Addressing Life Sciences Constantly Growing Data Challenges

The continued explosion of raw experimental data, the increased use of video, the growing adoption of new data retention practices, and the move to high throughput computational workflows are all placing new demands on the way life sciences organizations store and manage their data.

By Sal Salamone, Contributing Editor, *Bio•IT World*

BLUEARC®

www.bluearc.com

Addressing Life Sciences Constantly Growing Data Challenges

EXECUTIVE SUMMARY

Factors driving storage selection in the life sciences

- New lab equipment produces vastly more data per experiment

- Increased us of imaging and visualization is rapidly driving up storage demands

- Multi-core processor nodes require higher data throughputs to sustain computational workflows

- Embracement of virtualization is placing new demands on storage performance and provisioning

- Data management challenges continue to grow and must be addressed in new ways

- Storage solutions must now address performance, manageability, and energy efficiency issues.

FACING THE DATA DELUGE

Life sciences research is heavily compute and data intensive. That's no surprise. That's been the case particularly since the sequencing of the human genome project began. But now, the magnitude of the situation has changed radically. Today, researchers have much more powerful lab equipment that is capable of producing volumes more data than just a few years ago. This is driving the need for higher performance, cost effective storage solutions that can complement the ever-increasing computing power that is being applied to data analysis and imaging of that data.

Additionally, life sciences research is now typically conducted in a more operationally efficient manner. Rather than scientists cutting and pasting data from one spreadsheet or database into another for analysis, workflows are increasingly automated so that results from one analysis or computation are used as input to another. This again places demands on the storage systems and their interactions with HPC resources.

Simply adding raw storage capacity can help. But it is not the best solution. As the volume of data and power of HPC resources grow, the challenge is how to meet growing storage capacity and performance demands, while minimizing the burden of managing the expanding volumes of data.

DATA EXPLOSION GOES UNABATED

Each new generation of sequencers, mass spectrometers, microscopes, and other lab equipment produces a richer, more detailed set of data. A single experiment can produce hundreds of gigabytes (GB) of data. As a result, any organization running hundreds experiments a month or year quickly finds itself with a data management problem.

Newer equipment promises to exasperate the situation. For instance, new sequencers from 454 Life Sciences, Applied Biosystems (ABI), Illumina, and others are capable of enormous output.

And Helicos BioSciences recently burst onto the scene with its HeliScope sequencer. In an April 2008 *Bio-IT World* article[1] about new sequencers, Editor-in-Chief Kevin Davies noted the HeliScope can produce 24 terabytes of data in two full runs.

Some organizations have rooms filled with these machines that run 24x7. To accommodate the data, organizations must install enough storage capacity to hold the results of each experimental run. But frequently the larger challenge is to incorporate this data into analysis workflows. And that's where simply adding raw disk space fails. The data must be managed over the long-term, and readily available when needed for analysis or visualization.

Beyond sequencers, other lab instrumentation is pushing life sciences data growth upward at exponential rates. One area in particular is imaging. Virtually all of today's new microscopes produce digital image files.

In the recent past, a lab might examine individual microscope images, manually identify traits or characteristics, and add annotations to an image's file. Now, it is more likely that hundreds of images taken during a single experiment would automatically be analyzed using software to identify traits. In some cases, images representing slices of a cell might be combined into a 3D visualization; in other cases successive 2D images would be looped together to give researchers a view of the interaction at a membrane's edge or give clues some other physical phenomenon as it occurs over time.

More importantly, many labs are developing automated image analysis and virtualization pipelines and workflows. Such pipelines can place additional stress on the interplay between storage and HPC systems.

And like their sequencing counterpart, imaging equipment is also evolving rapidly. High-end microscopes are producing higher resolution images, which individually take up more storage capacity. And lower-end microscopes promise to open up imaging to more areas of research. For example, a July 2008 article[2] in MIT's *Technology Review* discussed a new "tiny microscope that employs the same kind of chip used in digital cameras [that] can produce high-resolution images of cells without the expensive, space-hogging lenses that have been part of microscope design for centuries."

The Caltech researchers who developed the microscope believe they could be mass-produced for about $10 each and "incorporated into large arrays, enabling high-throughput imaging in biology labs."

This potentially would bring microscopy to groups that could not afford the technology before. And it would simply add to the quantity of life sciences data generated in a lab that needed analysis or visualization.

OLD HABITS DIE HARD

As lab data that needs analysis or virtualization grows, other factors are adding to the total volume of data accumulated within a life sciences organization.

Even with the introduction of electronic lab notebooks and knowledge management systems, many life sciences researchers (like their counterparts in all industries) use their e-mail inbox as a poor man's collaborative environment and data repository.

The consequences for storage can be significant. For example, a large file with results of an experi-

> Lower-end microscopes promise to open up imaging to more areas of research.

ment or detailed analysis of that data could be posted to a content management system or uploaded to a shared Web site for all interested parties to download or view. But in many cases, researchers simply e-mail such a file to all those interested in the results.

This means rather than having, for example, one 10 megabyte (MB) file sitting on a server, there might be 50 copies of that file in coworkers' e-mail inboxes. Many of these workers will save the attached file in their e-mail inbox and download a copy to their hard disk drive. The result is instead of one shared 10 MB file, there could be the storage equivalent of 100 files of that size (about 1 GB) taking up valuable disk space.

It is easy to see how this routine operation of sharing files via e-mail can quickly add to the total storage capacity needed within an organization.

Exasperating this issue is the fact that organizations must now retain e-mail longer to meet regulatory and legal requirements. For instance, if an organization gets involved in a legal matter — a patent dispute or a worker harassment suit, for example — relatively new amendments to Federal Rules of Civil Procedure require that organizations quickly produce requested electronically stored information.

Recovering old e-mail messages to comply with these so called eDiscovery requirements has proven to be a costly and time consuming task. Companies in court cases have found that subpoenaed messages are often scattered on dozens to hundreds of incremental backup tapes. This is forcing many organizations to reconsider their e-mail archiving practices. And many are opting to store years of archived e-mail online so that retrieval would be easier in the event there is a need to produce it in a court case. This long-term storage must be factored into any storage capacity planning effort.

SERVER VIRTUALIZATION DRIVES PERFORMANCE REQUIREMENTS

Companies are increasingly turning to virtualization to reduced operating costs, consolidate servers, and simplify the deployment and management of applications.

Server cluster nodes based on multi-core processors are now commonly used in conjunction with virtualization software to enable dozens or more applications to run as virtual machines on a single physical server.

As this new architecture becomes more widely used, organizations must address several pointed issues related to matching and marrying storage to their HPC resources.

First, new performance issues crop up. A single server accessing a single storage device or a cluster running one application accessing a pool of storage drives is one thing. But it is an entirely more complex matter when a cluster runs dozens of applications as virtual machines, all of which require varying levels of access to data to feed their CPUs.

The key issue becomes how to match storage to servers in a virtualized environment. After all, the numerous virtual servers will all need access to storage all at the same time. And the storage will have to accommodate multiple concurrent workloads without degradation.

In particular, the rapid adoption of and migration to a virtual server environment requires that storage be flexible and capable of being dynamically grown to meet the capacity and performance requirements of the virtualized applications. Additionally, since multiple virtual servers will all be access storage simultaneously, storage must perform under multiple concurrent workloads without degradation. And since virtualized applications can be quickly and easily set up and torn down, the associated storage must support easy, dynamic provisioning.

Second, because virtual machines can be set up so quickly, provisioning of storage must be comparably fast.

Third, since many virtual machines will share storage resources, provisioning and addition of

> **The key issue becomes how to match storage to servers in a virtualized environment.**

4

new storage capacity must not involve taking systems offline or shutting them down.

A fourth point to consider is storage management specifically with regards to provisioning. One of the benefits of server virtualization is that new virtual machines can be set up with very little effort. Users do not have to configure a physical server, load the OS, drivers, and applications, add network connectivity, etc. Instead, a virtual machine can be installed and set to run in minutes.

This allows users to quickly create a virtual machine to test an application or help develop a new application. The downside to this easy setup is that can result in virtual machine sprawl and lead to a situation where many virtual machines are used to address an issue on the spot, and then are abandoned.

In this scenario, a problem can arise in the way storage is allocated. The traditional approach has been to set aside the desired storage capacity when a new server application is provisioned. When hundreds of virtual machines are added simply because they are easy to set up, storage would need to be allocated for each. With traditional provisioning, that allocated storage cannot be used by other applications. If a virtual machine is set up and abandoned, that storage capacity would be wasted.

And even if a virtual machine is used, it might not require all of the storage that was allocated on creation. In that case, the unused portion of the allocated storage would simply sit idly by where no other application could use it.

This means large volumes of storage must be purchased, added to a network, and managed even though much of the capacity might not be used. This is hugely expensive and a waste of corporate resources. The additional capacity must also be managed, taking valuable time from the IT staff. In today's tough business environment, such waste of resources is not acceptable.

These scenarios show that the explosive growth of server virtualization alters the traditional ideas about storage. Organizations must address storage performance, capacity, provisioning, and long-term management issues that virtualization raises.

OTHER FACTORS IMPACT RAW STORAGE NEEDS

As lab equipment spews out ever-growing amounts of data, other forms of data are exploding in organizations, in general, and life sciences organizations, in particular.

In addition to the increase in the volume of data, there is also an increase in data diversity due to the use of such things as video, voice over IP (VoIP), and RFID to tag and track samples and supplies in a lab. In a report released earlier this year, the consultancy IDC noted that this diversity complicates data management since the number of electronic information containers (files, images, packets, meta-tags, etc.) is growing 50 percent faster than the number of gigabytes of raw data.

Essentially, many elements of a life sciences organization's operations (e.g., phone conversations, tagging specimens, etc.) that were not digital in the past are now taking advantage of new technologies to add to the data mix.

One area of new data growth for life sciences organizations is the increased volumes of digital clinical trial data. In the past, much of this data was analog. For instance, patient diaries might be collected in a notebook or an X-ray would be captured on a sheet of film. Now, the growing adoption of Electronic Data Capture technology in clinical trials is allowing more of the information to be collected digitally for easier access, archiving, retrieval, and analysis. Fortunately, much of the data such as that collected in patient diaries can be contained in small files. But from a storage perspective, this data often must be retained for very long time periods.

At the other end of the file size spectrum one need only look at video. The advent of low cost

> One area of new data growth for life sciences organizations is the increased volumes of digital clinical trial data.

camcorders, video editing software bundled with Apple and Windows-based computers, and You-Tube is making video pervasive.

Some of the video is directly related to life sciences research. For instance, increased computer processing capabilities and finer spatial time resolution imaging equipment have enabled researches to routinely string together a sequence of images to produce action videos showing the real-time motion of an organism.

Other videos are more general. For example, self-produced educational and training videos are now within the reach of most organizations, again, thanks to the availability of low cost camcorders and free or inexpensive video editing software.

Regardless the purpose or nature of the video, the files must be stored somewhere. This can quickly drive up the storage capacity requirements of an organization. A two to five minute video can easily be several hundred MB to a couple of GB. Longer videos routinely can be in the 10 to 50 GB range each. And video animations of life science lab images from electron microscopes, MRI, or CAT scans, can be much larger.

What's the impact of the growing use of video? A July 2008 *eWEEK* article[3] noted: "Digital video storage is the single fastest-growing sector within the storage industry at the moment."

The impact of the growing use of video in the life sciences is much more than a pure storage capacity issue. If a video is going to be viewed by many parties, perhaps simultaneously, consideration must be given to the impact of drawing the video content off of the storage disks. Alone, this would likely require a high performance storage system and infrastructure. But when added to the data movement demands placed on storage systems when using HPC systems for analysis and visualization, serious consideration must be given to the performance between storage and cluster resources.

LONG-TERM DATA AVAILABILITY

Life sciences organizations typically have lots of data that is never modified after it is initially created. A prime example of this type of data is the data

files generated by a lab experiment. Certainly, that data is often analyzed or visualized, but the original data is not changed.

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

Ideally, researches want to keep all of the processed data on disk. The issue is not unique to the life sciences. Many industries such as financial services, oil and gas, and manufacturing must deal with data growth, archival storage issues, and matching storage performance to computational requirements.

In fact, most companies are finding they simply need to keep data for longer periods of time. Some industries do so to meet regulatory requirements for data retention. In the life sciences, data is often kept to support intellectual property claims or new drug application submissions.

In the past, most archival data 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. Additionally, many applications (particularly those that use Web 2.0 and Semantic Web approaches) are designed so the data is available all the time. That means the data is not likely to be taken off primary storage and archived to tape as other data is.

The combination of these factors means life sciences organizations must manage large volumes of data and have the ability to easily add more capacity as demand requires.

WHAT'S NEEDED PART 1? PERFORMANCE

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 ener-

gy issues when it comes to their storage systems.

For example, any type of storage drives can hold the data, but a low-performance storage system will likely impede the analysis work as data hungry HPC resources sit while data is moved between storage and CPUs.

At a minimum, this can prove to be an ineffective use of computing resources. If you've invested large sums of money in HPC gear, having it not run at its optimum speed wastes that investment.

Worse, the slowing due to poor data movement might slow an entire analysis and research workflow. For example, a choice as to which target to explore or which new chemical entity to focus on might hinge on the analysis of a sequencing or microarray run. If the operation is automated in an organized workflow, delaying any step impacts to

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

total time to reach a conclusion or make a decision.

Worse still, delays in the early stage research might allow a competitor to file their new drug application to the FDA sooner or bring a product to market faster.

Small delays may seem trivial, but the consequences can be severe. For instance, they might include:

- Loss of several million dollars in revenue for each day delayed with a blockbuster drug (i.e, a drug with sales of $1 billion per year or higher)
- When there are several similar drugs offered, the first to market often becomes the market leader
- If two companies are working on a patentable entity, being one day later than a competitor can lock a company out of the market for the life of the patent.

These examples are extreme. But they illustrate the need for speed in today's highly competitive life sciences research market.

Lab data needs to be processed, analyzed, and visualized to be of any value. Typically, this requires the use of high performance computing (HPC) clusters whose nodes must be constantly fed data. The attached storage systems must be capable of feeding cluster nodes CPUs in a timely manner.

WHAT'S NEEDED PART 2? MANAGEABILITY

More powerful HPC clusters, virtualization, and the growth of archives are placing new demands on the way life sciences data is stored and managed. IT staffs need solutions that are easy to manage and help with administrative chores over the long-term.

However, the complexity of the situation exasperates the management challenge.

Data must be managed over its lifetime. This includes addressing the following:

- Data associated with the current research needs to be on systems that have the performance to match the HPC systems that will be used for analysis and visualization.
- Data associated with virtualized server applications must be on systems that can handle the multiple simultaneous reads and writes that occur when dozens of virtual machines run on a single physical server.
- E-mail must be archived online for easier search and retrieval for research purposes and in case that data is subpoenaed in litigation. Naturally, this type of storage must be robust, but it does not need the high performance required for data in computational workflows.
- Storage volumes must be virtualized to simplify management and support the quick changes associated with applications running on virtual machines.
- Provisioning must be simple to match the rapid setup characteristics of virtual machine applications. A key element to consider is thin provisioning, where capacity is allocated at setup time, but that capacity is not used until data is written to it.

- Adding storage to meet data growth must be transparent. Storage systems cannot be taken offline since they typically serve multiple applications. Additionally, the remapping of drives and data pointers must be abstracted through a storage virtualization layer since IT does not have time to continuously perform these tasks.

WHAT'S NEEDED PART 3? ENERGY EFFICIENCY

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 need to run the systems and more power is needed to cool the densely packed (and hotter) units.

Organizations need to make more efficient use of existing storage capacity helps reduce electricity requirements. For example, a traditional way to allocate storage is to simply set up a dedicated device for each application. This often results in many physical storage devices running at low utilization rates. Unfortunately, each device will still need electricity to run and power for cooling. By virtualizing storage, applications can share volumes, reducing the number of physical devices and thus reducing the electricity needed to run the equipment.

Thin provisioning complements storage volume virtualization in that it allows capacity to be used more prudently.

BLUEARC AS YOUR TECHNOLOGY PARTNER

These varied needs require flexible, high performance, and easy to manage storage solutions.

To that end, BlueArc's Titan storage solutions deliver the hardware, storage management, and advanced management features required in today's data-intensive life sciences organizations.

Specifically, the BlueArc solutions offer dynamic

BlueArc Titan

scaling of performance and capacity. This is possible through its unique hardware architecture that is designed to support scaling without compromising performance. The company's software offerings work together to address the storage management challenges. For example, storage virtualization management features simplify many day-to-day administration tasks. Other management features can help virtualize storage servers to transparently shift workloads across various physical servers, thus improving overall performance.

To address long-term data management issues, the BlueArc solutions offer automated intelligent data migration allowing organizations to optimize the use of their higher performance storage. They also support snapshots to protect data in a timely manner.

As data volumes grow, the increased attention to energy efficiency is going hand in hand with simplified management and high performance as key criteria for storage systems to handle life sciences data. The BlueArc solutions offer energy efficient design and the ability to consolidate legacy storage infrastructures, dramatically increasing storage utilization rates and reducing total cost of ownership.

These features combine to enable life sciences organizations to expand the ways they explore, discover, and conduct research. The products replace complex and performance-limited products with high performance, scalable, and easy to use systems capable of handling the most data intensive applications and environments. ●

[1] "The DNA Data Deluge," *Bio-IT World*, April 1, 2008
www.bio-itworld.com/issues/2008/april/cover-story-dna-data-deluge.html/

[2] "Tiny $10 Microscope," *Technology Review*, July 30, 2008
www.technologyreview.com/Biotech/21147/?a=f

[3] "Movies, Security Drive Demand for Video Storage," *eWEEK*, July 9, 2008
www.eweek.com/c/a/Data-Storage/
High-Demand-for-Video-Storage/

BLUEARC®

www.bluearc.com