Bio·IT World Briefing On:

Next-Generation Data Storage

In this Briefing On supplement, we present a selection of stories that examine the technologies that are both creating and solving the next-gen data deluge ("The DNA Data Deluge").

Next-Generation Data Storage

The explosion in life sciences data in general, and next-generation sequencing in particular, is nothing short of remarkable. At the 2010 Bio-IT World Expo, the director of IT at a leading genome institute casually remarked that he had just ordered 1.1 petabytes of storage to try to cope with the extraordinary surge in data generation. This puts a premium on sophisticated IT solutions not only for storing data, with reliable and secure backup and retrieval systems. Organizing the data, much of which is generated and viewed once and once only, is a problem that will require more work in future.

In this Briefing On supplement, we present a selection of stories published in the past 12-24 months that have examined the technologies that are both creating and solving the next-gen data deluge ("The DNA Data Deluge").

Increasingly, the value of next-gen sequencing studies is being validated with the identification of rare disease genes in studies of small families ("Next-Generation Genome Sequencing Identifies Disease Genes"). It is possible that the arrival of third-generation sequencing solutions, such as that from Pacific Biosciences ("PacBio's Eleven"), will reduce the intense need to store image data, but IT managers aren't popping the champagne corks just yet.

One of the most interesting trends in next-gen data management is the potential of "on demand" Cloud computing ("The Dagdigian Doctrine"). Many genome centers and pharma companies are actively exploring the use of public and private clouds ("Warm Wellcome for Cloud Computing").

Handling next-gen data involves much more than sheer storage. Many organizations are developing innovative tools to federate and share genomic and clinical data ("Running tranSMART for the Drug Development Marathon," "No KIDding: Informatics in Reverse"). That even includes the U.S. Government ("Harnessing (and Securing) Meaningful Health Data").

One of the exciting trends in handling next-gen data is the arrival of proven IT vendors who have built their reputations in other industries and are now applying that expertise to the life sciences. Quantum is one noteworthy example, highlighted by the release in 2010 of the StorNext 4.0 system ("Quantum Leaps Into Life Sciences").

We hope you enjoy this Briefing ON supplement and continue to follow *Bio•IT World* and bio-itworld.com for the latest news on next-gen sequencing and data management.

Kevin Davies PhD
Allison Proffitt
Bio•IT World

The DNA Data Deluge

(Originally published April 2008)

On March 6, the employees of Expression Analysis (EA) in Durham, North Carolina, took delivery of the first ever commercial single-molecule DNA sequencing instrument, developed by Helicos BioSciences. Understandably, there was a lot of excitement when "that baby rolled through the door," says EA director of marketing Karen Michaelo.

The HeliScope weighs a little under 2000 pounds, and comes with a separate 32-CPU Dell PowerEdge server stack and a whopping 24 Terabytes (TB) of storage, which will be installed at EA by a pair of Helicos employees. "We're going to skin our knees for a while," says Michaelo, who anticipates offering sequencing services from the platform this summer.

During a tour of the Helicos production floor in Cambridge, Mass., senior VP of product research and development Bill Efcavitch stressed the platform's processing prowess. Customers are going to need it. Today, the HeliScope produces 25 Mb sequence per hour, but Efcavitch predicts a greater than tenfold increase within the next two years, putting the $1,000 genome firmly within reach.

University Challenge

The challenge facing next-generation platform users is how to manage the data glut pouring forth from 454, Illumina, and Applied Biosystems (ABI) instruments and the like. "Fundamentally what the community suffers from is, there's no best practices guide for setting up a lab with this equipment," says Geospiza founder Todd Smith.

For example, Smith estimates an Illumina Genome Analyzer produces 1500 times more data per run than an ABI 3730 capillary instrument. Even the major genome centers are scrambling. At the Broad Institute, a warehouse of 100 ABI 3730s produced 60 Gigabases (billion bases) of sequence last year, estimates director of informatics for sequencing platforms Toby Bloom. The institute's 20 Illumina instruments currently generate 20 Gigabases per week, and this could double in the near future.

"There's not even enough local storage on the data collection machine to complete a single paired end run, in the way we're using them, which means we have to be much more aggressive about moving data through the pipeline. So we've shifted to a pull model rather than a push model for moving the data along," says Bloom's colleague Matthew Trunnell, group leader of application and production systems. (See, "A Broad View")

A run on ABI's SOLiD instrument produces about two Gigabases, says ABI's Michael Rhodes, sounding somewhat awestruck. "I don't think I've really seen anyone who's first time with [SOLiD] who has not been somewhat overwhelmed by that amount of data. They may have rationally dealt with it in their head, but there really are challenges in just moving that amount of data around." Rhodes' team sometimes resorts to burning data on hard drives to ship between sites.

"A lot of smaller [organizations] might have only a few next-generation instruments, but they think about the science primarily," says BlueArc research markets director, James Reaney. "They have no ability to handle data of this size easily."

SGI's global higher education and research solutions manager, Deepak Thakkar, hears scientists complain about offloading Terabytes of data offline before they can start a new run. Says Thakkar: "Even if I don't keep 90% of the data, I still need to do the right thing before I decide to trash it. I don't need to keep every 5 or 10 TB, but I need to know what to keep!"

"The data glut is a huge problem," agrees Steven Salzberg, a bioinformatician at the University of Maryland. Salzberg says his lab (using Illumina instruments) "keeps the sequences and quality values, and throws away pretty much everything else almost right away. It's cheaper to re-sequence than to store the raw data for the long term."

With hundreds of next-gen instruments deployed in the past 1 to 2 years, and more platforms on the horizon, the deluge is just beginning.

Helping Hands

Illumina CIO Scott Kahn says the Genome Analyzer ships with "an instrument control PC that has adequate storage to collect all the images, so at the end of the run, they can be transferred and then processed with an offline pipeline that runs across a Linux cluster."

Kahn characterizes three broad groups of users. The most proficient customers, including genome centers such as the Broad, Wellcome Trust Sanger Institute, and Washington University, transfer data in real time as the runs are proceeding to gauge run quality and determine whether additional cycles are desirable. A second group of "quite sophisticated" users "will use mechanisms that we've provided to transfer data off the machine after the run has completed." The third camp doesn't have dedicated bioinformatics resources, and puts a premium on ease of use.

ABI offers a 12-core server with its SOLiD instrument providing 9 TB storage, which stores the images during the analysis and some of the results. "When you do the next run, you get rid of the pictures," says Rhodes. Problem solved!

Helicos' Efcavitch is realistic about the data handling challenges. "The amount of image data we're collecting is staggering," he admits. During the data acquisition phase, lasting a week or more, the HeliScope Analysis Engine cluster processes the image data in real time down to a sequence file. "To store that data would be just cost prohibitive. So we take the image data, we process it so all we're saving is 1 percent of the image data for diagnostic purposes and a sequence file."

The server stack can hold data from two separate runs. "We put enough room to save one run, start it up, transfer the data out, now you're ready for a new run," says Efcavitch. With some platforms, "you can't transfer off all the data, so they save all the data on a hard disk. You have to transfer that off [and] do that image processing."

Efcavitch says Helicos intends to make its software open source. Meanwhile, his colleague Kristen Stoops, director of informatics business development, is building a "Helicos IT pipeline" for customers. "It does require some effort on [the users'] part to think about who is going to use the data, and how it needs to be moved through various levels of access and storage," she says. (See, "Helicos' IT Pipeline")

Helping users to do that is a growing number of software firms and consultancies, who see an ever broadening niche to be exploited.

Core Solutions

Todd Smith's company, Geospiza, which builds IT infrastructures for core laboratories that provide centralized genomics and proteomics services applications, releases its new platform for next-gen labs this month and hopes to bring its first customers online shortly thereafter.

Core labs have to recognize that data management is for their users as much as them, says Smith. "Traditionally, core labs would send data or the researchers would download the sequence data, and use desktop software to analyze the data," says Smith. "Next-gen changes that. The first thing the core lab experiences is: How are we going to get the data to our researchers? How will they access the analysis tools and CPU clusters needed to consume the data we give them?"

And then there's the metadata. "Information about images, redundancy in datasets scattered in directories," says Smith. "So there's a lot of complexity within the data that people are trying to sort out."

Managing next-gen data is about collection and distribution, Smith reckons. "You have to be able to correlate the runs with different samples. One researcher might have one sample, another four samples." A next-gen instrument can "spit out a bunch of files. I need to link those files to the run and to the sample, and then make the data available to my end researcher."

Geospiza's FinchLab, "Next-Gen Edition," will be delivered as a hosted service (SaaS) or as a "drop-in" operational bundle that includes a LIMS for collecting the data, a Dell server, and a 7 TB Isilon clustered storage (scalable to 1500 TB) to make the data accessible. Seven TB may

A Broad View

The Broad Institute of Harvard and MIT is running 20 Illumina Genome Analyzers, three 454 GS FLX instruments in production, and three ABI SOLiDs, according to Toby Bloom. She manages the informatics pipelines for the Broad's sequencing platforms — old and new — for applications ranging from medical resequencing to epigenomics to pathogen genome sequencing.

Bloom says most next-gen vendors provide "fairly sophisticated pieces of software," much of which the Broad staff uses, including image processing, while also recommending improvements with certain vendors, for example on quality scoring. "We may come up with our own algorithms and feed that back to the vendors," she says. "Of course, for assemblies, alignments, mutation calling, we're looking at our own software as well."

Despite its considerable resources, Bloom's team has made sweeping changes to its data pipeline of late. "On the data management side, the old pipeline dealt with one read at a time," says Bloom. "Now, we deal with plate by plate or region by region or lane by lane. The data aren't stored in individual files but in batches." Another issue is that, "You're dealing with large numbers of small reads, not small numbers of large reads."

Store 24/7

Bloom says the core LIMS includes "added information about the new steps to help the lab track what their orders are. It's very different managing the lab to do large numbers of small projects. A mammalian genome would take several months to go through the lab using older technology… They now need more support for keeping track of everything."

To handle the storage demand, the Broad has 300 TB of Isilon high-speed parallel access storage, with more on the way. "We do a bunch of our work on SunFire 4500s, or Thumpers," says Bloom. These are reasonably inexpensive file-server units that have 15-20 usable TB per unit. "We actually use them to pull the images off the machines as they're being generated, so we don't have to stop the sequencers to do any processing on them between runs."

Bloom says the SunFires have "enough processing capability that we can do cycle by cycle processing." Once the image data are processed, the results are fed into the Isilon storage and core compute facility. Bloom says the images are stored "in case we need to go back to them, for a month or two. We leave them behind on the Thumpers — they never go anywhere else."

But even the Broad Institute can't store image files forever. "I don't think it's particularly useful; it's rare we'd ever go back to them," says Bloom. "What we do store forever is a sampling of the images on each run." Archiving a few images from each cycle enables troubleshooting of potential machine problems.

suffice for a year or so. "Once [users] get good at things and start getting creative, they start looking at 100 TB," says Smith.

The LIMS maps every step and order to a specific workflow. "You can't just park data on the instruments," says Smith. "Those files need to be moved to a central server, because the instruments have 10-13 TB storage, but when you think of all the image files to be processed, that storage is used up," says Smith. "The computers are only for the data processing, they're not part of the data management."

Smith advises moving the data to a data management system once, granting

Helicos' IT Pipeline

Kristen Stoops, Helicos' director of informatics business development, is building a federation of IT vendors to identify best practices that will help users manage the torrents of data they will be generating. Here she offers a glimpse of how those plans are progressing.

Bio•IT World: Why does the HeliScope require such an impressive computer server/Analysis Engine?
Stoops: Number one, the HeliScope Single Molecule Sequencer is very image intensive in its technology for generating sequences. We'll produce about 5 Terabytes of image data per day, which presents a daunting challenge from a storage perspective, and the perspective of moving the data from the instrument to external storage. So one of the things we're doing is to limit the amount of data our customers themselves have to store...

[Two,] the HeliScope Analysis Engine is a very high-performance image analysis platform and server that does the image analysis on the fly (and will delete those images as part of the process), and stores a digital representation of those images in an object table, which is roughly 1/10 the size of the full image pack. Every object in every image is represented in the object table... We strip out all the background, we're not saving that, but we are saving every object that we detect, then when we do our base calling, we apply some smart algorithms to figure out if something is an artifact or a real base. From there, we go ahead and form actual sequence data...

We also store 1% diagnostic images that we will save and make available to customers. Those diagnostic images represent random sets of fields of view for each of our channels on our flow cells. So if one of our customers wanted to do their own image analysis, base calling, and strand formation, they could test their own algorithms against those diagnostic images, because it does represent a full stack of images...

The 5 TB data is a full image stack. So for a HeliScope Sequencer run that represents 7 days, that's 35 TB data. So we'll save about 1 percent, or 350 GB data, far more tractable from a storage and management standpoint... It will also allow us to troubleshoot anything that went on during the run that we didn't expect.

How much data can the Analysis Engine store?
[There] is enough data storage for a minimum of two full runs worth of data. That's full runs, minus the full image stack — [i.e.] The diagnostic images, the 1 percent, the object table, the sequence data, and log files.... The data from that first run gets stored on the HeliScope Sequencer, and is completing processing while you start another run, and start saving that data onto the data store on the Analysis Engine. Before you start the third run, you take the data from the first run off the instrument, and delete that data from the Analysis Engine, thereby freeing up enough space to start the third run. So it makes it possible to run the instrument almost continuously...

How will you help customers manage all this data?
We're working with companies like BioTeam and GenomeQuest and a host of other vendors whose technologies can help mitigate the challenges of managing these data... This includes hardware vendors, networking technology, high-performance computing vendors, systems integrators, bioinformatics developers, and basic IT infrastructure companies, LIMS vendors, whose tools in an integrated way can support the pipeline needed to deal and analyze these large amounts of data...

Is your goal to name preferred solutions for these IT and computational needs?
This isn't about naming exclusive types of partners for our system. It's about validating this system... There are so many different factors in choosing the right storage environment. Today, one of the big drivers in choosing hardware is energy costs. You really have to understand the data lifecycle — how many people need to access data? How frequently? How quickly can the data be moved into persistent, less expensive storage, and then into archival storage mode?

What kind of data come off the Analysis Engine?
We've adopted the Short-Read Format as the HeliScope output. This format was developed by a workgroup started by people in the research and genome center community, including Sanger, as well as vendors of the short-read platforms, and is gaining adoption. It's been adopted as the standard for the 1000 Genomes project. NCBI has adopted it as the submission format for the short read archive. It all goes back to our belief in openness, and our open source strategy, and open format where nothing is hidden from the user.

Are potential customers ill prepared for the next-gen data deluge?
Speed every step of the way in dealing with this is going to be key in mitigating the data management challenges. Speed with efficacy — storing data, getting data off the instrument into a storage system, doing it in a manner that is lossless and not sacrificing speed for interruptions in data. Speed with which you can access the data. If you're a user, do you have the right data structures to support pulling out the data you need and just the data you need? Do you have the right tools to analyze the data and get the biology out of these sequences? Understanding just how much of the sequence data you need to accomplish your research goals but not overwhelm the whole data management pipeline.

customers access through a web interface to download data if necessary. "Once the data is in the data management system, cloud computing can play a big role in moving the algorithms to the data. We believe this will be far more practical and cost effective for researchers — that's our goal," says Smith.

For Canada's GenoLogics, there is an opportunity to find new users for its Geneus lab and data management system, which is primarily deployed in gene expression and genotyping settings. James DeGreef, VP market strategy, says that many of his customers have or are purchasing next-gen systems. "We took our core systems, and built it out to handle all the LIMS aspects of next-gen sequencing capabilities," says DeGreef. The goal is also to provide the bioinformatics capabilities of next-gen sequencing.

GenoLogics is currently developing its resource with the University of Pittsburgh core lab and the University of British Columbia Genome Sciences Center. DeGreef anticipates a full release later this year.

On a Quest

Geospiza recently joined ABI's next-generation software consortium. "We want to enable researchers to do a lot of downstream analysis," says ABI's Roger Canales, a senior program manager. "We provide [vendors] with data file formats and information about how to handle and process the data, to facilitate the development of these software tools."

Another member of the consortium is GenomeQuest. "Current IT architectures and components cannot be easily adapted to process the volume and scale of data," says president and CEO, Ron Ranauro. "The consequence is that next-generation sequencing is [otherwise] limited to a few leading organizations with the most advanced bioinformatics staffs."

Ranauro's company (see 2007 story) has spent several years developing a platform for biopharma customers to take advantage of the next-gen revolution. "Just as Google did for indexing the web, GenomeQuest assessed the existing IT components available for indexing and searching sequence data. The only way to achieve scalability and performance was to start from scratch," Ranauro says.

The GenomeQuest system manages the flow of data across large computer networks so that any one computer only operates on a small part of the search, while the central system coordinates the flow of algorithms and data and collates results. "The benefits are infinite, linear scaling of computation and fully managed data resources," says Ranauro. "We shorten the time it takes to turn next-gen sequence data into biological meaning."

The two main classes of workflows are reference-guided assemblies (or variant analysis) and all-against-all (for metagenomics and transcriptomics applications), which can be adapted to specific customer needs. "The system is open to allow scripted access or web linking," says Ranauro. "Either way, access to large scale computational and data management resources are completely virtualized. An all-against-all, metagenomics workflow takes about ten lines of code in our system."

Team Players

The BioTeam's managing partner, Stan Gloss, sees the next-gen field becoming 50% of his business in the next six months. "The marketplace is moving that quickly," he says. (See, "Next-Generation Sequencing Solutions," *Bio•IT World*, October 2007). BioTeam senior consultant Michael Cariaso says next-gen instruments provide basic software processes culminating in a directory of files, but "that's pretty much where every vendor is going to leave you. The more they do, the less likely it is to fit the way your lab works with data. Two next-gen machines sitting side-by-side have no knowledge of each other, and the vendor software does little to improve that."

And so BioTeam has developed a Wiki Laboratory Information Management System (see p. 12), which resembles a mediaWiki installation. Explains Cariaso: "Every time a [sequencing] run finishes, the same way it might write to a log file, it also logs the results into the Wiki. As we do that for multiple platforms, it becomes a one-stop shop for all devices as well as the pipelines you have in house. It quickly becomes an umbrella over every other workflow process and data source within the facility."

BioTeam has installed WikiLIMS with groups including Dick McCombie's team at Cold Spring Harbor Laboratory and Tim Read at the Naval Medical Research Center in Silver Spring, Maryland.

For a researcher working on a particular project, (s)he "hits one button on the wiki, it says, here are all the microarrays and next-gen runs you've done for this. It will build a nice table, with quality scores, matrices, graphs... It can look up projects and look across timelines. New data arrives automatically as it is generated. This provides centralized and up-to-date view of what's really happening in your lab. "

The WikiLIMS also provides the ability to integrate CGI scripting into the wiki. For example, Cariaso says the 454 software for base calling is good, but assembling multiple runs into a consensus requires someone to work at the UNIX command line. "We can come instead and say, 'Here's your project page and a list of all the runs you've done. Check which runs you want, hit the big red button, and launch the assembly, store the results back in the Wiki.' We can figure out the workflows in that environment and make them a single-button press."

Minding the Store

The rise of next-gen sequencing applications is fueling rabid demand for new storage solutions. Illumina's Kahn says the choice "is typically up to the specific environment, how coupled it is to Linux or Windows, the price point, the amount of storage needed, and the retention policies."

BlueArc's James Reaney sees the three key issues as storage bandwidth, computational analysis, and data retention/migration policies. "The architecture must be of robust enough design to solve all three bottlenecks," says Reaney.

As the Wellcome Trust Sanger Institute and Washington University in St. Louis can attest, BlueArc's Titan network storage platform provides workflow paralellization and easy upgrades. The new Titan 3200 platform doubles the storage of its predecessor (see p. 49), with 20 Gbps of fully bi-directional, non-blocking storage bandwidth. Reaney says the bladed modular architecture uses tiered storage, which can be configured to suit the user, offering a range of solid-state, fibre channel, and SATA disk, plus tape backup.

"The Titan 3200 is easily the performance leader," says Reaney. "Combined with a modular, upgradable architecture and now a maximum capacity of four usable petabytes, the Titan platform is also the most future-proof storage platform one could have."

According to SGI's Thakkar, most of SGI's deployment is with individual sequencing labs and pharma companies, rather than genome centers. Thakkar explains: "[Pharma] buys something for a particular program or application. Extra capacity is not something they usually like to keep on hand. They don't want to incur extra overheads."

SGI works with customers in three major areas. It offers a bioinformatics appliance to accelerate genomics and proteomics applications and high-performance compute power — hybrid computing with shared memory systems linked to clusters, where SGI can take sequencing information offline and use its large memory systems to compare against known databases. He says platform vendors would benefit if there was a "faster, more streamlined" way to take the data offline for high-throughput computational analysis.

But SGI's "mainstay is raw storage," says Thakkar. This could be three-tiered storage or storage to fit the needs of the data being produced. "Analysis is key to what you end up re-running: Do I go back and rerun the experiment, or do I just go ahead and save it and run another experiment? Having intuitive and efficient analysis systems are key," says Thakkar.

Thakkar anticipates more SOLiD installations this year, especially as users contemplate replacing some of their capillary instruments. He adds that "454 has done an excellent job of figuring out the entire workflow for the customer, maybe because they've been doing this for a bit longer. They've got many of the kinks figured out, especially on the storage end."

Scary Sequence

Asked about her biggest headaches, the Broad's Toby Bloom cites fault tolerance as the biggest problem. "We've always had a notion of a) storing our data for ever and b) having multiple places in our pipeline we could fall back to if we lost something. If something went down, we could queue up data behind it until the component was fixed. If something got corrupted, there was always a fallback going to the previous step. We can't do that anymore... We can't even afford to keep everything backed up on tape! All of that is the scariest part of all of this."

"As fast as Moore's Law is working to support us, we're still eclipsing it in our ability to generate sequence data," says Helicos' Stoops. Efcavitch, her colleague, hopes it stays that way. "With a simple improvement to our efficiency and error rate, we'll be at 90 Mb/hr, with the same hardware, simply changing the chemistry," he says. "And by increasing the density of molecules on the surface, we'll be at 360 Mb/hr, with the same imaging hardware." All within two years.

There could be an awful lot of drenched scientists and skinned knees by then.

Running tranSMART for the Drug Development Marathon

(Originally published Jan/Feb 2010)

If you sense a certain irresistible Lance Armstrong-like determination when meeting Centocor's Eric Perakslis, it is not a coincidence. Perakslis, the VP of R&D Informatics at Centocor R&D (and newly appointed member of the Corporate Office of Science and Technology at J&J) is a cancer survivor—he was diagnosed with stage III kidney cancer at 38—and has run marathons in support of Armstrong's LiveStrong Foundation.

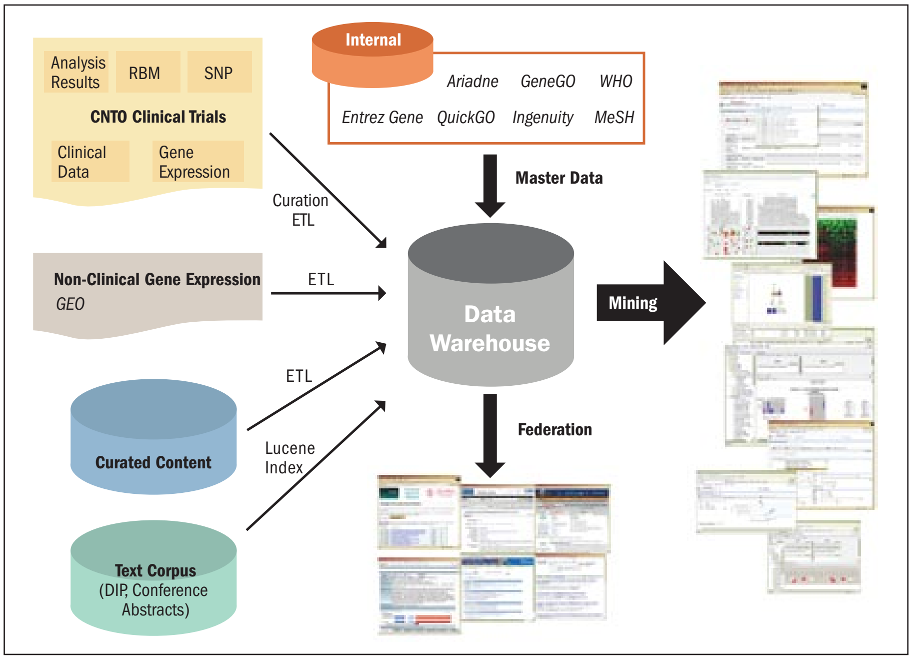
Once every other month, Centocor graciously lets Perakslis travel to Jordan where he volunteers as the CIO and head of Biomedical Engineering of the King Hussein Institute of Biotechnology and Cancer. It's a noble gesture given that Perakslis has his hands full at Centocor R&D—the biotechnology, immunology and oncology research arm of Johnson & Johnson (J&J). Trained as a biochemical engineer, Perakslis has spent the past decade in handling a variety of IT and lab responsibilities. "You've got to understand yours skills and weaknesses and then just kind of go where the fun is," he says.

Amazingly Advanced

The latest Perakslis project has been to spearhead the development of tranSMART (Translational Medicine Mart), a powerful new translational informatics enterprise data warehouse that went live at J&J in June 2009. TranSMART helps investigators mine drug target, gene and clinical trial data to aid in predictive biomarker discovery, chiefly in immunology and oncology. Perakslis says it is an "amazingly advanced" data warehouse that compares favorably with many such efforts he's seen in the pharma world.

TranSMART is a full translational medicine warehouse, says Perakslis. When he was considering the project, Perakslis insisted that a different approach was needed. "If I'm going to be the head of informatics, all the data has to be mine—I don't care where it is, what [J&J subsidiary] it's in, I don't care if its drug discovery or clinical or post-marketing—in order to do informatics well, all the data has to be in scope." Few pharmas follow that practice. "When I started recruiting, I could find some good bioinformaticists, but none of them had ever seen clinical data—because they weren't allowed!" It also meant changing the mindset that sees some R&D leaders more worried about their peers seeing, and potentially misunderstanding their data, than a rival pharma executive.

Breaking down those silos is like developing an internal informatics-without-borders ecosystem. "We take all the resources and capabilities and direct it toward the largest questions facing the organization." Instead of a traditional bioinformatician only supporting a process like target identification, Perakslis puts them all in one team to apply some scien-



TranSMART is a full translational medicine data warehouse, collecting a variety of molecular and clinical data sources suitable for mining.

tific and medical rigor to the key strategic decisions facing the development teams and R&D leadership.

The goal is an informatics package that combines all the different 'omics technologies and informatics disciplines to support or guide a decision. "It is still early days and, at first, we added a lot of value by ruling things out. I can't tell [the team leader] what indications to pick, but out of five disease indications we're considering for a novel therapeutic, I have evidence that three are a bad idea. How's that to start? We are then able to stop debating the five and focus on evolving the thinking about the two that were left."

Perakslis partnered with Recombinant Data to build tranSMART. He wanted a partner who knew how to assemble data warehouses. Head architect Bruce Johnson worked on clinical data warehousing at the Mayo Clinic, but before that, he worked on railroads. "He just understands data warehousing and doesn't care what sort of data it is," says Perakslis. The Boston firm has worked with Partners Healthcare in designing and building i2B2. "Technology-wise, tranSMART is very basic. We haven't brought in a lot of third-party software or heavy analytics. The system is about the data and the technology is based on open stuff. I don't owe people millions of dollars a year in contracts." The closest commercial competition is probably InforSense, says Perakslis. In addition, the system is the first production system from J&J to be hosted externally on the Amazon cloud.

"IT isn't always a profession that gets the respect it deserves. Sometimes, the MD on staff thinks he can do IT better than the IT staff, [but] I wanted to assemble a team that knew it down cold." For example, Sandor Szalma, the project manager for tranSMART, brings the needed mix of R&D knowledge, vision, software and social engineering experience and tenacity to the team. Perakslis is currently having discussions about a Red Hat approach where a lot of the advanced analytics is made open source. "I'm not running a commercial software company," he says. He's already offered the software to Guna Rajagopal, a colleague at the Cancer Institute of New Jersey. In addition, St Jude Children's Research Hospital, UCSF and other centers are all considering the adoption of the i2b2-based platform.

Unlike another prominent and excellent J&J informatics resource, ABCD (see, "ABCD: The Relentless Pursuit of Perfection," Bio•IT World, Nov 2007), Perakslis says is very focused on the chemistry and computational technologies, tranSMART handles large molecules and everything else, capable of providing access to every trial the company has ever run, and allowing clinicians to test hypotheses that enable the design of better clinical trials. In an effort to reinvent as little as possible, tranSMART integrates information from numerous existing vendors at Centocor/J&J, including Jubilant, Ariadne, Ingenuity and GeneGo. There are links to public repositories such as Gene Cards and PubMed as well as direct links to many key internal systems across J&J.

Perakslis has loaded several years of clinical trial data into tranSMART, which can be searched for screening data, lab data, you name it. "We have not had a lot of success in asthma yet but we feel the disease is strategic and we must keep trying. If the PI in asthma says, 'I just saw a new signature or a new gene in a paper. I wonder if that gene was amplified in any of the patients we've seen in our studies?' That's an instant search."

All kinds of data

The data warehouse can be searched by almost anything—compound, disease, gene, gene list, pathway, gene signature. Type in a gene, and a researcher can see every relevant clinical trial, gene expression datasets where expression of that gene was significantly altered, literature results, pathway analysis, and more.

Perakslis says he is incorporating all kinds of data into the warehouse. "When we add new trials, we use data diversity as a selection criteria. If a certain trial will bring a new and necessary data modality, great. We're growing the mart and growing the content." While it is meaningless and too expensive to do a whole genome on everybody right now, he says most trials are underpowered with respect to sampling for molecular profiling. "So what is just right? What is enough data from a trial to meet objectives but learn something about the biology or the compound that helps you design the next one in the event this one fails?"

Since tranSMART went live last summer, the accolades have been glowing, particularly from the clinicians and scientists who have changed their study designs because of the data warehouse. For example, J&J's Elliot Barnathan, executive director of clinical immunology research, says, "tranSMART gives the pharmaceutical physician/researcher the ability to use data already obtained internally or from the public domain at his or her desktop to ask and quickly answer questions with ease." Hans Winkler, senior director of oncology biomarkers, says tranSMART will transform the way his team analyzes data. "Cross-trial meta-analyses and combined pre-clinical and clinical data analyses are at our fingertips," he says. Direct cross-referencing with the literature and multiple visualization tools down-stream are all available instantly. This is clearly a quantum leap in our ability to extract knowledge from data."

A big reason for tranSMART's success is the social engineering—getting the datasets released, having a QA process, informed consent, etc. As a past CIO for Centocor R&D, Perakslis had written a lot of those policies, on records management, the FDA Gateway project, eCTD, and knew where a lot of that stuff was. But he admits: "The truth is, there's nothing you could do in this warehouse you couldn't do anyway if you had the time to do it."

Perakslis is trying to encourage the Google approach, spending a portion of one's time thinking creatively about something other than your job. "In pharma, when things are tough and people get busy, they tend to start acting like technicians. They run the models they know how to run... This is about efficiency and innovation. It does save time and money—if people are thinking differently and running better experiments, the quality of decision making should go up."

Perakslis praises J&J CIO John Reynders for being supportive and "getting out of the way and removing barriers, that's one of the best things he could do." As good a tool as tranSMART is, Perakslis says, "it's only good because we built all our processes around it. If you dropped that in an organization, if you don't have central agreement about data standards, data availability, compliance and protection of patient rights, it's not any use." ●

PacBio’s Eleven: Pacific Biosciences Announces Technology Partners

(Originally published February 2010)

Pacific Biosciences (PacBio), the next-generation sequencing company that intends to release the first version of its single molecule DNA sequencing instrument later this year, has announced a partner program featuring eleven leading technology providers. The Bay Area company believes that this program will help its future customers quickly adopt what it is calling a “third-generation” DNA sequencing solution into their research.

PacBio’s Chairman and CEO Hugh Martin said it has been a long-standing priority to establish “an entire ecosystem of complementary solutions.” In a statement, he said: “The overwhelming response from companies wanting to partner with us is extremely validating as we stand on the verge of introducing a breakthrough technology to redefine DNA sequencing.”

Members of the partner program, headlined by Amazon and featuring a spectrum of software, informatics, target-enrichment and sample prep providers (see below), will receive access to information and development tools, APIs, protocols, and potential co-marketing opportunities.

One segment not included in the program, at least for now, is the IT hardware/data storage community, but several members, including BioTeam, GenoLogics and Geospiza, serve as expert liaisons with the IT community. “PacBio has received immense interest from direct IT and storage providers that we will evaluate as the partner program matures,” said Edwin Hauw, PacBio’s senior product manager, software & informatics.

“We’re really excited,” said BioTeam managing partner Stan Gloss. “We think we provide a unique perspective to help PacBio integrate into existing environments and work with other systems that are already in place. I see our role as a system integrator and solutions architect, bringing together existing high-throughput sequence platforms, storage, HPC, and cloud technologies.”

Gloss added: “PacBio is trying to do something special here. This is all about building an ecosystem to support the platform.”

“Pacific Biosciences is changing the game with the real-time streaming of its sequencing data,” commented Geospiza president Rob Arnold. “This revolution is an ideal match for a cloud-based computing service and we look forward to working with PacBio to meet the informatics needs associated with this new era of DNA sequencing technology.”

CLC bio’s global head of PR and marketing, Lasse Goerlitz, said his company’s core strategy was to be an independent, cross-platform software provider, supporting all second-generation sequencers and now the emergence of 3rd generation sequencers. “We are very happy that we have been chosen as the first and presently only high-throughput sequencing data analysis software provider in Pacific Biosciences’ partner program,” he said. “We’re very excited about the prospects that single molecule real time sequencing brings.”

The full list of the eleven founding members of the PacBio program is as follows:

- Agilent Technologies – Provider of the SureSelect Target Enrichment System
- Amazon Web Services – Cloud computing, provider of the Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3)
- BioTeam – IT/informatics consultants with expertise in high-performance and cloud computing, storage, wiki media and DNA sequencing platforms
- Caliper Life Sciences – Tools and services for drug discovery/life sciences, including the LabChip XT fractionation technology
- CLC bio – Bioinformatics and sequencing analysis solutions, including CLC Genomics Workbench and CLC Genomics Server
- Fluidigm – Microfluidics solutions, such as the Fluidigm Access Array System for targeted enrichment
- GenoLogics – Informatics solutions for life sciences, including Geneus, a next-gen sequencing LIMS and data management system
- GenomeQuest – Sequence data management (SDM), provider of the GenomeQuest SDM platform and cloud services

- Geospiza – Cloud computing solutions for genetic analysis, such as the GeneSifter Lab and Analysis products
- NuGEN Technologies – Sample preparation, provider of the Ovation whole genome amplification, whole transcriptome amplification, and RNA-Seq kits
- RainDance Technologies – Microdroplet-based solutions for single-molecule and single- cell analysis, provider of the RDT 1000 system and sequence enrichment kits

PacBio has announced that the commercial launch of its first sequencing instrument will take place in the second half of 2010. The company says its SMRT Sequencing System will ultimately make it possible to sequence individual genomes as part of routine medical care, but the platform will enable many other applications in agriculture, clean energy, and global health. ●

Next-Generation Genome Sequencing Identifies Disease Genes

(Originally published March 2010)

Two studies published this week show convincingly that whole-genome sequencing of individual patients or affected families can reveal the one gene out of some 25,000 in the human genome bearing a deleterious mutation.

Writing in the New England Journal, Baylor College of Medicine's Jim Lupski, Richard Gibbs and colleagues show that by sequencing the whole genome of an affected individual—in this case Lupski himself—it is possible to identify the rogue gene for a recessive disease by filtering the variations in the coding genes to focus on just those that are novel and predicted to cause a significant phenotypic change.

Meanwhile, in Science, researchers at the Institute of Systems Biology (ISB) identified the gene for a rare Mendelian disorder called Miller syndrome by sequencing a family of four (parents and two children).

Jim Lupski is a widely respected clinical geneticist who played a key role in the interpretation of the genome of James Watson, published in 2008. Lupski is an expert on structural rearrangements in genetic disease, documenting one of the first such cases back in 1991. That disease was Charcot-Marie-Tooth (CMT) disorder, a peripheral neuropathy, which happens to affect Lupski and other family members.

Lupski disclosed the results of his personal genome sequencing project in a talk late last year at the American Society of Human Genetics convention. Despite studying CMT for more than two decades, Lupski could not attribute his disorder to any of the dozens of known CMT genes. Finally, Baylor genome chief Richard Gibbs offered to conduct a whole-genome sequencing experiment.

After the sequencing, using the SOLiD platform from Life Technologies (estimated cost less than $50,000), Lupski had to narrow the search. No copy number variants were found to affect any of the dozens of known CMT genes. But by focusing on coding variants in 40 known neuropathy genes, the errant gene, SH3TC2, was identified. One of the mutations appears to be associated with carpal tunnel syndrome.

While the Baylor team concedes that whole-genome sequencing was probably not essential in tracking down the faulty genes in this particular case, they conclude: "As a practical matter, the identification of rare, heterogeneous alleles by means of whole-genome sequencing may be the only way to definitively determine genetic contributions to the associated clinical phenotypes" in complex diseases.

Miller Time

The ISB study published in Science employed the human genome sequencing service of Complete Genomics. The results were among the first 14 genomes that Complete Genomics undertook. The two children in the family suffered from a rare craniofacial disorder called Miller syndrome, as well as a lung disorder called ciliary dyskinesia, which resembles cystic fibrosis. Neither parent had these conditions, suggesting the possibility that recessive mutations in two unrelated genes could be responsible.

Complete Genomics sequenced the four genomes to a depth of 51x to 88x, and called around 90% of the bases in each genome. Variations in four genes were consistent with recessive inheritance of rare variations. In a separate study on the same family, another Seattle group led by Michael Bamshad and Jay Shendure recently reported one of those genes, DHODH, as the cause of Miller syndrome. And mutations in one of the other candidate genes, DNAH5, were previously associated with ciliary dyskinesia. "We are convinced that this new kind of analysis, family sequencing, will be a remarkably powerful scientific and medical tool in the future," said David Galas, senior vice president of ISB.

Complete Genomics CEO Cliff Reid said his goal was "to provide large-scale complete human genome sequencing as a service that would enable our customers to make medically-relevant discoveries. We are delighted that ISB is already making breakthroughs of that caliber from its first study using our service. This is the type of positive disruptive influence that we want our technology to have on medical research."

Question Time

In a commentary accompanying the Lupski et al. New England Journal study, Yale University and Howard Hughes Medical Institute investigator Richard Lifton said the rapid progress in whole-genome sequencing raises profound medical and societal questions:

"Who will benefit from comprehensive sequencing? When in a person's life should sequencing be done? How should we deal with the many variants of uncertain clinical significance? How should we interpret changes found outside of genes? How should we effectively communicate the results to patients in ways that will improve health without inducing neurosis?"

Such questions are taking on added urgency as the trickle of published human genomes is on the brink of turning into a torrent. Another question is whether the early success in identifying Mendelian disease genes via whole genome sequencing will be repeated as these and other groups tackle other examples. ●

Harnessing (and Securing) Meaningful Data

(Originally published May 2010)

"There isn't going to be some massive database in the basement of the White House run by Sarah Palin," promised John Halamka, the CIO of Harvard Medical School, in his keynote at the Bio-IT World Expo. But there will be a "federated mechanism that enables us to send data from place to place for a whole variety of purposes for care and research."

Halamka serves as the Chair of the US Healthcare Information Technology Standards Panel. Of the $30 billion allotted to health care IT in the Obama Administration's stimulus package, most of it will be distributed to hospitals and clinics after they've put health care IT infrastructure in place and are using it wisely. The remaining $2 billion is being distributed by the Office of the National Coordinator for health care IT advances.

"Here's the strategy," Halamka said. "Give $2 billion in grants to accelerate the industry. Give the industry a set of standards that are unambiguous for everything from medications, to labs, to quality measurements, to both clinical care and population health... Declare how hospitals and doctors have to use this wisely, and then certify products as being good enough to have the features and functions and capabilities to make this whole thing work."

In the next five years as these standards are put into place, doctors and hospitals will be required to collect "meaningful" data and protect that data. "This is not using a word processor to record data!" Halamka clarified. "This is actually using codified mechanisms so that if you capture medications, problems, allergies, labs, etc. You could use them to inform drug discovery."

Meaningful data include requiring all orders to be electronic; recording medication and allergy lists for all patients; and recording and updating demographics and vital signs in a timely manner—all using consistent and controlled vocabularies.

There are rules that still need to be clarified, Halamka says, but gathering meaningful data will begin to enable smarter health care. Offices and hospitals will be able to mine their data and send targeted wellness reminders for preventative care. Patients will have access to a continuity of care document, delivered electronically, explaining treatment history, prescriptions, and diagnoses.

But with this wealth of data—and wealth of possibilities—comes a huge security responsibility. Halamka stressed that laptops and thumb drives must be encrypted. Patient privacy must be protected.

"I spend about $1 million a year just protecting the Beth Israel Deaconess [hospital] records against the nefarious internet. We're attacked every seven seconds, 24 hours a day, seven days a week," Halamka said. "Half of the attacks come from Eastern Europe; half of the attacks come from Eastern Cambridge. Every September, 1200 new hackers arrive – they're called freshmen!"

Emerging Data

Although about 20% of clinics and hospitals currently have electronic health records, all should by 2015. Public health applications are already beginning to emerge. Health data gathered can be aggregated regionally to look at public health trends or build doctor report cards. The SureScripts (representing pharmacies) and RxHub (representing payors) database includes de-identified drug information on 160 million people that can be used to check for drug interactions.

In another example, the Social Security Administration used to spend $500 million a year getting paper records. After moving to electronic records two years ago, a disability claim that took months to adjudicate can be handled in 48 hours.

There are even opportunities to gather rich patient data in the home. Halamka is testing a bathroom scale that calculates his lean body mass and body mass index and transmits the data via XML in real time to Google Health and Microsoft Health. But although an avid blogger and registered technophile, he said he declined the Twitter reporting feature. ●

No KIDding: Informatics in Reverse

(Originally published May 2010)

It is unlikely that too many software-as-a-service companies feature quotes from T.S. Eliot on their web site, but *Parthys* Reverse Informatics, based in Chennai, India, is the exception:

"Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?"

Parthys Reverse Informatics is the brainchild of Parthiban Srinivasan, a computational chemist by training. After various stints in academia and industry, Parthiban defined what he saw as a problem with the current approach to data analysis in life sciences:

"We're not working on bioinformatics, we're creating chaos! Everybody's worried about having too much data, but we're delivering data from existing knowledge. So informatics goes from data to information to knowledge. We [*Parthys*] go from knowledge to information to data. It's reverse informatics."

Parthiban worked for NASA in the late 1990s, applying his computational chemistry training to analyzing molecules in outer space. He later moved to the Weizmann Institute in Israel, working with Jan Martin, a recipient of the Dirac Prize. From Israel, Parthiban joined AstraZeneca back in India, working on pharmacophore and virtual screening for tuberculosis (TB). "I became more of a bio-IT person," he said.

In 2002, Parthiban joined Jubilant Biosys, which like other Indian bioinformatics companies was "facing the tempest of the industry to succeed," he recalls. "The Indian companies built products, but the products in the international market were superior and available for free. Nothing was moving." Parthiban had the opportunity to travel alongside some consultants from McKinsey. "They knew strategy, not science. I was living in a plane for a year." Jubilant turned around, and within a few months, Parthiban was tasked with recruiting some 400 scientists. "The key to success: Instead of focusing on pure bioinformatics, include chemistry," he said.

However, Parthiban questioned the business model, which saw scientists sitting in front of expensive computers designing compounds. "Why should a pharma company give you a biological target and ask you to design a compound?" he asks rhetorically. "It's such a sensitive matter. There's no business there. We were investing in too many Silicon Graphics machines!"

Instead, he chose to focus on small molecules with big opportunities. "We gathered all related information, built thematic databases and moved on the value chain. It was a phenomenal success." Jubilant was competing with the likes of Ingenuity, but with a greater focus on chemistry. "From the beginning, for 25-30 years, chemistry-based databases were all privately held, whereas life sciences databases are government supported. We couldn't make money out of it. We have to make money—sustainability was the question."

After three years, Parthiban joined another Indian informatics company, GVK Bio. He got an opportunity to make a new turn from informatics to patent analytics and he developed market intelligence reports for pharma research. An example was looking for ways to synthesize compounds without infringing intellectual property (IP). "It worked surprisingly well! What a success," he said.

Finally, Parthiban decided to strike out on his own. After tasting success in informatics and patent analystics, he moved to social networks, specifically building networks of key opinion leaders. His team worked with an American marketing firm to build a large database of some 20,000 key leaders in breast cancer for a major pharma company. Part of that project involved disambiguation of scientist's names. "We developed a web-spider that fetches relevant information from various sources, then did a 'de-duplication.'"

Reverse Flow

Parthiban coined a pet phrase to describe his company's business model: "KIDding is our business: Knowledge to Information to Data." The goal is to move from unstructured knowledge to structured data, he says. Computational science compliments experimental and theoretical approaches to science. "What is common in all the three modes of science is that you start with a question: you collect data, find an answer. Today, it's the reverse. You have answers everywhere. The secret of doing science is asking smart questions."

For example, say you want to know all the compounds that interact with Jun protein kinase. "You already have the answer, it is already there," says Parthiban. "But you are not able to access the information… If you hear a dog barking and it is raining, you might conclude that whenever a dog barks, it rains—that is a scientist's bias. We eliminate this bias and pull out the data from different studies—the naked facts."

Another buzz word in pharma, says Parthiban, is “interoperability” or “actionable data.” Finding data from one study that will interact with data from another is where the most value lies. “This is what we’re facilitating. Knowledge—Information—Data.” The key elements of reverse informatics, he says, are: Extraction; normalization; de-duplication; and integration.

The name of the company relates to India’s strict regulations over company names at the time of registration, which either had to include the name of the owner and/or describe what the company was about. The complete name of the company Parthiban was obliged to register is “*Parthys* Reverse Informatics Analytics Solutions Private Limited.” (Parthiban thinks he might be a contender for the Guinness Book of World Records.) Eventually, the Indian government relaxed its rules, but Parthiban was already promoting the name.

Networking at conferences, including the Bio-IT World Expo, Parthiban started to win contracts from some small companies. He weathered the economic crisis in 2009 and has built the company up to 100 staff.

Parthys focuses on informatics, social networking, and IP. For example, patent search services are now done by *Parthys* in the Cloud, no less. In another project with Collaborative Drug Discovery, led by Barry Bunin, *Parthys* built a thematic database capturing biological assay details, activity data and chemical information with a focus on disease. Part of that work involves 25 chemists “sitting in India looking at life sciences and chemistry information from patents across the globe in languages we don’t read.”

Another new project is construction of a next-generation pharmacodynamics database, which will be *Parthys*’ first product. “A leading PK/PD software company is inquiring about this. Fingers crossed”

Parthiban also uses the term “*in litero*” (in line with *in vivo*, *in vitro* and *in silico*) to describe his belief in the value of literature searching. He recounts a story of a big pharma who filed a patent application on a new technology with the U.S. Patent and Trademark Office. “They said, ‘Bad news, for innovation, there is already a patent. The good news: you’re the owner.’ They didn’t know they had the patent! It’s not just what you’re competitors are doing. Tracking back is the biggest problem. You don’t know what your company has already patented.”

Big pharmas don’t know how many compounds they’ve patented, and nor do the scientists. “I would build a beautiful database,” for one of them, he says. ●

The Dagdigian Doctrine

(Originally published May 2009)

In the opening keynote of the 2009 Bio-IT World Conference & Expo, Chris Dagdigian delivered a candid assessment of the best, the worthwhile, and the most overhyped information technologies (IT) for life sciences.

Some formerly hyped technologies were now mainstream, including virtualization and storage. Others were over hyped but still on balance worthwhile, such as green IT and utility computing. And in the future, Dagdigian saw major benefits accruing from trickle-down best practices and federated storage.

Dagdigian, a founding partner for the BioTeam IT consultancy, is a regular speaker at the annual Bio-IT World Expo. This year, he brought his trademark "trends in the trenches" talk to the opening plenary session. Offering his customary disclaimers, he said he was "comfortable with deliverables, but not comfortable with being a talking head or a pundit." The audience begged to differ.

Already Mainstream

Dagdigian divided his talk into three sections: reviewing what he called "old news," discussing "currently exciting" technologies, and those that could be exciting in the future. But, he warned, there is "a ridiculous list of stuff on the 'hype meter.'"

In the "already mainstream" category, Dagdigian tagged virtualization and the "bio-IT storage tsunami." Most of his recent work has been in helping "people getting buried by instruments." The community had been talking about the "data deluge" for the past four years. It is "still buried," said Dagdigian, "but the problem domain is understood. Our smallest customers aren't losing hope, and our biggest customers are staying ahead."

Last year, Dagdigian saw the first "100-Terabyte (TB) single namespace project." But, he cautioned, he had also witnessed for the first time a "10 TB catastrophic data loss" with consequent job losses. The loss, which occurred in a government lab, resulted primarily from double-disk failure in a RAID5 volume holding SAN F5 metadata. Dagdigian said he used to be a "huge fan of RAID5," but abandoned it last year. "The statistical probability [of failure] is too high—it's going to happen!" Everything is now RAID6, with maximum attention paid to monitoring and maintaining data integrity.

In a welcome development, data triage discussions are spreading beyond cost-sensitive industry organizations. "Everyone has come to the realization that data triage is a given," said Dagdigian. Displaying his favorite slide, Dagdigian showed a screen shot of an 82-TB folder on a Mac. Even more impressive was a 1-Petabyte (PB) output mounted on a Linux system output.

While storage is cheap (and getting cheaper), operational costs, staff, tape and backup costs are still fairly constant. Users, enamored with the plummeting price of TB storage devices from Costco, do not understand enterprise IT and backup requirements. "IT organizations need to set expectations, because the electronics market is skewing expectations," said Dagdigian. Many next-generation sequencing grants simply didn't budget for storage, let alone a 100+ TB storage system. Meanwhile, Dagdigian noted, the Broad Institute already has more than 1 PB of storage.

"Unlimited data storage is over," noted Dagdigian. It's simply not possible to back up all data, keep it safe, secure, and so on. "Sometimes," he said, "it's better to go back to the -40 F freezer and repeat the experiment."

In short, storage is no longer a major bottleneck—rather, that falls to chemistry, reagents, and human factors. Customers are starting to trust instrument vendor software more. "The problems are not as scary as they once seemed," he said. Dagdigian also noted that storage devices are running more 3rd party software, such as Ocarina Reader software on Isilon, and Ocarina Optimizer on BlueArc.

Virtualization offered the "lowest hanging fruit," said Dagdigian. The tipping point, Dagdigian said, was the live migration of a VMS (virtual memory system) without requiring a proprietary file system underneath. In 2009, he helped build and design a virtual collocation facility for an academic west coast campus, which was experiencing limits imposed by electrical power and air conditioning. "400 servers are currently virtualized on a lightweight simple platform," said Dagdigian. Large numbers of physical servers have been shut down, realizing significant savings from de-duplication, compression and thin provisioning, not to mention electricity. Moreover, scientists now have full administration control.

Coming soon, said Dagdigian, "virtualized cluster head nodes." Not coming soon: grids and clusters distributing entire VMS for task execution. It isn't practical, argued Dagdigian but rather a case of "marketing winning out over practical stuff."

Green IT

In the category of "hyped beyond all reasonable measure—but still worth pursuing," Dagdigian said green IT could deliver real electrical savings. "Use green IT for political cover," he urged the audience. A deployment of a Nexan SATABeast had led to a 30% reduction in power draw with no impact on cluster throughput. One of his best 2009 moments came in deploying a Linux HPC cluster for a west coast organization. The system interface talks to Platform LFS, powering nodes up and down and sending automatic email alerts to management such as: "Hello, I've saved $80K in facility costs this year."

Utility (or cloud) computing is "not rocket science, but fast becoming mainstream," said Dagdigian. "Amazon web services [EC2] is the cloud," he said. "It's simple, practical, and understandable," and enjoyed a multi-year head start on the competition. The rollout of features are "amazing," such as Hadoop and applications for short-read sequence mapping. "I drank the EC2 Kool-Aid: I saw it, I used it, I solved real-world problems," said Dagdigian.

The biggest problem is ingesting data into the cloud. "There is no easy solution," he said. How does one push 1TB/day into Amazon? "Have patience," said Dagdigian, expressing "100% confidence" that Amazon is working on the problem. "If we can solve data ingestion problem, I see a lot of scientific data taking a one-way trip into the cloud. Data would rarely, if ever, move back. If I take it back, it's going to be really obnoxious—big data pipes or people driving minivans of USB drives."

Worth Watching

In his "worth watching" category, Dagdigian cited federated storage and the trickle -down of best practices from Amazon, Google, and others. Amazon, Google, and Microsoft had all been computing at such a scale that it was too much of a trade secret, he complained.

But there are signs, such as a recently released video of the Google data center circa 2004, that their best practices will eventually trickle out, benefiting the entire community. ●

Warm Wellcome for Cloud Computing

(Originally published November 2010

With a sequencing output approaching 500 Gigabases/week and electricity consumption a whopping 3 MegaWatts, the Wellcome Trust Sanger Institute (WTSI) is a prime contender for cloud computing. Guy Coates, group leader of informatics systems at WTSI, heads a quasi internal consulting team that sits between the hardcore systems teams and the research scientists. Kevin Davies caught up with Coates at Bio-IT World Europe to discuss the institute's early experiences with cloud computing.

Bio • IT World: What is the attraction of cloud computing for the Wellcome Trust Sanger Institute?

Coates: We have these very 'spiky,' very agile, very diverse workloads. We get ambushed when data arrives with very little notice. So one of the things we've been looking at is, can we use the cloud as emergency compute? Also, the cloud as a remote data store is interesting—we've got so much important data now that we need to think about data recovery and continuity. Traditionally we stuck things on tape and sent them to a warehouse or bunker somewhere. But the economics of doing that on tape now don't stack up. If it goes into the cloud, we can do really cheap long-term data storage because of economies of scale.

Third is data sharing. We're a net data exporter—we make all our data publicly available. But getting to it on a remote network is quite difficult. We collaborate with BGI-Shenzhen on the 1000 Genomes Project data, but moving those data backwards and forwards without resorting to a truck full of hard disk drives is difficult. If we need to make data available, do we park it in Amazon and use them as a content distribution network, because they have data centers and we can ask them to replicate data? Rather than people having to get to us, maybe they just need to have fast data paths to their local Amazon data center, which might be a more economical way of doing things…

We started doing this on a small scale. Amazon has a program to make public datasets available. If you want to compute across the Ensembl genome dataset, that's all there in Amazon, so when you spin up your virtual machines, you don't have to worry about downloading a copy of Ensembl and uploading it back into Amazon—it's already there. That model is quite appealing.

What's the biggest IT challenge given the explosive rate at which you're churning out data?

Everyone has the standard exponential graph, which tracks sequencing output, disc storage, compute cores. The biggest IT headache is access to compute cycles and storage together. You've got the data pipelines, such that we don't keep data in the early parts of the pipeline longer than it needs to be processed. But the disk pool is constant—I might need all 5000 compute cores for a meeting tomorrow, with hundreds of other users all queued up. Trying to cope with that is hard. We've been looking at some feasibility studies—can we take the Illumina pipeline, take the raw data set, chuck it on Amazon, compute it, and get the same answers as we get internally? Can we get it to run in a sensible amount of time? Can we do the downstream analysis? Can we align the data against a reference genome?

Is your use of 'the cloud' synonymous with Amazon Web Services?

Yes, Amazon primarily, but we've been looking at other providers as well. We've been using Amazon's hub in Dublin because it's reasonably close to us in Cambridge, U.K. We go across the public Internet—JANET, the U.K. academic network. In theory we have a 2-Gigabit dedicated link onto JANET, which has very fast links between the various network hubs. Trying to get to Dublin, you end up peering out through Telehouse, London, and onto the public Internet, many different network providers. We found the big limiting factor is getting data in and out of the cloud. We realized 5-10% of our theoretical bandwidth. It's really hard to trace down what's going wrong.

The first step is to use data distribution tools, which know about wide area networks (e.g. the standard tools (scp/rsync)). They really don't work well across the wider Internet. There's a lot of work coming out of the grid communities on more intelligent software. There are also commercial companies who will sell you software to solve these problems. Even using Grid FTP, it's better, but we're still not seeing the performance we should. It's a complicated software stack as well… Finding easy ways to move data to other institutions that haven't got dedicated IT staffs is going to be really challenging.

We're chasing down this problem, but how do you fix the Internet?! We have control until it goes outside our borders, and Amazon has control of its space, but in the middle it's really hard to track down who is responsible for that piece of network infrastructure.

What have you done in the cloud so far?

We've taken an Illumina dataset, run the analysis pipeline and done the alignments. Getting the actual software running inside the cloud, we got it to work eventually but it wasn't trouble-free. We had a very good collaboration with Amazon and their technical people… The big problem, unsurprisingly, was that the same problems we have with I/O and disk inside sequencing data centers, you basically have in the cloud. The storage infrastructure behind Amazon is quite different from what you might find in a data center. We have these big Unix file systems, whereas in Amazon, you can create these Unix systems, but it doesn't go particularly quickly or scale well. We have problems once you go above 8 nodes trying to do traditional NFS client-server operation, it really doesn't work.

Amazon's response is: 'Don't do that, then.' They provide a different storage model called S3, which is web interfaces where you pull blocks of data. You don't have a Unix file system. It's much more scalable, and you get good performance across lots of nodes. It's not traditionally how HPC applications have behaved. We spent time taking all the code that expects to read and write to a file system and get it to talk to this S3 storage layer instead. That way, we were able to get the [desired] performance. There are all sorts of tricks to make that migration easier, if not as efficient. It's a trade-off between time to rewrite your code and time to just run the stuff in a slightly inefficient manner.

Are there any security concerns with your data in the cloud?

As far as data security goes, Amazon will give you the earth—sets of legal t's and c's saying what they'll do to the data and who it's invisible to, you can request to lock down sets of machines that can't talk to each other, and get firewalled off. If you need something tighter beyond what they offer by default, because of extra regulatory targets say, they're open to discussion. For most of our stuff and the things we're thinking about, our stuff will be publicly available anyhow, so putting it in the cloud is an easy jump for us.

Fair to say you'll be doing more work in the cloud in the future?

We're still feeling it out. Penguin Computing has Penguin On-Demand—a slightly different model. Instead of having virtual machines, you get time on a big cluster with a big cluster file system behind it… If you really want fast dedicated networking, they have one or two datacenters [in the U.S.], but for the moment we just ship them hard discs. The nearer you are, the more likely you are to get a more dedicated link. If we wanted to pay for a dedicated link into Dublin, we could do that. But those things are not cheap—you'd have to lease fiber from a telecom company. The great thing about cloud and computing as a service is that, if cloud provider A is suddenly cheaper than cloud company B, then you can just migrate everything across. But if you're tied to the physical infrastructure, you've changed from an on-demand model to a long-term partnership. The physics of communications may be that you just get forced down that road anyhow. If the data is just too difficult to move, do we just spin up more compute services to allow people to run VMs to compute against our silos of data? That's an interesting question. ●

Quantum Leaps Into Life Sciences Data Management With StorNext 4.0

(Originally published January 2010

With one eye firmly on the soaring data demands in life sciences and next-generation sequencing, San Jose-based Quantum Corp. has released a version 4.0 of its scalable StorNext data management software, which it believes offers a solution for the storage needs imposed by the remarkable growth in unstructured data and rich file formats.

Quantum provides data storage solutions for a variety of industries, including digital media, oil and gas, financial and the sciences. The major focus is on data protection, archiving, backup and recovery, using virtual tape libraries and other products to share high-speed content for files on disk, with petabyte volumes par for the course.

The StorNext product line, which has been around for about a dozen years, started as a high-speed, shared SAN file system and as the product evolved, transparent data movement for tiered storage and archiving was added with the Storage Manager option.

According to Shawn Klein, director of software partner development, Quantum views the life sciences is an "adjacent market" but Quantum hasn't needed to make many changes to the product. The move beyond media and entertainment is relatively new, says Klein.

Quantum's strength is in managing data storage, from disc-based and tape libraries used for data backup to tape libraries used for archiving. Klein says the "secret sauce" is that the software works across multiple platforms, compatible with most types of disk drive, tape libraries, and operating system.

What Happens Next

The StorNext file system is a heterogeneous, high-performance file system for large volumes of data. It makes the SAN storage look like a local hard drive. "You just point the application in Windows to the D drive and it looks like a local hard drive to that box," says Klein. "We don't care whether it's a satellite image or MRI. We know how to handle images real well."

The storage engine moves data automatically. Users can set whatever data policy they like: for example, move any file in a certain directory into a second tier, or move to tape if unopened in 10 days or 30 days. "After you render an image for a couple of weeks, you can move it to a lower tier of storage. Because that file isn't being accessed with the same frequency, you can save cost and better manage your environment by moving to a lower tier of storage," says StorNext product marketing manager Chris Duffy. "And yet, the applications and users don't know the files have been moved. It looks like the file is still in your D drive."

This provides flexibility for long-term archiving, providing a large number of distributed targets. Klein says that whereas competitors have to introduce a third-party data mover, StorNext has it embedded. "There's no third-party application to plug in to lift the information and place it somewhere else," says Klein.

Among the new features of the version 4.0 is support for file-based replication from one environment to another. The data deduplication feature is fully integrated resides in the file system. As Klein explains, "you can drop a file into a "dedupe" directory, and once it is there, the engine takes over and deduplicates the data."

Klein calls the data dedupe feature nothing short of revolutionary, like "compression on steroids." Says Klein: "Dedupe has been hot and putting it into a primary file system is a really big deal. Replication has been around for a long time, and we've been using third parties for that functionality." Now, the policy engine for replication drives the dedupe feature. There's a lot of cool use cases for that combined functionality."

The new version of StorNext also has a web services GUI. "Managing petabytes of data isn't easy, so the GUI is simplified, providing monitoring and web services XML available for application partners," says Duffy.

So far, the most visible life science case study is at Baylor College of Medicine (BCM). Perhaps not coincidentally, Baylor is based in Houston, where Quantum has a strong presence given its relationship with the oil and gas industry. Quantum won the Baylor deal last summer after a relatively short sales cycle. Quantum partnered with storage company DASDI, which originally identified the opportunity for StorNext. Quantum has also done deals in Singapore and Austria.

"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 BCM Human Genome Sequencing Center. Morgan expects the BCM genome center's storage requirements to push past two PB.

"We can scale performance and capacity independently. If somebody needs to add more sequencers but their capacity remains unchanged, we can help them increase performance without adding mandatory capacity," says Klein. "Similarly, if you just need another 1 PB data, you don't have to buy additional appliances, you can just grow capacity. We're vendor agnostic when it comes to disk. If the customer loves HP today but wants to put in EMC or IBM, we can put those all together and make them look like one big hard drive and virtualize that. So we really help the customers keep costs down." ●

Quantum®

Q&U CASE STUDY

StorNext & Scalar Series

It's All About DNA: Genome Sequencing Center Relies on StorNext Data Management

There are few tasks more data intensive than sequencing the 3 billion chemical building blocks that make up the DNA of the 24 different human chromosomes. Sharing, managing and storing that data was a frustrating challenge until the Baylor College of Medicine's Human Genome Sequencing Center (HGSC) deployed Quantum's StorNext data management software.

LEGACY TECHNOLOGY INFRASTRUCTURE IMPACTS RESEARCH

As one of three U.S. federally funded centers driving the rapid growth of knowledge about genetic influences on human disease, HGSC has just under 200 employees, including approximately 40 research scientists who spend their time analyzing DNA sequencing data. With high volumes of data generated daily and the need for hundreds of terabytes of data to be accessible for analysis at any time, HGSC's piecemeal technology infrastructure that had been built over time was becoming a barrier to the important health research work under way.

In August 2008, Geraint Morgan was brought in as Director of Information Systems when it became apparent that the rate at which sequence data could be produced and analyzed was outpacing the ability of the technology on which it was residing to keep up.

HGSC has 32 genome sequencers, including 20 Applied Biosystems SOLiD Sequencing Instruments, 2 Illumina Genome Analyzers and 10 Roche/454Genome Sequencers. The most productive of these can generate up to approximately 1TB of raw data per day.

Morgan entered an environment where primary data from DNA sequencing was initially written locally to attached storage units from a variety of vendors and then moved to a centralized repository. There it was further processed by clusters of compute nodes and made accessible to the researchers through simple network file system mounts.

"The volume of data put a strain on the network infrastructure and limited accessibility to the data, which is critical in a research organization such as ours," says Morgan. "It also contributed to the huge management overhead needed to ensure that no issue could impact the tail-end of the researchers' sequencing pipeline operation."

To expand HGSC's storage capabilities, an additional data center was created. However, Morgan still faced the challenge of finding a way to centrally manage a complex heterogeneous environment of servers, networks and storage technology. Because funding for technology is often limited in such grant-based institutions, he needed to accommodate existing servers and storage arrays, rather than starting from scratch with an entirely new approach.

"I needed to find a solution that could not only utilize this existing hardware but also would be easily scalable to accommodate a predicted 20 petabytes of sequenced data over the next two years," Morgan says. "The solution had to require minimal changes to how the environment as a whole would be managed as the environment grew."

SOLD ON STORNEXT'S REPUTATION IN HIGH-PERFORMANCE, MULTI-PETABYTE ENVIRONMENTS

Morgan began his search with a number of vendors and consultants but found that many had difficulty understanding the complexity of the operation or could not provide adequate answers to how their solution would cope with the expected growth.

HGSC
HUMAN GENOME SEQUENCING CENTER

"By combining high-speed data sharing and cost-effective content retention in a single solution, StorNext has enabled our researchers to access the data they need quickly and easily and eliminated the significant management overhead we incurred with our legacy system."

Geraint Morgan
Director of Information Systems

SOLUTION OVERVIEW

- StorNext File System
- StorNext Storage Manager
- Scalar i2000 tape library system

KEY 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 20PB over next 2-3 years

"Given the constraints of existing hardware, limited budget and the ability to accommodate significant data growth, there were surprisingly few solutions that I felt would be a good fit," says Morgan, reflecting on his three-month search for an answer. "However, one solution that kept coming up and impressed us was Quantum's StorNext data management software."

Morgan liked the fact that there were a number of companies with similar high-performance data processing needs and large multi-petabyte storage requirements that were already successfully using StorNext. He also realized that the expectation that an open-source solution would be least expensive and least intrusive was wrong, once support contracts and other fees were factored in.

"StorNext offered the scalability we needed, support for existing storage hardware with no significant investment needed for additional hardware, and an easy to manage system," says Morgan.

HGSC purchased both the StorNext File System and Storage Manager to enable file sharing across multiple operating environments and automated data movement across storage tiers.

THE FOUNDATION OF A MASSIVELY SCALABLE SOLUTION

The StorNext implementation was straightforward, according to Morgan. HGSC currently has 2 metadata controllers and 8 StorNext File System SAN gateways. These gateways are connected to storage arrays via a 4 Gbp FC network, and the compute nodes access storage via StorNext Distributed LAN Clients over dual 4×10 Gbp Ethernet uplinks through the SAN gateways. The system started with 110TB of storage and an additional 560TB has been recently added.

Following the ingest of data on the local genome scanner devices and some initial pre-processing, the data is copied via NFS to a centralized StorNext File System. The pipeline analysis is then done by multiple genome analysis applications running on top of the StorNext Distributed LAN Client, which connects to the centralized storage to process the data in parallel.

In addition, HGSC uses StorNext Storage Manager for automatically moving data between different disk systems and a Quantum Scalar i2000 tape library, thereby protecting content at lower costs. Older genome projects can also be archived automatically on the Scalar i2000, freeing up the fast primary disk storage for newer sequencing workflows.

Since deploying StorNext, Morgan has been very pleased with the benefits it has provided from both a research and IT perspective.

"By combining high-speed data sharing and cost-effective content retention in a single solution, StorNext has enabled our researchers to access the data they need quickly and easily and also eliminated the significant management overhead we incurred with our legacy system," he says.

Looking toward the future, Morgan says that StorNext will serve as the foundation of a massively scalable solution that's needed to cope with expected storage requirements of two petabytes over the next 12 months and up to 20 petabytes over the coming two or three years.

"Because of the nature of genomics research—where data generated today might not have obvious value but could lead to important discoveries in the future—we preserve all the data generated at HGSC," says Morgan. "This creates an ever-growing archive, and StorNext will play a critical role in helping us to manage this growth. The exponential data growth is also one of the reasons we plan to leverage the data deduplication feature offered through StorNext—it will enable us to optimize the amount of storage capacity needed for archiving."

> "The exponential data growth is also one of the reasons we plan to leverage the data deduplication feature offered through StorNext—it will enable us to optimize the amount of storage capacity needed for archiving."
>
> **Geraint Morgan**
> Director of Information Systems

ABOUT THE HUMAN GENOME SEQUENCING CENTER

Baylor College of Medicine's Human Genome Sequencing Center, founded in 1996, is a world leader in genomics. The fundamental interests of the HGSC are in advancing biology and genetics by improved genome technologies. One of three large-scale sequencing centers funded by the National Institutes of Health, the HGSC's location at the heart of the Texas Medical Center provides a unique opportunity to apply the cutting edge of genome technologies in science and medicine.

ABOUT THE RESELLER

Disk Array and System Design, Inc (DASDI) designs HPC storage clusters using Quantum's StorNext software. These systems provide customers a competitive advantage in price/performance without compromising data availability or integrity and can scale to over a PB of usable capacity.

Q&U

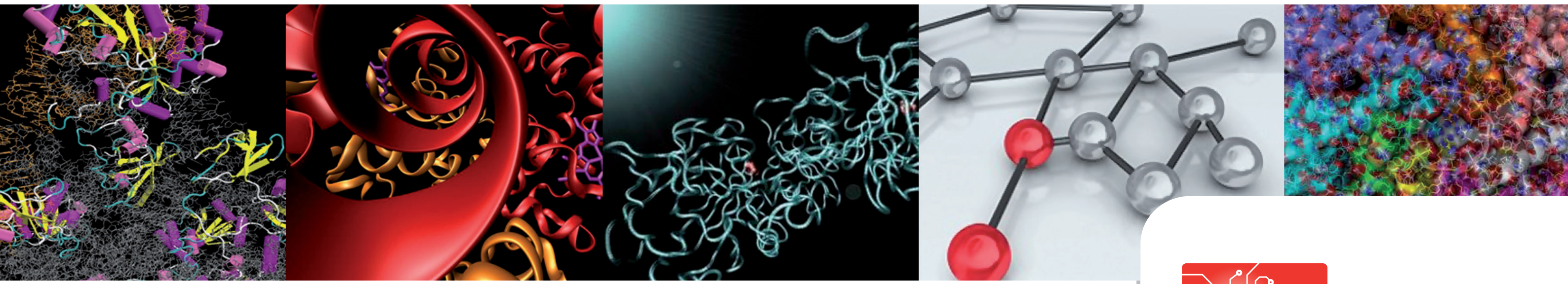
What is Q&U?

Our goal is to preserve the world's most important data. Yours. Q&U is Quantum's approach of collaborating with you to address your specific data protection and retention challenges. It's about sharing our insights and expertise, giving you the resources to make educated choices, and delivering comprehensive solutions that help you tackle today's challenges while preparing for tomorrow. When Q&U come together, great things happen.

Quantum®

Preserving the World's Most Important Data. Yours.™

To contact your local sales office, please visit www.quantum.com



StorNext

The Swiss Institute of Bioinformatics Reduces Cost of Multi-Petabyte Storage by 50% with Quantum StorNext Software

When The SIB Swiss Institute of Bioinformatics was faced with spiralling data growth arising from next generation sequencing, it deployed a hierarchical storage management (HSM) solution centered on Quantum StorNext data management software and HP hardware. This provided high performance file sharing and data protection and reduced SIB's total cost of storage by 50%.

HUGE DATA VOLUMES PUT PRESSURE ON STORAGE INFRASTRUCTURE

SIB is a federation of bioinformatics research and services groups from leading Swiss universities and the Swiss Federal Institutes of Technology. The Vital-IT Center is the heartbeat of SIB, acting as a high-performance computing (HPC) joint venture between academic and industrial partners. Dedicated to life sciences, the Center supports software development and optimisation along with HPC and data storage for biology and medicine.

Bioinformatics—the process of applying information technology to biological problems—creates an ever-increasing volume of data. An example is the DNA sequence analysis needed to support the worldwide Human Genome Project.

Next generation, ultra-high throughput sequencing allows genomes to be sequenced in an unprecedented manner. A single experiment produces up to 743,000 files per run, with each run sized at an average of 2 TB and performed every 3.5 days. The performance requirements relating to these scans are equally large: a 600 MB/sec aggregate read and 400 MB/sec aggregate write are mandatory to allow data analysis.

Over time, the data per run had grown significantly at SIB. In 2007, 1 TB of raw and processed data was produced during each week — by 2009 this had risen to 7TB per week and was continuing to grow. This put tremendous pressure on the Vital-IT Center to store, protect, and manage the data as existing storage capacity and budget were no longer sufficient.

"We needed to create a storage infrastructure capable of scaling to multiple petabytes and managing hundreds of millions of files," says Roberto Fabbretti, IT Manager at the Vital-IT Center. "We wanted to eliminate the need for a separate backup and provide a comprehensive disaster recovery solution. All of this needed to be wrapped within a cost-effective storage environment."

The Center considered traditional storage area network (SAN) and network attached storage (NAS) options. A SAN would have enabled SIB to store data centrally and access that data quickly; however the costs of deploying a fiber channel-connected infrastructure were found to be prohibitive. A NAS solution was also eliminated from consideration because of the costs involved, as well as the limited performance and scalability it would have provided.

In addition, both options would still not have provided the required data protection - keeping track of original data for up to 20 years is particularly important for the pharmaceutical and biotechnology industries.

Rejecting the two options, Vital-IT discussed the situation with partners and decided to choose an HSM solution that would move data automatically between hard disk arrays and tape storage and would provide flexible, high speed access to many users. StorNext emerged as the clear choice.

STORNEXT MEETS HPC FILE SYSTEM AND HSM REQUIREMENTS

"We examined a number of solutions," explains Fabbretti. "StorNext met our combined HPC file system, HSM, and data protection requirements. This transport- and hardware-independent solution offered scalable performance and easy resizing of volumes, as well as water-tight disaster recovery protection for the 400 TB of data we were managing."

The Vital-IT Center deployed a StorNext solution comprising StorNext File System and StorNext Storage Manager. StorNext is currently integrated within an HP

Vital-IT
High Performance Computing Center

"When it comes to efficient file sharing, transparent tiered storage, and cost-effectiveness, StorNext has fulfilled all our requirements."

Roberto Fabbretti
IT Manager, the Vital-IT Center,
part of The Swiss Institute of
Bioinformatics

SOLUTION OVERVIEW

- Quantum StorNext File System
- Quantum StorNext Storage Manager
- HP BL480 metadata servers
- Three HP BL680 SAN gateways
- HP disk cache with 160 TB of extensible storage
- HP tape library scalable to 570 TB

KEY BENEFITS

- Reduced total cost of storage by 50%
- Improved sequence tag productivity by 20%
- Provided operational cost savings, including cooling and power
- Enabled high-speed data sharing with cost-effective content retention
- Consolidated resources and enabled workflow operations to run faster
- Moved data between storage tiers transparently for simplicity, scalability, and economy
- Eliminated vendor lock-in through StorNext platform independence
- Ensured all files were easily accessible to all hosts

disk and tape library environment which includes two HP BL480 StorNext metadata servers, three HP BL680 SAN gateways, an HP disk cache with 160 TB of extensible storage and an HP tape library that can scale to 570 TB.

"STORNEXT HAS TRANSFORMED OUR DATA STORAGE INFRASTRUCTURE"

Since deploying StorNext, Fabbretti and others at the Center have been very pleased with the results.

"StorNext has transformed our data storage infrastructure," he says. "Our first file system spans 22 million files, and the second file system we recently purchased already has 2.5 million files."

The StorNext File System provides high performance access to the central storage pool for multiple users, and because it is heterogeneous, applications on different operating systems can be used to collect, store and analyse the data, all at the same time.

File system performance is optimised through the use of StorNext "affinities" which organise the writing of data from one physical disk to another.

Access is provided over GbE using industry standard protocols (NFS / CIFS) and also over SIB's InfiniBand network providing high performance and resilience. In the future, SIB is also likely to deploy the StorNext performance protocol, Distributed LAN Client, to provide even higher levels of performance access for key users.

When dealing with large data sets, it isn't cost-effective to hold petabytes on high performance disk, and yet the management and administration costs of moving data to cheaper storage can be prohibitive unless an automated system like StorNext can be utilised. Figure 1 shows the cost of storage in relation to purchasing and ongoing operational costs whilst Figure 2 shows data in relation to its value in terms of frequency of access.

In the case of DNA sequencing, high performance storage is required to collect and process sequence data. However once results have been generated, the data is suitable for storage on more cost-effective tape and secondary disk.

As data is stored in the file system, StorNext Storage Manager also copies it to other tiers of storage based on predefined policies. Over time, unused data is removed from the primary storage, leaving just the data held on the other tiers. File stubs are left in the file system, so users can access data from the same place that it was originally stored.

Data is continuously protected by making copies as it is written to primary disk, thereby eliminating backup window requirements. Also, there is no need for additional backup hardware or software, saving further costs.

In the event of a disaster, SIB doesn't have to restore the complete 400 TB of data before it can be used, which would take months. Instead, the system is brought up in a nearline state with project data available as and when needed.

StorNext has also allowed researchers to keep the raw data that most other sequencing centers have to discard—saving 20% on sequence tags as new algorithms are run on original data.

"By standardising on StorNext, we have reduced our total cost of storage by 50%," Fabbretti reports. "When it comes to efficient file sharing, transparent tiered storage, and cost-effectiveness, StorNext has fulfilled all our requirements."

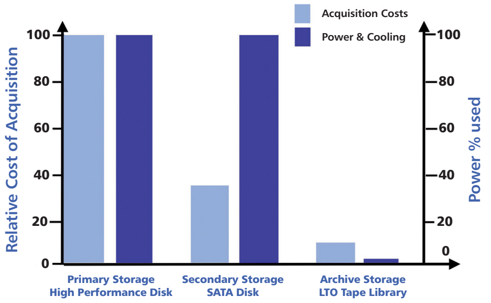


Figure 1

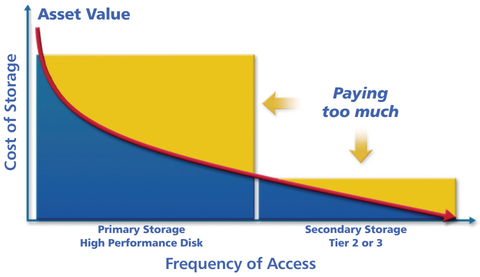


Figure 2

What is Q&U?

Our goal is to preserve the world's most important data. Yours. Q&U is Quantum's approach of collaborating with you to address your specific data protection and retention challenges. It's about sharing our insights and expertise, giving you the resources to make educated choices, and delivering comprehensive solutions that help you tackle today's challenges while preparing for tomorrow. When Q&U come together, great things happen.

> "By standardising on StorNext, the Swiss Institute of Bioinformatics has halved its total cost of storage."
>
> **Roberto Fabbretti**
> IT Manager, the Vital-IT Center, part of The Swiss Institute of Bioinformatics

ABOUT SIB—THE SWISS INSTITUTE OF BIOINFORMATICS

The SIB Swiss Institute of Bioinformatics is an academic not-for-profit foundation federating bioinformatics activities throughout Switzerland. Its two-fold mission is to provide world-class core bioinformatics resources to the national and international life science research community in key fields such as genomics, proteomics and systems biology; as well as to lead and coordinate the field of bioinformatics in Switzerland. It has a long-standing tradition of producing state-of-the-art software for the life science research community, as well as carefully annotated databases.

The SIB includes 29 world-class research and service groups, which gather more close to 400 researchers, in the fields of proteomics, transcriptomics, genomics, systems biology, structural bioinformatics, evolutionary bioinformatics, modelling, imaging, biophysics, and population genetics in Geneva, Lausanne, Berne, Basel and Zurich. SIB expertise is widely appreciated and its infrastructure and bioinformatics resources are used by life science researchers worldwide.

Preserving the World's Most Important Data. Yours.

To contact your local sales office, please visit www.quantum.com