insideHPC

The insideHPC Guide to

Genomics

BROUGHT TO YOU BY

DELL | intel

Introduction

There are times when a convergence of technologies happens that can benefit a very large number of humans in order to improve their well-being. A number of technological innovations are coming together that can greatly enhance the recovery from life-threatening illnesses and prolong and improve the quality of life.

With a combination of faster and more accurate genomic sequencing, faster computer systems and new algorithms, the movement of discovering what medicine will work best on individual patients has moved from research institutions to bedside doctors. Physicians and other healthcare providers now have better, faster, and more accurate tools and data to determine optimal treatment plans based on more patient data. This is especially true for pediatric cancer patients. These fast-moving technologies have become the center of a national effort to help millions of people overcome certain diseases. "Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes — and to give all of us access to the personalized information we need to keep ourselves and our families healthier," stated President Obama in the 2015 State of the Union speech in January. This was followed up with a White House speech on more of the specifics of the program. From the White House fact sheet on this proposal, *"Building on President Obama's announcement in his State of the Union Address, today the Administration is unveiling details about the Precision Medicine Initiative, a bold new research effort to revolutionize how we improve health and treat disease."*

Personalized (or "precision," per President Obama) medicine holds the key to innovative approaches to manage diseases on an individual level. Various decisions regarding the management of healthcare to each individual will be customized, based on the knowledge of the person's genetic or cellular information. Diagnosis of diseases, treatment and drug use can be tailored for each person. However, a number of challenges exist as this scientific field moves forward, such as regulatory oversight, intellectual property rights and patient privacy.

Contents

Benefits to Organizations

Besides the obvious benefits to individuals, who will receive more targeted diagnosis and treatment, organizations that implement or contribute to this cause can expect a number of benefits.

> A well-designed and well-tuned architecture of the compute, storage, networking and access of large-scale resources can contribute to easy end-user access to HPC resources, as well as making maintenance and support easier for the IT organization.

INNOVATION

IT departments of an organization that can provide for seamless access to the latest systems equipment will accelerate new discoveries. Researchers can focus on their scientific research without having to delve into IT issues. A well-designed and well-tuned architecture of the compute, storage, networking and access of large-scale resources can contribute to easy end-user access to HPC resources, as well as making maintenance and support easier for the IT organization.

LEADERSHIP

An organization that makes better use of its leading-edge IT infrastructure draws additional opportunities for new grants and improves its ability to recruit new researchers. State-of-the-art facilities, including the latest computing and storage systems, allow for researchers to concentrate on their new research with the latest tools. Organizations that are confident in their use of modern IT infrastructure can help other institutions as needed, by providing consulting on best practices as well as assisting in research opportunities.

PATIENT OUTCOMES

An organization that works directly with patients can demonstrate and quantify better patient outcomes, not only helping patients, but also increasing the organization's visibility. Using the latest techniques in personalized medicine leads to better outcomes for patients, which in turn raises the bar for all competing organizations. As precision medicine enables treatment based on individual symptoms and diagnosis, there is less likelihood of a misdiagnosis and more options for the treatment of rare diseases. (The Center for Rare Childhood Diseases defines "rare" as fewer than five in a 10k population.)

> Advanced technology enables clinicians to work with larger amounts of data, create more realistic models and determine better treatment paths for individuals. This can also lead to the sharing of information and insights with a variety of organizations.

ENABLEMENT

Researchers usually are consumers of all of the available computing power and storage capacity. When a more optimal solution is available for them to use, more complete simulations can be performed and more outcomes can be simulated. This not only allows for more personalized decisions to be made, but empowers experts in many fields to move their science forward. Advanced technology enables clinicians to work with larger amounts of data, create more realistic models and determine better treatment paths for individuals. This can also lead to the sharing of information and insights with a variety of organizations.

Challenges

A number of challenges exist for both the wider adoption of technologies that can impede personalized medicine workflows and the implementation of such systems.

PERCEPTION

The perception exists that a genome analysis can only be done on hundreds of nodes or an expensive supercomputer. However, optimized systems that include the right hardware and software, architected by experts from leading vendors, can bring genomics analysis to a broad base of researchers and users. In many cases, IT departments look for an immediate ROI, or will quickly look at utilization of the compute/storage cluster. However, it is possible to start small and grow as the needs grow. Careful planning for this expansion allows for servers and storage to be added incrementally. As projects become more complex or the number of users increases, servers can be added to increase the overall capacity of the compute and storage cluster.

If a turnkey-type solution were available with minimal IT expertise needed, departments or smaller companies would be able to take advantage of current and future technologies.

ROI FOR SMALL USE CASES

Small organizations that require significant IT resources may avoid updating their infrastructure to serve the needs of the users. Aside from the confusion as to the scalability of starting with a small system and growing as needs grow, these organizations may not have the staff to investigate a number of alternatives or implement a piece-by-piece purchase path. They may be resigned to using their older technology rather than upgrading, due to a fear of the IT unknown. However, if a turnkey-type solution were available with minimal IT expertise needed, departments or smaller companies would be able to take advantage of current and future technologies.

FDA AND CLIA COMPLIANCE

FDA approval (compliance) is required for devices used to treat and diagnose patient diseases. Clinical uses must endure a safety period. However, the FDA has in place a number of regulations and safety assurances that must be followed when working with patient health, such as certifications when working with lab instruments, appliances, and technology that are used to facilitate patient health. These safety checks can be daunting, thus the need to work with experienced vendor services teams.

SECURITY

Patient data is obviously very valuable and must be kept secure. Genomics is no exception; it is actually even more important to provide security attention and resources to patient health record data. Special tools, processes and products must be used for patient data and must be compliant with all federal requirements.

CLINICIAN PRACTICES

Clinicians using electronic medical records and imaging archives must abide by procedures and protocols in order to comply with the Health Insurance Portability and Accountability Act (HIPAA) and best practices. These practices include various consultations with specialists and experts that may or may not be a part of the existing IT infrastructure. Due to the fact that these records can assist in treatment or diagnosis, they must be accurate and available quickly to those involved in the treatment of patients.

DATA MANAGEMENT

Large chunks of data must be managed in a genomics solution. A single genome is approximately 200GB to 300GB. Even though the data consists of just four letters (with TGAC as its building blocks), there are approximately 3 billion of these nucleotide bases in a single person. The data from the sequencer is a very large data file that must be accessed, stored and acted upon. Analyzing the genome magnifies the need for nearby storage, scratch storage, archival storage and network bandwidth.

Successes

NEUROBLASTOMA AND MEDULLOBLASTOMA TRANSLATIONAL RESEARCH CONSORTIUM (NMTRC)

The NMTRC is a group of 18 universities and children's hospitals headquartered at the Helen DeVos Children's Hospital in Grand Rapids, Michigan. The group offers a nationwide network of childhood cancer clinical trials. These trials are based on the research from a group of collaborative investigators that are linked with laboratory programs and developing novel therapies for high-risk neuroblastoma and medulloblastoma.

> It's a team-based approach that includes bioinformatics, genomics, oncology and pharma in a way that really delivers on the promise of improved outcomes in the lives of children that participate in our studies."
>
> *– Giselle Sholler, MD MSC chair, NMTRC*

"Working with partners like Dell, TGen and NMTRC, we're seeing an entirely new reality in patient care, starting with clinical trials," says Giselle Sholler, MD MSC chair, Neuroblastoma and Medulloblastoma Translational Research Consortium and endowed director of the Haworth Innovative Therapeutics Clinic at Helen DeVos Children's Hospital. "In this new model, information technology is the bridge that connects all the clinical disciplines for truly personalized patient care. It's a team-based approach that includes bioinformatics, genomics, oncology and pharma in a way that really delivers on the promise of improved outcomes in the lives of children that participate in our studies."

A leading research clinic studies a wide range of areas, including biomedical engineering, cancer biology, cellular and molecular medicine, genomic medicine, immunology, molecular cardiology, molecular genetics, neurosciences, ophthalmic research, pathobiology, stem cell biology and regenerative medicine, and oncology research.

> The results of using the Dell cluster were very positive. Initial run time for a methylation status analysis was reduced from 20 hours to four hours. False discovery rate calculation was reduced from one week to 15 hours.

This research institute first discovered that storage was needed beyond a common desktop computer, and thus added a petabyte storage system. The next step was to implement a computer system that could respond to the most demanding computational problems. The customer turned to Dell and Intel® to provide a solution that consisted of the latest Intel® Xeon® processors in Dell PowerEdge™ servers, which provided multiple teraflops of performance with many terabytes of high-performance storage. Software tools included CentOS Linux, Bright Cluster Manager®, OpenMPI library, GNU Compiler Collection (GCC), Simple Linux Utility for Resource Management (SLURM), Intel Solutions for Lustre® software and the Intel Math Kernel Library (Intel MKL). This organization was able to scale their infrastructure and translate the clinical needs to actionable workflows that help patients.

The results of using the Dell cluster were very positive. Initial run time for a methylation status analysis was reduced from 20 hours to four hours. False discovery rate calculation was reduced from one week to 15 hours. Multiple runs were now made in a week, as compared with months. Other examples exist of reducing run times from weeks or days to hours. Multiple runs are now possible, which generate the data needed for correlation with cancer types.

In addition to genomic analysis, other science domains use the Dell cluster to enhance their own research and patient records. This includes natural language processing and free text patient notes. Physicians' hand-written notes are now able to be scanned and converted into text and made part of the electronic health record. Techniques typically associated with structural mechanics such as finite element analysis are being used on the Dell cluster to perform volume simulations on bones, and simulate passive flexion of the knee joint. Run times are reduced from 20 hours to one hour and by 75 percent in another case. Thousands of simulations that were not able to be run previously can now be run.

TGEN

TGen helps fight cancer and other diseases through the use of genomics. TGen realized that speed and precision are key to a patient's survival. To achieve this speed, they found they needed high-performance computing (HPC) to quickly run very complex algorithms. Terabytes of genetic and molecular data are available from patient and research databases. Custom treatments are needed based on the patient's genome and other biological information.

In order to improve the turnaround time for genomic analyses and create a more customized treatment plan, TGen turned to Dell to deploy an HPC cluster, which would accelerate the time to get results. The Dell Genomic Data Analysis Platform (GDAP solution consists of Dell PowerEdge servers with Intel® Xeon® processors, storage arrays and management software.

Time is critical when diagnosing and creating a customized treatment plan. With the Dell DGAP, the time needed for genetic sequencing has been reduced considerably, as well as the analytical processes that facilitate custom treatment from seven days to four hours.

"I'm not aware of any other solution on the market that's like the Dell Genomic Data Analysis Platform. It's optimized for genomic workflows out of the box, and within a few days, you can install, configure and launch it into production," says James Lowey, vice president of technology, Translational Genomics Research Institute. "Today, we help save more lives because researchers spend less time waiting for HPC resources. And it's also easy for us to scale and customize our Dell Genomic Data Analysis Platform to support our unique requirements."

Center For Rare Childhood Diseases (C4RCD)

The TGen Center for Rare Childhood Disorders (C4RCD) harnesses the latest technologic leaps in genome sequencing to pinpoint the causes of rare childhood disorders that largely remain a mystery to modern medicine.

"By using the Dell GDAP platform, C4RCD is able to process genetic samples quickly," said James Lowey, TGen Vice President of Technology. "This is important as many of the families of these children have been on a diagnostic odyssey, often going years without a clear answer about what is causing the condition of their child. By taking advantage of a system designed to process NGS data, researchers can focus on exploration and discovery, instead of IT infrastructure."

Dell and Intel Solution

Dell has teamed with Intel to create innovative solutions that can accelerate the research, diagnosis and treatment of diseases through personalized medicine. The combination of leading-edge Intel® processors and the systems and storage expertise from Dell create a state-of-the-art solution that is easy to install, manage and expand as required.

> Dell (systems vendor) has teamed with Intel (CPU vendor) to create an optimized solution that takes advantage of the latest technology to deliver an efficient and easy-to-use set of technologies. This solution has been architected, benchmarked and packaged to deliver a solution to a range of users.

Labelled the Dell Genomic Data Analysis Platform (GDAP), this solution is designed to achieve fast results with maximum efficiency. The solution is architected to solve a number of customer challenges, including the perception that implementation must be large-scale in nature, compliance, security and clinician uses.

- **Data explosion –** A single genome will produce between 200GB and 300GB of data. This data must be readily available to the computer systems that will need to decode it. Databanks are doubling in size every few months.
- **Big compute requirements –** With the massive amounts of data arriving in such a short amount of time, the expectation is that the results from computational analysis will arrive in shorter amounts of time as well.
- **Cumbersome infrastructure –** If a system is cobbled together as needs grow, there will likely be a mismatch of the most optimum components. Old systems will have to be networked with newer systems, and a conglomeration of patches, storage mis-matches, etc. will surely surface. The IT department will ultimately have to manage these incompatibilities and deal with lack of expected cluster performance. In smaller organizations that do not have enough dedicated IT skilled administrators, chaos and lack of confidence in the computing systems will surely become an issue.
- **Shareware, middleware, favorite tools** will make their way into the software stack. A defined system will have to deal with these specific applications or middleware.

Dell (systems vendor) has teamed with Intel (CPU vendor) to create an optimized solution that takes advantage of the latest technology to deliver an efficient and easy-to-use set of technologies. This solution has been architected, benchmarked and packaged to deliver a solution to a range of users. More patients' scenarios can be diagnosed and treated due to the following selection of best-in-class components and technologies:

- Intel® Xeon® processors, which become more powerful and energy-efficient with every generation
- HPC servers designed by Dell to deliver fast compute at a lower energy footprint
- Storage hardware from Dell combined with leading software for maximum throughput and data retrieval
- Standardized software, which has been tested and tuned to the hardware selected
- Dell services to work with scientists and clinicians to implement and maintain the solution as needed or act as consultants

Dell GDAP

The Dell GDAP is a complete, integrated genomic processing infrastructure. It is designed to meet the needs of researchers and clinicians, and includes all of the components necessary to reduce turnaround time from days to hours. A detailed report of the solution can be found in reference 1.

Chassis

The solution uses a Dell PowerEdge APC AR3300 Netshelter rack, which includes Intel® Xeon® processors, chosen for its ease of mounting Power Distribution Units and simple cable management.

Login nodes

The solution includes four of the Dell PowerEdge R420 servers with Intel® Xeon® processors which are in the solution as either login nodes for job submittal and monitoring, or head nodes to run the Bright Cluster Manager® for deploying, provisioning and monitoring the servers.

Fat memory node

The Dell PowerEdge R820 with Intel® Xeon® processors is included for certain applications that require a significant amount of memory. This four-socket, 2U-high node has significant memory bandwidth and is available for the mode memory-intensive applications.

Compute nodes

The computational nodes are the Dell PowerEdge M420 Blades with Intel® Xeon® processors which are housed in a Dell PowerEdge M1000e chassis. The PowerEdge M420 Blades are the only quarter-height blades on the market, and have been chosen for their performance per watt and performance per U. (1 U = 1.75 inches in height.)

Storage

The storage component contains the following:

- NFS high-availability solution with raw capacity of up to 180TB
- Dell Terascala High-Performance Storage Solution (HSS) based on Lustre with up to 360TB available in a single name space.
- Dell PowerEdge R320 as the CIFS gateway

Networking

Two options are available, one with a 10-Gb Ethernet component and the other with an InfiniBand® FDR component. With the InfiniBand version, only FDR10 is supported with the Dell PowerEdge M420 blades in a non-blocking mode, and in a 2:1 blocking FDR10 connectivity to the top of the rack switch.

Software

The following software components are contained in the Dell GDAP:

- Bright Cluster Manager for provisioning, monitoring and managing the Dell cluster. Two Dell PowerEdge R420 servers are deployed as the head nodes for running the Bright Cluster Manager.
- Intel Cluster Studio for the development of applications. Components include:
  - C++ Composer XE
  - Fortran Composer XE
  - Math Kernel Library (MKL)
  - Integrated Performance Primitives (IPP)
  - Threading Building Blocks
  - MPI Benchmarks
  - Trace Analyzer and Collector
  - Debugger

Customizations

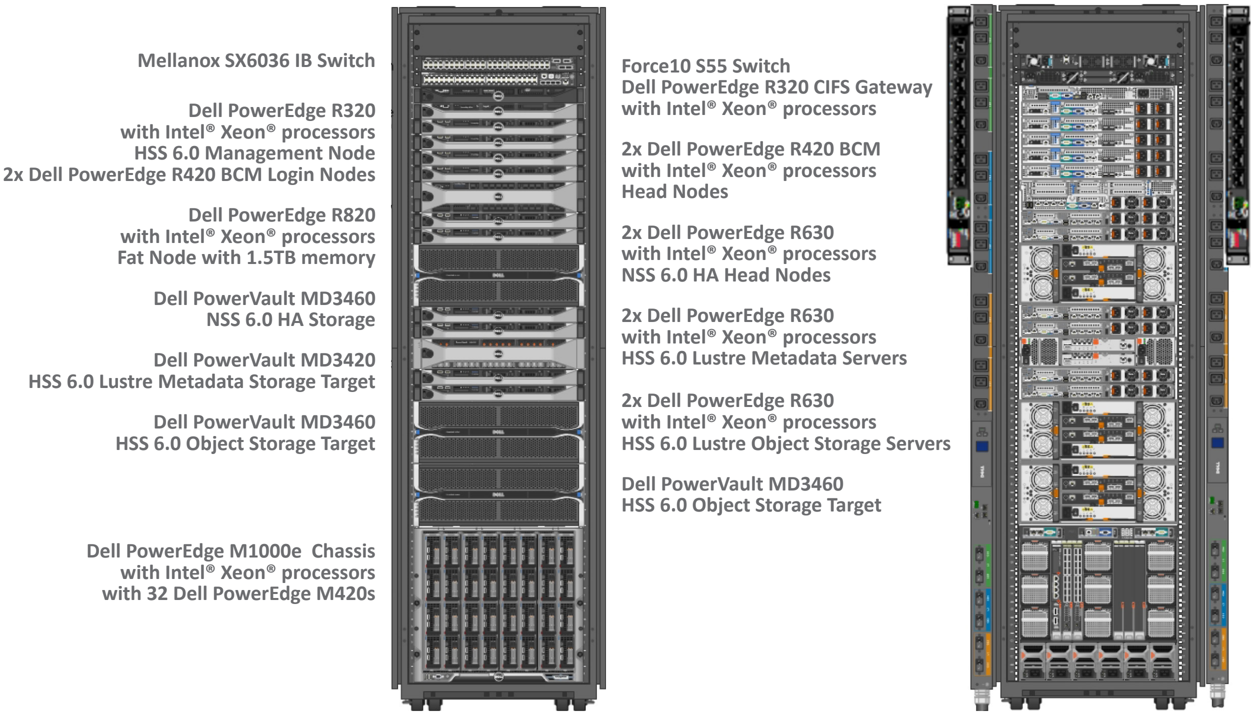
The Dell solution can be modified from the baseline list of components to satisfy varying workloads.

Dell solution benefits

The Dell GDAP can bring a number of benefits to those running genomics applications:

- High availability
- Improved time to insight
- Scalability
- Energy efficiency
- Plug-and-play model

The Dell GDAP has been designed to be flexible, yet with optimized components. Benchmarks have been run that demonstrate the whole genome analysis workflow phases with excellent results of being able analyze up to 37 genomes per day. In addition, the power used over time was monitored, demonstrating excellent performance/power, as expressed in kWh.


Mellanox SX6036 IB Switch
Dell PowerEdge R320 with Intel® Xeon® processors HSS 6.0 Management Node
2x Dell PowerEdge R420 BCM Login Nodes
Dell PowerEdge R820 with Intel® Xeon® processors Fat Node with 1.5TB memory
Dell PowerVault MD3460 NSS 6.0 HA Storage
Dell PowerVault MD3420 HSS 6.0 Lustre Metadata Storage Target
Dell PowerVault MD3460 HSS 6.0 Object Storage Target
Dell PowerEdge M1000e Chassis with Intel® Xeon® processors with 32 Dell PowerEdge M420s
Force10 S55 Switch
Dell PowerEdge R320 CIFS Gateway with Intel® Xeon® processors
2x Dell PowerEdge R420 BCM with Intel® Xeon® processors Head Nodes
2x Dell PowerEdge R630 with Intel® Xeon® processors NSS 6.0 HA Head Nodes
2x Dell PowerEdge R630 with Intel® Xeon® processors HSS 6.0 Lustre Metadata Servers
2x Dell PowerEdge R630 with Intel® Xeon® processors HSS 6.0 Lustre Object Storage Servers
Dell PowerVault MD3460 HSS 6.0 Object Storage Target

References

1) **Dell Genomic Data Analysis Platform:** http://i.dell.com/sites/doccontent/business/solutions/brochures/en/Documents/Brochure-Genomic-Data-Analysis-Platform.pdf

2) **TGEN:** http://www.dell.com/learn/us/en/vn/corporate~case-studies~en/documents~2014-tgen-10013443-scalable-hpc-data-ceter-consulting.pdf

3) **TGEN:** http://i.dell.com/sites/doccontent/corporate/case-studies/en/Documents/2014-tgen-10013443-scalable-hpc-data-ceter-consulting.pdf

4) **NMRTC:** http://onlinelibrary.wiley.com/doi/10.1002/cam4.436/pdf

5) **NMRTC:** http://beatnb.org/about/our-mission/

About the author: Michael A. Schulman

Michael is an experienced writer and marketing professional in High Performance Computing. His interests lie in the areas of how HPC technologies can be used to produce new insights in various technical domains, as well as new HPC technologies that make access easier. His experience includes working at Silicon Graphics, Inc., and Sun Microsystems and other HPC organizations. He has a B.S. and an M.S. from Cornell University. Michael is the Features Editor for insideHPC.

Intel, the Intel logo, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation in the U.S. and/or other countries.