STELLUS

SOLUTION BRIEF: GENOMICS
Unlock the Targeted Therapies of the Future

Stellus Data Platform Supercharges the Data Pipeline for Genomic Analysis

In the race to develop tomorrow's drug therapies, the world's most advanced research organizations face a number of challenges in incorporating genomic data in the overall R&D process. It's not that they lack the desire or even the tools. In fact, the cost to sequence a genome has dropped from about $10 million in 2007 to less than $1,000 today. At this price, we should be seeing data from millions of samples used to develop more targeted and effective pharmaceuticals—and we probably would if the systems analyzing all that data could keep up. Today it takes about 20 hours to run a single genome through a typical analysis pipeline. Even when running multiple workloads concurrently, 24x7 on HPC clusters, laboratories routinely face backlogs of weeks, even months.

Now, Stellus is rewriting the rules for life sciences computing. The Stellus Data Platform (SDP) is built from the ground up for storing and processing unstructured data. It enables organizations to dramatically accelerate application performance for genomic analysis. Researchers can now process more workloads in far less time and take concrete steps to enable personalized medicine therapies for the future.

Breaking the Performance Bottleneck

Modern life sciences environments contend with a veritable firehose of genomic data, typically acquiring 2-4 terabytes every day, all of which must be processed through a multi-step analysis pipeline. To analyze that data as quickly as possible researchers rely on a variety of acceleration techniques. Specialized processors (GPUs, TPUs, FPGAs), smarter networks and SSDs, software-based parallelization (from tools like GATK, ADAM, and others) can all boost performance—at least, in theory. In practice, since so many stages of the pipeline are read/write-intensive, these accelerators end up shifting the bottleneck to the storage systems feeding data to the compute complex. Effectively, they change genomics workloads from being CPU-bound to being I/O-bound.

The barrier to faster performance is antiquated hardware and legacy file systems. Modern components like NVMe interconnects and NAND flash media are capable of order-of-magnitude I/O improvements. The decades-old software and file systems run by even the newest storage systems still cannot exploit the most modern storage media components. These aging architectures waste significant resources on processes like the following:

- Converting data between file and block I/O, which gets more resource-intensive as data grows
- Maintaining global data maps at scale as the number of files grows exponentially
- Ensuring global cache coherence across multiple nodes in a large cluster

These processes were useful when storage primarily meant working with HDDs and block I/O on structured data. But those approaches quickly become significant challenges when storing today's vast amounts of unstructured genomics data on native flash storage systems, diverting resources that could be used to service I/O requests.

The New Standard in Life Sciences Performance: Stellus Data Platform

Stellus created the SDP to address the problems of legacy storage architectures. With a new software stack built from the ground up to exploit the latest Compute, network and memory infrastructures, the platform sets a new standard in I/O performance for life sciences applications and unstructured Genomics data.

The Stellus Data Platform replaces block stores, data maps, and data caches with high-performance Key-Value Stores, Key-Value-over NVMe, and algorithmic data placement. At the same time, it provides file access across standard protocols like NFS and SMB, as well as newer object storage access methods like S3—all in a scalable, enterprise-ready storage system.

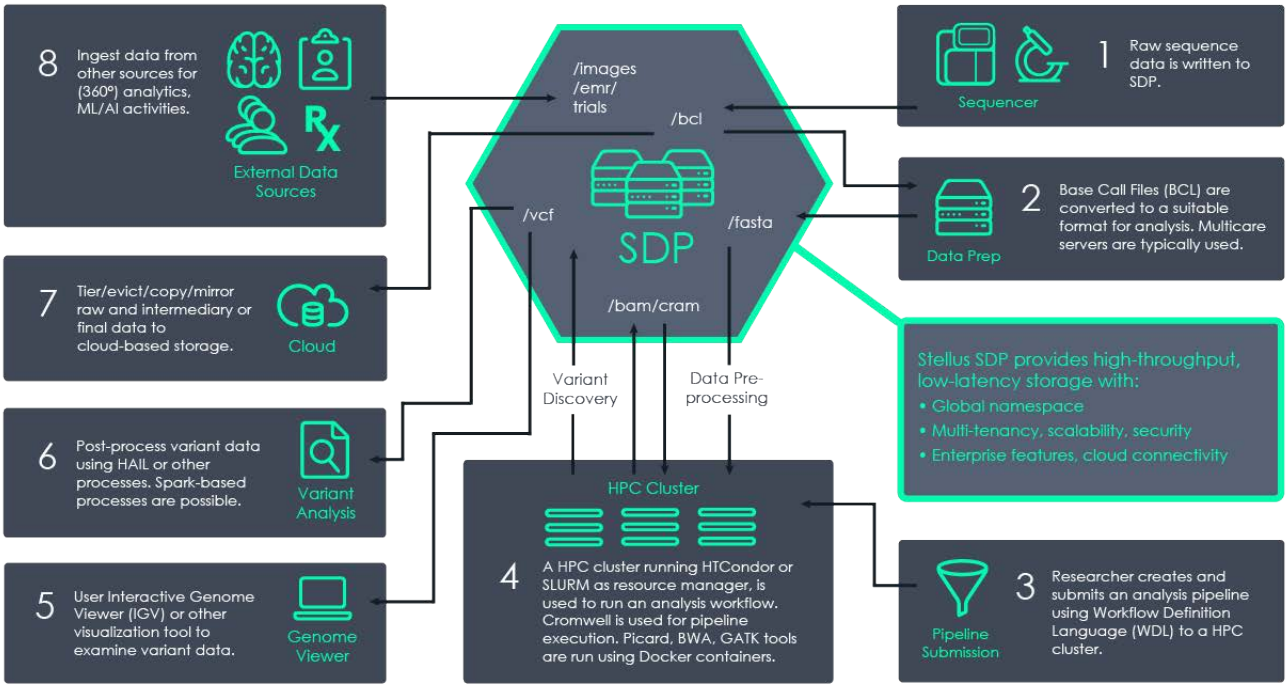
The Stellus Data Platform delivers these key benefits:

- **Composable platform flexibility**—In most life sciences organizations, HPC/Multi-GPU clusters are a shared resource supporting a range of applications, with requirements that change all the time. Today, if organizations want to add more capacity or performance, they have to buy a new node—paying for more compute, network, and storage, even if they need to increase only one of those dimensions. With the SDP, organizations can add performance (throughput) and capacity independently to cost-effectively scale as requirements evolve. Increase throughput by adding new Data Manager (DMs). Add capacity by adding drives to the Key-Value Store (KVS) layer. As labs scale up to processing hundreds of genomes per day, that flexibility will be essential to keeping costs predictable and compute investments aligned with actual needs.
- **Software-defined storage**—To keep budgets under control, organizations need to get maximum use out of their IT resources. That means augmenting rather than replacing storage hardware whenever possible. The smartest way to do that is with software-defined storage. Most system intelligence resides in software, which can be changed with relative ease, rather than locked within the hardware. Unfortunately, most storage vendors still rely on legacy hardware-based models to deliver high performance, often requiring forklift overhauls to take advantage of new capabilities. The Stellus Data Platform is a software-based file system that is able to deliver perfroamnce in Cloud, Core & Edge environments.
- **User-mode file system**—Life sciences organizations have many more options today to achieve high-performance storage. To deliver it though, most solutions require custom client software, specialized controllers, or Linux kernel customization. Those strategies can work on smaller scales, but in large HPC environments with hundreds of machines and thousands of cores, they just can't scale and quickly become a nightmare to maintain. The Stellus Data Platform runs as strictly user-mode software on top of standard Linux—no kernel hacks, special client software, or custom controllers required.

Changing the Game for Genomic Data Analysis

New processors and software-based parallelization, combined with faster and less expensive networking, should be unleashing a new world of genomics insights for drug discovery and development. Now, Stellus is helping research organizations make this vision a reality.

At multiple stages in the genomic data analysis pipeline the SDP can accelerate data storage transformation and analytics. SDP helps organizations to continually speed data through the pipeline to break through the backlog and process more genomes in less time.



Stellus Data Platform for the Life Sciences Workflow

With these capabilities, researchers can reduce the costs and time-to-market for new drug discovery. Combined with state-of-the-art genomics analysis and new AI/ML applications—all of which can also benefit from Data Platform I/O improvements—researchers can usher in the age of personalized medicine. They can help build a future where clinicians draw on fine-grained genomic and imaging data to target the right patient, with the right therapy, in the right dosage, at the right time.

Stellus Technologies
3833 North First Street
San Jose, CA 95134

www.stellus.com