Overcoming Storage Barriers in Life Sciences Research with IBM's Next Generation Sequencing Solutions

Sponsored by IBM

Srini Chari, Ph.D., MBA
October 2011

mailto:chari@cabotpartners.com

Cabot Partners Group, Inc. 100 Woodcrest Lane, Danbury CT 06810, www.cabotpartners.com

Cabot Partners
Optimizing Business Value

Executive Summary

Life sciences in general and genomics in particular have advanced at a rapid pace with the advent of Next Generation Sequencing (NGS). There is an explosion of research data generated by new instruments, new research platforms, complex genomic applications, and experiments. Life sciences research requires high-end compute, memory, and large amounts of storage for analysis, future reference and collaboration across researchers. Data access by a growing number of users who have widely varying computing needs poses a daunting IT challenge. The shortening of discovery pipelines, greater emphasis on productivity and collaboration, and the need to do more science with fewer scientists are some of the today's realities. Budgetary pressures on hardware, software, facilities and IT administration costs make it imperative to have smarter, energy- and cost-efficient scalable IT solutions that can deal with not just more data but very large data.

Life sciences research such as NGS demands aggressive performance from the underlying IT resources for higher scale, cost-efficiency and speed of execution. At the same time, organizations need to balance research demands and the ensuing power usage needs with energy and space constraints of their IT facility.

IBM's Next Generation Sequencing Solution (NGS Solution) combines scalable and energy efficient compute and the memory muscle of System x servers (eX5, iDataPlex, BladeCenter) and high density storage (DSC3700) with scalable GPFS architecture for efficient management of large amounts of data using a single point of control. Compared to traditional server environments, IBM's iDataPlex servers provide five times the computing power by doubling the number of servers per rack and at the same time reducing energy consumption by 40%. With its intelligent server cooling technology, IBM servers require less air-conditioning energy and space, given their compute power and density. Through smart customization, the IBM NGS Solution uses a mix of compute, network, and high density energy and space efficient scalable storage nodes that enables customers to meet application requirements across verticals and industries. In addition to Web 2.0 and SaaS workloads, IBM intelligent servers and storage solutions are successfully deployed in financial risk analysis in banking, life sciences, and several of today's internet-scale high performance computing data centers.

This paper describes IBM's HPC Storage solutions deployed at some leading life sciences research enterprises and overviews some ISV solutions that harness IBM technology for NGS. We discuss how IBM collaborators such as Accelrys and others use IBM's HPC storage solution to address the stringent needs of life sciences research and analysis involving scale, performance, efficiency, cost-effectiveness and robustness.

Target audience: Executives, business and technology leaders, life science researchers, and decision makers who are considering state-of-the-art storage solutions that can complement their computing environments to enable smarter Life Science research.

Overview

Nearly a decade after completion of the first Human Genome Sequencing project, genome informatics still lags behind the biology[1]. Each new scientific discovery results in a significant jump in the amount of research data and processes that refer to new findings. IBM addresses life science research IT needs through Next Generation Sequencing (NGS) Solutions featuring IBM's System x eX5 and Intel multi-core powered iDataPlex Systems, high-density DSC3700 HPC Storage, GPFS based SONAS and Hierarchical Storage Management (HSM) along with IBM System Storage TS3500 Tape Libraries that supports NGS research environments comprising sophisticated NGS algorithms and complex software platforms from leading life science vendors such as Accelrys and others.

[1] CLC Bio – Genomic Informatics frameworkhttp://www.businesswire.com/news/home/20110609005065/en/CLC-bio-reveals-Genomics-Gateway---genome

High Performance Computing Challenges in Life Sciences

Compute, storage and network are the three pillars of the NGS IT infrastructure. Previously, CPU was at the center of the compute universe for compute intensive activities such as NGS, business intelligence, analytics, and similar activities. Today data explosion is at the center of the research universe, demanding more storage and better management (see Fig. 1).

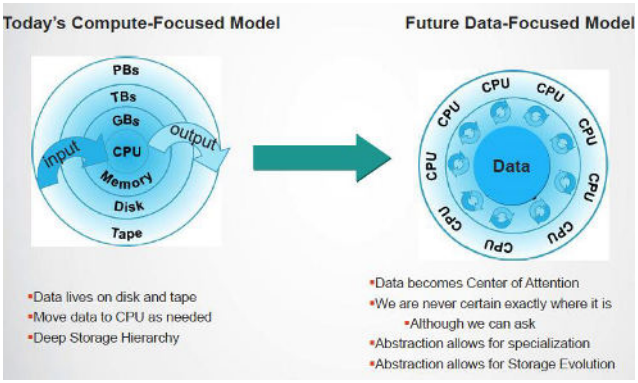

Figure 1: Data Centric Approach (Source: Matthew Drazhal - IBM)

The overall storage market today is growing at more than 50% year-on-year in terms of total capacity delivered[2]. In life sciences, companies such as Genomic Health today store about 50 GB (gigabytes) of combined genomic and transcriptome data per patient. This amounts to 12.5 TB (terabytes) per day or 1 PB (petabyte) per quarter. It would take 15 EB (exabytes) to sequence every person in the US, and 3.5 ZB (zettabytes) for the world population. But the aggregate storage manufactured in 2010 was only 600 EB[3].

It is not just the size of storage but also data management which is a key pain point for NGS. Figure 2 shows some of the key IT infrastructure challenges in NGS.

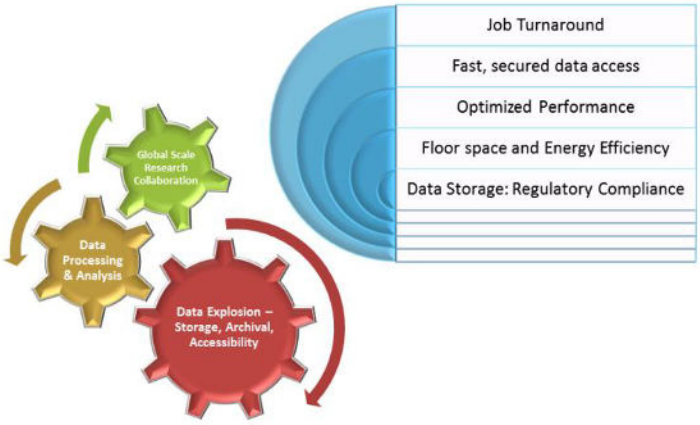

Figure 2: NGS - Key IT challenges

Some of the main drivers of IT infrastructure in life sciences research are genomics (mammalian genomes, HTPS), faculty programs for analysis, visualization and collaboration in wet and dry labs, imaging, proteomics, and research involving human subjects. An estimate[4] by Jackson Laboratory (Fig. 3) shows that the projected investment in storage and compute far exceeds that in network, visualization, collaboration, security, and disaster recovery combined. To deal with storage effectively, experts recommend tiered storage to address varying needs of life sciences for scalable, manageable and cost-effective information management.

---

[2] How Storage is becoming even bigger…http://www.asiacloudforum.com/content/idc-report-storage-shipments-keep-surging
[3] Chris Aldridge, CIO of Genomic Health at Bio IT World, 2011http://www.bio-itworld.com/BioIT_Article.aspx?id=106818
[4] Gregg TeHennepe, Research Liaison , IT, The Jackson Laboratory http://www.giiresearch.com/conference/bio-itworld11/track1.shtml

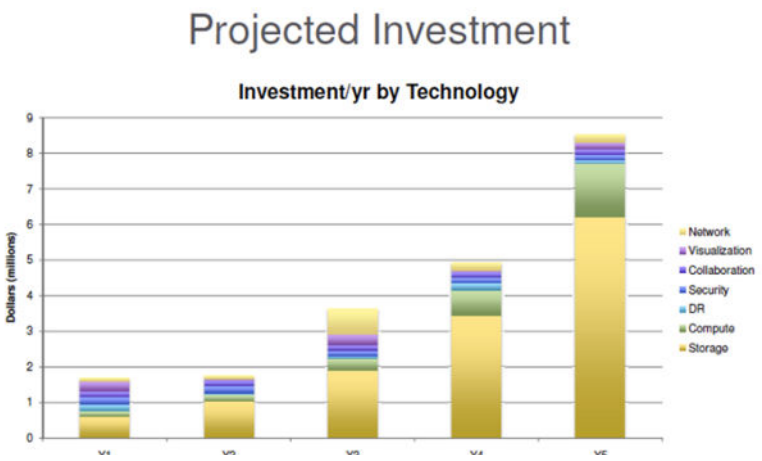

Figure 3: Storage vs. Compute investment over 5 yrs. (*Source*: Jackson Laboratory)

This paradigm shift in life sciences research is fuelled by emerging technologies such as High Performance Computing (HPC) and High Throughput Sequencing, and is enabled by the digitization of data. The result is an exponential growth in research data leading to the demand for extremely high scale storage solutions. Figure 4 shows some current trends in life sciences research IT resource usage. Given the pace of scientific advances and the amount of data, the underlying IT systems are required to have multi-petabyte scale capability, large memory, and top-notch compute performance while using a general purpose file system that can help research collaboration and ease data management woes.

Figure 4: Trends in IT Infrastructure usage by Life Science Researchers

As the genomic revolution drives NGS at a rate faster than Moore's law, commoditization and proliferation of research instruments, newer lithe applications and platforms are spurring the rate of research data generation. Today, unstructured research data consumes over 35% of storage and is doubling every 10 months. Fuelled by the significant fall in associated sequencing costs and time, the demand for Next Generation Sequencing is growing rapidly, opening new avenues for breakthrough research and medicine (see Fig. 5).

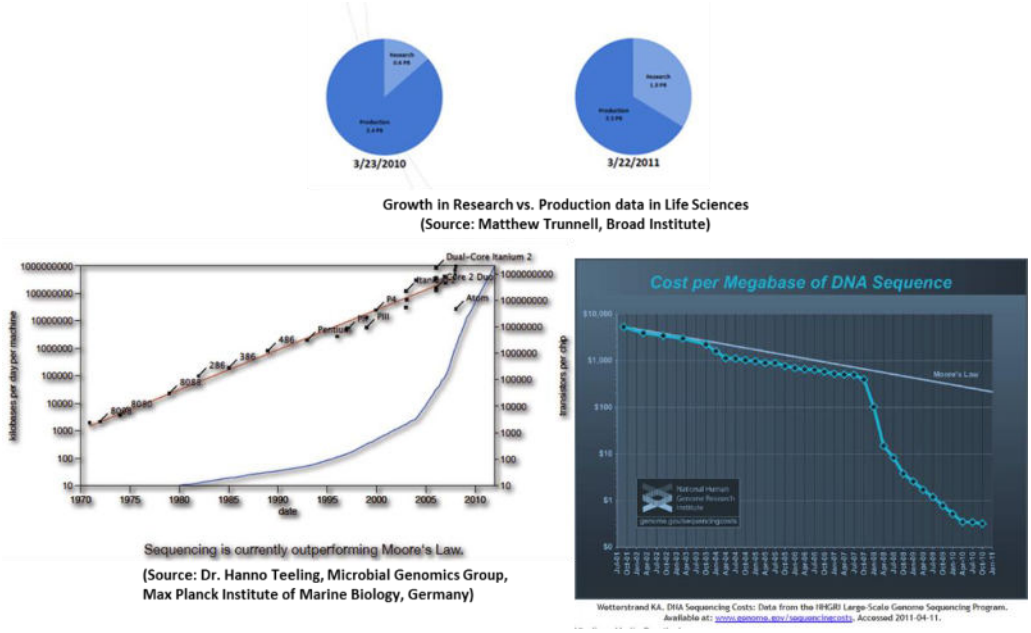

Figure 5: Exploding NGS Data, Plummeting Costs vs. Moore's Law

IBM NGS Solution: Powering Life Science Research

IBM addresses key pain points of NGS and life sciences researchers (see Fig. 6) through its Smart Next Generation Sequencing Solution (NGS Solution) comprising a fully integrated configuration of IBM servers, storage, tape, and associated software. These solutions have been successfully deployed at several bioinformatics and life sciences research facilities such as Sanger and Beijing Genomics Institute. And many leading NGS ISVs use these IBM solutions to deliver better performance and value to their clients. IBM servers, such as the iDataPlex feature the latest versions of the Intel Xeon Processors and fully harness the power of the new applications in life sciences. In addition to the flexible, scalable and robust compute and storage hardware, IBM solutions also provide scalable, integrated data management software through Global Parallel File System (GPFS). This allows life sciences researchers to focus on research while their IT administration is simplified in a cost-effective and flexible manner. IBM's solution supports hierarchical and tiered storage management that can seamlessly transfer data across storage tiers depending on the data moving policies in the organization.

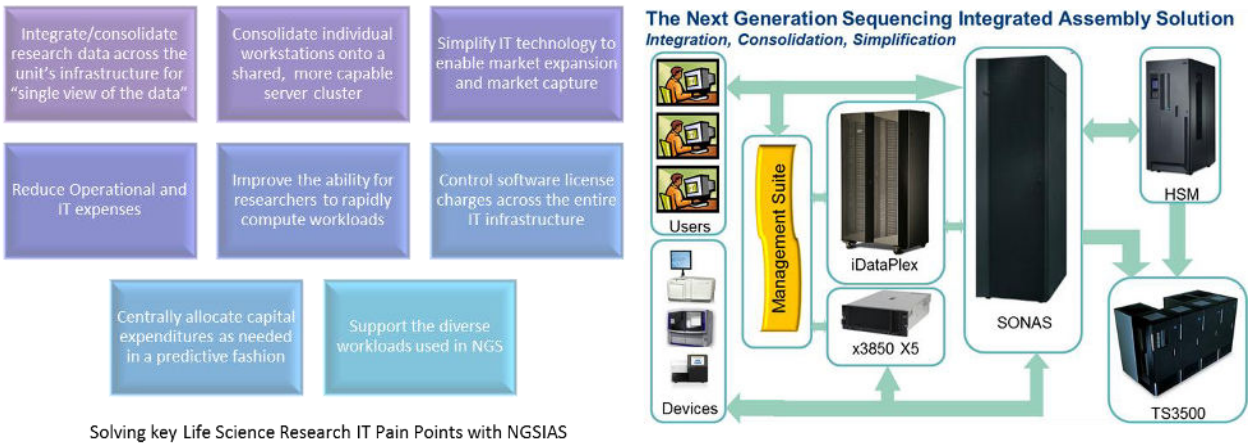

Figure 6: IBM's NGS Solution

Solution Components

The IBM NGS Solution consists of IBM's Intelligent servers- System x rack servers, iDataPlex and BladeCenter servers, high density storage such as DCS3700, IBM System Storage TS3500 Tape Library, and GPFS powered SONAS (Scale Out Network Attached Storage) with policy based tiered storage management (Fig. 7). All components of the solution are assembled and tested by IBM, shipped fully-integrated and ready for deployment at the customer site with an added advantage of a single point of contact at IBM for worldwide support. IBM servers are designed for power and cooling efficiencies, rapid scalability, and usable server density. The iDataPlex uses 40% less power and cooling, allowing customers to pack more racks per square foot of an existing data center floor space area using the existing power and cooling infrastructure. The iDataPlex is designed to support dense server rack configurations with improved energy efficiency and is ideally suited for today's internet scale data centers and massive scale cloud computing environments.

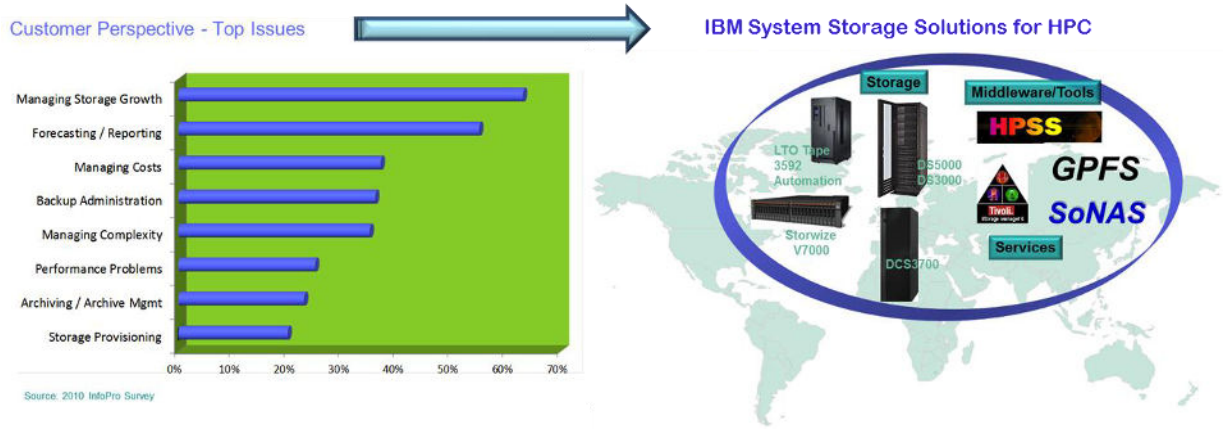

Figure 7: Addressing HPC Storage Issues with IBM Storage Solutions

To deal with data explosion cost effectively, the IBM NGS Solution helps life sciences researchers with data management use cases, for example, separating important data from the not-so-important data, data protection (backup, replication, data integrity), and preservation of data in a user transparent, intelligent and effective manner. Other research needs such as scheduling, data management, collaboration, sharing data, moving large files, global namespace across different tiers of storage, and reducing dependence on high performance storage tiers for peak demand are also addressed. Storage tiering is particularly attractive as customers consider the use of solid state disk (SSD). Customers need to balance the performance of SSD with its cost. The IBM NGS Solution's policy-based hierarchical storage management for life sciences provides a way to ensure that the most active data is always on SSD, thus enabling customers to buy just 10% of primary storage capacity as SSD. With the use of the IBM GPFS technology and SSD, it is possible to scan 10 billion files in 43 minutes. This can significantly help life sciences organizations to manage their rapidly growing data stores, address speed of analysis, and reduce storage costs by moving lesser used or archived data to less expensive tape storage.

The IBM Next Generation Sequencing Value Proposition

Because of decades of deep industry experience and integration capabilities, IBM is uniquely positioned to help researchers enhance their insights, analysis, and results. The IBM NGS Solution helps to significantly accelerate time-to-solution through compute density, higher memory capabilities, flexible architecture and cost effective tiered storage management. At the same time it reduces execution risk and eases management. IBM helps clients to grow revenue, improve efficiency, and better manage regulatory-driven challenges through innovations and technological advances at a much higher pace than the average compute infrastructure refresh cycles at a typical organization today. Total cost of ownership studies prove the value of flexible system design and automated storage life cycle management of IBM's solution for life sciences research.

Start Small, Scale up Easily with High Density Storage DCS3700

Low-latency performance in applications such as life sciences, real-time analytics, rich media, seismic processing, weather forecasting, telecommunications, and financial markets require high-performance storage architectures. In addition, organizations involved in these areas often need to improve operational efficiency while maintaining the same data center footprint, quality of service, and high availability. DCS3700 (Fig. 8) is IBM's densest storage offering, and provides 30 TB per U in a rack. Through a 'building block' approach, IBM servers, storage and software can be deployed for life sciences

research, especially NGS use cases, and the IT capacity can be easily scaled up when needed. DCS3700 is a 6 Gbps SAS high density storage system delivering scalable capacity at an affordable price point and is ideally suited for addressing high storage demands of NGS and life science applications. Combined with IBM's best-in-class GPFS, the new DCS3700 can help organizations optimize the flow and management of large file-based data while retaining ease of data access. With its simple, efficient and flexible approach to storage, the DCS3700 is a cost-effective, fully integrated complement to IBM System x servers – eX5 and iDataPlex, IBM BladeCenter-- and IBM Power Systems.

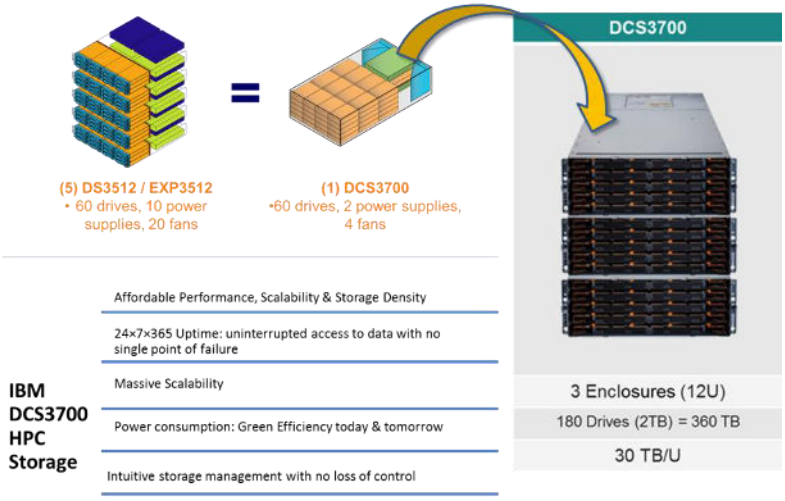

Figure 8: IBM DCS3700 Highlights

Meeting Data Explosion Challenges Cost-effectively with SONAS

There is a lot of network attached storage (NAS) in use in life sciences, especially in NGS. Dealing with islands of storage is a challenge given the scale and rate of data generation. Compute, high amount of memory to run complex NGS algorithms such as Velvet, and accessing storage containing genome datasets is critical in such cases. The GPFS architecture eliminates latencies associated with file metadata access and makes indexing extremely fast – that is the key to dealing with access, backup and storage of multi-billion files in life sciences environments. IBM's SONAS (Fig. 9) is based on IBM GPFS (Fig. 10). The information lifecycle management (ILM) function in GPFS acts like a database query engine to identify files of interest.

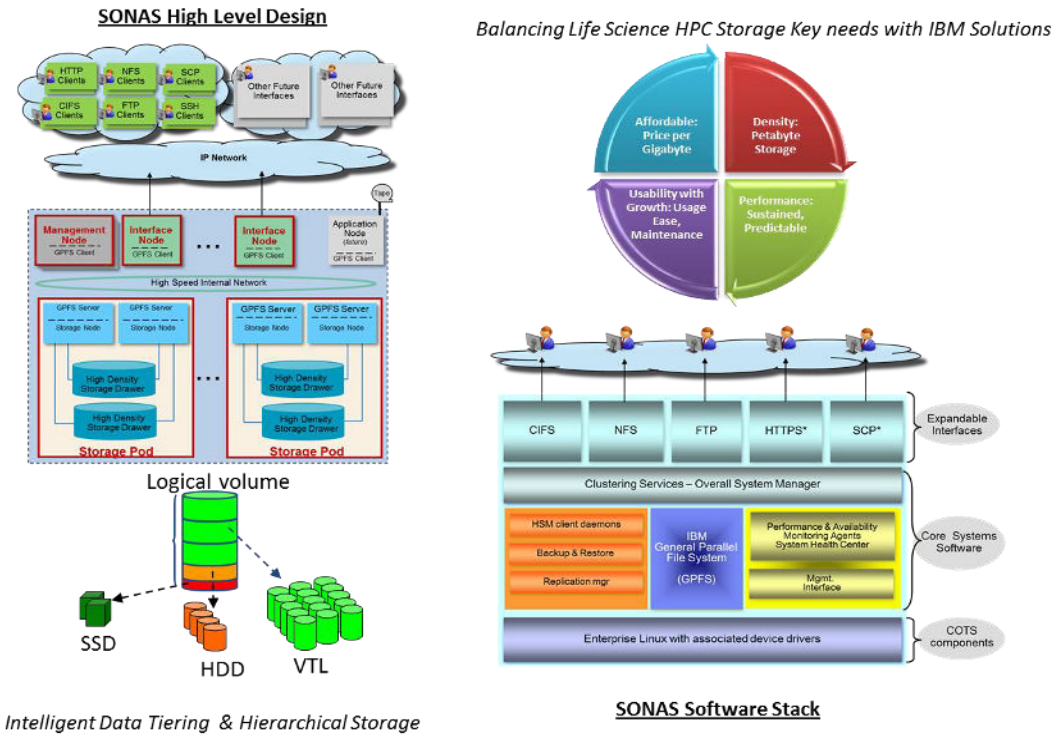

Figure 9: IBM's SONAS Architecture and benefits

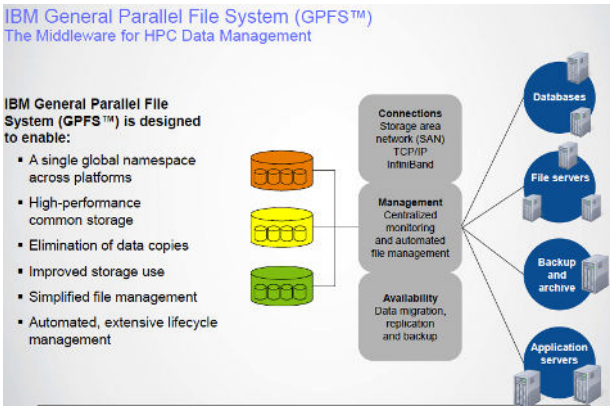

Figure 10: GPFS Salient Features

IBM GPFS runs in parallel and scales out as additional resources are added. Once the files of interest are identified, the GPFS data management function uses parallel access to move, back up, or archive user data. GPFS tightly integrates the policy-driven data management functionality into the file system. This high-performance engine allows GPFS to support policy-based file operations on billions of files. The rate at which GPFS can scan a billion files is important to deal with the explosion of data especially in life sciences research. This can tremendously help researchers who have to manage that data to identify what to backup, what to replicate for disaster recovery, and to determine what data is appropriate for storage tiering. They have to scan files to figure out what changed. The IBM NGS Solution with GPFS technology makes this simple. The IBM GPFS is a true distributed, clustered file system. Multiple servers are used to manage the data and metadata of a single file system. Individual files are broken into multiple blocks and striped across multiple disks and multiple servers, thus eliminating bottlenecks. Information Lifecycle Management (ILM) policy scans are also distributed across multiple servers and this enables GPFS to quickly scan the entire file system, identifying files that match specific criteria. GPFS forms the base of highly scalable, highly available and performing hierarchical storage management features of the IBM NGS Solution.

Figure 11: IBM Storage Solution for Life Sciences - Feature benefits

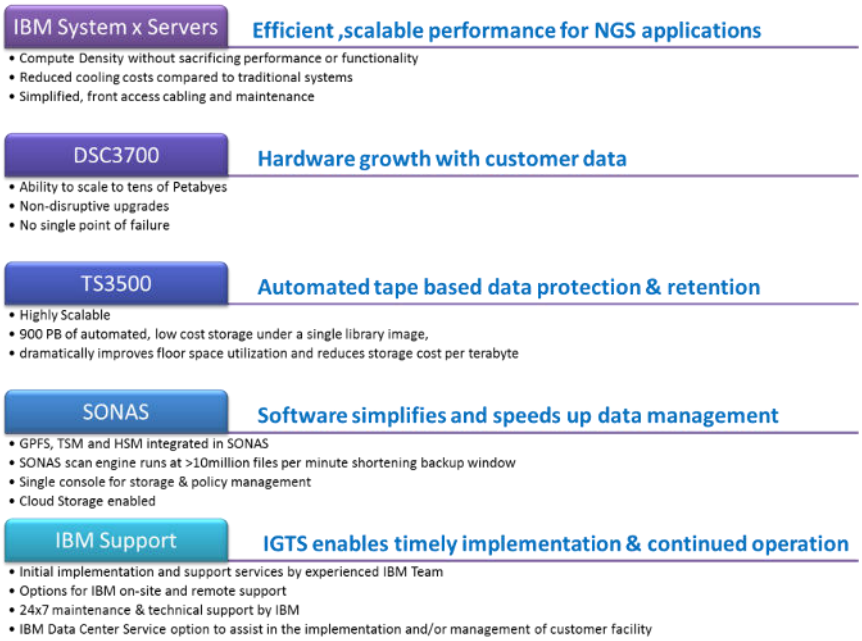

Figure 12: IBM Solutions for Life Science Research

Case Studies

Pharmaceutical companies, research organizations and universities use IBM solutions in many life sciences areas, such as drug discovery, bioinformatics, and biotech research. IBM System x servers are powerful and affordable. These systems offer customers the flexibility of choosing to deploy selected parts or completely integrated systems. Among the IBM System x servers, iDataPlex provides flexible design, significant power and cooling efficiencies, and compute density while eX5 enterprise systems provide maximized memory, minimized cost, and simplified deployment. We highlight some instances where IBM iDataplex, GPFS, and SONAS have made significant impact in life sciences research, especially NGS.

Accelrys: Complete Human Genome Mapping in Hours, not Days

Accelrys is a leading scientific enterprise R&D software and services company. It produces various molecular modeling and simulation software such as Discovery Studio for both life sciences and materials science research. Accelrys Pipeline Pilot platform enables NGS researchers to access, organize, analyze and share scientific data in unprecedented ways, ultimately enhancing innovation, improving productivity and compliance, reducing costs and speeding time from lab to market.

Business Challenge

Accelrys Next Generation Sequencing (NGS) Collection for Pipeline Pilot lets users analyze and interpret the massive datasets generated by the most current DNA sequencing instruments. Built for use with the Pipeline Pilot informatics platform, the NGS Collection comes with a comprehensive assortment of NGS data analysis pipelines. Sophisticated NGS platforms such as Accelrys Pipeline Pilot demand extreme IT performance including compute, memory, and hundreds of terabyte scale data management solutions. For NGS research IT support staff it is a daunting challenge to keep up with the compute and memory requirements of NGS analysis methods against petabytes of data while simultaneously keeping pace with rapidly evolving algorithmic best practices.

Solution

Each day's delay in bringing a drug to market costs millions of dollars. Bringing down research cycle times and IT infrastructure related delays can significantly impact the bottom line for scientific research by reducing wastage of the most important resource – the researcher's time. Accelrys NGS Collection can run faster and more reliably using IBM iDataPlex hardware and IBM SONAS data management solution. IBM partners with Accelrys, powering its Pipeline Pilot platform with an NGS Solution consisting of iDataplex and GPFS based SONAS for tiered storage, and faster computing. This not only saves money but also takes researchers closer to the goal of $1000 for a human genome. Medical researchers believe that the target of $1000 for a human genome will make next generation sequencing affordable and transform preventive and prescriptive medicine.

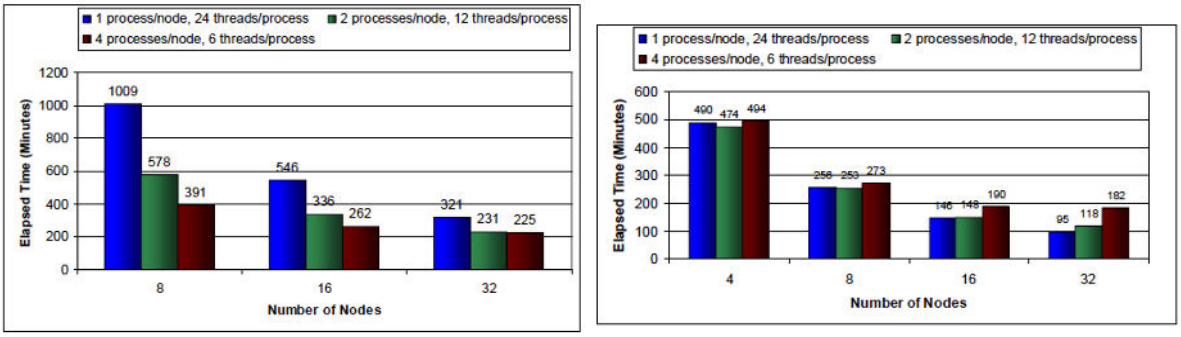

Figure 13: BWA Algorithm (left) and Bowtie (right) with the IBM iDataplex and SONAS enabled storage infrastructure powering the Accelrys Pipeline Pilot Platform

Typically, it takes about two to three days to complete human genome mapping with typical 30x coverage with the widely-used BWA algorithm on a single node. With Pipeline Pilot BWA Mapper using IBM iDataplex based compute architecture and SONAS for data management, researchers have achieved complete mapping in less than 4 hours using 32 iDataPlex system nodes with GPFS. Similarly, it takes a few days to complete human genome mapping with typical 30x coverage using the open source algorithm Bowtie on a single node. Running the Bowtie mapping pipeline on Pipeline Pilot in an IBM iDataPlex based compute architecture, and SONAS for data management, researchers have achieved complete mapping in approximately 1.5 hours with 32 iDataPlex system nodes with GPFS (see Fig. 13).

Benefits

The key benefits gained through IBM's NGS Solution for running Accelrys NGS Collection for Pipeline Pilot are:

- iDataPlex unique design allows customers to double the number of servers that can run in a single rack for better space utilization
- iDataPlex nodes use up to 40% less energy while increasing data center computing power five times
- Despite high compute density of iDataPlex, its better cooling technology can help customers reduce air conditioning of IT data center significantly

Sanger: Breaks the Gene-Sequencing Memory Barrier

Wellcome Trust Sanger Institute, based in Hinxton, Cambridgeshire, UK, is a non-profit British genomics and genetics research institute primarily funded by the Wellcome Trust. A leader in the Human Genome Project, Sanger is focused on understanding the role of genetics in health and disease with the aim to provide results that can be translated into diagnostics, treatments or therapies that reduce global health burdens.

Solution components:

- IBM System x3850 X5
- IBM System x MAX 5
- DDN SFA10000
- DDN ExaScalar File System

Business Challenge

The number of gigabases sequenced per month at Sanger for genomic research analysis projects grows exponentially. The rate of data generation by high-throughput genomic research equipment far exceeds the capability to analyze it. The process and workflows required to analyze the data are extremely complex and resource-intensive. The Sanger Institute aims to be at the leading edge of genome scale scientific research, and high throughput sequencing is at the core of their work. At a generation rate of over a terabase of sequences per week, the key requirement at Sanger from the high performance computing IT infrastructure is downstream meta-analysis of sequencing data. The rate of increase in sequencing technologies puts a lot of pressure on the HPC infrastructure. Every six months, Sanger has a *twofold* increase in their compute and storage requirements as the rate of data output from their sequencers double. This not only poses challenges for data access, storage and analysis but also on the facilities power and cooling capabilities.

Solution

Sanger deployed the IBM NGS Solution consisting of System x3850 nodes for its high performance cluster and SONAS for tiered storage management. At Sanger, a typical ABySS run using a dataset belonging to an African Male individual sequencing study and "read" size of 68bp consumed 180 GB of system memory on a previous-generation IBM x3950 M2 [72334MG] system with 8 Intel quad-core 2.4 GHz Xeon E7400 processors and 512 GB memory. This was a Debian/Lenny OS and Lustre 1.8 File System based setup. In another experiment that used Plasmodium Falciparum – the Malarial parasite sequencing dataset – the peak memory consumed by Velvet (kmer =21) running on the same IBM system was 334 GB. This is far higher than most competitor systems can hold. For the same dataset, SOAP used only 7 GB of peak memory. Some of these runs, especially Velvet with kmer=21, would not even complete on non-IBM systems that were not equipped with large memory, as in IBM's eX5 systems.

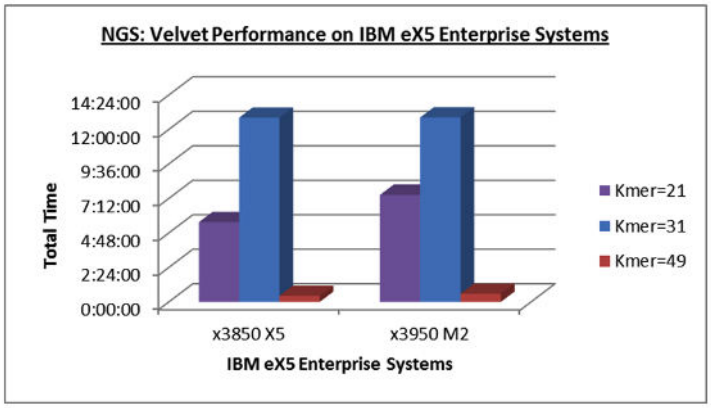

Figure 14: Velvet performance improvement with IBM hardware

Benefits

The IBM solution helped Sanger run Velvet based next generation gene sequencing experiments that require extremely high memory. Also, the large amounts of research data were better managed in a cost-effective and reliable manner through IBM SONAS with GPFS technology.

- MAX 5 memory drawer for NGS extreme memory requirements
- Compact architecture for simplified management
- Green, optimal floor space and energy efficient

Conclusions

Over the last decade, with the widespread penetration of industry-standard clusters in life sciences research, HPC (high performance computing) capital expenses as a percentage of IT spend have decreased even with the research data explosion. Through global scale collaboration and newer research tools, scientists have improved HPC research productivity. But associated operational expenses to manage these higher computing density HPC data centers and the research storage islands have escalated primarily because of increased costs in systems administration, energy, and facilities. IBM's NGS Solution for life sciences research consisting of System x servers and SONAS (Scale Out Network Attached Storage) using GPFS (Global Parallel File System) technology can lower these operational expenses while reducing capital expenses by supporting greater compute density with significantly lower energy and floor space footprint.

With Tiered Storage Management, SONAS helps deal with data explosion in life sciences research in a cost-effective manner, enabling researchers to store and access data as per requirements through SSD (solid state disk), NAS (network attached storage), and Tape Storage, bringing down data storage and management costs and decoupling them from the rate of storage growth. With Smart SONAS and GPFS enabled data management, the IBM Smart NGS Solution is a cost-effective, high performing, scalable, flexible, and robust solution, and efficiently meets the stringent compute, storage and memory requirements of NGS in particular and in life sciences research in general.

Copyright ® 2011. Cabot Partners Group. Inc. All rights reserved. Other companies' product names or trademarks or service marks are used herein for identification only and belong to their respective owner. All images were obtained from IBM or from public sources. The information and product recommendations made by the Cabot Partners Group are based upon public information and sources and may also include personal opinions both of the Cabot Partners Group and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. The Cabot Partners Group, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.