

Coping with the Explosion of Data in Life Sciences Research

How storage is managed will either advance—
or slow—the pace of biomedical discovery

“As scientific researchers acquire data at faster and faster rates, optimizing the analysis of that data with scalable storage solutions is essential...Ocarina provides a cost-effective way to maximize storage capacity without sacrificing performance.”

— Dr. David Lifka, Director, Cornell
Center for Advanced Computing

The emergence of genomics and advanced gene sequencing techniques has made the collection and storage of data a centerpiece of biomedical research. It was not that long ago that the human genome project first sequenced the human genome as part of an international cooperative research effort. That first sequence took up about 750GB in 2000 – an amount of data that would fit on a single disk today. However, genomics research has rapidly moved past the first basic sequencing of the human genome, and now research advances are made through increasingly sophisticated sequencing machines and technologies. Today, research institutions, universities, pharmaceutical companies and even hospitals generate genomic data almost continually.

For example, Cornell University's computational biology service unit, which supports life sciences across its many research facilities and hospitals across New York State, often collects as much as a terabyte a day from each of its many sources. Putting the data onto tape backups is not ideal, as many researchers need immediate, fast access to a “hot copy” of the gene sequencing data they are analyzing.

“As scientific researchers acquire data at faster and faster rates, optimizing the analysis of that data with scalable storage solutions is essential,” said Dr. David Lifka, Cornell Center for Advanced Computing director. “Despite advances in disk technology, storing research data remains an expensive proposition,” he explained. “Ocarina provides a cost-effective way to maximize storage capacity without sacrificing performance.”

This is a field where technology is advancing very quickly, and the next generation of machines from leading companies like Illumina and Affymetrix will generate even richer data and require even more storage to hold that data. Because knowledge in the field is moving so rapidly, the value in the data may not be completely understood now. Keeping the data long term for analysis could hold great value for research. Unfortunately, as the amount of data generated grows, the burden on biomedical researchers to capture it and store it puts them at the center of a problem facing many parts of IT – coping with massive data growth.

For the most part, life sciences researchers are not storage experts, nor do they have a long history of running the world's largest data stores. They are being put in this position by the fast increase in the amount of rich data being generated by gene sequencers, ChIP sequencers, and other advanced technology. What's daunting is that the next generation



of analyzers, sequencers and other genomics technology will generate even more data.

In fact, storage is such a crucial piece of the puzzle that it is entirely possible that the pace of genomics research will be slowed by the inability of researchers to deal with the onslaught of data. This could mean a slowdown in finding cures and treatments to the world's most pressing medical crises like; cancer, heart disease, and many other diseases and conditions. Money to purchase storage, staff to manage it, data center space to keep it, and energy to power and cool it will all become important factors in overall research budgets – money that might otherwise be spent on research itself.

Backups Present Further Challenges

Another challenge with the overwhelming

amount of data growth is the strain it puts on traditional backups. When data comes in at 10 terabytes or more per day, backing up to tape the old way is becoming unfeasible. Data reduction with content-aware compression and dedupe offers other alternatives for data protection and retention. Once the primary copy of the data has been processed and reduced down to one-tenth its original size, it may make sense to create a replica of that data on another storage platform, rather than trying to back it up using legacy backup tools or tape.

The replica can be stored in another location, on cheaper storage than production data. This serves the purpose of protecting data and making all data accessible in the event of a data loss on primary storage, but uses a simpler workflow and one better suited for the volumes of data biomedical IT is now facing.

“Next generation sequencing techniques produce vast quantities of data that must be quickly processed and stored online for short periods of time,” noted Dr. Jaroslaw Pillardy, a senior researcher at Cornell University’s Computational Biology Service Unit. “For example,” explained Pillardy, “one Solexa sequencing run produces 0.5 terabytes of raw data, and a single sequencer may be used multiple times per week.” Pillardy expects that new sequencing techniques will soon generate data at a rate of 0.1 terabyte per hour.

These needs include such services as making it easy for doctors to bring up genetic data on patients while they are visiting a hospital or clinic, in real time.

As a result, one of the most promising areas of storage technology is data reduction for

New breakthroughs in content-aware dedupe and compression now makes it possible for genomics and other life sciences data sets to be reduced dramatically in size shortly after they come off the originating devices and have been analyzed.

life sciences data sets. New breakthroughs in content-aware dedupe and compression make it possible for genomics and other life sciences data sets to be reduced dramatically in size shortly after they come off the originating devices and have been analyzed. While generic compression can make some headway in reducing data set size, content-aware dedupe, algorithms that are aware of the specific types of patterns found in life sciences data, get much better results. For research data of this sort, it is also absolutely key that any compression be completely lossless. Data would lose its value if even a single bit were changed by the compression/decompression process. Therefore, compression technologies specific to life sciences and genomics and validated as lossless against specific biomedical test data sets have the advantage.

Ocarina Networks: The Cure for the Common Storage Problem

| File Type | Device Source | Savings |
|------------------|----------------------|---------|
| .dat, .arr, .cel | Affymetrix | 50% |
| .tiff, .txt | Illumina | 45%–85% |
| .fa | | 77% |
| .srf | Sequence Read Format | 40%–75% |
| .dat | | 60% |
| .img | | 55% |

Ocarina Networks is the industry leader in biomedical specific data reduction technologies. Ocarina's ECOsystem is a solution that inserts transparently as part of a data center's storage infrastructure. It is transparent to applications, users, and devices. The ECOsystem processes life sciences data files after they have been stored on disk, reducing them in size by as much as 75% on the first pass. When a user or application accesses the optimized file, the ECOsystem automatically expands the file, providing the bit-for-bit original data to the requesting user or application. The ECOsystem includes multiple life sciences-specific data compressors for the types of files most commonly found in the research environment. The following table summarizes some of the most important data types supported, but the overall solution includes over 100 algorithms that support over 600 file types.

Deduplication is the next phase of data reduction for life sciences data. Content awareness is key. While there are many data deduplication solutions for backup data, they get their results based on the repetitive nature of backups. You back up your data every day, and consequently, there will be a lot of duplicate data from day to day. Backup dedupe solutions find those duplicates and remove them, saving on the space needed to store your backups. For online data sets, duplicate information is harder to find. To find redundant information a solution must understand the data format of files it processes. In life sciences, there are repetitive patterns of information – whether base proteins, common sequences, or other patterns. These repetitive pieces, or chunks of information, might not be duplicate blocks of data on disk, but they are duplicate chunks of information. A

While generic compression can make some headway in reducing data set size, content-aware dedupe, algorithms that are aware of the specific types of patterns found in life sciences data, get much better results. For research data of this sort, it is also absolutely key that any compression be completely lossless. Data would lose its value if even a single bit were changed by the compression/decompression process.



About Ocarina Networks

Ocarina Networks is the leader in content-aware compression and dedupe for online storage. Designed specifically for online storage, the patented ECOsystem solution gets 5–10 times better data reduction results than generic compression or the back-up dedupe solutions.

Already installed at Fortune 500 companies and Top 10 web sites, the Ocarina ECOsystem easily integrates your existing storage and processes. We can create new free space on your existing storage, or give your next storage purchase 10 times as much capacity for the dollar.

Based in San Jose, California, Ocarina is privately-held and financed by leading investors.

Contact Us

General Inquiries: info@ocarinanetworks.com

Press: press@ocarinanetworks.com

Sales: sales@ocarinanetworks.com

www.ocarinanetworks.com

info@ocarinanetworks.com

408.512.2966

42 Airport Parkway, San Jose, CA 95110



content-aware deduplication solution can recognize some of these patterns and replace them with codes that represent the same information but are more efficient. To do this requires understanding a given file format – such as Sequence Read Format – and being able to look into SRF files, find the duplicate information, and process it to store it more efficiently.

If the first pass of content-aware dedupe and compression can achieve results of up to 60% data reduction on some life sciences data sets, dedupe can improve those results when applied across multiple sets of data. In a genomics, biomedical, or life sciences data archive, content-aware dedupe may take overall space savings over 60%.

Storage does not need to get in the way of research advances. Institutions can spend budget furthering science instead of on disk purchases and storage administration overhead. Going forward, as the data generated in biomedical research becomes richer and richer, and as new generations of equipment generate amounts of data that would be hard to imagine today, having the infrastructure in place to deal with data growth efficiently is going to be a cornerstone of biomedical data management. Ocarina Networks designed the ECOsystem solution specifically for this market. Ocarina is working with leading device manufacturers as well as leading research institutions to understand and analyze the most important current and future life sciences file and data types, and to deliver a complete end-to-end data reduction solution for biomedical and genomic research.

In addition to having production-ready solutions for today's data already installed in some of the world's best-known research data centers, Ocarina is also interested in partnering with leaders in biomedical research – labs, device manufacturers, and pharmaceutical companies. A goal is to understand forthcoming emerging technologies and to keep advancing the state of lossless data reduction for the biomedical industry. If you are seeing large data growth in your research environment, or are working on a data intensive new technology, contact Ocarina Networks to see how we can work together to address the cost and management of data growth in life sciences.