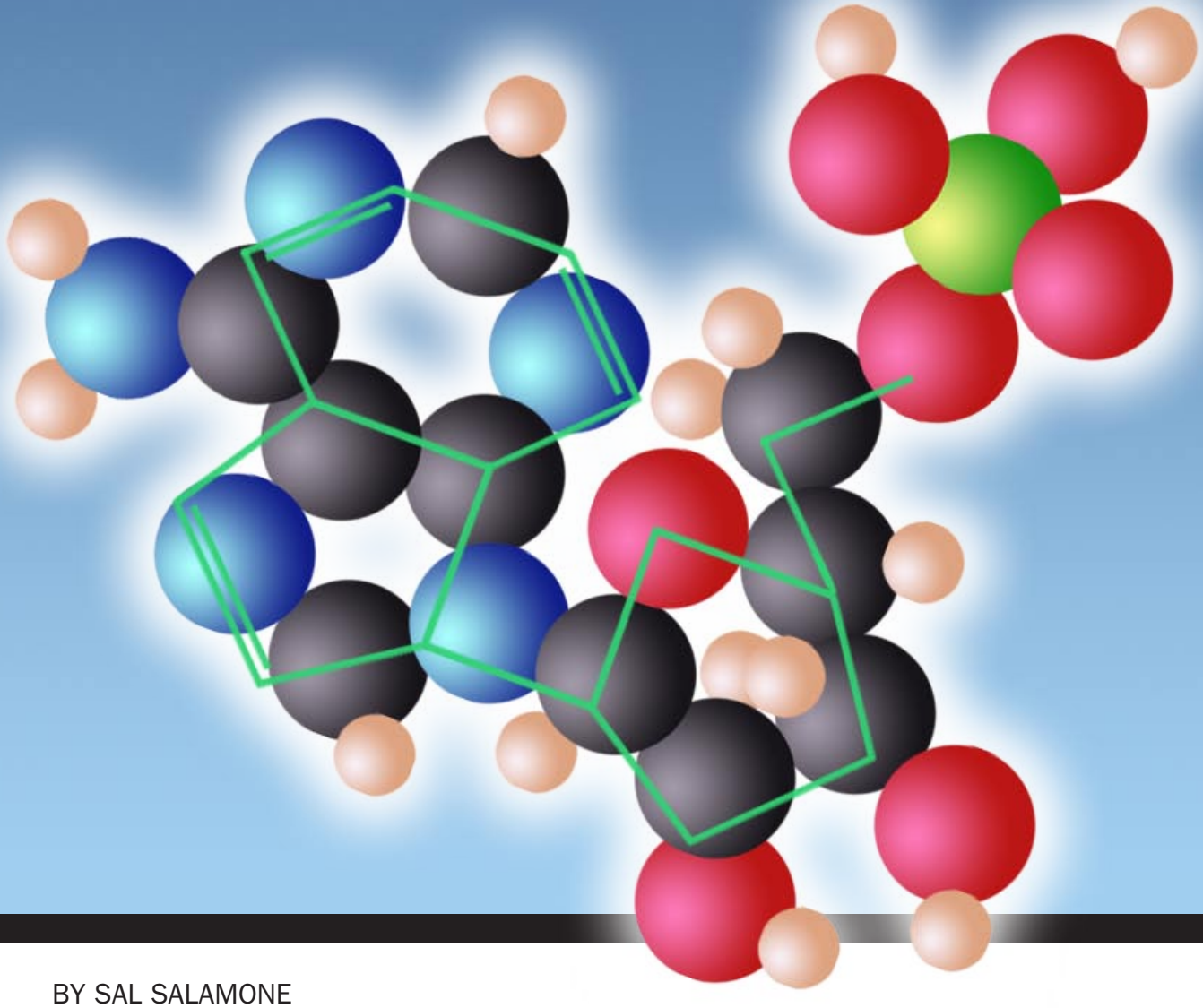
Next Gen Data Management for Next Gen Life Sciences



BY SAL SALAMONE

Produced by Cambridge Healthtech Media Group Custom Publishing

Next Gen Data Management for Next Gen Life Sciences

INTRODUCTION

Next-generating sequencing and new medical imaging systems are causing a radical change in the life sciences.

Specifically, next-gen lab equipment is providing significantly more insight into basic biological processes and chemical interactions, which in turn aids in the search for new drug candidates. Similarly, new imaging and sequencing technologies are being applied to basic research and in biomarker analysis, ADME/TOX studies, and clinical trials, helping to better identify targeted treatments and promising drugs before more money is invested to develop them and before they are submitted for FDA approval.

One consequence of these changes is the unprecedented amount of data generated in life sciences organizations. New sequencing and imaging systems can easily generate an order of magnitude more data than their predecessors. And as a result, many life sciences organizations are finding they must add large volumes of raw storage capacity.

Unfortunately, this can be quite a costly proposition. Beyond the capex spending for the raw storage capacity, there are operational costs that can easily dwarf the initial purchase price. In particular, the data must be managed, retained, backed up, and in many cases, safeguarded — all of which increases the storage system's total cost of ownership.

What's needed is a storage management solution that helps organizations get the most efficient use out of their storage investment while automating the many labor-intensive tasks associated with data management.

EXECUTIVE SUMMARY

- **New sequencers and lab equipment are generating orders of magnitude more data**
- **As data volumes grow, lifecycle management can consume great amounts of staff time**
- **Optimized use of storage resources is essential to contain costs, improve operations, and keep data analysis workflows humming**
- **Quantum solutions offer simplified management of data through tiers to ensure high performance workflows are not interrupted and storage resources are used most efficiently**

NEW LAB GEAR DRIVES DATA GROWTH

Life sciences organizations must deal with a glut of data and the volumes involved are poised to explode as new sequencing and imaging equipment enter the market.

To put the issue into perspective, consider the developments in sequencing equipment used in primary research and new drug discovery.

Recently, the market has seen a flurry of activity

with the introduction of next-generation sequencing systems from 454 Life Sciences (a Roche company), Applied Biosystems, Helicos BioSciences Corporation, Illumina Inc., Pacific Biosciences, and others.

According to a 2008 article in *Nature*,[1] interest in next-generation sequencers is high because they have the potential to "dramatically accelerate biological research by enabling the comprehensive analysis of genomes, transcriptomes, and interactomes."

The newer sequencers increase the number of consecutive bases that can be read (a parameter often referred to as the read length). And the systems run faster.

Essentially, the newer sequencing systems produce more data per run and complete each run in a shorter time than previous generation equipment. The more data per run simply increases the volume of data that must be stored. This data is then analyzed and visualized using third-party tools.

The faster sequencing completion times means more experiments or runs can be completed in a week or month. Thus the data piles up faster than it ever has before. Both attributes simply add to the data management challenge life sciences organizations face today.

What's the specific impact of these new systems? One run on Illumina's HiSeq 2000 sequencer "could recreate the Human Genome Project in a week," according to a 2010 *Bio•IT World* article.[2] When applied to more conventional research, the article noted that one run on the system "could sequence thousands of bacterial genomes simultaneously or unravel 16 transcriptomes in a mere four days."

Consequently, a single new sequencer can typically generate terabytes of data a day. (That represents an order of magnitude more data than what was generated with previous generation equipment.) And as a result, a lab with multiple sequencers is capable of producing petabytes of data in a year.

The new sequencers are also responsible for a secondary data explosion.

As is the case with other competitive, technology-driven markets, the cost of sequencing (including the cost of equipment and the chemicals to perform each experiment) is dropping. In particular, some industry experts estimate the cost of sequencing has dropped by a factor of about 100 in about three years.

For organizations already using sequencing, the lower cost means they can install more systems to speed research efforts. They also can extend sequencing to more groups within the organization.

The lower cost also is making sequencing available to organizations that previously could not afford the technology. "The market is not only growing nicely, but diversifying strongly into labs that have not previously been involved in sequencing," according to a 2009 *Bio•IT World* article.[3]

> Life sciences organizations must deal with a glut of data and the volumes involved are poised to explode as new sequencing and imaging equipment enter the market

IMAGING ON THE RISE

Similar to sequencing, changes in the use of imaging technology is producing new volumes of data that must be analyzed, stored, and managed.

The use of imaging comes into play in several places.

First, there is the use of imaging in basic research and early drug discovery. In particular, imaging is expected to be a major source of knowledge in cell biology, protein interactions and behavior, and systems biology. The technology has progressed to a point where "we have the opportunity to observe *in vivo*, *in situ*, in the cell directly, the expression of genes," according to a 2009

[1] "Next-generation DNA sequencing," Nature Biotechnology 26, 1135–1145 (1 October 2008) www.nature.com/nbt/journal/v26/n10/full/nbt1486.html

[2] "Illumina's HiSeq 2000: Secrets and Buys," *Bio•IT World*, January 22, 2010 www.bio-itworld.com/BioIT_Article.aspx?id=96512

[3] "Sequencer Market Heats Up," *Bio-IT World*, November 10, 2009 www.bio-itworld.com/BioIT_Article.aspx?id=94999

Bio•IT World article.[4]

An example is the growing use of microscopy in many phases of life sciences research. Comparable to what is happening with sequencers, new electron and confocal microscopes deliver higher resolution images and more automation. This in turn creates large data files and more of them.

And as is the case with sequencing data, imaging data is now routinely analyzed and visualized as part of computational and analytic workflows. More importantly, many labs are developing automated image analysis and visualization pipelines. Such pipelines can place additional stress on the interplay between storage and the high performance computing (HPC) systems that perform the analysis and rendering.

Another aspect of the imaging explosion comes in the form of an increased use of microarrays for single-nucleotide polymorphism (SNP) detection and gene expression profiling. SNPs, which are DNA sequence variations occurring when a single nucleotide, and the level of gene expression are detected by imaging the array using a laser or light source, capturing the image digitally, and then analyzing the image. In many labs, such images are routinely generated and must be stored for later analysis.

An additional research area where imaging has been widely adopted is in the growing use of gel electrophoresis systems for DNA and RNA analysis. As is the case with microarrays, images of the gels are saved and later analyzed.

The growing interest in basic research and in areas such as biomarker analysis has helped fuel the interest in these areas. And as these fields mature and their benefits come to be, there will be even more wide-scale interest in the gel and microarrays technologies.

Finally, there is the rapid expansion of imaging's use in clinical trials. Essentially, many life sciences organizations are embracing imaging earlier on in the clinical trial process, as a tool for a faster identification of more promising compounds, according to a 2009 Bio-IT World article[5]. In fact, it is now quite common for early stage clinical trial work to include the 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.

Systems based on higher resolution imaging technology are constantly coming to market. Similar to the impact of new sequencing technology, the newer systems are driving down the costs of what was previously state of the art equipment. And as such, this opens up the market to many labs and organizations that previously could not afford the technology or limited its use.

[4] "Imaging Informatics," *Bio•IT World*, February 18, 2009 www.bio-itworld.com/BioIT_Article.aspx?id=87782

[5] "Pharma Sees a Bigger Role for Imaging in Trials," *Bio•IT World*, May 4, 2009 www.bio-itworld.com/2009/05/04/trial-imaging-perceptive.html

CASE STUDY:

Human Genome Sequencing Center, Baylor College of Medicine

Key Challenges

- The research center needed to access, share and manage hundreds of terabytes of DNA sequencing data for analysis at any time.

Project Objectives

- Centrally manage complex heterogeneous environment of servers, networks and storage technology
- Expand Baylor College's Human Genome Sequencing Center's data storage capabilities

Quantum Solution

- Quantum StorNext File System and Storage Manager
- Quantum Scalar i2000 tape library

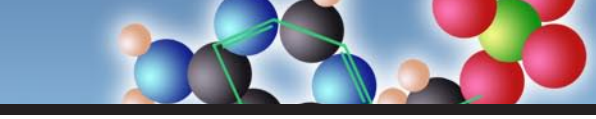
Benefits

- Enabled simultaneous access to huge volumes of data without impacting system users
- Provided cost-effective content creation through automated data management
- Allowed centralized management of heterogeneous environment
- Protected prior investments by integrating legacy resources
- Provided scalable foundation to meet anticipated storage growth of up to 20 PB over next 2-3 years

WORKFLOWS BECOME THE NORM

As noted, lab data is growing exponentially. Each new generation of sequencers, mass spectrometers, microscopes, and other lab equipment produces a richer, more detailed set of data.

For years, the way to handle data growth was to simply throw raw storage capacity at the problem. But that approach no longer works. Besides dealing



with capacity challenges, life sciences organizations must also deal with performance, management, and energy issues when it comes to their storage systems.

Lab data needs to be processed, analyzed, and visualized to be of any value. Typically, this requires the use of HPC clusters whose nodes must be constantly fed data. Moving the data on and off of storage devices to the cluster nodes becomes the challenge.

In particular, most life sciences organizations have developed sophisticated production-quality computational and analysis workflows and pipelines. A single performance bottleneck, such as data being fed too slowly to waiting cluster nodes, has a ripple effect. Not only is that analysis or computation for that particular job slowed, but all work queued up behind that job is also delayed.

Within these environments, applications require high performance and scalable I/O subsystems. That means the storage and the file system must meet the application's requirements in terms of throughput, as well as in capacity and scalability.

One solution would be to store all data on high performance storage, but that is not an economical use of resources (more on this later in the paper). And it may not even solve the performance issue today.

In the past, staging data that was to be analyzed or visualized on high performance storage would help feed the CPUs used to perform these operations. But in today's research organizations, there is often a need for shared access to data. In particular, many workflows require that multiple servers and workstations have access to the same data to perform different calculations or analyses.

What's needed is an intelligent way to share data across servers and workstations from a single storage source without requiring migration to various pools of storage, as well as selectively moving data between different tiers of storage, each with different price/performance characteristics.

The data must be on the highest performance systems, accessible to researchers across disciplines for analysis and discovery when it is part of a workflow, stored on more cost-effective systems to retain it for additional review and retrieval, backed up and, if necessary archived to tape. If done manually, management of these processes can be time consuming adding to the total cost of ownership of the storage and IT systems.

Cost of another type is now starting to become an issue as well. It takes electricity to run and cool IT equipment and storage systems.

As data volumes explode and more disk-based storage is added to accommodate it, energy consumption rises. And like most IT and data center equipment, storage devices have, over the years, increased in performance while physically shrinking in size. While the combination of higher performance and higher densities helps meet the capacity and computational requirements for life sciences research, it also means more electricity is needed than ever before. More power is needed to run the systems and even more power is needed to cool the densely packed (and hotter) units.

> Life sciences organizations need a storage management solution that helps them get the most efficient use out of their storage investment while automating the many labor-intensive tasks associated with data management

Quite interestingly, tape systems offer an advantage in energy savings over disk-based storage. Spinning disks consume electricity, whereas data once written to tape does not.

And whether or not tape is used, the main point to keep in mind is that organizations should choose a solution that is flexible enough to archive long-term data on the storage tier of their choice be it disk, tape, or NAS appliances.

LONG-TERM DATA MANAGEMENT CHALLENGES GROW

When an experiment is run, the data needs to be stored on a system that has the appropriate performance capabilities to support whatever analytic or visualization workflows are used to process the information. After a while, decisions must be made as to how to cost-effectively manage this data.

In the past, once the analysis was done, data

StorNext Data Management Software

High Performance File Sharing Features	Enterprise Data Management and Protection Features
• **SAN File System:** Delivers high-performance • **Distributed LAN Client:** Provides NAS-like scalability to thousands of server nodes • **Shared File System:** Offers simultaneous file access across platforms • **Platform Independence:** Supports Windows, Linux, Mac and UNIX • **Storage Vendor Agnostic:** Supports all major disk and tape systems	• **Replication:** Enables flexible data protection and data distribution • **Nearline Deduplication:** Reduces storage requirements, optimizes capacity and cost of Tier 1 storage • **Management Console:** Simplifies data management complexities and reporting • **Storage Manager:** Drives transparent tiered storage and archiving • **Distributed Data Mover (DDM):** Improves access performance and scalability of storage tiers

would be moved off of online storage systems and retired to tape and eventually deleted. But today, a large portion of data must remain available online and much of the data must be retained for very long periods.

Two factors are driving the long-term data storage requirements in the life sciences.

First, whether applying for a patent on a new chemical entity or for FDA approval of a new drug, some raw data associated with early research and development experiments and clinical trials must be retained. And given that it typically takes 10 to 15 years today for new drugs to be approved, that data must not only be stored for all those years, it must be stored in a manner where it can be easily retrieved if needed.

A second factor impacting life sciences long-term data storage is the increased interest in finding new uses for existing drugs. Specifically, since many blockbuster drugs are about to go off-patent and new drug pipelines are shrinking, many life sciences organizations seek to expand their potential patient market by submitting drugs for use as treatments for different diseases or conditions than when the drug was originally approved. This practice, known as indication expansion, is becoming very common.

An indication for a drug refers to the use of that drug for treating a particular disease. (For example, diabetes is an indication for insulin.) There are many drugs with multiple approved indications, meaning they can be used to treat more than one disease. For example, Eli Lilly and Company's Cymbalta is approved for the treatment of major depressive disorder, generalized anxiety disorder, and for the management of neuropathic pain and fibromyalgia.

More than three quarters of the 50 top selling drugs have had at least one additional indication approved since their initial launch in the US, according to industry experts. And this is a trend that is likely to grow.

This has great implications on data retention strategies. Rather than re-doing many basic experiments, data associated with a drug approved years ago could help make the case to pursue a new indication. So it makes sense to retain as much data as possible for the longest time. Besides retaining data for additional indications, over time new analysis tools and techniques can often provide further insights through the re-examination of old experimental data.

Obviously, not all of this data needs to be kept online. But there might be hesitation about moving certain data to tape if retrieval years later would be a laborious and time-consuming task.

MEETING TODAY'S DATA CHALLENGES

To meet today's life sciences storage and data management needs requires using a variety of technologies. And the solution that is employed must offer easy day-to-day management, a means to make the most efficient use of storage resources, and help with long-term data retention.

Quantum is focused squarely on this market today. For years, the company was known for its tape

technology and systems that automated data management, backup, recovery, and archiving. It has leveraged this expertise and moved beyond tape with solutions that include disk-based backup, software, security, and services optimized for tiered storage.

For example, Quantum offers a full range of price/performance tape solutions that include high capacity media, autoloaders, tape libraries, enhanced security including encryption, and advanced media monitoring and diagnostics capabilities. Additionally, it offers disk-based backup/recovery appliances.

Most importantly, the heart of its offerings to address today's life sciences data challenges is Quantum's StorNext data management software.

StorNext software is comprised of two core components. The first is the StorNext File System. This is a shared file system that is operating system independent, enabling concurrent shared access to a single pool of data across heterogeneous operating systems so that organizations do not need to keep separate copies of large files or move one file to different systems for workers using computers with different operating systems. Essentially, one copy of the file can be saved and all users can access it simultaneously.

The file system supports data access by LAN and SAN clients. In particular, SAN clients can directly access files over high-speed Fibre Channel or iSCSI connections to enable high-performance computing workflows.

Additionally, LAN-connected clients can achieve a performance boost using StorNext Distributed LAN Client (DLC), which uses a proprietary protocol that significantly reduces the overhead typically found with NFS and CIFS clients. The DLCs access data over IP via a single or multiple gateway servers. The high line rates and efficiencies achieved when using this protocol means the solution can scale to thousands of compute or analysis nodes, outperforming many standard NAS solutions.

These features can simplify data management in a life sciences organization. For example, when used in genomic research, a bench chemist running a Windows PC, an IT manager running a Linux desktop, and a database specialist running a UNIX workstation can all access a single copy of a data file concurrently, while maintaining file integrity.

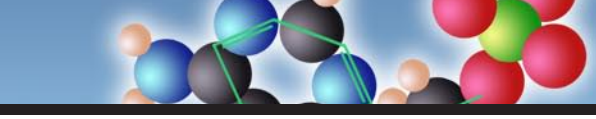
The second part of StorNext data management software is the Storage Manager. A primary capability of Storage Manager is that it supports tiered storage and transparent data movement. This capability can be used to move data off of old and onto new storage systems, as well as migrating data between storage tiers. So as new storage systems offering higher performance, more capacity, and lower energy consumption are added, older systems can be gracefully retired. Similarly, as data progresses through its lifecycle, it can be moved off of the highest performance systems after analysis, to more economical systems for easy access, and to tape for long-term and even more economical storage.

StorNext Management Console View

During these processes, StorNext Storage Manager helps ensure that high levels of access are maintained. From the user's perspective, the data movement is completely transparent — the files always remain in the same namespace, regardless of physical location on disk or tape tiers. This allows life sciences organizations to keep projects and workflows intact since, for example, third party applications do not need to be modified to access the moved data.

The current version of StorNext software (StorNext 4.0) includes additional data management features to help organizations get optimal use of their storage resources. The features include:

Replication: StorNext replication enables research data to be safely copied to a remote location to safeguard against primary site downtime, as well as for data distribution models so that valuable



data can be shared amongst geographically disbursed researchers. Replication supports flexible deployment options, such as distributing data from one primary site to several secondary sites or consolidating data on many remote sites to one consolidated central site. A one-to-many replication scenario could be used to send the most recent version of a curated or annotated database to researchers in multiple sites to support distributed workflows or collaborative projects. And the many-to-one mode can help consolidate data generated in geographically distributed labs to a central site so in-site tech staff can more easily back it up and safeguard it.

Deduplication: StorNext nearline deduplication enables enterprise data management to reduce primary storage requirements, thereby decreasing costs associated with managing and procuring costly primary disk. StorNext offers directory-level deduplication within the primary file system based on time, size, and file type. Quantum's patented deduplication technology ensures that only unique data blocks are stored, resulting in lower storage requirements. In addition, the deduplication feature is tightly integrated with the replication policy engine. This allows the option to dedupe the data at the source location prior to replication, thereby reducing bandwidth requirements to transfer data, as well as storage requirements at the remote location.

StorNext's deduplication can be used to optimize primary storage capacity and free up Tier 1 storage, thus slowing the need to add more high-performance storage capacity. Deduplication can offer significant savings. Some users claim they have achieved 30-to-1 deduplication ratios and as a result they were able to reduce storage capacity needs by more than 95 percent.

Distributed Data Mover: Some new high-performance storage solutions on the market today aimed at the life sciences deliver incredible performance, but require an abandonment of installed equipment. StorNext's Distributed Data Mover (DDM) technology allows use of existing equipment and boosts file access allowing quicker retrieval from tiered storage including from tape. With DDM, multiple clients can be placed on a SAN to spread I/O throughput across multiple units. This improves access times, thus speeding any particular workflow. But perhaps more importantly, DDM brings a higher level of performance to data retrieval off of tape systems. This makes it easier to migrate data to tape since, depending on the level of parallelism deployed, users will not notice any difference as they work.

Combined, these features will help life sciences organizations better manage the data explosion in their labs. That is the case at the Human Genome Sequencing Center(HGSC) at the Baylor College of Medicine.

"With such high volumes of data generated daily from DNA sequencing, and the need to access hundreds of terabytes of data at any given time, StorNext offers the scalability and support we need"

— Geraint Morgan, director of Information Systems at Baylor's Human Genome Sequencing Center

"With such high volumes of data generated daily from DNA sequencing, and the need to access hundreds of terabytes of data at any given time, StorNext offers the scalability and support we need," says Geraint Morgan, director of Information Systems at the Human Genome Sequencing Center. Morgan expects the genome center's storage requirements to push past two petabytes in the next two to three years.

The bottom line is that as data volumes grow, Quantum solutions can be used to address the concerns that life science organizations may have about data management and protection. Specifically, Quantum solutions meet the need for simplified management of data through tiers to ensure high performance workflows are not interrupted and storage resources are used most efficiently.

For more information, visit: www.quantum.com/StorNext

Quantum. StorNext.