Curing Life Sciences Data Management Challenges with Scalable Storage

By Sal Salamone

Produced by Cambridge Healthtech Media Group Custom Publishing

Curing Life Sciences Data Management Challenges with Scalable Storage

INTRODUCTION

As the calendar turned to 2010, many reflected on the changes that occurred over the last 10 years. Looking back on the life sciences, the biggest event was the sequencing of the human genome, which caused a data explosion that continues today.

Specifically, the growing use of sequencing and medical imaging is producing a constant increase in the volume of life sciences data that must be used to develop new drugs. This growth is forcing most life sciences organizations to re-evaluate their storage solutions. What's needed is a highly scalable solution that delivers the various levels of data throughputs required in today's R&D analytic workflows. Such solutions must be affordable, easy to manage and offer a way to address long-term data storage issues such as data migration and protection against media errors in high-capacity drives.

This paper will look at the drivers forcing organizations to rethink their traditional storage solutions, the data storage and management challenges life sciences organizations face today, and what's on the horizon. The paper will also examine how cost-effective scalable storage solutions are becoming essential to meet these challenges and help enable the search for new pharmaceuticals.

EXECUTIVE SUMMARY

- **New lab equipment is producing orders of magnitudes more data**
- **Traditional storage systems cannot meet performance demands of today's R&D analytic workflows**
- **The life sciences data explosion is forcing organizations to re-evaluate their storage options**
- **New storage solutions must address performance, data management and TCO issues**

SEISMIC SHIFT UNDERWAY

The transformation of life sciences to a data-centric field can be attributed to the work done sequencing the human genome. In early 2001, when *Nature* and *Science* published details about the techniques used and the first analysis of the sequence, information technology's role in life sciences research rose to a level comparable to the underlying science and lab work itself.

That prominent role has simply continued to grow in importance as each new generation of lab equipment has increased the volume of data generated in each test or experiment.

For example, over the last few years, so called next-generation sequencing has emerged as a critical tool in investigating biological systems at the genetic level. According to a 2008 article[1] in *Nature*, next-generation sequencers are prized because they have the potential to "dramatically accelerate biological research by enabling the com-

prehensive analysis of genomes, transcriptomes and interactomes."

The notable difference with these new sequencers over previous generations is that they increased the volume of lab-generated data by orders of magnitude. In particular, next-gen sequencers from companies including 454 Life Sciences, a Roche Company, Applied Biosystems Inc., Illumina Inc., and others can generate terabytes of data a day and hundreds of terabytes per year in a lab. With many users deploying more than one sequencer, data growth can therefore quickly escalate to multiple petabytes.

As with many new technologies, the next-generation of lab instruments have driven down the cost of DNA sequencing (by some estimates over two orders of magnitude). This cost reduction made sequencing available to many more organizations and even individual researchers. So while sequencers were traditionally only used in major genomic research centers in the early to mid-2000s (and essentially treated like mainframe computers of old), over the last few years organizations of all sizes now have access to the technology.

Unfortunately, users of the new sequencers have to deal with the growing volumes of data the equipment generates.

Similar to sequencing, imaging's role in the life sciences has greatly expanded, thus adding to the data explosion.

In early stage drug discovery, there has been a significant increase in the use of microarrays for SNP detection and gene expression profiling. SNPs and the level of gene expression are detected by imaging the array using a laser or light source, capturing the image using photomultiplier tube detection or a charge-coupled device camera and then analyzing the image. Thus many images are routinely generated and must be stored for later analysis.

Additionally, gel electrophoresis is increasingly being used in labs for DNA and RNA analysis.

[1] "Next-generation DNA sequencing," Nature Biotechnology 26, 1135–1145 (1 October 2008)

[2] "Pharma Sees a Bigger Role for Imaging in Trials," *Bio•IT World* (1 May 2009) www.bio-itworld.com/2009/05/04/trial-imaging-perceptive.html

> To achieve the most efficient use of such HPC clusters requires a high performance storage solution that can feed data to satiate the awaiting computing processors

Here again, images of a gel are saved and later analyzed.

And finally, there has been a push in recent years toward earlier use of medical imaging in clinical trials, according to a 2009 *Bio•IT World* article[2]. This push comes from life sciences organizations who are embracing imaging, earlier on in the clinical trial process, as a tool for a faster identification of more promising compounds. This trend has led to broader use of techniques such as PET, MRI, and single-photon emission computed tomography (SPECT) – all of which are image-based and generate very large data files per use.

CENTER STAGE: DATA MANAGEMENT

The life sciences data explosion is forcing organizations to re-evaluate their storage options.

What's absolutely needed is a scalable solution that allows organizations to quickly add capacity with minimal disruption. However, the situation is more complicated than simply installing more disk drives.

Much of the data generated in life sciences labs must be analyzed and visualized so that decisions about which drug candidates to pursue can be made in a timely manner.

Matching the advances in sequencing and imaging, new generations of high-performance computing (HPC) servers and clusters have made supercomputing processing power available to organizations of all sizes. This has allowed the creation of high-throughput life sciences computational data and image analysis workflows.

To achieve the most efficient use of such clusters and the speed they offer, workflows require a high performance storage solution that can feed data to satiate the awaiting computing processors.

This complicates the choice in a storage solution.

Traditional storage solutions often cannot deliver the throughput required to keep the workflows running in an optimal manner. Thus, these solutions become a bottleneck that slows research and the decision making process when deciding which new drug candidate is worth further effort. To achieve the best results, the data involved in such workflows must reside on a storage solution that has the performance and throughput to match the diverse range of clients that it serves — from HPC clusters to traditional desktop systems.

Beyond meeting the capacity and performance requirements, a storage solution must have other key attributes to efficiently handle the Petascale volumes of data that are generated in drug discovery and clinical trials work.

First, the solution must scale with no downtime. Many traditional and clustered storage solutions require taking an array offline to add capacity. In today's labs, where experiments and analysis run around the clock, this is not acceptable. The addition of new capacity needs to be transparent to the users and the applications. As new units are added, they must automatically be recognized and made part of the existing array.

Second, the solution must have features that simplify management. With traditional solutions, capacity is normally grown by adding new, separate storage systems. And typically, this leads to numerous silos and multiple copies of data spread throughout an organization. As project data grows beyond its allocated storage capacity, IT managers, or users, must make changes to redirect applications to new volumes. With the amount of data generated in today's fast-paced research environments, this approach is time-consuming. A practical storage solution for life sciences therefore needs to scale to tens of petabytes in a single instance of the file system.

Third, the data storage solution must be future-proof in order to provide a reliable upgrade path. Similar to Moore's Law in computing whereby processor technology increases speed and reduces cost every 18 months, data storage media consistently increases its capacity density and decreases its cost per terabyte. Providing a seamless upgrade path to newer, higher density and less expensive hardware is therefore essential to protect current investments. However, higher capacity drives come with an increased risk of media errors. Innovative solutions are therefore required to address this increased risk in order to maintain a reliable storage solution, without adding huge incremental costs.

> A data storage solution must be future-proof in order to provide a reliable upgrade path

Fourth, any storage solution introduced into an organization must fit into the current environment. There must be support for data migration from the old systems to the new. And the new system must work with existing backup and recovery solutions.

And finally, the solution must be affordable both in terms of acquisition cost and total cost of ownership.

To meet these requirements, life sciences organizations have a variety of network attached storage (NAS) choices today. (See Table 1: A Comparison of Storage Architectures.)

Traditional NAS is easy to install and manage as an individual file server, but, due to limitations on file systems and the maximum number of files, it becomes highly complex to administer as the storage capacity scales. As more NAS systems are deployed, the system administration overhead increases faster than the number of NAS systems deployed. Administrators must constantly load balance each NAS system as well as migrate data between systems. Additionally, single file servers become a performance bottleneck as the number of clients accessing the server grows. This performance aspect is a particular bottleneck for high-

TABLE 1

A Comparison of Storage Architectures

Feature	Benefit	Traditional NAS	Clustered NAS	Panasas Parallel NAS
Out-of-the-box appliance	Painless installation	✓	✓	✓
Single management interface	Simplified system administration	✓	✓	✓
Fully compatible with existing back-up infrastructure	Requires no workflow upheaval	✓	✓	✓
Volume failover & storage network redundancy	High availability of data	✓	✓	✓
Global shared file system scalability >1PB	Low system management overhead	✗	✓	✓
Aggregate bandwidth performance to tens of Gigabytes/second & >100,000 IOPS performance	Scalable performance for demanding workflows and strong small file performance for many concurrent clients	✗	✓	✓
Concurrent, parallel client access to large files	Ability to parallelize I/O to accelerate large-file performance	✗	✗	✓
Global shared file system scalability to tens of Petabytes and beyond...	Low system management overhead from 10TBs to extreme scalability	✗	✗	✓
Aggregate Bandwidth Scalability >50+GB/sec	Leading performance for most demanding applications	✗	✗	✓
Tiered Parity architecture protecting against likely unrecoverable read errors in high-capacity drives	Reliability for deployments with today's high-capacity drives	✗	✗	✓

performance clients running complex tasks such as sequence alignment.

Clustered NAS solutions virtually aggregate file servers and present a global shared file system view to all clients. This provides a better way to scale up to a couple of petabytes than traditional NAS solutions. However, once you exceed this capacity, similar management overheads enter the environment. Clustered NAS also fails to address the performance requirements of diverse clients, in particular that of high-performance computing systems. Previously, it was common for life sciences organizations to deploy a clustered NAS solution in parts of the workflow, but to deploy a parallel storage solution for the more performance-demanding applications.

Parallel storage solutions can scale to tens of petabytes in a global shared file system image. Furthermore, providers such as Panasas Inc. incorporate multiple levels of performance and pricing under a single management layer that is exceedingly easy to manage and scale. Founded by Dr. Garth Gibson (co-inventor of RAID technology) in 1999, Panasas has entered the life sciences market at an inflection point where the industry demands the rich and unique feature-set that it has developed.

To meet the new demands created by high-throughput analysis workflows and the data explosion, a number of life sciences organizations

including the UC Berkeley Center for Integrative Genomics, Uppsala University (Sweden), EBI/EMBL (UK), The National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), and others use Panasas storage today.

A DEEPER DIVE ON PARALLEL STORAGE SOLUTIONS FROM PANASAS

Panasas parallel storage is serving data to over half-a-million servers in thirty-three countries. Its systems are in production at customers addressing over 12,000 core processors and serving data at 55 gigabytes/sec, in a single instance of the file system. PanFS theoretically scales to 96 yottabytes (a "yotta" being 1,000 exabytes). Even Panasas' largest customer today, operating with tens of petabytes of data, still has a lot of headroom for growth. With computing only recently entering the petascale realm, it will take some years before compute, storage and networking hardware is working in harmony at the exascale, never mind the yottascale level.

More importantly, Panasas solutions include the PanFS parallel file system tightly integrated with the Panasas ActiveStor storage modules. Both the software and hardware are tightly coupled, resulting in a solution that is both feature-rich and easy to manage.

Scalability: Panasas storage solutions scale capacity, performance and clients with zero downtime. Starting from 10 terabytes to tens of petabytes in a single instance of the file system, users scale their Panasas solution simply by introducing new Panasas modules into their existing array.

When adding Panasas storage arrays, the file system, PanFS, recognizes the new capacity and provisions it online with zero downtime. This task is completely transparent to both user and application. Furthermore, when new capacity is introduced, PanFS automatically load-balances the system by targeting new data writes to the new capacity and migrating as much of the existing data as required. The solution therefore always maintains a balanced performance.

Panasas Storage: Modular, Easy to Scale

Management Simplicity: Panasas storage arrays are based on a blade architecture that includes all the functionality of a disk array, controller, storage server(s) and metadata server(s). There is no requirement for additional storage server nodes or controllers. PanFS and its management layer, with a single user interface, is integrated into each Panasas storage array. Administrators have a single management interface for the whole storage system. With massive scalability across the number of users, files, directories, volumes and the file system itself, Panasas is extremely easy to administer within both large and small data center environments.

Investment and Data Protection: Panasas provides a data migration feature that allows its customers to move data seamlessly from lower-density Panasas modules to its latest higher-density solutions. Unlike alternative solutions, Panasas includes its unique Tiered Parity architecture, which delivers superior reliability and data integrity. Tiered parity fully protects against media errors in high capacity drives by applying parity at the block-level so that users never have to worry about media errors.

Compatibility with Existing Life Sciences Workflows: Panasas solutions are built on industry-standards, allowing them to integrate fully into

Panasas High-Performance, Scalable Storage

current life sciences workflows. Back-up clients can access data on a Panasas system via a standard NFS mount or leveraging Panasas' parallel protocol, DirectFLOW. The Panasas solution is also NDMP compliant and is therefore compatible with industry-standard back-up solutions. Panasas users today can integrate with various back-up solutions such as NetBackup, TSM and many others.

Affordability: During 2009 Panasas introduced a range of models with highly competitive acquisition costs to fit any IT budget. The new product line provides a range of solutions with at different price/performance points which allow a single system to be optimized for all of the needs across the company's life sciences customers' workflows. Panasas solutions also reduce total cost of ownership by:

- Lowering management costs through ease of administration
- Lowering capacity requirements by removing silos and unnecessary copies of data
- Protecting investment with a seamless upgrade path to higher capacity drives

All Panasas products include the storage hardware, client software and parallel file system as an integrated solution. Multiple Panasas models can reside under a single instance of the PanFS file system. The product line extends from full-featured entry level systems to high-end systems that integrate solid-state disk (SSD) technology. The lines include:

- **ActiveStor Series 7:** A full-featured, entry-level system with attractive acquisition costs that scales to tens of petabytes in a single instance of the file system. This system is easily upgradeable to the series 8.
- **ActiveStor Series 8:** Delivers higher performance than the Series 7 and includes high-availability and snapshot features
- **ActiveStor Series 9:** Builds upon the performance and functionality of the Series 8 by adding significantly higher IOPS (I/Os per second) performance across diverse workloads, much lower latency, maximum data availability, and integrated tiered storage capabilities. It integrates SSDs for maximum system speed.

In addition to extending performance across a wider range of applications and environments, Panasas systems include a full complement of data management and protection features.

For example, all Panasas systems support automatic tiered storage, which helps fully automate storage optimization. The storage is self-managing and supports two tiers of storage (DRAM & SATA) on a single Panasas ActiveStor Series 7 or 8 storage blade and three tiers (DRAM, SATA & SSD) on Panasas ActiveStor Series 9. The Panasas system uses these three tiers to accelerate application performance without user intervention.

To help protect data, all Panasas systems support asynchronous replication. This enables data to be replicated from the primary Panasas system, over a LAN or WAN, to a backup/disaster recovery Panasas system. The system continually compares successive data snapshots to assess what data has changed and sends only the changes for maximum speed and efficiency of data transfer.

To ensure that organizations get a complete solution, Panasas has partnerships with key systems vendors, software vendors, backup solutions providers, networking equipment companies, resellers, cloud services partners, supercomputer vendors and others.

For more information, go to: www.panasas.com

panasas