows. OneFS is a multi-protocol file system which can receive and send data over multiple protocols like SMB, NFS, HDF, and S3 for a wide variety of genomics workflows, users and analysis environments. A single storage cluster can host multiple node types including Flash, Hybrid, Archive and PowerScale nodes. The All-Flash and Hybrid nodes are ideal for performance workflows like mapping and alignment of NGS data while the Archive nodes provide high- density, low cost, long-term storage for raw data and analysis results. Isilon nodes come with 40 or 10 GbE network connectivity to accommodate instrument traffic, HPC analysis, and data transfer to other storage end-points. Moreover, Dell EMC PowerScale SmartConnect efficiently balances client connections between an HPC cluster, users and instrumentation.

PetaGene genomic data compression is a software toolkit that works seamlessly with Dell EMC solutions. PetaGene's lossless compression algorithms are highly optimized for NGS data delivering 60-90% reduction of file sizes without any loss of information or performance.

Reliable

Many Life Sciences IT organizations operate with an aim that sequencing should never stop. PowerScale delivers the highest level of data protection. In the event of system failure, OneFS is designed with capabilities to minimize system downtime. OneFS supports up to four simultaneous device failures (N+4) without compromising data reliability and availability. Protection levels can be modified on the fly or set by policy to match the value of the data. Dell EMC PowerScale SnapshotIQ protects against application or user errors while Dell EMC PowerScale SyncIQ provides high-performance data replication between an on-prem and remote cluster. Organizations can also take advantage data management tools like DataIQ to move data to object storage like Dell EMC ECS or public cloud vendors with Dell Technologies Cloud Storage for Multi-Cloud.

Easy to Manage

Responding to the dynamic workflow requirements throughout the NGS data life-cycle can be challenging and time consuming. The PowerScale Storage Portfolio systems offer one easy-to-manage storage system. The OneFS file system offers a comprehensive set of command line and GUI based system administration tools that simplify data management and free IT staff to focus on higher priority projects. For example, IT managers can set up Dell EMC PowerScale SmartQuotas for users to actively manage their data growth and storage consumption. If additional storage is needed, a node can be added to a production cluster in 60 seconds without any downtime.

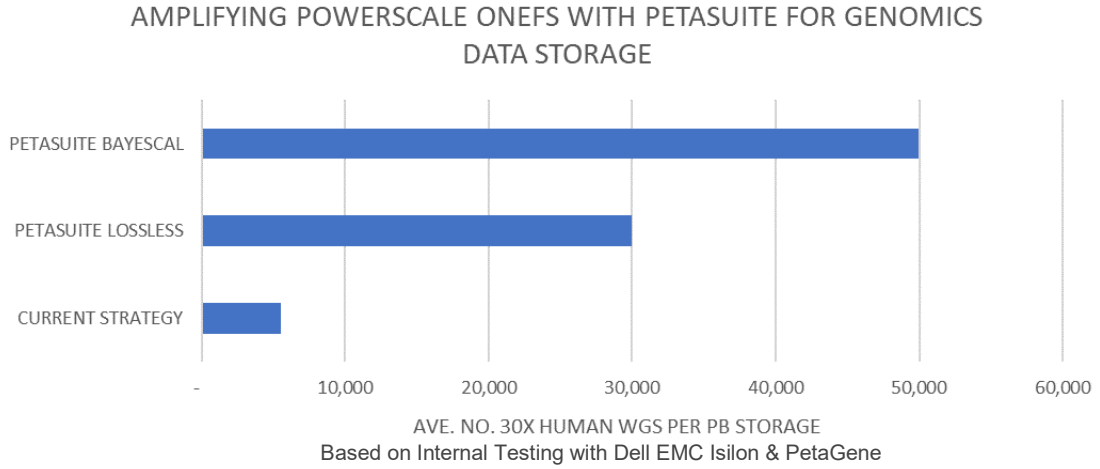
Smart Allocation of Resources Throughout the NGS Lifecycle

As the output of sequencing machines continues to grow beyond terabases/run, the capacity and throughput of storage systems needs to keep up with demand. Source data is usually captured to high performance storage systems for primary processing and secondary analysis. Derivative data and results may be moved to lower-cost analysis or archive tiers, depending on institutional data governance and policies. Dell EMC PowerScale SmartPools provide policy-based movement of data through tiers within a PowerScale or Isilon cluster. Institutions can setup rules where data can remain on higher performance F and H series nodes during the analysis phase, then automatically tier data and results down to the Archive tier (A), post analysis. Infrequently accessed data can be tiered further to public clouds with PowerScale for Multi-Cloud or object storage systems like Dell EMC ECS using tools like DataIQ and Dell EMC PowerScale CloudPools.

PetaGene compression amplifies the capacity of all storage tiers for genomic data, avoiding pipeline bottlenecks due to storage bandwidth. When it is time to move data, PetaGene genomic data compression and streaming is a good option to reduce the latency and speed up transfers. PetaLink improves readback and processing of NGS data by 2-3x while improving data movement speeds up to 10x. Unlike other compression options, PetaGene improves storage and data performance.

Devoted to Enabling Life Sciences Workflows

As a leader and trusted partner in life sciences workflows, Isilon is used by more than 350+ life science organizations worldwide, including leading genome centers, pharmaceutical companies, and academic research centers. Dell EMC also participates in industry organizations such as the Global Alliance for Genomics and Healthcare (GA4GH) and the IRODS Consortium with an aim eliminate the complexities of storage so life scientists can focus on research.



Based on Internal Testing with Dell EMC Isilon & PetaGene