SGI® StorHouse™ for Life Sciences at the National Center for Biotechnology Information

Helping the National Institutes of Health Prevent, Control, and Manage Disease

Key Facts

Organization:
National Center for Biotechnology Information

Location:
Bethesda, MD

Application:
Biomedical Research

The information age is an exciting and challenging time to conduct biomedical research. The excitement stems from the potential to discover breakthrough knowledge for managing disease and promoting global wellness. One of the biggest challenges to unlocking this potential is finding a way to integrate the surge of new biomedical information with advanced computing technology. Transforming raw information into useable knowledge requires a highly accessible and reliable technology infrastructure to archive, access, analyze, manage, and link the overwhelming volume of molecular-level data currently being generated by diverse research laboratories worldwide. Without this infrastructure in place, fragmented data access could lead to incomplete or misleading results.

What is NCBI?

The National Center for Biotechnology Information (NCBI) embraces both the excitement and the challenges of the bio-information age. NCBI was founded in November of 1988 through legislation sponsored by the late Senator Claude Pepper. It is a division of the National Library of Medicine at the National Institute of Health (NIH), the largest biomedical research facility in the world.

The NCBI mission is to combat health problems and disease by developing new information technologies and infrastructures that help researchers study fundamental, molecular-level biomedical problems such as gene organization, sequence analysis, and structure prediction. NCBI responsibilities encompass:

- Coordinating national and international efforts to collect molecular biology, biochemistry, and genetics research information
- Designing automated systems for archiving and analyzing this data
- Providing access to NCBI repositories to a global medical research community

To achieve this challenging mission, NCBI develops sophisticated applications to create, maintain, and access many interconnected databases related to genomic research. Researchers use the NCBI Entrez system to search these repositories and retrieve related sequences, structures, references, and graphical views of chromosome maps.

The NCBI Sequence Read Archive Application

Since 2007, NCBI has also been responsible for what is considered to be the most exciting project in the bioinformatics world today – Sequence Read Archive (SRA) – an application that captures, stores, shares, and protects all genomic sequencing data developed through funding from the United States Government. SRA data is produced using massively parallel, next-generation sequencing technologies such as Illumina®, 454 (Roche), HeliScope, SOLiD™ (Applied Biosystems™), and Complete Genomics platforms, which can generate hundreds of gigabytes of SRA data in a single instrument run. And, because these next-generation technologies are constantly evolving to higher levels, more and larger public domain sequences are continually being generated and downloaded to the SRA. As a result, NCBI expects the SRA repository to grow from its current 2000 terabyte size to a daunting 12 petabytes within the next two years. Even more mind-boggling, NCBI expects the total SRA archive size to approximate 30 petabytes over time.

SRA is a key enabler that connects many bio-scientific projects. It is a public service endeavor where researchers can upload, download, and share important genomic information. Architecturally, SRA uses an innovative column-oriented database design with information organized in a hierarchy of studies, experiments, samples, and their corresponding runs. Studies may contain multiple experiments, which describe the sequence data and the sequence method. Each experiment consists of one or more instrument runs with the results, or reads, from each spot in the run.

It is easy to access, download, and upload SRA data. From the SRA Web-based interface, researchers can access sequence data from various experiments, search for specific records, or browse all records by study, sample, or experiment. In addition, they can use the NCBI Entrez interface to perform keyword searches on SRA data, which can be viewed graphically and/or filtered and downloaded in FASTA or FASTQ formats. Using fasp™ protocol from Aspera, researchers can download data from the SRA homepage or from a link on study, sample, and experiment records. Data submissions occur through a web-based interface or an automated pipeline.

Prior to 2009, the SRA application stored sequence data on whichever RAID product was more economical to purchase at the time. However, due to anticipated archive growth, NCBI recognized an urgent need to upgrade the existing RAID infrastructure to a more cost-effective, full-featured, automated storage and data management architecture. With a 12-petabyte database looming on the horizon, NCBI began a search for new technology products that provided affordable scalability, sustained performance, automated system management, and a storage virtualization platform that supported both traditional and alternative media types such as tape, which is more economical for storing less frequently accessed large sequence records.

StorHouse for Life Sciences

Cambridge Computer Services, Inc., a national integrator with proven expertise in data storage and data protection solutions, was familiar with NCBI SRA requirements and a variety of technology products in the marketplace. Based on its industry experience and knowledge, the company recommended that NCBI consider StorHouse® software for Life Sciences to manage its growing SRA repository.

StorHouse for Life Sciences is a comprehensive and versatile storage virtualization and data management platform developed to specialize in large-scale data management and information governance solutions. By storing, protecting, backing up, and managing vital biomedical information from private, government, and university research institutes and laboratories, the product is a major enabling technology component in the race to collect, study, link, and analyze critical biomedical information. The goal of such research and analysis is to foster universal health and well-being by preventing and managing disease.

StorHouse software was designed for Life Sciences to administer the terabytes to petabytes of structured and unstructured fixed content generated by today's highly scalable and extremely dynamic biomedical and bioinformatics applications. Example data types include genetic sequencing, drug discovery, and clinical research information. The product combines industry-leading traditional and alternative storage devices and open system processors with storage management, relational database management, and file system interface software components. No application program interfaces are required.

Based on StorHouse Release 5.6 – a proven code base installed at prestigious customer locations worldwide – StorHouse for Life Sciences has many unique features that support the sustained performance, virtually unlimited scalability, storage virtualization, and centralized system administration requirements that life sciences applications demand. These distinguishing features include:

- High-performance direct reads from tape to accommodate very large files and random access to archived data stored on tape
- Throttling to enable continuous ingest at the maximum possible rates required for large volumes of genomic sequencing data
- Support for Linux® platforms, very large file systems, and tape libraries most commonly used in life science applications
- Scalability to trillions of files and multiple petabytes of managed data with no performance degradation
- Virtualized storage that cost-effectively uses traditional and alternative storage media to lower the total cost of data ownership and provide a measurable, cost-correct ROI
- Automated system, storage, and data management, including storage allocation and control as well as data migration, replication, backup, recovery, and retention
- Automatic content validation and repair processes to ensure data integrity and archive reliability
- Easy migration to newer and more advanced technologies as they becomes available with no performance degradation or system downtime to ensure future accessibility of current data
- User-friendly, browser-based management tools that lower administration overhead by automating many routine system management tasks

StorHouse for Life Sciences also includes StorHouse/RFS, the StorHouse file system interface, an industry-standard relational file system layer that archives, retrieves, and/or backs up any format of unstructured data in native file format

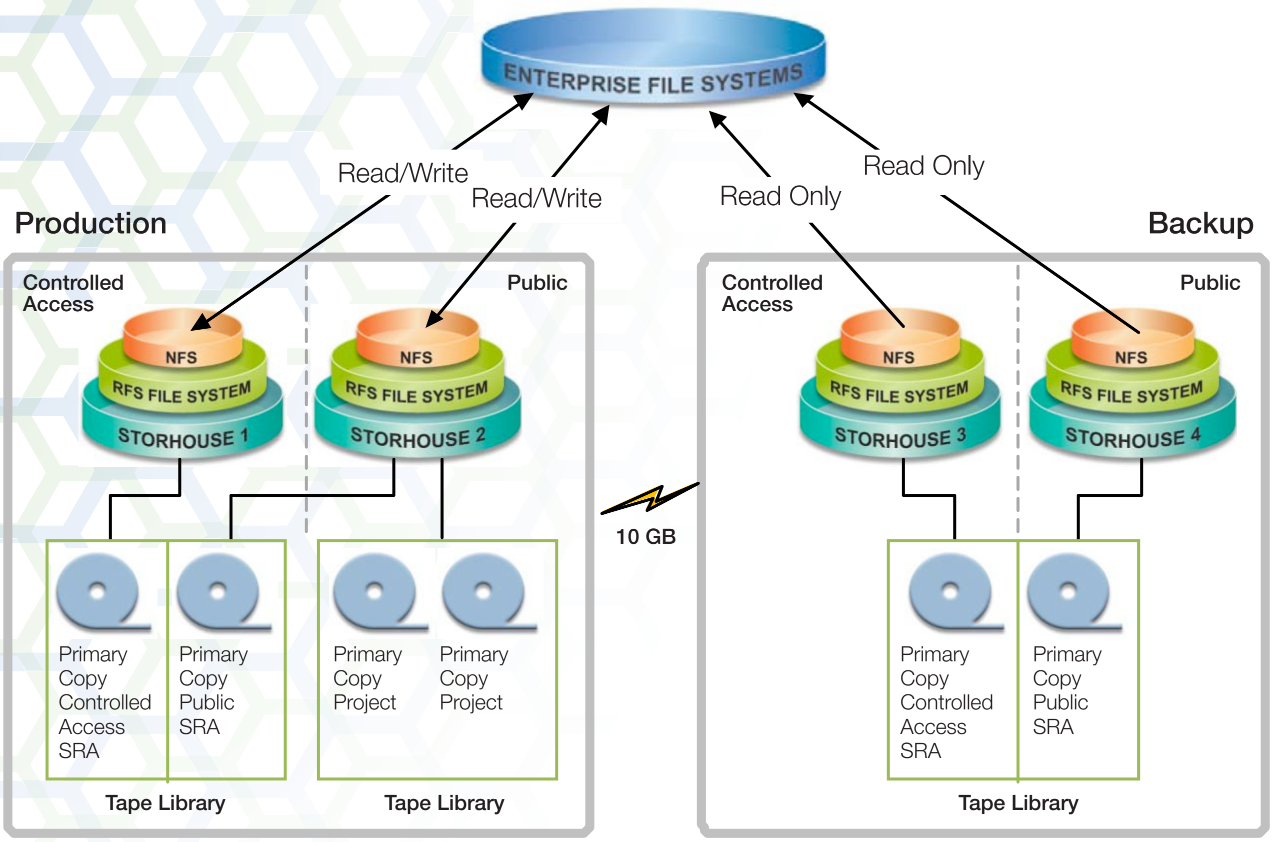

Figure 1: StorHouse Architecture at NCBI

with no need for application modifications. With StorHouse/RFS, StorHouse simply appears on a network as a single, unified file share. Users and applications can access the StorHouse file share through traditional drive letter mapping (V:\) or a server-oriented file path. Organizations currently use StorHouse and StorHouse/RFS to create enterprise archive solutions that can grow to trillions of files without impacting the integrity or stability of the archive.

Deploying StorHouse at NCBI

After careful consideration and thorough evaluation and testing, NCBI chose StorHouse for Life Sciences to manage its SRA data. With minimal input and assistance from the experienced StorHouse system engineers, NCBI deployed two production StorHouse systems and two backup/disaster recovery (DR) StorHouse systems at geographically dispersed locations in the Washington, D.C. metropolitan area. NCBI uses the duplex and replication features to write data between the production and DR sites through a 10 gigabyte Ethernet connection.

System Architecture

As Figure 1 illustrates, StorHouse for Life Sciences at NCBI uses a modular, building-block design. Each production and DR system represents a StorHouse instance. Each instance propagates the same comprehensive StorHouse features and benefits across a network of StorHouse deployments. The current two-year plan is to expand the existing StorHouse instances to dozens of new StorHouse sites to accommodate the anticipated 12-petabyte SRA size.

Each production system has a separate StorHouse/RFS file system interface and manages a separate tape library. StorHouse 1 archives and manages controlled access SRA data, which is available to a subset of scientists and researchers. StorHouse 2 archives public SRA data and SRA project data from the National Library of Medicine. The public SRA data and the SRA project data have no access restrictions.

The backup and recovery SRA site consists of two additional StorHouse instances, StorHouse 3 and StorHouse 4, which mirror StorHouse 1 and StorHouse 2. Similar to the two production systems, StorHouse 3 and StorHouse 4 have their own StorHouse/RFS file system interfaces and a public and controlled access side.

Sustaining Data Integrity

One feature that distinguishes StorHouse from other archive and backup systems is the ability to proactively monitor system operation, detect errors, and immediately notify customers and the company's support personnel of potential problems, often before they occur. The benefits are ensured data integrity and enhanced system availability and accessibility.

StorHouse error detection proved to be an essential service at NCBI when the software automatically detected a multiple-bit CRC error caused by silent corruption of a critical genomics data file. StorHouse automatically sent Call Home alerts to NCBI and to the company to report the problem. Recovery simply entailed deleting the corrupted primary file and automatically writing, or replicating, the backup copy at the NCBI DR site to the primary site. Because no code modifications or extra administrative cycles were required, StorHouse found, reported, and fixed the problem through replication with no interruption to normal system operation.

System Statistics

The StorHouse for Life Sciences platform at NCBI is a dynamic and efficient system that currently stores, manages, and protects 1.7 terabytes of genomic data and sustains a 3- terabyte-per-day load rate. In terms of subject matter, SRA stores information for over 1000 single organisms, with approximately half representing human data. In the coming months, NCBI plans to expand the SRA content by adding data for the alignment of reads to reference genomes. By deploying modular instances to propagate the infrastructure, StorHouse for Life Sciences can easily accommodate this and other future SRA growth.

Summary and Benefits

In 2007, NCBI deployed what is considered to be the most exciting application in the bioinformatics world – the SRA application for next- generation sequencing data. Originally, SRA used RAID systems as its storage infrastructure. However, with anticipated system growth expected to reach 30 petabytes over time, NCBI needed a more economical and full-featured storage and data management platform to ensure future SRA data integrity, accessibility, availability, and longevity. In 2009, after evaluating many products, NCBI chose StorHouse for Life Sciences – proven technology – to store, protect, and manage SRA data.

StorHouse for Life Sciences provides NCBI with many benefits, such as affordable scalability, sustained performance, automated system management, data reliability, and a cost-effective storage virtualization platform that includes disk and tape. Moreover, StorHouse facilitates SRA data loading and retrieval with:

- High-performance direct reads from tape to accommodate random access to very large files and direct reads of SRA data on tape
- Throttling to enable continuous SRA data record ingest at the maximum possible rates
- A comprehensive file system interface for storing and accessing SRA data
- The platforms and tape libraries used most frequently by life science applications
- Simple migration to new and improved storage devices, media types, and operating systems as they become available

Shepherding the SRA repository is an enormous responsibility considering NCBI will ultimately be accountable for multiple site locations, an unprecedented data volume, and all associated system, storage, and data management requirements. StorHouse for Life Sciences makes it easier for NCBI to accomplish these difficult tasks. Together, the SRA application and StorHouse advance access to critical genomic information, thereby unlocking the potential for scientists and medical researchers to transform raw data into breakthrough knowledge – knowledge that will one day lead to advancements in treating disease and improving global health.

Global Sales and Support: sgi.com/global