Dell HPC System for Genomics v2.0

A Dell Reference Architecture
HPC Infrastructure for Next Generation Sequencing Analysis

Dell HPC Engineering
November 2015

Nishanth Dandapanthula
Kihoon Yoon

Revisions

Date	Description
August 2015	Updated Dell Genomic Data Analysis Platform to version 2.0
November 2015	Naming updated to Dell HPC system for genomics v2.0

Contents

Executive Summary

Because of the significant cost reduction in Next Generation Sequencing (NGS), the usage of this technique has become widespread from both an academia and an industry standpoint. The cost of NGS data analysis has become directly proportional to the size of the NGS data. Therefore, there is high demand for a plug-and-play solution incorporating massive compute, storage, and networking capabilities to handle this data more cost-effectively.

In August 2013, Dell HPC system for genomics v1.0[1] was introduced to tackle this and many other challenges[1] faced by the lifesciences community. Since then, there have been several improvements to the technology and the software which comprised the solution. A few of them are: the introduction of 13th generation servers of Dell which include the latest Intel E5-2600 v3 (code name: Haswell) processors, updated server portfolio, improved memory, and storage subsystem performance. HPC system for genomics v1.0 was capable of processing 37 genomes per day as measured in our benchmarking. HPC system for genomics v2.0 is aimed at answering the following question: At the pace at which requirements for NGS analysis in terms of compute/storage/networking are growing, will the technological advances deliver the required performance?

This whitepaper describes the architectural changes and updates to the follow-on of HPC system for genomics v1.0, the v2.0. It explains the new features, demonstrates the benefits, and shows the improved performance. HPC system for genomics v2.0 is capable of processing up to 163 genomes per day while consuming 2 Kilowatt-hour (kWh) per genome. That is an almost 4.5x improvement from HPC system for genomics v1.0.

We are very thankful to Dr. Brad Chapman from the Harvard School of Public health for his valuable time and expertise in helping us run the whole genome analysis.

Audience

This document is intended for organizations interested in accelerating genomic research with advanced computing and data management solutions. System administrators, solution architects, and others within those organizations consititute the target audience.

1 Introduction

Next generation sequencing (NGS) has been adopted as a standard method in the life sciences domain. Thanks to the low costs of sequencing, less than $2,300 per genome today[2], there is an explosion of sequencing data requiring complex bioinformatics analysis. State-of-the-art, high performance computing solutions are essential to analyze this data for meaningful results, while keeping up with the speed of data generated. Along with being reliable (system uptime) and easy to maintain, these solutions must provide high throughput processing. However, deploying the required resources, compute and storage infrastructure is not a simple and straightforward procedure.

HPC genomics system of Dell was introduced as a solution to these challenges faced by the contemporary life sciences industry. Some of these challenges are listed here. It was designed to be a plug-and-play turnkey solution so that researchers could spend more time working on matters in their domain rather than concerning themselves with the computer science aspect of getting the system to function, which deals with cluster deployment and maintenance.

- Advancements in low-cost genome sequencing[2]
  - This resulted in a meteoric rise in the volume of data generated, and with it, rose the compute and storage infrastructure required to effectively analyze genomic data.
- Data explosion
  - The size of each genome sample can range from a tens of gigabytes to the order of terabytes. The challenge is in accessing, managing, migrating, and archiving data of this magnitude.
- Compute requirements
  - The increase in complexity and advancements to the pipelines (set of applications) used to process this genomic data has been immense. The requirement in a clinical setting is to drastically reduce the time to insight. This lead to a need for a faster compute infrastructure.
- Managing the Infrastructure
  - Deploying and managing these compute, network, and storage components is challenging without the proper know-how. Cloud computing solutions are starting to catch up, but are not yet able to handle the confidentiality and privacy concerns because of the personal nature of human genomic data.

After the success of the first iteration of HPC genomics system (August 2013), the solution was updated to include the latest advances in the high performance computing industry. HPC genomics system v1.6 moved the compute component from Intel's E5-2400 (Sandy Bridge EN) processor to Intel's E5-2400 v2 (Ivy Bridge) processor. It also incorporated several memory and storage subsystem updates. HPC genomics system v2 is the most recent version of this class of genomic data processing solutions. Further sections in this technical white paper describe the architecture, the updates, and performance characterization of the solution HPC genomics system v2.0.

Note: Coverage of detailed genomic analysis is outside the scope of this document.

1.1 The Solution

HPC genomics system v2.0, similar to its predecessor, is a pre-integrated, tested, tuned, and purpose-built platform, leveraging the most relevant of Dell's High Performance Computing line of products and best-in-class partner products[3]. It encompasses all the hardware resources requied for genome analysis while providing an optimal balance of compute density, energy efficiency, and performance from Enterprise server line-up of Dell.

Figure 1 provides an insight into the updated solution with all its components. A detailed account of all the changes to HPC genomics system v2 from HPC genomics system v1 are listed in section 2.1. HPC genomics system v2.0 provides a 480 TB Intel Enterprise Edition Lustre® file system (IEEL[0]), which acts as the fast scratch space for the solution; and, 240 TB of Network File System (NSS[2.2.2.1]) accessible storage, which acts as primary storage for user or home directories and application data. The data from next generation sequencing instruments can be moved into the Lustre file system for processing via the Common Internet File System (CIFS) gateway. The solution comprises of 40 x Dell PowerEdge FC430 quarter width sleds (1120 cores with a theoretical peak performance of 34 TFLOPS in 10U) representing best-in-class performance/Watt and performance/U rating. These constitute the compute platform for the solution. After these FC430 sleds process the data from the NGS, results can be moved from the scratch space to the primary storage for analysis. Also, the solution includes a Dell PowerEdge R930 with 1.5 TB of memory for customers performing genomic assemblies. All these features make this infrastructure capable of handling the compute and storage requirements of genome analysis workflows.

The software components used in the solution are Bright Cluster Manager (BCM)[2.2.4.1] and Biobuilds from Lab7[2.2.4.2]. BCM is used to deploy, manage, and maintain the solution's head nodes, login nodes, and compute infrastructure. Biobuilds is a collection of opensource bio informatics tools prebuilt for Linux on x86 which is primarily maintained by Lab7.

The above description is of a fully loaded configuration. Some of these components are optional or can be tailored to meet individual customers' requirements. More details of which components can be deleted down are in Table 3. In Figure 1:

- MDS: metadata servers
- MDT: metadata targets
- OSS: object storage servers
- OST: object storage targets

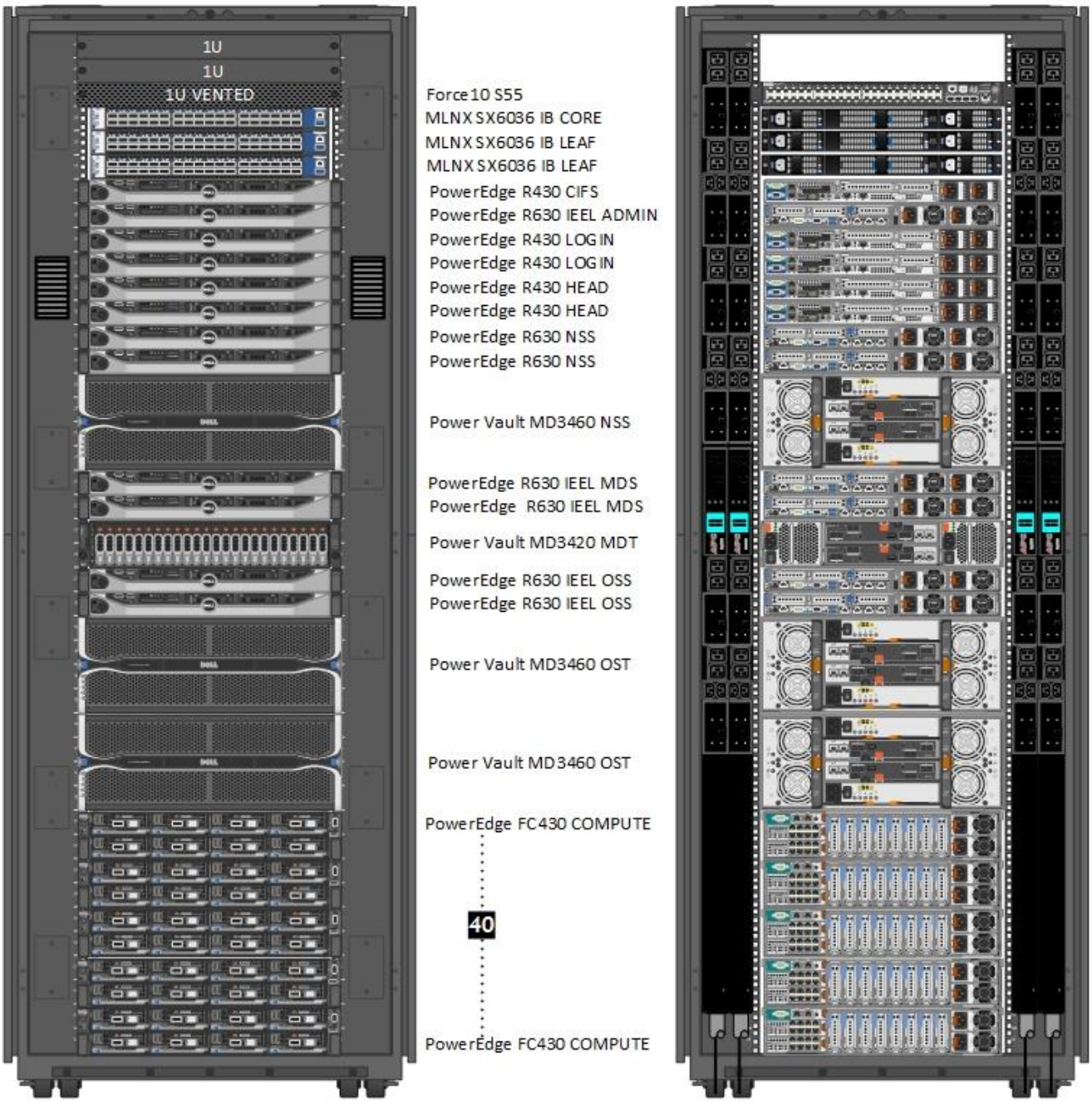

Figure 1 HPC genomics system v2

Note: The PowerEdge FC430 sleds have IB connectivity out the front. The Mellanox InfiniBand switches face the cold isle and have their airflow configured from ports to PSU.

2 Solution Overview

This section describes the architectural changes in v2.0 of the solution from the previous version. It describes the various components used and the rationale for the components' selection that make them optimal for a solution targeted towards genomics. The primary changes are:

- The move from 12th generation hardware of Dell to 13th generation enterprise server portfolio
  - Upgrade to Intel Haswell processors
  - Support for higher wattage processors
  - Upgrade to 2133 MHz memory
- The improvements to the storage subsystems which include NFS and Lustre file systems
  - Updated 12 Gbps SAS controllers
  - Updated to Intel's Enterprise Edition Lustre
  - Updated to NSS6.0–HA
- The updates to the high speed interconnect
  - Updated InfiniBand FDR network
- The updates to the power infrastructure

2.1 HPC system for genomics v2.0: Comparison to v1.0

Table 1 Comparison between HPC genomics system v1.0 and v2.0

Switching component	HPC genomics system v1.0	HPC genomics system v2.0
InfiniBand	1 x Mellanox SX 6036 FDR switch	3 x Mellanox SX 6036 FDR switch
	FDR10 connectivity among compute nodes with a 2:1 Blocking	FDR connectivity among compute nodes with a 2:1 Blocking
Compute Platform	PowerEdge M1000e chassis with 32 x PowerEdge M420 Blades	5 x PowerEdge FX2 chassis with 8 x FC430 servers per chassis
	2 x E5-2470 (2.30 GHz) @ 8c	2 x E5-2695 v3 (2.30 GHz) @ 14c
	48 GB (6x8 GB at 1600 MHz)	128 GB (8x16 GB at 2133 MHz)
Switches in Compute Chassis	2 x Mellanox M4001T FDR10 IB switches in slots B1 and C1	1 x F410T 10 GB I/O Aggregator per FX chassis
Storage	Dell NFS Solution (NSS 4.5 HA) 180 TB raw space	Dell NFS Solution (NSS 6.0 HA) 240 TB raw space
	Intel Enterprise Edition for Lustre Software 360 TB raw space	Intel Enterprise Edition for Lustre Software 480 TB raw space

	Dell PowerVault MD3460 with 6 GB SAS controllers	Dell PowerVault MD3460 with 12 Gbps SAS controllers
Login & Head Nodes	4 x PowerEdge R420s	4 x PowerEdge R430s
Fat Node	PowerEdge R820 with 1.5 TB of memory	PowerEdge R930 with 1.5 TB of memory
	2 QPI links per socket	3 QPI links per socket

2.2 Architecture

HPC genomics system v2.0 provides more flexibility in the number of options for the solution. In the previous generation of the HPC genomics system, the platform was available in two variants, depending on the cluster interconnects selected, which can be either 10 Gigabit Ethernet or InfiniBand®(IB) FDR. In this version, the following options are available:

- PowerEdge FX2 compute subsystem with IB FDR fabric
- PowerEdge FX2 compute subsystem with 10 GigE fabric
- PowerEdge C6320 compute subsystem with IB FDR fabric
- PowerEdge C6320 compute subsystem with 10 GigE fabric

Table 1 shows a bird's-eye view of the updates made in HPC genomics system v2.0 when compared to v1.0. Figure 1 shows the components of a fully loaded rack that is using the PowerEdge FX2 chassis as the compute subsystem and InfiniBand as the cluster high speed interconnect. The solutions are nearly identical for the InfiniBand and the 10 Gigabit Ethernet versions, except for a couple of changes in the switching infrastructure and network adapters. These differences are outlined in section 0. The solution ships in a deep rack enclosure, which was chosen because of its ease of mounting PDUs and for effortless cable management. This rack houses the compute, storage, and networking modules of the solution. Also, there are software modules which deploy, manage, and maintain the cluster.

2.2.1 Compute and Management Components

The compute infrastructure consists of the following components:

- Dell PowerEdge FX2 chassis with 8 x FC430 chassis each
- Dell PowerEdge R930
- Dell PowerEdge R430

2.2.1.1 Dell PowerEdge FX2 chassis with 8 x FC430 chassis each

There were several studies[4] which determided the choice of the compute workhorse of this solution and other details such as the BIOS tuning options[9], and the Intel processor model[5]. The conclusion was that, for a solution targeted at genomic workloads, it is required to optimize the performance per U (density) and performance per Watt (Energy efficiency) rather than focusing just on pure performance. HPC genomics

system v1.0 could accommodate 512 Intel Sandy Bridge compute cores into a 10 U rack space, where as the updated version can pack 1120 Intel Haswell cores in the same rack space.

The move from the Dell PowerEdge M420 to the FC430 servers was complelling for many reasons.

- More density.
    - A single FX2 chassis takes up 2U of rack space and houses 8 servers. 10U of rack space can now accommodate 40 servers instead of the 32 M420 servers in HPC genomics system v1.0.
    - The PowerEdge M420s have a limit on the supported processor wattage (<= 95 W). The FC 430 compute sleds can accommodate up to 120 W processors. This improves the flexibility customers have in terms of picking the processor.
    - The FC430s can also accommodate faster DIMMS (2133 MHz).
- Flexibility.
    - Customers can get compute nodes in multiples of 8. The solution, by default, is configured with 5 x FX2 chassis. If all 5 are not requried, there is an option to delete down and configure the solution with fewer compute servers which helps with flexibility in terms of cost and rack space.
- Improvements to Intel's E5-2600 v3 processor architecture compared to Intel E5-2400 and E5-2400 v2 architectures.
    - QPI speeds
    - Memory controllers

Embarrassingly parallel integer-based sequence alignment codes do not require the most beefy and power hungry processors, which are underutilized in most situations. The 40 PowerEdge FC430 servers configured with Intel Xeon E5-2695 v3 processors provide a sustained 29.6 TFlops of high performance linpack performance putting them at an efficiency of ~87% for 1120 Haswell compute cores, which fulfills the requirement of most applications. Within a chassis, the individually serviceable sleds take advantage of the shared infrastructure by sharing the power and networking components.

2.2.1.2 Dell PowerEdge R930

The Dell PowerEdge R930 is a 4-socket, 4U platform, equipped with the Intel Haswell E7-8860 v3 processors (16 cores per socket – 64 cores in server) and is dubbed "the fat node", because of its 1.5 TB of memory capacity. This processor has 3 QPI links per socket. In HPC genomics system v1.0's fat node, the sockets were connected in a "ring" so that farthest sockets had two hops. The three QPI links in the PowerEdge R930 allow each CPU socket to be connected to its neighbors in the ring and diagonally across.

All these features benefit applications or methodologies, such as DE novo, Velvet, and Velour, which are targeted at genome sequence assemblies. The size of the data on which these applications operate is massive. Hosting this data in memory with 64 cores operating on this data eliminates the overhead caused by interconnects, disk look-ups, and swapping, resulting in a speedup in time-to-results. This server is an optional component and can be added to the solution. One FX2 chassis hosting eight FC430 sleds will be removed to accommodate this 4U server in the HPC genomics system v2.0 rack.

2.2.1.3 Dell PowerEdge R430

The solution includes four Dell PowerEdge R430 servers. Two of these servers are designated as login nodes. Users can log in to these nodes and submit, monitor, or delete jobs. The other two nodes function as redundant head nodes for the cluster, which are used by Bright Cluster Manager®[2.2.4.1] for the purpose of deploying, provisioning, managing, and monitoring the cluster in a high availability (HA) configuration. The head nodes are in an active–passive HA state and use the NSS6.0-HA[2.2.2.1] solution as shared storage.

2.2.2 Storage Components

The storage infrastructure consists of the following components:

- NFS storage solution with HA (NSS6.0-HA)[2.2.2.1]
- Intel Enterprise Edition Lustre solution[0]
- Dell PowerEdge R430 as the CIFS Gateway

Note: The Lustre solution can scale to the order of petabytes. NSS6.0-HA can scale up to 480 TB of raw disk space in a single namespace and up to 960 TB of raw disk space if two namespaces are used. Increasing the capacities will also require more rack space.

2.2.2.1 NSS 6.0 HA

The solution uses NSS6.0-HA[6] as the primary storage for user or home directories and for application data with a raw storage capacity (disk space) of 240 TB.

NSS6.0-HA is Dell's high performance computing network file system (NFS) storage solution optimized for performance, availability, resilience, and data reliability. The best practices used to implement this solution result in better throughput compared to non-optimized network file system. It uses a high availability (HA) cluster to provide a highly reliable and available storage service to the HPC compute cluster; the HA cluster consists of a pair of Dell PowerEdge R630 servers and a network switch. The network switch in this case is the Dell Force10 S3048-ON partitioned with untagged Virtual LANs.

The two Dell PowerEdge R630 servers are used as NFS servers which are in an active-passive mode. A Dell PowerVault MD3460 dense storage enclosure provides storage for the file system. The solution uses a total of 60 x 4 TB, 7.2K near line SAS drives, which amount to 240 TB of raw disk space in 4U. There are 6 virtual disks (VDs). Each VD consists of 10 hard drives (HDDs), and is configured in RAID6 (8+2).

The NFS server nodes are directly attached to the dense storage enclosures via 12 Gbps SAS connections. NSS6.0-HA provides two network connectivity options: InfiniBand FDR and 10 Gigabit Ethernet. NSS's active and passive nodes run Red Hat Enterprise Linux 7.0 with Red Hat's Scalable File System (XFS) and Red Hat Cluster Suite for implementing the HA feature. A detailed study of the design and performance optimizations used in NSS6.0-HA can be found in Reference 5.

2.2.2.2 Intel Enterprise Edition for Lustre Software

Intel EE for Lustre Software[7] is used as the parallel file system in the solution for computational scratch space with a raw capacity of 480 TB under a single name space.

Dell partners with Intel to provide the Dell Storage for HPC with Intel EE for Lustre software solution, a Lustre file system-based storage appliance consisting of a management interface, Lustre metadata servers, Lustre object storage servers, and the associated backend storage which can scale, both in capacity and performance. The management interface provides end-to-end management and monitoring for the entire Lustre storage system.

Figure 2 shows an overview of the components and their connectivity used in the Lustre appliance as a part of the Dell Dell HPC system for genomics solution.

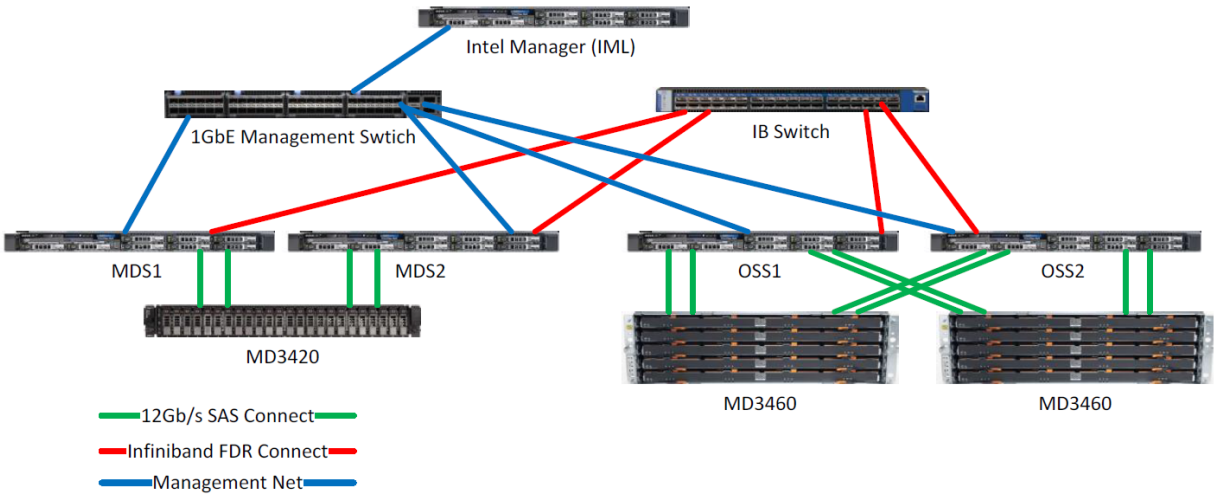

Figure 2 *Intel Lustre appliance components* and logical connectivity diagram

The solution uses two Dell PowerEdge R630 servers as the metadata servers, in an active-passive high availability configuration. The metadata servers are directly attached, via 12 Gbps SAS connections, to a Dell PowerVault MD3420 storage array populated with 24, 15K 600 GB SAS drives, which acts as the metadata target. The meta data target is configured with 22 drives in RAID 10 with two hot-spares. This amounts to 13.2 TB of raw storage capacity, of which, the Lustre solution provides about 6.1 TB of usable disk space for metadata. The metadata servers are responsible for routing file and directory requests to the appropriate object storage targets. These requests are handled across LNET (Lustre Networking Layer) by either InfiniBand FDR or 10 Gigabit Ethernet links, depending upon the type of solution.

The solution uses two Dell PowerEdge R630 servers as Object Storage Servers (OSS) in an active-active configuration. The OSS servers are directly attached to two Dell PowerVault MD3460 dense storage enclosures with 240 TB of raw capacity (each), which provide the Object Storage Targets (OST). Each storage array is fully populated with 60, 4 TB, 3.5", 7.2K near-line SAS drives. Each Dell PowerEdge R630 acts as the active node to one Dell PowerVault MD3460 disk array and as the passive node to the other array,

resulting in a resilient HA configuration. Each OSS has two dual port 12 Gbps SAS controllers, which are connected to each controller of the two Dell PowerVault MD3460 dense storage arrays as shown in Figure 2. Each storage enclosure is configured with six OSTs by dividing the storage array into 6 x RAID 6 virtual disks. Each of the virtual disks consists of eight data and two parity disks, using two disks per tray. Each OST provides ~29 TB of formatted object storage space. There are 12 OSTs per single pair of OSSs.

A Dell PowerEdge R630 acts as the management server and is connected to the rest of the Lustre solution through an internal one Gigabit Ethernet link to the switch. The management server handles management tasks for the Lustre file system via a web graphical user interface (GUI) as shown in Figure 3, which provides an intuitive and user friendly managing and monitoring for all the components. Performance and health metrics for the Lustre file system and servers can be collected for extended periods of time and presented to the user in an easy-to-use, interactive GUI. The GUI allows detailed monitoring of the file system and clients usage of it. The analytics section provide's information about possible problems and even useful trend analysis, which can be used to predict if the scratch space will be exhausted, and how soon it may happen.

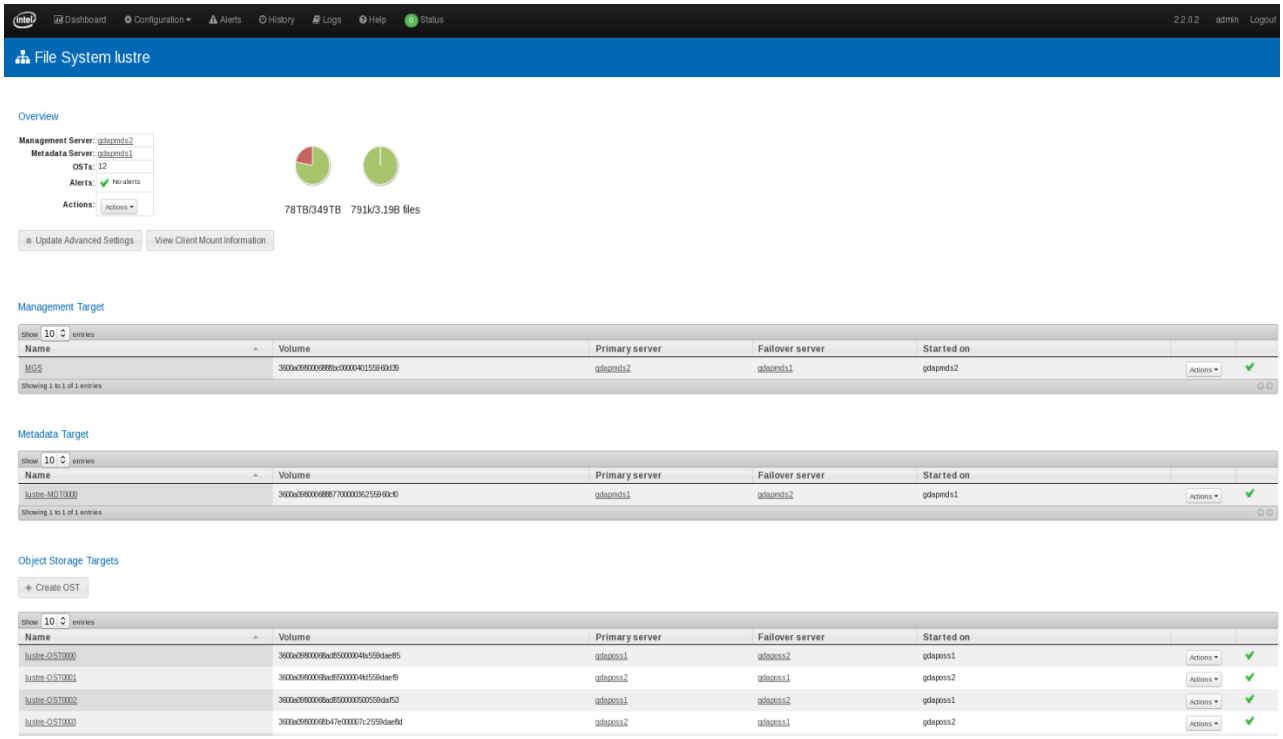

Figure 3 IEEL Management GUI

A Dell PowerEdge R430 is used as the CIFS gateway for transferring data generated by the next generation sequencing machines into the Lustre file system.

2.2.3 Network Components

The Dell HPC system for genomics is available in two variants—InfiniBand FDR and 10 Gigabit Ethernet. The differences in switching infrastructure between the two are shown in Table 2. There is also a Force10 S3048-ON Gigabit Ethernet switch which is used in both configurations whose purpose is described here. In the InfiniBand version, the Dell PowerEdge FC430 sleds have 2:1 blocking FDR connectivity to the top of rack FDR switch.

2.2.3.1 Force10 S3048-ON switch

In the InfiniBand configuration, the Force10 S3048-ON switch ports are split into multiple untagged virtual LANs to accommodate multiple internal networks.

The port assignment of a Force10 S3048-ON switch for the InfiniBand version of the solution is as follows.

- Ports 0–23 are assigned to the cluster's private management network to be used by Bright Cluster Manager®.
    - The FC430 server's Ethernet and iDRAC consititute of a majority of these ports.
- Ports 24–29 are used for the private network associated with NSS6.0-HA.
- The rest of the ports from 30–47 are allocated to the Lustre solution for its private management network.

For the 10GbE configuration, the deployment and management of the cluster is done over the 10 Gigabit Ethernet network by using the Dell Force10 S4820T switch. So, the first virtual LAN on the S3048-ON, from ports 0–16, is not used. The other two virtual LANs are still used for the same purpose as in the InfiniBand configuration.

Table 2 Differences in Switching Infrastructure Between InfiniBand and 10GbE Configurations

Switching component	InfiniBand FDR	10 Gigabit Ethernet
Top of Rack switch	3 x Mellanox SX 6036 FDR switch 1 x Force10 S3048-ON 1 GbE switch	1 x Force10 S4820T 10GbE switch 1 x Force10 S3048-ON 1 GbE switch
Switches/IOAs in Dell PowerEdge FX2 chassis	1 x FN 410T 10GB I/O Aggregator 1 Link per chassis to Force10 S3048-ON	2 x FN 410T 10GB I/O Aggregator 6 links up to Force10 S4820T and 2 links for stacking.
Adapters in Login nodes, head nodes, NFS servers, Lustre Metadata servers and Object storage servers, CIFS gateway	Mellanox ConnectX-3 InfiniBand FDR adapter	Intel X520 DA SFP+ DP 10 GigE low profile adapter
Interconnect on Dell PowerEdge FC430 sleds	Mellanox ConnectX-3 FDR Mezzanine adapter	10 GigE LOM

2.2.4 Software components

Along with the hardware components, the solution includes the following software components:

- Bright Cluster Manager 7.1®
  - MLNX OFED 2.4-1.0.4
  - Red Hat Enterprise linux 6.6 (RHEL 6.6)
  - Lab7 Bio-Builds

2.2.4.1 Bright Cluster Manager® 7.1

Bright Cluster Manager® (BCM)[8] for Dell is a comprehensive solution for provisioning, monitoring, and managing Dell clusters.

Two Dell PowerEdge R430 servers are deployed as head nodes in a HA active-passive configuration by using the NSS6.0-HA solution as shared storage. The active head node is responsible for deploying and monitoring the 40 Dell PowerEdge FC430 sleds, the Dell PowerEdge R930 (if used) and the other Dell PowerEdge R430 servers which act as the login nodes. In a scenario where the active head node fails, Bright Cluster Manager® provides an option of automatically failing over to the second head node, or a failover can also be done manually in case the active head node requires servicing. The BCM image includes Mellanox OFED and Red Hat Enterprise Linux version (RHEL) 6.6, with which, the head nodes and compute nodes are deployed. The Bright Cluster Manager® 7.1 can be used to perform several day-to-day tasks, a few of which are:

- Monitoring made easier with on-the-fly graphs, rack view, multiple clusters, and custom metrics
- Parallel shell to run commands on all or a subset of the nodes effortlessly
- Powerful automation: thresholds, email alerts, and actions
- Integration with key workload managers such as SLURM, PBS Pro, Moab, Maui, Torque, Grid Engine, LSF, and OpenLava
- Simplified patching and updating OS to keep the system secure

2.2.4.2 Lab7 BioBuilds[13]

BioBuilds is a well maintained, versioned and continuously growing collection of open-source bio-informatics tools from Lab7. They are prebuilt and optimized for a variety of platforms and environments. BioBuilds tries to solve the software challenges faced by the life sciences domain.

- Imagine a newer version of a tool being released. Updating it may not be straight forward and would probably involve updating all the dependencies the software has as well. BioBuilds includes the softwares and its supporting dependencies for ease of deployment.
- Using BioBuilds among all the collaboraters can ensure reproducibility since everyone is running the same version of the software.

In short, it is a turnkey application package. More information about Lab7 and Bio-Builds can be found at Reference 13.

2.3 Customizations

The solution described in this paper is a fully-loaded configuration and can be customized to accommodate individual user requirements. Customizations and their respective consequences are listed in Table 3. For example, a fully loaded configuration of HPC genomics system v2.0 has 2 login nodes. They can be deleted down to 0 or 1, if the customer does not have a requirement for login nodes. The storage subsystem (NSS and IEEL), however, is not an optional component. So, Table 3 does not mention any delete-down options for them.

Table 3 Delete Down Options From Fully Loaded Configurations

Component	Standard	Deleted down to	Consequence
Login nodes	2	0 or 1	If no login nodes are present, the head nodes must be used as submission hosts. If a single login node is present, there is loss of HA capability at the login node level.
Head nodes	2	1	Lose HA capability
R930	0	1	If limited by rack space at customer site, will lose an FX2 chassis to accommodate the fat node. This means, loss of 224 cores.
FC430	40	Variable	Density/U varies depending on the number of FX2 chassis blades used. Processors and memory can be customized.

2.4 Application Workflow

The following is the workflow outlining how application data moves through various components of the solution are:

1. Next generation sequencing machines output the data generated by the sequencing operation.
2. This data, residing on a Windows machine, is transferred into the computational scratch space (.

3. Intel Enterprise Edition for Lustre Software) of the Dell HPC system for genomics by using the CIFS gateway.
4. The compute nodes (Dell PowerEdge FX2 chassis with 8 x FC430 chassis each) process the data.
5. The final results and user files are moved to the primary storage which is the NFS Storage Solution (NSS 6.0 HA) and are accessible by researchers for further investigation and analysis via NFS.

2.5 Benefits

This section, summarizes the benefits of the Dell Dell HPC system for genomics Platform solution.

- **High Availability**: The solution has HA at various levels to provide improved resiliency to failures. HA features are present on the head nodes, login nodes, NSS6.0-HA servers, and Lustre metadata servers and object storage servers. There is also HA on the PDU configuration.
- **Improved Time to Insight**: The compute, network, and storage components of the solution have been chosen specifically for genome analysis workloads, thereby substantially decreasing the time required to process sequencing data.
    - High speed Interconnect among compute nodes and distributed storages provide enough bandwidth to handle large amount of data simultaneously.
    - Large memory configuration will cover memory requirements for various genomic data analysis ranging from alignment to assemblies. For example, the minimum requirement to perform a De novo assembly of a mammalian genome is about 512 GB[12].
- **Scalability**: Compute and storage infrastructure can be scaled further based on future requirements. Lustre (performance storage) is highly scalable.
- **Fully customizable solution**: It is simple to change the configuration from one mode to other. For example, it is easy to convert from a configuration supporting high memory applications to a configuration for the application requiring large disk I/Os.
- **Energy Efficiency**: With the usage of lower wattage mid-range processors for the life sciences domain, the solution strikes a critical balance between performance and performance/Watt.
- **Plug and Play Model**: The solution is already customized in terms of servers, memory, storage, network, management software, and deployment tools, and is pre-integrated and tested before shipping in a single 42U rack, making it a no-hassle turnkey solution.

3 Test Configuration

The test configuration used for the results shown in the following sections is the InfiniBand version, using the PowerEdge FX2 chassis as the compute workhorse of the Dell HPC system for genomics. Each component and its details are shown in Table 4. The BIOS settings on all the Dell PowerEdge FC430 blades are set to DAPC (Dell Active Power Controller) profile, which is the default power profile, to optimize performance and energy efficiency[9] .

Table 4 Dell HPC system for genomics Test Configuration

Component	Usage	Details
4 x Dell PowerEdge R430s	2 x Login Nodes 2 x Head Nodes	2 x Intel Xeon E5-2470 v3 (2.30 GHz) 64 GB (8x8 GB 2133 MHz)
40 x Dell PowerEdge FC430 sleds	compute nodes	2 x Intel Xeon E5-2695 v3 (2.30 GHz) 128 GB (16x8 GB 2133 MHz)
5 x Dell PowerEdge FX2 Chassis	Enclosure	1 x Dell PowerEdge FN 410T I/O Aggregator each
3 x Dell PowerVault MD3460	1 x NSS backend storage 2 x Lustre backend storage	60 x 4 TB, 3.5", 7.2K, NL SAS drives
1 x Dell PowerVault MD3420	Lustre Metadata storage	24x600 GB drives (2.5")
4 x Dell PowerEdge R630	Lustre Metadata servers Lustre Object storage servers	2x Intel Xeon E5-2660 v3 (2.6 GHz) 256 GB memory (16x16 GB 2133 MHz)
2 x Dell PowerEdge R630	NSS servers	2x Intel Xeon E5-2697 v3 (2.6 GHz) 128 GB memory (16x8 GB 2133 MHz)

Component	Usage	Details
1 x Dell PowerEdge R630	Lustre Management server	2xE5-2660 v3 (2.6 GHz) 64 GB memory (8x8 GB 2133 MHz)
1 x Dell PowerEdge R430	CIFS gateway	1x E5-2670 v3 (2.30 GHz) 48 GB (6x8 GB 2133 MHz)
Software Component	Operating system	RHEL 6.6
	OFED Distribution	Mellanox OFED 2.4-1.0.4
	Cluster manager	Bright Cluster Manager 7.1®
	Workload Manager	Torque with MAUI scheduler

4 Test Methodology

The solution is stressed and tested with respect to performance and power consumption. The idle power consumed by the solution is measured over a period of time when the solution has completed the boot up process and is not under any workload. To obtain the power and energy consumption at a whole rack level, a Fluke 1735 Three-Phase Power Logger is used[10]. The maximum, minimum, and average power consumption are recorded over the course of the benchmarking effort every 1–5 minutes.

Note: These tests aim at creating new metrics which are relevant to the life sciences domain, rather than using the traditional GFLOPS. The new metrics obtained as a part of this effort are Kilowatt-hours/genome and Number of genomes analyzed/day.

4.1 Whole Genome Pipeline Analysis

The software pipeline framework used to run the whole genome analysis is called bcbio-nextgen[11]. It is a python toolkit, providing best-practice pipelines for fully automated high-throughput, variance calling analysis. A typical variance calling pipeline consists of two major steps; aligning sequence reads to a reference sequence and identifing regions contain mutations/SNPs. In the tested pipeline, Burrows-Wheeler Aligner (BWA) is used for the alignment step and Genome Analysis Tool Kit (GATK) is selected for the variance calling step. These are considered as standard tools for aligning and variance calling in genomic sequence analysis. The reference genome used is the GRCh37 (Genome Reference Consortium Human build 37). For the purpose of benchmarking, we used the 10x coverage whole genome human sequencing data from the Illumina platinum genomes project, named ERR091571_1.fastq.gz and ERR091571_2.fastq.gz[11]. Further tests with 50x coverage whole genomes was done to check for scalability of the solution as the data size increases. The details of the dataset are mentioned in Table 5.

Table 5 The details for read sequence files for benchmarking

Fastq files	Coverage Depth	Number of Raw Reads	Read Length in Nucleotides(nt)	Total Nucleotides
ERR091571_1.fastq	10x	211,437,919	102	21,566,667,738
ERR091571_2.fastq	10x	211,437,919	102	21,566,667,738
ERR194146_1.fastq	50x	813,180,578	102	82,944,418,956
ERR194146_2.fastq	50x	813,180,578	102	82,944,418,956

Since the application (bcbio-nextgen) is optimized to run multiple genome samples in parallel, 78 genomes were run by using a total of 1092 compute cores. The data points obtained from this test are:

- The time taken to analyze 78 genomes. If the time taken to analyze 78 genomes is "x" hours, the number of genomes analyzed/day is calculated by using $\frac{24\ hours * 78\ genomes}{x\ hours}$
- Energy consumed by the whole solution while analyzing 78 genomes. If this value is "y" Kilowatt-hours (kWh), the number of kWh/genome can be calculated as $\frac{y\ \text{kWh}}{78\ genomes}$.
- Time spent in the various analysis steps such as alignment, alignment post-processing, variant calling, and variant post-processing.

5 Results and Analysis

5.1 Whole Genome Analysis

Bcbio-nextgen runs through several phases for the purpose of analyzing the whole genome. The first four phases are common for most genome analysis pipelines and thus are considered for the purpose of this testing. They are the alignment phase, alignment post-processing phase, variant calling phase, and variant post-processing phase. The alignment and the variant calling phases are mostly CPU and memory intensive, whereas the post-processing phases are very I/O intensive.

Table 6 lists the results in terms of performance and power consumption of the whole solution while running the genome analysis pipeline. The metrics introduced in section 4 are also shown in this table. HPC genomics system v2.0 provides (10x data)

- ~4.4X improvement in terms of genomes/day than HPC genomics system v1.0.
- ~3.75X improvement in terms of kWh/genome than HPC genomics system v1.0.

When comparing the metrics HPC genomics system v2.0 while using 10x data vs 50x data,

- ~3x more 10x samples/day than 50x samples.
- ~4x more energy consumed/50x sample than 10x sample.

Table 6 Performance and Power Consumption - Whole Genome Analysis

Parameter	V1.0 (30 Samples 10x)	V2.0 (78 Samples 10x)	V2.0 (78 Samples 50x)
Time taken for analyzing samples	19.5 Hours	11.45 Hours	34.32 Hours
Energy Consumption for analyzing samples	222.7 kWh	154.5 kWh	637.3 kWh
kWh/Genome	7.42 kWh/Genome	2 kWh/Genome	8.2 kWh/Genome
Genomes/day	37	163	54
Idle Power	~7.2 kW	~11.05 kW	~11.05 kW

Peak power during genome analysis	14.35 kW	20.9 kW	20.9 kW

During the course of the pipeline's run, a maximum power consumption of 20.9 kW was recorded during the alignment phase and a minimum power consumption of 11.5 kW was recorded during the variant post-processing phase.

Figure 4 and Figure 5 show the time spent and the energy consumed by the whole solution during each phase of the pipeline while analyzing 78 10x samples by using HPC genomics system v2.0 (blue bar) and compare that to HPC genomics system v1.0 while running 30 10x samples (red bar).

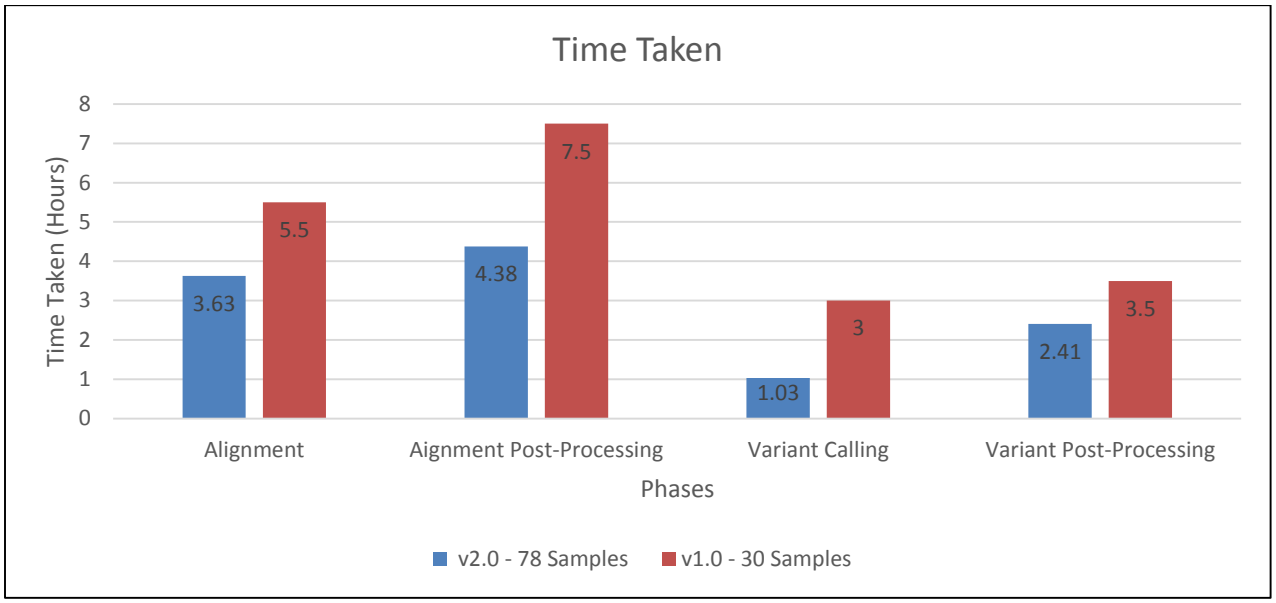

Figure 4 Comparison of time taken at various phases

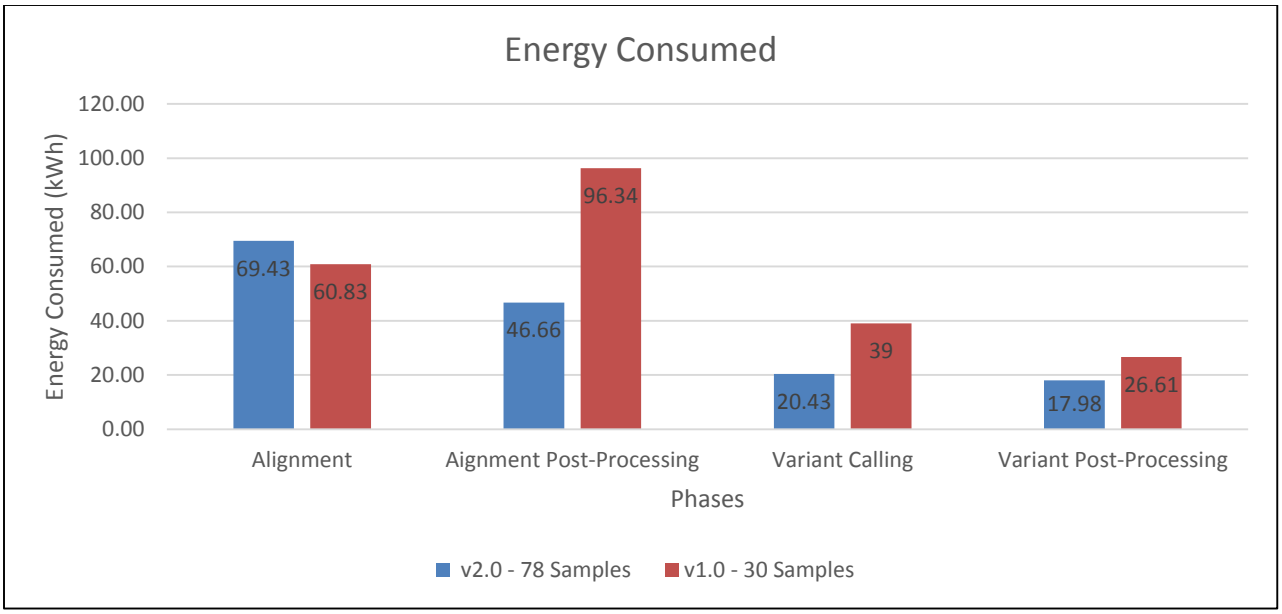

Figure 5 Energy Consumed Comparison

5.2 CIFS Gateway Testing

Actual data files used in the previous benchmarking effort from the Illumina platinum genome project were transferred from a Windows Active Directory (AD) machine to the Lustre file system using the CIFS gateway, and the bandwidth was measured. The details of the experiment are as shown in Table 7.

16.2 GB of data was transferred from the AD machine to the Lustre FS through the CIFS gateway. The file took 233 seconds to transfer.

Table 7 Experimental Setup for CIFS Gateway

Component	Description
Data file used	ERR091571_1.fastq.gz from Illumina platinum genome project
File size	16.2 GB
Lustre FS stripe count	12
Lustre FS stripe size	4 MB

6 Conclusion

This technical white paper demonstrates the improvement in performance because of the architectural changes or updates in HPC genomics system v2.0. The solution was tested by using the whole genome analysis pipeline and the results were compared against HPC genomics system v1.0.

The upgraded solution is capable of processing 163 genomes per day (~4.4X compared to HPC genomics system v1.0) while consuming ~2kWh per genome (~3.75X lower energy compared to HPC genomics system v1.0). The solution also scales almost linearly when the size of the data increases. While using 50x data, the HPC genomics system v2.0 solution can process ~3x more 10x samples/day and take up ~4x more energy/sample.

Most of this improvement can be attributed to the updated server platform in Dell's 13th generation servers. This includes the improved core counts, processor architecture, improved memory speeds and bandwidth, updates to NSS and Lustre solutions, and so on.

Future plans include further experimentation with the Lustre file system's performance tuning, performance of other pipelines. to obtain best practices for the Dell HPC system for genomics. Results from these follow-up experiments with the solution and the genome analysis pipeline will be posted as technical blog posts.

A References

1. Dell HPC system for genomics
   http://i.dell.com/sites/doccontent/business/solutions/whitepapers/en/Documents/active-infrastructure-for-hpc-life-sciences-wp.pdf
   http://www.dell.com/learn/us/en/555/hpcc/high-performance-computing-life-sciences

2. Advancement in genome sequencing
   http://www.genome.gov/sequencingcosts/
   http://www.genome.gov/10000368
   https://www.scienceexchange.com/services/illumina-ngs

3. Dell's work with Translational Genomics Research Institute (TGen)
   http://i.dell.com/sites/doccontent/corporate/case-studies/en/Documents/2012-tgen-10011143.pdf

4. Performance analysis of HPC applications on several Dell PowerEdge 12th generation servers
   ftp://ftp.dell.com/Manuals/Common/Dell PowerEdge-r420_White%20Papers96_en-us.pdf

5. Haswell Processor Studies
   http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2014/10/06/comparing-haswell-processor-models-for-hpc-applications

6. Dell NFS Storage Solution with High Availability
   http://i.dell.com/sites/doccontent/business/solutions/whitepapers/en/Documents/dell-nfs-hpc-storage-solution-nss6-0-ha.pdf

7. Intel Enterprise Edition for Lustre Software
   http://i.dell.com/sites/doccontent/business/solutions/whitepapers/en/Documents/DellHPCStorageWithIntelEELustre.pdf

8. Bright Cluster Manager
   http://www.brightcomputing.com/News-Bright-Computing-Collaborates-with-Dell-for-Genomic-Analysis.php

9. Optimal BIOS settings for HPC
   http://www.dellhpcsolutions.com/assets/pdfs/Optimal_BIOS_HPC_Dell_12G.v1.0.pdf
   http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2014/09/23/bios-tuning-for-hpc-on-13th-generation-haswell-servers

10. Fluke 1735 Three-Phase Power Logger

http://www.fluke.com/Fluke/usen/Power-Quality-Tools/Logging-Power-Meters/Fluke-1735.htm?PID=56028

11. Bcbio-nextgen whole genome analysis
    https://bcbio-nextgen.readthedocs.org

12. Memory requirement for assemblies
    http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3029755/

13. BioBuilds by Lab7
    http://www.lab7.io/
    https://biobuilds.org/