CASE STUDY

Accelerating Genomics Analysis with Mellanox InfiniBand Solutions

BACKGROUND

One of the biggest catchphrases in modern science is Human Genome: the DNA coding that largely pre-determines who we are and many of our medical outcomes. By mapping and analyzing the structure of the human genetic code, scientists and doctors have already started to identify the causes of many diseases and to pinpoint effective treatments based on the specific genetic sequence of a given patient.

The field of precision medicine cannot, however, take flight without high performance technology behind it. One such technology is Dell's Genomic Data Analysis Platform (GDAP), which is capable of assimilating and processing the strings of billions of letters that make up each person's genome. GDAP searches for specific sequences within the string of letters to understand the underlying molecular pathways for diseases. For some diseases, every hour closer to discovery and result can mean the difference between life and death.

REAL-WORLD APPLICATIONS

With genomic analysis, doctors can more precisely offer treatment decisions, at times when no other clinically relevant treatment options exist. With the advanced data that such analysis provides, doctors can offer more targeted strategies for potentially terminal patients. Plus, there are side benefits such as improved cost structures, stabilized insurance models, and higher patient satisfaction.

This has become especially useful in treating pediatric cancer patients. In certain aggressive types of cancer, a patient would have little chance of survival without a precise reading of the genomic code behind the disease, and the tailored treatment options that result from such analysis. One such cancer is neuroblastoma, which commonly affects children younger than age 5. Neuroblastoma often arises in and around the adrenal glands or in areas where groups of nerve cells exist. It is very difficult to treat effectively and has a high rate of recurrence, unless the genetic coding of the cancerous cells can be pinpointed and treated accordingly.

Figure 1. *Genomes are billions of letters long, creating hundreds of gigabytes of raw data*

HIGHLIGHTS

Upon implementing GDAP with Mellanox InfiniBand interconnect, the resulting time-to-analysis was as little as one hour.

A performance boost of over 160X was achieved by implementing high performance computing combined with lightning-fast interconnect to overcome data transfer bottlenecks from storage and the file system to the compute cluster, as well as RDMA to avoid multiple copies of the sequencing data.

BIOINFORMATICS

The analysis of the human genome is a very specialized, expert skill that falls into the interdisciplinary field of Bioinformatics, which uses today's best technology to retrieve, store, organize, and evaluate biological data in order to solve practical issues, for example, medical conditions. This process is currently very expensive and very time-consuming, but thanks in large part to partnerships between the scientific community and the leaders in the technological industry, ther have been great advances toward Bioinformatics becoming a mainstream capability.

These advances include standardized best practices, hundreds of open source analytics tools, added automation, upgraded computational performance, and even funding to bring the cost of Next Generation Sequencers (NGS) below $1,000.

One such partnership between high performance computing and Bioinformatics has been that of Dell's GDAP with the Translational Genomics Research Institute (TGen). TGen uses NGS technologies to retrieve human genomes of pediatric neuroblastoma patients, while GDAP provides the computing power to analyze the data into relevant results.

However, there are a number of technological challenges that the project has faced, and Mellanox has been instrumental in participating in overcoming these challenges.

DATA EXPLOSION

Originally, sequencing hardware would take a single tissue sample and used chemicals to break down the bonds so that the DNA could be sequenced. The process took a long time and was highly prone to error. Today's NGS hardware can take multiple tissue samples in order to improve accuracy, and, using light, chemicals, and integrated circuits, convert them into binary code.

Depending upon the number of samples run at once, the data set from a single genome can range from 100GB (for 10X samples) to as much as 500GB (for 50X samples) of code. However large that may seem, it pales in comparison to the amount of raw data produced when the code is then analyzed. A single genome can require up to 5TB of raw data storage for analysis. Moreover, this is just scratching the surface in terms of data collection. The National Center for Biotechnology Information Genebank database is doubling in size every 10 months because of the explosion of genomic data.

INFRASTRUCTURE AND NETWORKING CHALLENGES

The computational and storage infrastructure that is necessary to handle the storage and processing of genomic data can be exceptionally cumbersome for clinical institutions. The upfront costs are very high, and there is significant ongoing maintenance required. It is therefore not uncommon for shared resources to be used, but that, too, comes with its own issues. Shared resources often mean limited access, and there can be confidentiality and security concerns as well.

Besides the physical cluster requirements, there is a need for massive computational and networking capabilities to handle the large amounts of data and the complex genomic analysis. This requires high-performance computers, large storage capacity, large memory, high bandwidth to bridge storage and compute resources, and low latency to enable lightning fast multi-node computation.

TIME CONSTRAINTS

A full genomic analysis using traditional networking solutions can take up to a week. With certain cancers, however, even a few hours can make a huge difference in a patient's chances of survival. Something needed to be done to speed up the process to give precision medicine a chance at being effective.

MELLANOX SOLUTION

Dell offers two cluster fabric interconnect options with the Dell Genomic Data Analysis Platform, which a client can choose based on the site's needs. One option is 10Gb Ethernet and the other is Mellanox InfiniBand. In benchmarking trials run using the BC-BIO Python framework for genomics, the InfiniBand protocol, supplied by Mellanox significantly outperformed 10Gb Ethernet.

As such, TGen opted for the Mellanox InfiniBand as the interconnect protocol for its GDAP implementation. Mellanox helped to sculpt a network that included:

- Mellanox InfiniBand Host Channel Adapters (HCA) in the compute cluster
- Mellanox SX6036 InfiniBand switches
- RDMA based-storage
- An InfiniBand connection for the NFS appliance via I/O servers with HCAs
- HCAs for metadata servers and object storage server for connection to the Lustre file system
- Performance optimizations for the pipeline

7 DAYS TO 1 HOUR

TGen set a goal to reduce its time-to-analysis within its existing sequencing system. Upon partnering with Dell, the goal was set at changing the analysis time from the existing standard of one week down to five days. This would provide an extra couple days in which doctors might provide a pinpointed treatment plan for patients with aggressive time-critical illnesses.

Incredibly, upon implementing GDAP with Mellanox InfiniBand interconnect, the resulting time-to-analysis was as little as one hour! With high performance computing combined with lightning fast interconnect that overcomes bottlenecks in the transfer of data from storage and the file system to the compute cluster, as well as RDMA to avoid multiple copies of the sequencing data, there was a performance boost of over 160X.

CONCLUSION

Specifically for pediatric cancer patients, every day that treatment is delayed is potentially deadly. By identifying specific vulnerabilities within the cancer cells and offering customized treatment plans to fight them, Genomic analysis provides hope for such patients.

TGen is leading the way toward making such genomic analysis mainstream in the hopes of significantly improving survival rates for pediatric oncology patients. It therefore partnered with Dell to apply the Genomic Data Analysis Platform to its Next Generation Sequencing cluster, with the goal of reducing time-to-analysis from one week to five days.

With Mellanox InfiniBand introduced as the interconnect protocol of choice, and Mellanox-supplied InfiniBand switches and Host Channel Adapters, the resulting boost to performance actually reduced time-to-analysis from one week to about an hour.

Mellanox is proud to have accelerated the fight against pediatric cancer and helped bring precision medicine to a point where it can be clinically relevant.

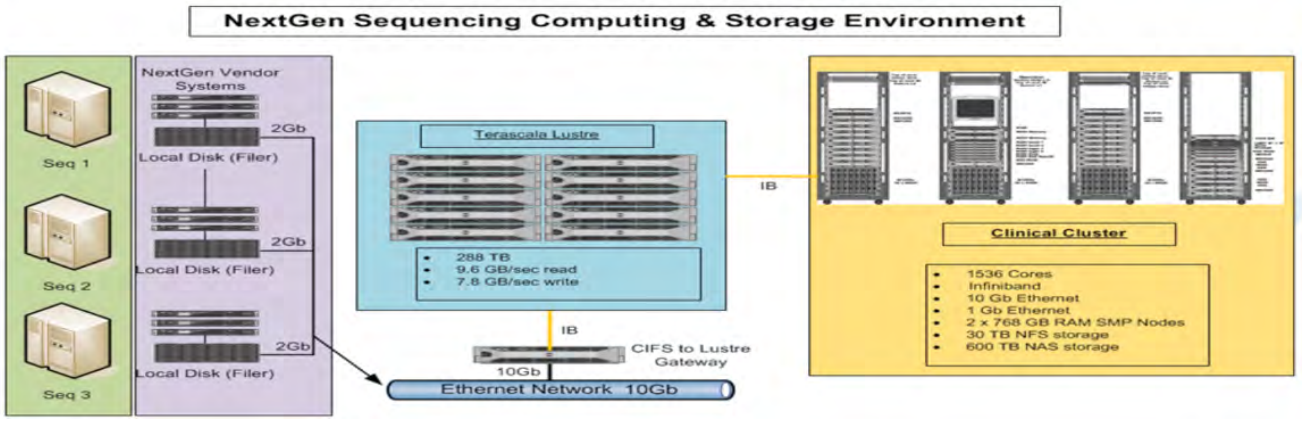

Figure 2. *Dell's Genomic Data Analysis Platform solution for TGen, featuring Mellanox InfiniBand*

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

15-4083CS
Rev 1.1