Genomics at scale

Sequence genomes faster with Lenovo Distributed Storage Solution for IBM Spectrum Scale (DSS-G)

Summary

One of the greatest challenges in genomics today is scale, especially now that an entire human genome can be sequenced in less than a day at a cost under $1,000. Companies in the healthcare industry require genomics storage solutions that can accommodate such exponential data growth. Lenovo Distributed Storage Solution for IBM Spectrum Scale™ (DSS-G) is the leading implementation of one of the most proven and robust software-defined storage architectures available in the industry. In fact, Lenovo DSS-G won the HPC Wire 2017 Readers Choice Award for Best HPC Storage Product or Technology.

DSS-G offers a modular, building-block approach to IBM Spectrum Scale with performance and capacity that can scale near linearly. Lenovo DSS-G can be tuned specifically for genomics, and enables customers to scale their storage on the fly as demands change. Even with storage hardware located at multiple sites, DSS-G still maintains a global namespace that enables users to view everything as a single storage cluster. This capability vastly reduces genomics data silos and enables users to increase their genomics data flow and data sharing – all while improving security and management efficiency.

In 2003, just one human genome had been sequenced completely. Today, 15 years later, there are more than 500,000 human genomes. By some estimates, by 2025 the number of human genomes sequenced is expected to reach up to 2 billion. That amount alone represents up to 40 exabytes of data.[1] However, this estimate doesn't even account for all the animal, plant, bacterial, viral, cancer, archaeal, and other eukaryotic genomes that will also be sequenced by that time. Such an explosion of genomics data requires widescale genomic data storage to match it.

With DSS-G Lenovo has developed a scalable storage and data management solution, used by some of the world's largest companies and many of the world's fastest supercomputers. The underlying file system, created by IBM, is called Spectrum Scale (formerly known as GPFS).

Spectrum Scale provides a shared file system across a global namespace for simultaneous file access from multiple nodes, high recoverability and data availability through replication, and simplified administration even in large environments.

Lenovo DSS-G also features Spectrum Scale RAID, providing native de-clustered RAID for maximum reliability and performance, end-to-end data integrity, as well as performance neutral drive-rebuilds.

Recently, Spectrum Scale development optimized key tuning parameters in the Spectrum Scale file system specifically for genomics data pipelines.[2] However, a comprehensive, integrated solution implementation of this bold vision was still lacking, especially for the industry-leading Intel® x86 architecture.

1. Erika Check Hayden, "Genome researchers raise alarm over big data" Nature (July 7, 2015)

2. Joanna Wong et al., "IBM Spectrum Scale Best Practices for Genomics Medicine Workloads" IBM Red Book (April 2018)

The DSS-G vision for genomics storage

Lenovo, with its own widely adopted line of servers, storage and networking, has created a fully integrated solution aligned with the Spectrum Scale software vision optimal for genomics storage.

Lenovo Distributed Storage Solution for Spectrum Scale (DSS-G) leverages the best in class software and hardware components to provide a complete genomics storage solution. Moreover, DSS-G ensures large, complex data sets and multiple applications run at a speed that accelerates workflows and reduces bottlenecks.

DSS-G enables customers to manage the exponential rate of data growth and the subsequent need to store large amounts of both structured and unstructured data. With Lenovo DSS-G, genomics researchers and organizations can realize a host of items on their data storage wish lists. Gone are the days of intractably large files that proliferate at unmanageable speeds, many containing unstructured data. Users are freed from legacy silos of storage and are free to innovate.

Lenovo DSS-G is a single, integrated solution built using Intel® Xeon® Scalable family based Lenovo ThinkSystem servers, Lenovo Storage and IBM Spectrum Scale software. DSS-G allows genomics organizations to start small and build via incremental 'building block' additions, providing expanded capacity and bandwidth with each additional DSS-G solution.

DSS-G fully implements the Spectrum Scale specifications tuned for genomics data. The file system is based on the tuning recommendations from the Broad Institute Genome Analysis Toolkit (GATK) guidelines. These specifications and guidelines were developed around the assumption of rapid data ingestion from genome sequencers and cryo-electron microscopes. The GATK guidelines allow for the sharing and access of genomics data by technicians and physicians across sites and institutions.

Specifically, DSS-G is tuned to enable:

Acceleration of the genomics pipeline via Spectrum Scale for genomics data tuning guidelines.

Secure high-speed data access for analysis on the compute cluster.

External Data transfer via NFS and SMB to access data from sequencers, microscopes, and other equipment for broad-scale sharing across research sites and facilities.

End-to-end checksum ensuring data integrity from compute to storage cluster.

Scale-out architecture capable of genomics data storage many terabytes to hundreds of petabytes.

Data management GUI for configuring and maintaining storage resources.

Lenovo DSS-G is fulfilled by Lenovo Scalable Infrastructure (LeSI), which leverages decades of engineering experience and leadership to reduce complexity of deployment, and deliver an integrated, fully-supported solution that matches best-in-industry components with an optimized solution design. This enables maximum system availability and rapid root-cause problem detection throughout the life of the system.

DSS-G utilizes Lenovo Storage D3284 high-density enclosures with 84x large capacity NL SAS drives (3.5-inch form factor) or Lenovo Storage D1224 enclosures with 24x high performance SAS or Solid-State Drives (2.5-inch form factor). These storage enclosures and drive options allow for a wide choice of technology for a use-case tailored design within an integrated, delivered solution.

DSS-G is also compatible with all the leading HPC compute fabrics, including the latest Intel® Omni-Path® Architecture solutions. The coming “exascale” genomics data influx demands a software-defined storage solution custom-tuned for its demanding size and scale, and DSS-G marries a best-in-class software-defined storage framework with a trusted and leading-edge storage hardware system. Unlike competing solutions, configured generically to host data of all kinds, DSS-G can be tuned specifically to the Broad Institute’s GATK guidelines for optimal genomics data storage and throughput.

Genomics is a rapidly changing field that in the years ahead will be built upon an unprecedented magnitude of files, containing unprecedented quantities of unstructured data. Your data storage solution deserves to be ready to handle it, ensuring your researchers are free to perform world-class science powered by world-class data.

Lenovo DSS-G storage system features

DSS product overview

Distributed Storage Solution configurations for IBM Spectrum Scale

DSS-G100	DSS-G201	DSS-G202	DSS-G204	DSS-G206	DSS-G210	DSS-G220	DSS-G240	DSS-G260
NVMe	SSD	SSD or SAS HDD		SAS HDD	Low cost entry	Low cost entry	Performance optimized	
NVMe Support No JBODs Needed				SR650				SR650
				SR650				SR650
			SR650	D1224			SR650	D3284
			SR650	D1224			SR650	D3284
		SR650	D1224	D1224		SR650	D3284	D3284
	SR650	SR650	D1224	D1224	SR650	SR650	D3284	D3284
	SR650	D1224	D1224	D1224	SR650	D3284	D3284	D3284
SR650	D1224	D1224	D1224	D1224	D3284	D3284	D3284	D3284

- Scalable building block approach to High Performance and Technical Computing storage.
- Industry leading IBM Spectrum Scale and Spectrum Scale RAID software defined storage.
- Best-in-class data integrity, reliability with tuned performance for genomics workloads.
- Solution support provided by Lenovo's world-class, industry recognized services organization.
- Lenovo ThinkSystem SR650 with powerful Intel® Xeon® processor Scalable family CPUs.
- Lenovo D3284 high-density storage enclosures with 84x high capacity NL SAS drives (3.5-inch form factor).
- Lenovo D1224 storage enclosures with 24x high performance SAS drives or SSDs (3.5-inch form factor).
- Choice of 10GbE/25GbE/40GbE/100GbE, FDR/EDR InfiniBand, Intel® Omni-Path® Architecture.

To discover how Lenovo DSS-G can drive your sequencing and genomics analysis storage needs, contact your trusted Lenovo solutions partner today.

Contact your Lenovo representative