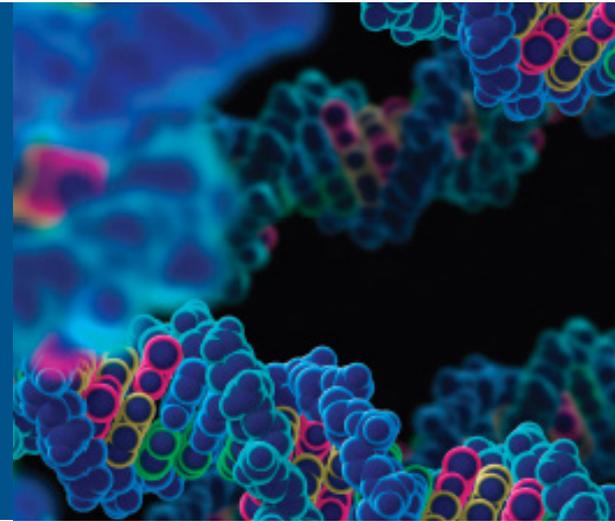


# CRAY HPC STORAGE SOLUTION FOR LIFE SCIENCES



High-performance computing (HPC) is at the center of modern life sciences research. From genomic sequencing to large-scale genomic analysis to precision medicine, HPC plays a critical role in advancing many life sciences disciplines.

The success of next-generation sequencing has had a profound impact on biomedical and biological research. The ability to produce whole genomes in a short period of time has made genomic sequencers more affordable. According to data from the National Institutes of Health (NIH), the cost of generating a high-quality whole human genome sequence in 2015 fell below \$1,500 and will only continue to drop. The reduction in cost for genomic sequencing has allowed many labs and healthcare providers to conduct sophisticated analyses on genomes, transcriptomes and interactomes, the results of which have enabled many advanced initiatives such as the development of precision medicine and disease prevention for targeted regions and populations.

For example, pharmacogenomics, part of precision medicine research, uses both pharmacology and genomics to develop effective, safe medications and doses tailored to variations in a person's genes. As modern biology and instrumentations go digital, new tools that can support more sophisticated analysis and workflows are in greater demand.

## KEY BENEFITS

### **FASTER TIME TO DISCOVERY AND TREATMENT**

Enabling faster insights from genomic analysis and effective cures for disease with performance-optimized parallel storage solutions

### **HIGHER OPERATIONAL EFFICIENCY**

Optimal storage efficiency and mixed I/O support lead to less management complexity and higher HPC productivity

### **LOWER TCO**

More performance delivered with smaller hardware deployment reduces TCO

## Problem

While the success of next-generation sequencing has enabled many whole genomes to be sequenced in a short period of time, this new method has also generated huge amounts of data in diverse formats. To put things in perspective, a whole human genome has the size of roughly 6 billion letters. A genome sequence (a string of those letters) is about 700 megabytes. A simpler version of the sequence would be approximately 125 megabytes of data.

As the cost of sequencing continues to drop, bigger studies are being rolled out — consider the data and storage requirement for the United Kingdom's 100,000 Genomes Project. And there are even bigger projects underway. In addition to the vast amounts of genomic data, more affordable genetic testing is now combined with real-time imaging and other digital instruments to carry out more comprehensive diagnosis, further flooding the genomic dataset. To further complicate things, genomic analysis workflows have varying characteristics — some can be I/O intensive, some CPU bound.

Properly storing, processing and analyzing this volume of data while delivering optimal application performance regardless of its I/O characteristics is a challenge. HPC users need a storage solution that provides sufficient capacity and faster performance, offers flexibility to support mixed workload requirements, runs efficiently, and is easy to manage.

## Solution

Cray® ClusterStor™ storage systems are designed to provide optimal performance, efficiency and scalability to tackle the life sciences' storage I/O challenges faced by the HPC users in life sciences. ClusterStor storage is purpose-engineered, performance-optimized HPC storage solution designed, produced and supported by Cray's holistic approach to system design. These systems enable up to 80 percent of the raw performance from the underlying media, compared to other legacy storage solutions which struggle to obtain more than 50 percent. Leveraging and maximizing the efficiency of the underlying media ultimately optimizes performance, management and cost.

ClusterStor storage systems have enabled life sciences organizations across the globe to gain faster insights from the mountains of genomic and other large-scale datasets. For example, one of the leading cancer research centers in Europe uses ClusterStor storage to accelerate new computational models for breast cancer genomic research. The research revealed new genes and mutations involved in the disease, delivered new insights into the cause of the breast tumors and demonstrated that breast cancer genomes are highly individual.

To provide a complete, end-to-end solution for life sciences organizations, Cray's ClusterStor portfolio includes a rich set of products that solves life sciences' biggest HPC storage challenges.

**I/O Bottlenecks:** Many life sciences codes are I/O bound. Given the size and number of the intermediate files involved in the modern genomic analysis workflow, the requirement for I/O performance becomes more demanding. Traditional storage solutions such as NAS and NFS-based file systems lack the scalability and performance needed to solve the I/O bottleneck. High-performance parallel file systems such as Lustre® are best suited to deliver optimal I/O results.

**Mixed Workloads:** Life sciences workloads are often mixed. Small random file reads and writes as well as large sequential I/O jobs will be in the same workflow, making performance tuning for the overall workflow complex. The ClusterStor L300N solution accelerates mixed I/Os on the same platform by monitoring and directing small, random workloads to flash and large, sequential I/O jobs to the disk storage. Unlike the traditional burst buffer approach, which requires either a separate file system tier or proprietary client software, the ClusterStor L300N solution manages different storage tiers transparently within the same file system.

**Faster Time to Insight and Treatment:** The tightly integrated, performance-optimized design of ClusterStor storage solutions accelerates both traditional HPC and a new generation of data-intensive life sciences workflows, which enables faster genomic discovery and more effective

treatment for disease. Cray is the only storage vendor who can deliver an end-to-end solution that is purpose engineered, rigorously tested, and performance optimized to meet the most demanding storage needs for the life sciences.

**Higher Operational Efficiency:** ClusterStor storage systems maximize the performance efficiency offering organizations better performance with smaller hardware deployments and fewer management headaches. Life sciences users can fully exploit a hybrid storage environment by supporting mixed I/O workloads on the same platform.

**Reduced TCO:** With superior efficiency and a smaller hardware footprint come less management complexity, and reduced power, cooling and floor space costs. ClusterStor systems help life sciences HPC users achieve reduced total cost of ownership including both capital and operating expenditures without compromising the overall performance.

### CRAY CLUSTERSTOR AT A GLANCE

ClusterStor storage systems are built on Cray's robust enterprise-class storage designs from the device level up. Powered by industry-leading parallel file systems including Lustre, ClusterStor storage systems help customers accelerate their most demanding data intensive workloads at scale.

Cray Inc.  
 901 Fifth Avenue, Suite 1000  
 Seattle, WA 98164  
 Tel: 206.701.2000  
 Fax: 206.701.2500  
**WWW.CRAY.COM**

