Accelerate Your Research with a Turn-key HPC Solution for Bioinformatics Analysis

PSSC LABS

HPC wire

Introduction

Fast analysis of terabytes (TBs) or more of bioinformatics data is essential to research progress and scientific advances. This need for speed is critical for all types of research including basic biological research, new drug discovery, translational medicine, and clinical genomics.

As a result, boinformatics research today typically requires easy access to supercomputing capabilities. Researchers in organizations of all sizes ranging from a small lab to a large commercial R&D group need high-performance computing (HPC) solutions optimized to handle workflows where TBs or petabytes (PBs) of data can be quickly accessed, manipulated, and analyzed. Such solutions enable new areas of research and have the potential to increase the pace and scale of future bioinformatics analysis.

> "Most IT infrastructures cause genomic analysis workflow problems."

Fortunately, thanks to the availability of more powerful processors, lower cost and higher performance memory and storage (SSDs), new cluster compute technology (e.g., Hadoop), faster interconnects, and object storage for accessing and managing enormous datasets, it possible for everyone to have the computational capabilities to conduct innovative research.

However, assembly of the right HPC system and the tuning of the system to optimize its performance for a specific research agenda requires extensive expertise in HPC technologies. Unfortunately, most scientists, labs, or life sciences organizations either do not have the needed internal skills or the time to carry out these tasks.

That is why there is a growing interest in using turn-key HPC systems that are pre-assembled, include the needed hardware and bioinformatics analysis software, easily connect to data sources and lab equipment that produce bioinformatics data, are optimally tuned for performance, and are easy to manage.

Life science developments that are increasing the need for HPC

Two factors have contributed to the rapid pace of change in the life sciences over the last decade. Improvements in next-generation sequencing (NGS) systems have steeply driven down the cost of sequencing. And improvements in processors, memory, interconnect technology, and storage have made significantly more computational power available to organizations of all sizes.

Let's first look at the impact of NGS. Today's NGS equipment produces more data per run and can perform more runs in a given time than comparable equipment available just a few years ago. The data coming from NGS systems needs rapid analysis to make intelligent decisions about next steps to take in research, drug discovery, or personalized medicine.

Like any technology product, each new generation of NGS equipment not only offers higher performance, but the cost of using the technology is much lower. In fact, the cost of sequencing has followed a Moore's Law type of drop comparable to what has happened with computing systems.

To put the lower cost into perspective and give a sense of its role in generating much more data, consider what has happened in the last decade. One parameter often cited is the cost to sequence a number of base pairs. For years, the National Institutes of Health's National Human Genome Research Institute (NHGRI) has tracked its costs to sequence a million base pairs (referred to as the cost of determining one megabase). Since 2008, when the first of the next-generation sequencing systems came to market, the cost per megabase has dropped by half roughly every five months.i Specifically, the rate has fallen from about $10 to $100 per megabase in the 2008 time frame to less than a penny a megabase today. There have been comparable reductions in cost to run other life sciences laboratory systems used in drug discovery and basic research.

> "Traditional clusters might not be the best solution to conduct bioinformatics research today."

As the costs have dropped, use of sequencing has dramatically increased and applied in new areas. Genomic and bioinformatics analysis is expanding from genetic discovery that determines the sequence of an entity to include identification of biomarkers of specific diseases and more clinical applications such as looking for genomic variants that provide guidance on treating diseases in specific populations of patients. A bit further down the line in time, there should be many translational research applications such as the use of genomic information in personalized medicine and bench-to-bedside models of healthcare delivery.

Many of these applications are based on bioinformatics analysis workflows that involved large volumes of data. In fact, major NGS centers output many terabytes of sequencing data every day from each machine, and there can be dozens of machines all running concurrently.

And thus, there is a potential problem. Today, most life sciences organizations must be able to manipulate large datasets easily and derive insights from that data as fast as possible.

As the use of genomic analysis grows and spreads to new application areas (e.g., personalized medicine and clinical applications), organizations will need more compute power and be adept at collecting, storing, transmitting, managing, and ultimately archiving petascale amounts of data.

Unfortunately, the compute infrastructure available in many organizations is not adequate for such workloads. Most researchers rely on nothing more advanced than a small Linux cluster. They need highly-scalable cost-effective compute capacity; advanced techniques to split up the data for analysis (e.g., Hadoop-style workflows); and better ways to store, access, and manage data.

To put the issues into perspective, consider the work done in most labs and research centers today. Advances in NGS have significantly lowered the cost to produce a sequence allowing organizations of all sizes to perform many experiments in a month. It is not unusual for an organization to be managing workflows with TBs of data for a single sequencing run and PBs for a lab.

Analysis of other lab data has similar issues. For example, the growing use of light sheet microscopy results in the production of more than a TB of data in an hour for a single experiment. And as organizations move into translational and precision medicine application areas and use genomic analysis in research and clinical settings, there is increased use of whole Exome sequencing (WES). Such work can involve PBs of data that require fast analysis to make "real-time" decisions about customized treatments and therapies.

What's needed is an HPC solution optimized to handle the workflows where such volumes of data are quickly analyzed. Moving to such a solution will enable new areas of discovery and research. Adoption of the right HPC solution will drive more sophisticated bioinformatics use cases such as translational analysis and clinical genomics.

Common limitations and complications

While newer genomic analysis application areas have great potential, most life sciences organizations today have inadequate compute infrastructures. Quite simply, most IT infrastructures cause genomic analysis workflow problems.

For example, Linux clusters, while widely used in life sciences research, have limitations. They are well-suited for highly-parallelized applications such as BLAST, where calculations can easily be split to run on separate nodes simultaneously. However, other types of applications that require manipulation of an entire dataset need a different kind of system.

Beyond computational issues, most storage systems have shortcomings when it comes to newer bioinformatics analysis applications. The reason: With modern analysis, TBs and PBs of data must routinely be moved from storage systems to compute nodes and vice versa. This issue is currently a challenge for many organizations, and it will continue to grow. Modern sequencers already produce up to a few hundred TB per instrument per year, and that's expected to grow 100-fold as capacities increase and more annotation data is captured. Such large amounts of data will cause a problem because most storage solutions cannot match the performance requirements of such modern workflows. Scalability in capacity and performance is critical, and many storage infrastructures are not up to the task.

A further complication is data management. Administering storage systems and protecting the raw data drain IT resources requiring staff to dedicate a significant amount of time to this one aspect of operations.

New application areas are placing even higher demands on organizations. For example, there is a growing use of whole human genome sequencing in life sciences organizations today. A binary alignment/map (BAM) file — which contains the sequences, base qualities, and alignments to a reference sequence — for a 30x whole genome is about 80-90 gigabytes in size. The BAM files for a modest sample size (1,000) might consume 80 terabytes of disk space.ii

> "HPC solutions for bioinformatics research are complex and need lots of fine-tuning."

A compute solution must have a storage infrastructure sized to handle such data volumes cost-effectively without impacting performance. And since much of the data must be retained, a system must ensure data integrity over the lifetime of the data.

Additionally, the traditional approach to accommodating data growth by adding raw storage capacity does not address the analytics workflow performance issues found in life sciences organizations today. Complicating matters is the fact that the richer datasets from today's next-generation lab equipment make the experimental data useful to researchers from different disciplines. For example, it is quite common for scientists specializing in bioinformatics, genomics, proteomics, systems biology, and other research areas to all use the same experimental data. Each discipline uses different analysis tools, each with varying compute, throughput, and IOPs requirements.

The bottom line is that traditional clusters consisting of commodity servers and storage might not be the best solution to conduct bioinformatics research today. The compute power may not scale, the storage systems may not match the performance requirements of modern workflows, and data management tasks may drain IT resources requiring staff to dedicate an impractical amount of time to this one aspect of operations.

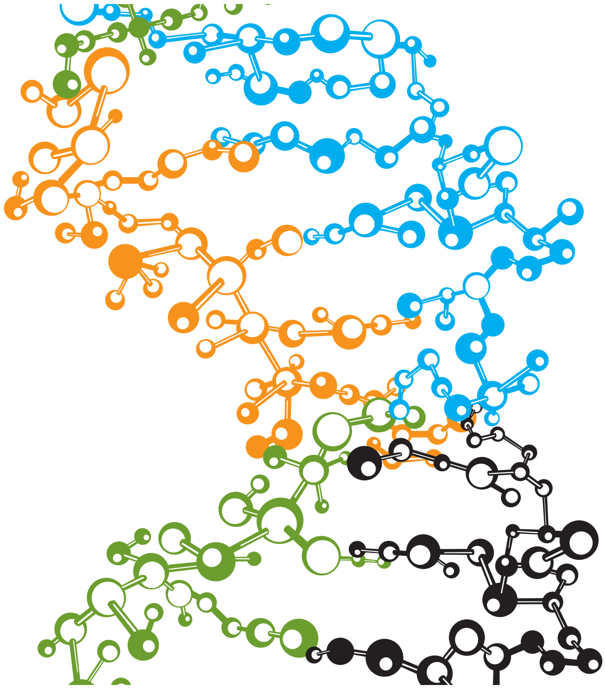
What is needed?

Organizations will need to re-examine their compute capabilities to accommodate current and future life sciences research and clinical analysis workloads compute capabilities. While every organization will have different requirements, several common elements will be needed. They include:

Access to highly scalable HPC resources: Typical life sciences workloads today require enormous amounts of raw compute power. Speed-to-results is critical whether it a commercial setting where an organization is racing to identify a new drug candidate or in a clinical setting where doctors are trying to identify genetic markers to personalize a patient's treatment.

An ability to run workloads with widely varying compute, memory, storage, IO, and throughput requirements: Certainly, some life sciences organizations will have a standard set of workflows that need to be supported. If that is the case, IT infrastructures can be optimized for those workflows. However, given the rapid changes in genomics research and the increasingly multi-disciplinary nature of the work, it is more likely that organizations will have to support a mix of very different workflows. Some may be highly-parallelized and lend themselves to run on distributed systems; others may require solutions with large shared memory.

High-performance networking and storage: Many organizations have invested in processing power over the years, adding more compute nodes to clusters or installing higher performance systems. This additional raw processing power alone will not necessarily improve bioinformatics analysis workflows. With the large volumes of data that are routinely analyzed, the network infrastructure must be improved to ensure data is fed to the processors promptly. Similarly, storage solutions must offer the IO and throughput to ensure workflows are not limited by the rate at which data is moved or read and written to a drive.



Cost-effective scalability: Organizations keep adding capacity to support the massive data volumes and more robust compute requirements of today's genomic analysis workflows. Obviously, over-provisioning on both the compute and storage sides would help improve any workflow. However, most organizations have limited budgets and cannot indiscriminately add capacity. Organizations must find solutions that deliver higher performance per dollar with an eye on reining in operating costs. Fortunately, many newer compute and storage solutions include extensive built-in monitoring and systems management capabilities, take less data center space, and use less electricity for power and cooling.

Simplified data management and protection: The volumes of data being generated by next-generation lab equipment and used in genomic analysis workflows are driving up storage capacity requirements. Much of the data is used at different times (by different researchers or groups) depending on the type of work carried out by an organization. This means that once data is generated by an NGS, imaging system, or other lab equipment it must be saved to disk and over time staged on high-performance storage for analysis, made accessible for future analysis, and archived and protected for the long term.

IT infrastructures that incorporate these characteristics and address these issues will enable innovative life sciences research and offer a competitive differentiator. Committing to such capabilities can help an organization lead the field in discoveries and attract top talent.

Selecting the right technology partner

Bioinformatics research today is greatly aided by the availability of new technology. HPC systems can be assembled that incorporate the latest generation of ever-more powerful processors, faster memory, high-performance storage drives and arrays, faster interconnect solutions, and more sophisticated bioinformatics analytics software.

Organizations have many choices in these technology areas to build a system that matches their needs. Unfortunately, most researchers, labs, and even large life sciences organizations do not have the expertise or time to assemble a system and manage it over time. Additionally, HPC solutions required to conduct bioinformatics research today are complex and need lots of fine-tuning to optimize performance.

What's needed is a technology partner that offers both HPC and bioinformatics expertise. This is an area where PSSC Labs can help. PSSC Labs has years of experience providing HPC solutions that meet the most demanding bioinformatics workload requirements.

> "There is growing interest in HPC systems that are optimally tuned and easy to manage."

PSSC Labs leverages partnerships with leading hardware and software companies to design, build, and install systems that deliver the processing power, storage capacity, and data management capabilities needed by bioinformatics researchers.

The company specializes in bringing HPC to small and medium organizations. It has expertise in:

- Helping researchers and organizations prepare for today's big data bioinformatics workflows
- Setting up and optimizing Hadoop clusters
- Using object storage to accommodate today's large bioinformatics datasets and simplify data management.

Complementing its expertise in designing and installing HPC systems, PSSC Labs has years of expertise in bioinformatics through customer engagements and partnerships with companies including 454 Sequencing, CLC bio, and BioSoft Integrators (BSI).

The BSI partnership illustrates some of the significant benefits offered by using a PSSC Labs system. The company was founded by Henry Marentes and Stu Shannon, who between them have over 40 years in the fields of software, hardware, and laboratory integration. Both were early thought leaders at Illumina in early 2000, responsible for the design, integration, and implementation of the "Bead Lab" systems and Infinium LIMS and Automation.

"We can offer a white-glove, turn-key system," said Marentes. "And we know it will not need constant attention." He notes that BSI has shipped more than 30 PSSC systems and has only had one disk problem in all the systems over their entire operating lifetime. Not having to do service calls lowers BSI's operating costs and most importantly, gives researchers uninterrupted access to their HPC resources.

Marentes noted that installation of a complete system (hardware, software, boinformatics analysis software, etc.) is a relatively fast process. PSSC and BSI configure and test everything at PSSC Labs. Once delivered to a site, the system can be up and running in a day.

Recently, the two companies introducediii the PowerWulf Bio Titanium Cluster, a plug-and-play supercomputing solution with proven compatible with all leading sequencing platforms. The cluster features:

- Latest Intel Xeon processors
- Memory for large datasets
- High-performance network backplane providing up to 100 Gbps speeds
- Large storage capabilities ranging from 20TB to multiple PBs.

"Our partnership with PSSC Labs over the last six years has allowed us to provide the fastest and most reliable computing solutions to our customers across the globe," said Stu Shannon Co-Founder and COO of BSI. "PSSC's expertise in high-performance computing components and server architecture allowed us to configure a best-in-class turn-key solution for our worldwide customer base to support rapid time to answer, high availability, and complete customer satisfaction."

Summary and Call to Action

Bioinformatics is an increasingly data-driven endeavor today. HPC requirements for rapid analysis are high and will continue to grow as more data is used and as the use of bioinformatics analysis expands from traditional areas like new drug discovery to clinical areas, biomarker identification, and translational medicine.

While compute technology has made such research possible, most scientists and labs do not have the time or expertise to assemble a suitable HPC solution. As a result, there is a growing interest in the use of turn-key solutions that combine the right hardware, software, and bioinformatics analysis software into an easy to use system. The ideal solution will tightly integrate with the most commonly used next-generation lab equipment. And the entire solution will be tuned to optimize today's boinformatics computational workflows.

For more information about turn-key HPC solutions optimized for today's bioinformatics and genetics analysis workloads, visit: *https://www.pssclabs.com/*

Sources

[i] http://www.genome.gov/sequencingcosts/

[ii] http://massgenomics.org/2014/11/brace-yourself-for-large-scale-whole-genome-sequencing.html

[iii] https://www.pssclabs.com/company/press/pssc-labs-partners-with-biosoft-integrators-to-debut-specialized-genetic-research-cluster-at-ashg/