DELL EMC HPC SYSTEM FOR LIFE SCIENCES V1.1

Designed for genomics sequencing analysis, bioinformatics and computational biology

ABSTRACT

Designing flexible HPC architecture requires the best technologies and clever strategies. Dell EMC HPC System for Life Sciences is the result of Dell EMC's ongoing effort to provide customers the most suitable and cost-effective solution. Improving variant analysis performance by more than 17% from Dell EMC HPC Solution for Genomics v2.0 is evidence of Dell EMC's long-standing commitment to the global HPC community.

January 2017

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In October 2015, Dell announced the Genomic Data Analysis Platform (GDAP) v2.0 to address the growing necessity of rapid genomic analysis due to the availability of next-generation sequencing (NGS) technologies [1]. Upon the successful implementation of GDAP v2.0, which is capable of processing up to 163 genomes per day[1] while consuming 2 kilowatt-hour (kWh) per genome, we started to explore the life science domains beyond genomics. In addition to NGS, other instruments used in life science research, e.g. mass spectrometers and electron microscopes also produce much more data than before. The analyses required to extract information from these various data formats are highly variable and, thus, it is difficult to optimize a single system architecture to suit all the different use cases. The Dell EMC HPC System for Life Sciences v1.1 is a flexible high performance computing environment designed to address the computational challenges in genomic sequencing analysis, bioinformatics and computational biology.

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 constitute the target audience.

INTRODUCTION

In the last decade, modern society has continued to improve the quality of life through better healthcare, producing and consuming sustainable food and energy, and protecting our environment. All these societal advancements are tightly related to progress in the life sciences domain. Bioinformatics emerged from the massive amount of data that is now available in these fields and the advancement of HPC takes a key role to finding solutions.

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data [2].

Here are some examples how HPC has been helping the life sciences;

- **NGS data analysis**: Affordable genome sequencing and 'omics' technologies have driven the need for new bioinformatics methods and HPC.
- **Understanding biochemical reactions in biomolecular systems**: HPC is also a key component in advancing our understanding of biochemical reactions and discovering new therapeutic molecules using molecular dynamics/mechanics and quantum mechanics.
- **Modeling/simulating biological networks**: HPC is required to model biological networks and simulate how networks behave in response to perturbations. This is the area frequently referred as pathway analysis, and Boolean networks are used as a simulation platform. The results can help identify adverse outcomes for various disease treatments.
- **Constructing the 3D structural images of biomolecules from electron microscopy**: Cryo-electron microscopy (Cryo-EM) is gaining popularity in structural biology due to the recent improvement in its resolution. This is a device that uses electron beams to photograph frozen biological molecules and generates terabytes' worth of data. HPC is an essential tool to reconstruct images of biomolecules from the large volume of data.
- **Simulation of human organ functions**: HPC is used to integrate diagnostic images, DNA/RNA profiles and protein expression data into organ function simulation.

[1] 163 genomes per day was obtained when 'UnifiedGenotyper' in a previous version of Genome Analysis Tool Kit (GATK). In GATK 3.5 used in this study, 'HaplotypeCaller' is recommended. The pipeline was updated according to the 'GATK Best Practices and Beyond' guideline. GDAP v2.0 performed 133 genomes per day on the updated BWA-GATK pipeline.

The most accurate simulation of the human brain to date has been carried out in a Japanese supercomputer, with a single second's worth of activity from just one percent of the complex organ taking one of the world's most powerful supercomputers 40 minutes to calculate *[3].*

Not only could multiple sub-fields of life sciences benefit immensely from the massive growth in HPC system performance, but someday life sciences as a whole will use HPC as a tool to integrate all the cellular data, biochemistry, genomics, proteomics and biophysics into a single frame of work.

With this overarching goal, the HPC System for Life Sciences is designed to be a plug-and-play turnkey solution that allows researchers to spend more time working on scientific matters in their domain rather than HPC cluster deployment, tuning and maintenance.

SOLUTION OVERVIEW

HPC in life sciences requires systems to support a flexible architecture to accommodate the different needs of the various applications in life sciences. The Dell EMC HPC System for Life Sciences was created to meet this need.

The HPC System for Life Sciences is a pre-integrated, tested, tuned, and purpose-built platform, leveraging the most relevant of Dell EMC's high performance computing line of products and best-in-class partner products [4]. It encompasses all the hardware resources required for various life sciences data analysis, while providing an optimal balance of compute density, energy efficiency and performance.

Architecture

Dell EMC HPC System for Life Sciences v1.1 provides more flexibility than previous generations. The platform is available in five variants, determined by the compute platform and the cluster interconnect selected, which can be either 10 Gigabit Ethernet (GbE), Intel® Omni-Path (Intel® OPA), Mellanox InfiniBand® (IB) EDR or IB FDR. In the current version, the following options are available:

- PowerEdge C6320 compute subsystem with Intel® OPA fabric
- PowerEdge C6320 compute subsystem with IB EDR fabric
- PowerEdge C6320 compute subsystem with 10 GbE fabric
- PowerEdge FX2 compute subsystem with IB FDR fabric
- PowerEdge FX2 compute subsystem with 10 GbE fabric

PowerEdge C6320 compute subsystem can cover more user cases in addition to NGS data analysis, while PowerEdge FX2 compute subsystem is dedicated for NGS data analysis. The different interconnect options will allow customers to choose the most cost-effective solution for their needs. For example, if a small number of compute nodes is required for a particular workload, a high-speed interconnect, like Intel® OPA, IB EDR or IB FDR, will not improve performance. These high-speed interconnect options are useful when designing a large system with different workloads. IB FDR is suitable for a medium-size workload, less than 165 genomes per day, while Intel® OPA and IB EDR are better for larger workloads. Since there are no notable performance differences between Intel® OPA and IB EDR, the decision for these options should depend on other factors, such as customers' budget, integration with existing infrastructure or preferences for a particular technology. In addition to the compute and network options, there are several other components that perform different functions in the HPC System for Life Sciences. These include the CIFS gateway, the Dell EMC Ready Bundle for HPC NFS Storage, the Dell EMC Ready Bundle for HPC Lustre Storage and other management components. Each of these components in described in detail in subsequent sections. The IB and the 10 GbE versions are nearly identical to Figure 1except for a few changes in the switching infrastructure and network adapters as outlined in the Network Components section. The complete solution ships in a deep 48U rack enclosure, which was chosen for its ease of mounting PDUs and effortless cable management. There are software modules which deploy, manage and maintain the cluster.

Figure 1 shows the components of a fully loaded rack using the PowerEdge C6320 rack server chassis as a compute subsystem and Intel® OPA as the cluster high speed interconnect.

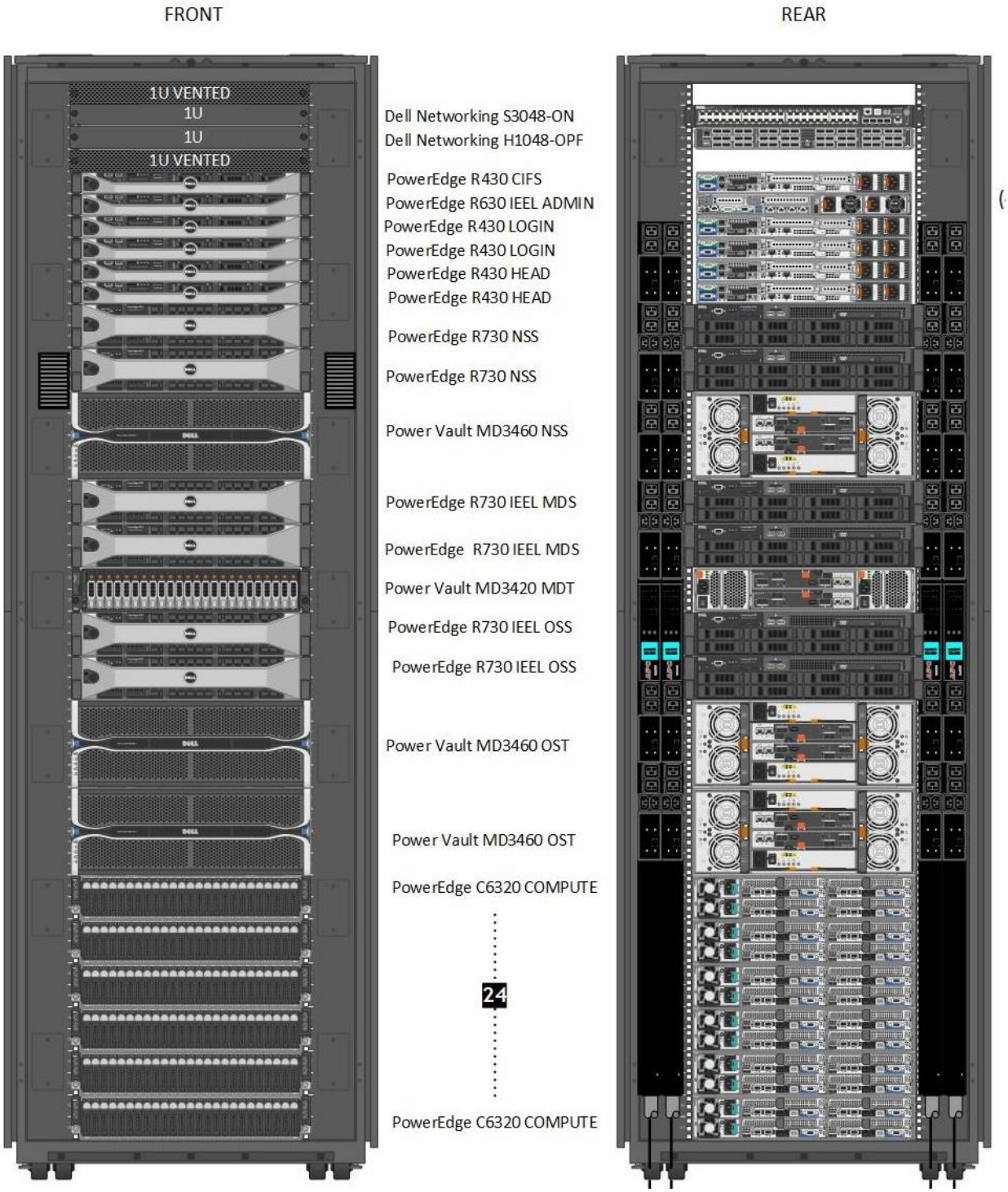


Figure 1: Dell EMC HPC System for Life Sciences with PowerEdge C6320 rack servers and Intel® OPA fabric

Compute and Management Components

There are several considerations when selecting the servers for master node, login node, compute node, fat node and accelerator node. For master and login node, 1U form factor PowerEdge R430 is recommended. Master node is responsible for managing the compute nodes and optimizing the overall compute capacity. Login node is used for user access, compilations and job submissions. Usually, master and login nodes are the only nodes that communicate with the outside world, and they act as a middle point between the actual cluster and the outside network. For this reason, high availability is provided for master and login nodes in the example solution illustrated in Figure 1.

Ideally, the compute nodes in a cluster should be as identical as possible, since the performance of parallel computation is bounded by the slowest component in the cluster. Heterogeneous clusters do work, but it requires a careful execution to achieve the best performance. We recommend PowerEdge C6320 as a compute node due to its density, a wide choice of CPUs and high maximum memory capacity.

PowerEdge R930 is an optional node with up to 3TB of memory. This node is recommended for customers who need to run applications requiring large memory.

Accelerators are used to speed up computationally intensive applications, such as molecular dynamics simulation applications. Although there are five different configurations for four NVIDIA K80 GPUs on PowerEdge C4130, we tested configuration C for this solution.

The compute and management infrastructure consists of the following components.

- Compute
  - Dell EMC PowerEdge C6320 rack server with 4 x C6320 servers and either Intel® OPA fabric or IB EDR fabric
  - Dell EMC PowerEdge FX2 chassis with 8 x FC430 chassis each and IB FDR interconnect [1]
  - Dell EMC PowerEdge C4130
  - Dell EMC PowerEdge R930
- Management
  - Dell EMC PowerEdge R430

Dell EMC PowerEdge C6320 for compute node

High-performance computing workloads, such as scientific simulations, seismic processing and data analytics, rely on compute performance, memory bandwidth and overall server efficiency to reduce processing time and data center costs. The next-generation Dell EMC PowerEdge C6320 provides an optimized compute and storage platform for HPC and scale-out workloads with up to four independent two-socket servers with flexible 24 x 2.5” or 12 x 3.5” high capacity storage in a compact 2U shared infrastructure platform [5]. C6320 supports up to 512GB of memory per server node, for a total of 2TB of memory in a highly dense and modular 2U solution.

Dell EMC PowerEdge C4130 for accelerator node

The PowerEdge C4130 provides supercomputing agility and performance in an ultra-dense platform purpose-built for scale-out HPC workloads [6]. Speed through the most complex research, simulation and visualization problems in medicine, finance, energy exploration, and related fields without compromising on versatility or data center space. Get results faster with greater precision by combining up to two Intel® Xeon® E5-2600 v3 processors and up to four 300W dual-width PCIe accelerators in each C4130 server. Support for an array of NVIDIA® Tesla™ GPUs and Intel Xeon Phi™ coprocessors, along with up to 256GB of DDR4 memory, gives you ultimate control in matching your server architecture to your specific performance requirements. This server is an optional component for molecular dynamics simulation applications.

Dell EMC PowerEdge R930 for fat node

The Dell EMC PowerEdge R930 is a 4-socket, 4U platform, equipped with the Intel® Xeon® E7-8890 v4 (24 cores per socket – 96 cores in server) and is dubbed “the fat node,” because of its 6 TB of memory capacity. This processor has 3 QPI links per socket. The particular features are needed for *de novo* genome assembly application like Velvet. Hosting large genomic data sets in memory with 64 cores operating on it eliminates the overhead caused by interconnects, disk look-ups, and swapping, thereby resulting in decreased run time. This server is an optional component that can be added to the solution.

Dell EMC PowerEdge R430 for master node, login node and CIFS gateway

The solution includes four Dell EMC 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® to provision, manage and monitor the cluster in a high availability (HA) configuration. The head nodes are in an active–passive HA state and use the NSS7.0-HA solution as shared storage.

Storage Components

The storage infrastructure consists of the following components:

- Dell EMC Ready Bundle for HPC NFS Storage (NSS7.0-HA)
- Dell EMC Ready Bundle for HPC Lustre StorageDell EMC PowerEdge R430 as the CIFS Gateway

NSS 7.0 HA

NSS 7.0 HA is designed to enhance the availability of storage services to the HPC cluster by using a pair of Dell EMC PowerEdge servers with PowerVault storage arrays, Red Hat HA software stack [7] and a network switch. The two PowerEdge servers have shared access to disk-based Dell EMC PowerVault storage in a variety of capacities, and both are directly connected to the HPC cluster using OPA, IB or 10GigE. The two servers are equipped with two fence devices: iDRAC8 Enterprise, and an APC Power Distribution Unit (PDU). If system failures occur on one server, the HA cluster will failover the storage service to the healthy server with the assistance of the two fence devices and also ensure that the failed server does not return to life without the administrator's knowledge or control.

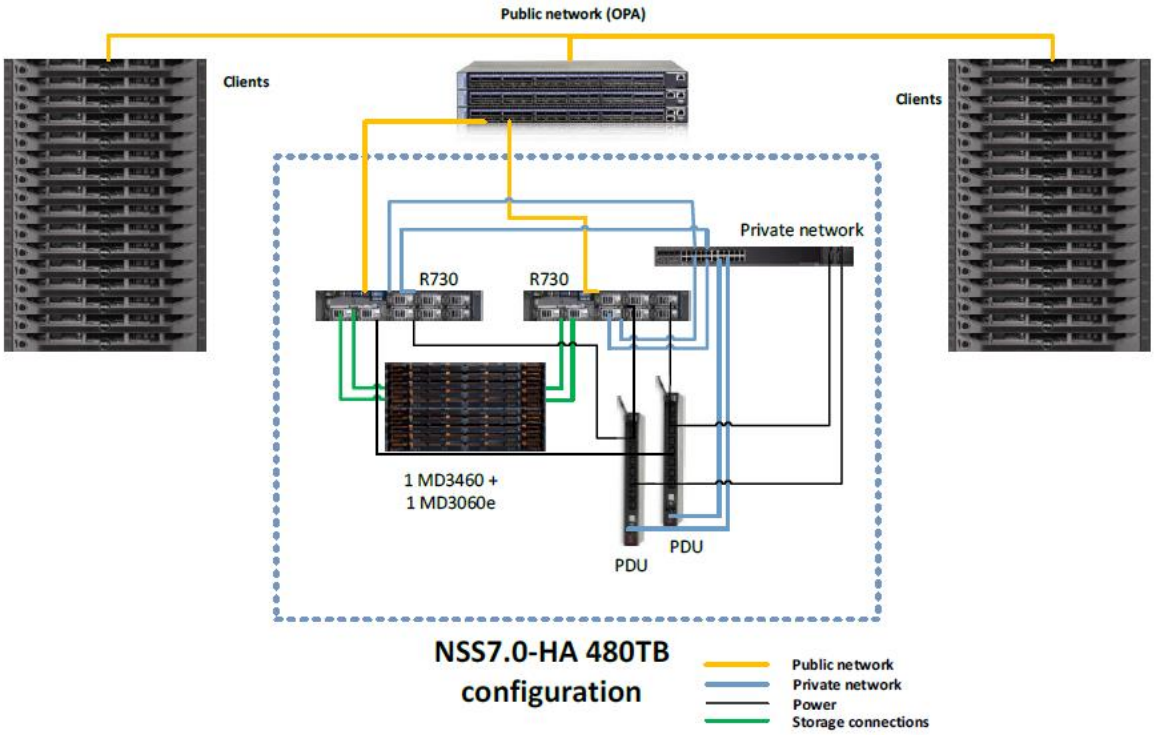


Figure 2: NSS7.0-HA test bed

Dell EMC HPC Lustre Storage Solution

The Dell EMC Ready Bundle for HPC Lustre Storage, referred to as Dell EMC HPC Lustre Storage is designed for academic and industry users who need to deploy a fully-supported, easy-to-use, high-throughput, scale-out and cost-effective parallel file system storage solution. The solution uses the Intel® Enterprise Edition (EE) for Lustre® software v.3.0 [8]. It is a scale-out storage solution appliance capable of providing a high performance and high availability storage system. Utilizing an intelligent, extensive and intuitive management interface, the Intel Manager for Lustre (IML) greatly simplifies deploying, managing and monitoring all of the hardware and storage system components. It is easy to scale in capacity, performance or both, thereby providing a convenient path to expand in the future.

The Dell EMC HPC Lustre Storage solution utilizes the 13th generation of enterprise Dell EMC PowerEdge™ servers and the latest generation of high-density PowerVault™ storage products. With full hardware and software support from Dell EMC and Intel, the Dell EMC Ready Bundle for HPC Lustre Storage solution delivers a superior combination of performance, reliability, density, ease of use and cost-effectiveness.

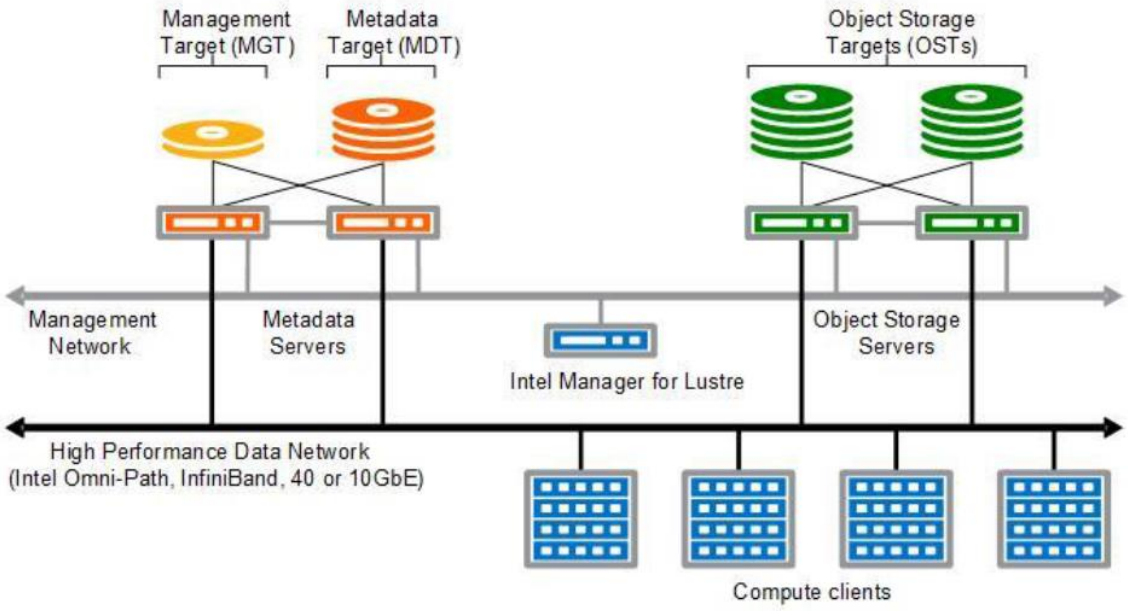


Figure 3: Lustre-based storage solution components

Dell EMC PowerEdge R430 as the CIFS Gateway

A Dell EMC PowerEdge R430 is used as the CIFS gateway for transferring data generated by NGS instruments into the Lustre file system.

Network Components

The Dell EMC HPC System for Life Sciences platform is available with four different high performance networking options — Intel® OPA, IB EDR, IB FDR and 10 GbE. The differences in switching infrastructure are shown in Table 1. There is also a Force10 S3048-ON GbE switch, which is used in both configurations whose purpose is described here. In the IB version, the Dell EMC PowerEdge FC430 sleds have 2:1 blocking FDR connectivity to the top of rack FDR switch and the Dell EMC PowerEdge C6320 sleds have either 1:1 non-blocking EDR connectivity to the top of rack EDR switch or 1:1 non-blocking Intel® OPA connectivity with H1048-OPF switch.

Force10 S3048-ON switch

In the Intel® OPA or IB configurations, the Force10 S3048-ON switch ports are split into multiple untagged virtual LANs to accommodate multiple internal networks.

- The port assignment of the Dell EMC Networking S3048-ON switch for the Intel® OPA or IB versions of the solution is as follows.

- Ports 01-04 and 27–52 are assigned to the cluster's private management network to be used by Bright Cluster Manager® connecting master, login, CIFS gateway and compute nodes. The PowerEdge C6320 server's ethernet and iDRAC constitute a majority of these ports.
- Ports 06–09 are used for the private network associated with NSS7.0-HA.
- The rest of the port 05 and ports 12–26 are allocated to the Lustre solution for its private management network
- Port 10 and 11 are used for the PDUs.

For the 10GigE configuration, the deployment and management of the cluster is done over the 10 GbE network by using the Dell EMC 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 Intel® OPA or IB configuration.

Table 1 Differences in Switching Infrastructure among Intel® OPA, IB FDR/EDR and 10GbE Configurations

Switching component	Intel® OPA with C6320	IB EDR with C6320	IB FDR with FC430	10 GbE with FC430
Top of Rack switch	1x Dell EMC Networking H1048-OPF switch 1 x Force10 S3048-ON 1 GbE switch	1 x SB7700 EDR switch 1 x Force10 S3048-ON 1 GbE switch	3 x Mellanox SX 6036 FDR switch 1 x Force10 S3048-ON 1 GbE switch	1 x Force10 S4820T 10 GbE switch 1 x Force10 S3048-ON 1 GbE switch
Switches/IOAs in Dell EMC PowerEdge chassis	N/A	N/A	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	Intel® Omni-Path Host Fabric Interface (HFI) 100 series card	Mellanox ConnectX-4 IB EDR adapter	Mellanox ConnectX-3 IB FDR adapter	Intel X520 DA SFP+ DP 10 GbE low profile adapter
Interconnect on Dell EMC PowerEdge sleds	Intel® Omni-Path host fabric interface (HFI) 100 series card	Mellanox ConnectX-4 IB EDR adapter	Mellanox ConnectX-3 FDR mezzanine adapter	10 GbE LOM

Dell EMC Networking H-Series OPA Switch

Intel® Omni-Path Architecture (OPA) is an evolution of the Intel® True Scale Fabric Cray Aries interconnect and internal Intel® IP [9]. In contrast to Intel® True Scale Fabric edge switches that support 36 ports of InfiniBand QDR-40Gbps performance, the new Intel® Omni-Path fabric edge switches support 48 ports of 100Gbps performance. The switching latency for True Scale edge switches is 165ns-175ns. The switching latency for the 48-port Omni-Path edge switch has been reduced to around 100ns-110ns. The Omni-Path host fabric interface (HFI) MPI messaging rate is expected to be around 160 million messages per second (Mmps) with a link bandwidth of 100Gbps.

Dell EMC Networking Infiniband FDR and EDR Switch

Mellanox EDR adapters are based on a new generation ASIC, also known as ConnectX-4, while the FDR adapters are based on ConnectX-3. The theoretical uni-directional bandwidth for EDR is 100 Gb/s versus FDR which is 56Gb/s. Another difference is that EDR adapters are x16 adapters while FDR adapters are available in x8 and x16. Both of these adapters operate at a bus width of 4X link. The messaging rate for EDR can reach up to 150 million messages per second compared with FDR ConnectX-3 adapters which deliver more than 90 million messages per second.

Software Components

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

- Bright Cluster Manager 7.2®
- Red Hat Enterprise Linux 7.2 (RHEL 7.2)
- CUDA 7.5

- MPSS 3.6.1
- MLNX OFED 3.2
- OM 8.3 and DTK update
- Lab7 Bio-Builds
- Molecular dynamics simulation

Bright Cluster Manager® 7.2

Bright Cluster Manager® (BCM) for Dell EMC is a comprehensive solution for provisioning, monitoring, and managing Dell EMC clusters [10].

Two Dell EMC PowerEdge R430 servers are deployed as head nodes in a HA active-passive configuration by using the NSS7.0-HA solution as shared storage . The active head node is responsible for deploying and monitoring the 24 Dell EMC PowerEdge C6320 (assuming one rack design), the Dell EMC PowerEdge R930 (if used) and the other Dell EMC 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) 7.2, with which, the head nodes and compute nodes are deployed. The Bright Cluster Manager® 7.2 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

Lab7 BioBuilds

BioBuilds is a well maintained, versioned and continuously growing collection of open-source bio-informatics tools from Lab7 [11]. They are prebuilt and optimized for a variety of platforms and environments. BioBuilds solve the software challenges faced in 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 include the software and its supporting dependencies for ease of deployment.

- Using BioBuilds among all the collaborators 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 BioBuilds can be found at Reference 10.

PERFORMANCE EVALUATION AND ANALYSIS

Aligner scalability

This is a base-line test in order to obtain information useful to set up fair performance tests. Burrows-Wheeler Aligner (BWA) short sequence aligner is tested here since it is a key application in variant analysis pipelines for whole genome sequencing data.

Aligner scaling test configuration

A single PowerEdge C6320 server is used to generate baseline performance metrics and ascertain the optimum number of cores for running BWA [12]. Two different configurations, Intel Xeon Processor E5-2680 v3 with DDR4-2133 and Intel Xeon Processor E5-2690 v4 with DDR4-2400 are tested. For Broadwell CPUs, snoop mode is set to “opportunistic snoop broadcast” while “early snoop” mode is chosen for Haswell CPUs. The total 256 GB of memory is used for the both test cases.

BWA shows stable scalability

Figure 4 shows the run times of BWA on various sequence data sizes ranging from 2 to 208 million fragments (MF) and different number of threads. Oversubscription is avoided to ensure each thread runs on a single physical core. As shown in Figure 4 and Figure 5, BWA scales linearly over both input size and the number of cores.

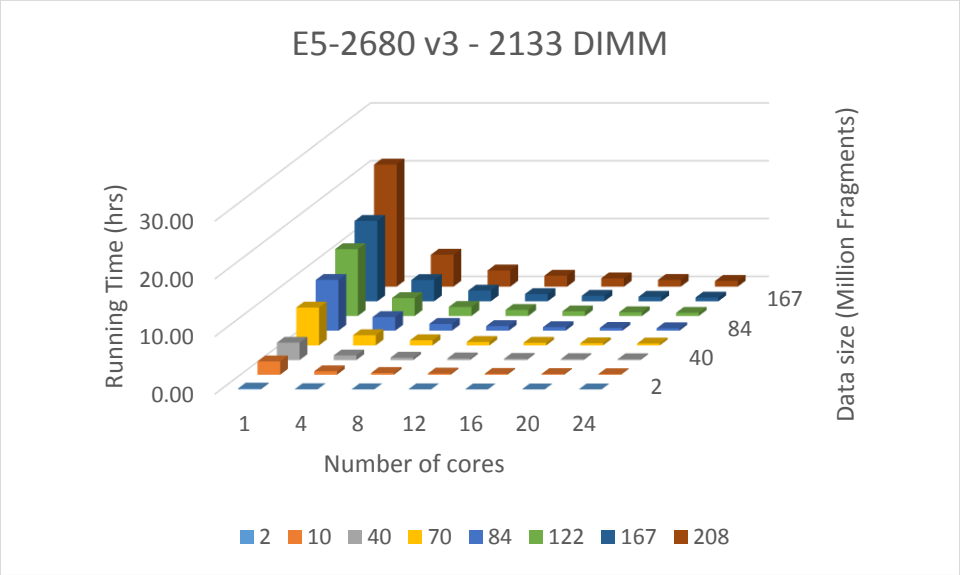


Figure 4: BWA's scaling behavior on Haswell CPU with different input data size

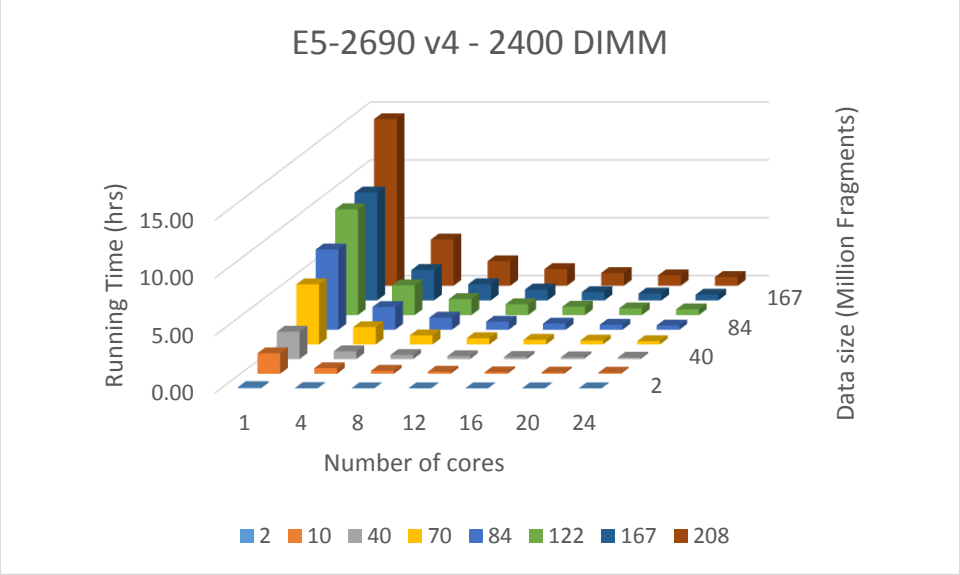


Figure 5: BWA's scaling behavior on Broadwell CPU with different input data size

Table 2 and Table 3 show speed-up due to increasing core count. The results indicate that the optimum number of threads for BWA is in between 10 - 16. Based on this observation, 13 cores are used for BWA processes throughout the tests.

Table 2: Speed-up by increasing parallelism on E5-2680 v3/DDR4-2133

Speed-up		Sequence Data Size (Million Fragments)							
		2	10	40	70	84	122	167	208
Number Of Cores	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	3.84	3.84	3.77	3.72	3.73	3.74	3.77	3.82
	8	7.38	7.59	7.26	7.43	7.46	7.44	7.50	7.53
	12	10.14	11.30	10.72	10.97	11.09	10.96	11.06	11.15
	16	12.62	14.88	14.05	14.41	14.57	14.39	14.58	14.67
	20	14.20	18.25	17.34	17.71	17.95	17.65	17.64	18.07
	24	15.35	21.72	20.37	20.64	21.12	20.84	21.24	21.36

Table 3: Speed-up by increasing parallelism on E5-2690 v4/DDR4-2400

Speed-up		Sequence Data Size (Million Fragments)							
		2	10	40	70	84	122	167	208
Number Of Cores	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	3.48	3.70	3.49	3.48	3.52	3.55	3.53	3.59
	8	6.26	6.86	6.57	6.64	6.63	6.63	6.64	6.75
	12	8.47	10.16	9.54	9.59	9.88	9.72	9.85	9.96
	16	10.54	13.36	12.56	12.82	12.95	12.53	13.03	13.01
	20	11.68	16.40	15.37	15.67	15.97	15.64	15.69	15.94
	24	12.71	19.20	17.95	18.53	18.77	18.47	18.83	18.93

More importantly, the gain in speed up by using E5-2690 v4/DDR4 2400 is at least 8 % as shown in Table 4. This speed up is likely due to the higher clock speed of CPUs and faster memory.

Table 4: Speed-up by E5-2697 v4/DDR4-2400 in comparison to E5-2680 v3/DDR4-2133

Speed-up		Sequence Data Size (Million Fragments)							
		2	10	40	70	84	122	167	208
Number Of Cores	1	1.31	1.30	1.26	1.26	1.26	1.27	1.49	1.47
	4	1.19	1.25	1.17	1.18	1.19	1.20	1.40	1.38
	8	1.12	1.17	1.14	1.12	1.12	1.13	1.32	1.31
	12	1.10	1.17	1.12	1.10	1.12	1.12	1.33	1.31
	16	1.10	1.16	1.13	1.12	1.12	1.10	1.33	1.30
	20	1.08	1.17	1.12	1.11	1.12	1.12	1.33	1.29
	24	1.09	1.15	1.11	1.13	1.12	1.12	1.32	1.30

However, our choice of CPUs for the solution is Intel® Xeon® Processor E5-2697 v4 with 18 cores since larger core count CPUs are preferable in most life science applications.

Genomics/NGS data analysis performance

A typical variant calling pipeline consists of three major steps 1) aligning sequence reads to a reference genome sequence; 2) identifying regions containing SNPs/InDels; and 3) performing preliminary downstream analysis. In the tested pipeline, BWA 0.7.2-r1039 is used for the alignment step and Genome Analysis Tool Kit (GATK) is selected for the variant calling step. These are considered standard tools for aligning and variant calling in whole genome or exome sequencing data analysis. The version of GATK for the tests is 3.5, and the actual workflow tested was obtained from the workshop, 'GATK Best Practices and Beyond'. In this workshop, they introduce a new workflow with three phases.

- Best Practices Phase 1: Pre-processing
- Best Practices Phase 2A: Calling germline variants
- Best Practices Phase 2B: Calling somatic variants
- Best Practices Phase 3: Preliminary analyses

Here we tested phase 1, phase 2A and phase 3 for a germline variant calling pipeline. The details of commands used in the benchmark are in APPENDIX A. GRCh37 (Genome Reference Consortium Human build 37) was used as a reference genome sequence, and 10x whole human genome sequencing data from the Illumina platinum genomes project, named ERR091571_1.fastq.gz and ERR091571_2.fastq.gz11 were used for a baseline test [13]. Further tests with 50x coverage whole genomes were done to check for scalability of the solution as the data size increased. In addition to human genome, other mammal and plants genomes from cow, pig, rice and corn were also tested. The details of the datasets are mentioned in APPENDIX B.

It is ideal to use non-identical sequence data for each run. However, it is extremely difficult to collect non-identical sequence data having more than 30x depth of coverage from the public domain. Hence, we used a single sequence data set for multiple simultaneous runs. A clear drawback of this practice is that the running time of Phase 2, Step 2 might not reflect the true running time as researchers tend to analyze multiple samples together. Also, this step is known to be less scalable. The running time of this step increases as the number of samples increases. A subtle pitfall is a storage cache effect. Since all the simultaneous runs will read/write roughly at the same moment, the run time would be shorter than real cases. Despite these built-in inaccuracies, this variant analysis performance test can provide valuable insights to estimating how many resources are required for an identical or even similar analysis pipeline with a defined workload.

Variant Analysis test configuration

Two different configurations, Intel® OPA and IB FDR were used to benchmark BWA-GATK pipeline variant analysis. In Table 5, the configuration of PowerEdge C6320 with Intel® OPA and FC430 with IB FDR solutions are summarized. Processors on both solutions are nearly identical, except Intel® Xeon® Dual E5-2697 v4 has more cores, higher max turbo frequency and better max memory bandwidth. The same versions of all applications were installed on both platforms before testing.

Table 5: Test configurations for Genomics/NGS data analysis

Component	C6320/OPA	FC430/FDR
Server	40x PowerEdge C6320	40x PowerEdge FC430 in FX2 chassis
Processor	Total 1440 cores; Intel® Xeon® Dual E5-2697 v4 - 18 cores Processor base frequency: 2.30 GHz Max turbo frequency: 3.60 GHz Max memory bandwidth: 76.8 GB/s	Total of 1120 cores: Intel® Xeon® Dual E5-2695 v3 - 14 cores Processor base frequency: 2.30 GHz Max turbo frequency: 3.30 GHz Max memory bandwidth: 68 GB/s
Memory	128GB-8x16GB DIMM, 2400 MT/s, dual rank, x4 data width	128GB - 8x 16GB RDIMM, 2133 MT/s, dual rank, x4 data width
Storage	480TB IEEL (Lustre)	480TB IEEL (Lustre)
Interconnect	Intel® OPA	IB FDR
OS	Red Hat Enterprise 7.2	Red Hat Enterprise 6.6
Cluster Management tool	Bright Cluster Manager 7.2	Bright Cluster Manager 7.1
Snoop Mode	opportunistic snoop broadcast	early snoop
Short Sequence Aligner	BWA 0.7.2-r1039	BWA 0.7.2-r1039
Variant Analysis	GATK 3.5	GATK 3.5
Utilities	sambamba 0.6.0, samtools 1.2.1	sambamba 0.6.0, samtools 1.2.1

Each node in the C6320/OPA solution has 36 cores whereas FC430/FDR solution provides 28 cores for each node. For a fair comparison, two BWA-GATK pipelines are loaded on each node and any sub process uses 13 cores at most. Hence, a maximum of 26 cores are used at any one moment in each node on both solutions. In addition to CPU usage, an equal amount of memory is allocated for every process.

The pipeline tested here is not optimized for any system, and default values are used when it is allowed. A total of 80 processes of BWA-GATK pipelines are run simultaneously in each test and results output to Lustre storage.

The throughput of Dell EMC HPC System for Life Sciences

Total run time is the elapsed wall time from the earliest start of Phase 1, Step 1 to the latest completion of Phase 3, Step 2. Time measurement for each step is from the latest completion time of the previous step to the latest completion time of the current step as illustrated in Figure 6.

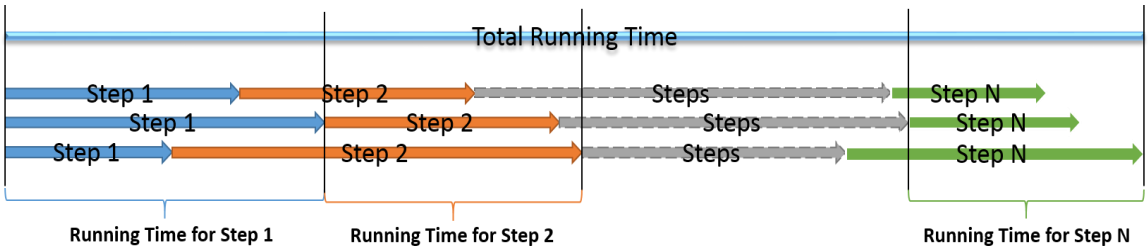


Figure 6: Running time measurement method

Feeding multiple samples into an analytical pipeline is the simplest way to increase parallelism, and this practice will improve the throughput of a system if a system is well-designed to accommodate the sample load. In Figure 7, the throughput in total number of genomes per day for all tests are summarized. As expected, it is clear that the C6320/OPA combination outperform the FC430/FDR combination due to CPU higher memory bandwidth, faster memory and a faster interconnect. Although the run time depends heavily on the data, the C6320/OPA solution shows better performance throughout the whole range of genome data sizes. The detailed run time measurements are in APPENDIX C, Table 8.

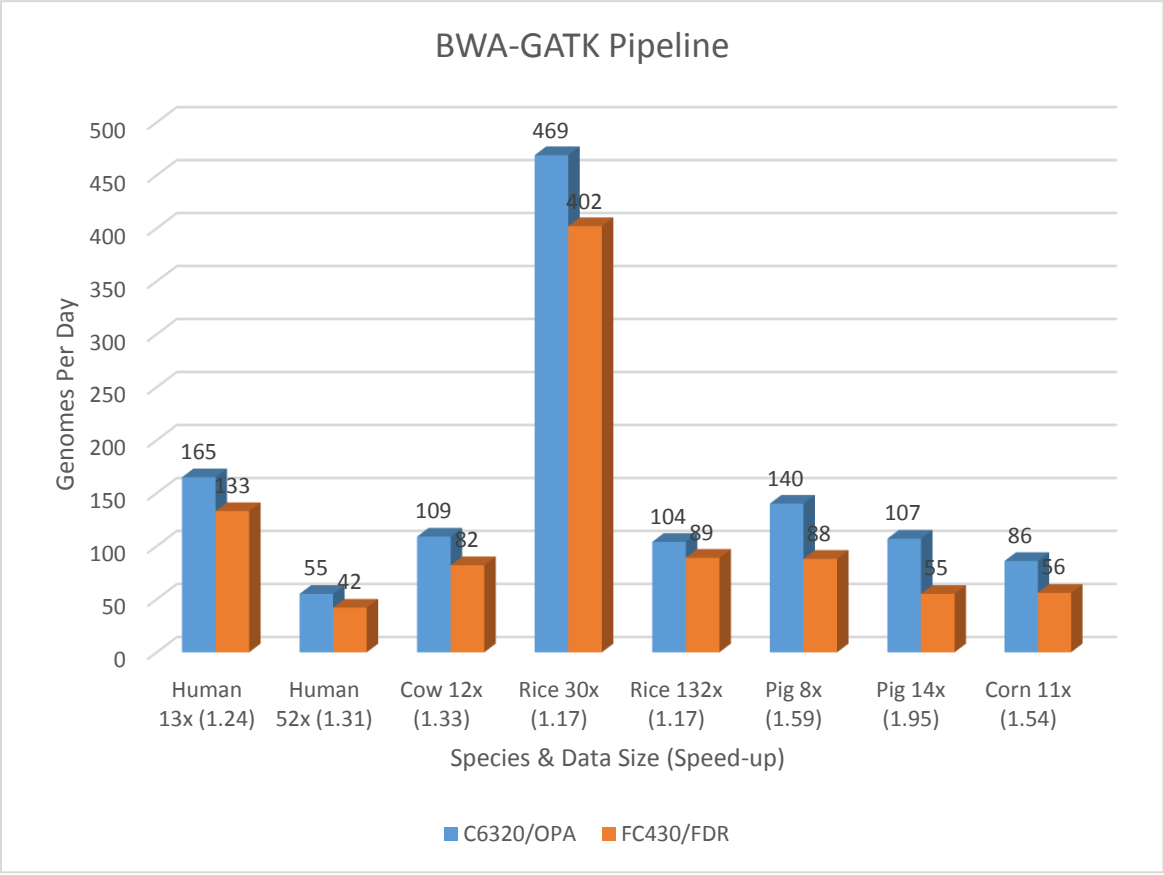


Figure 7: Number of genomes processed through BWA-GATK pipeline: the comparisons are between PowerEdge C6320 with Intel® OPA and PowerEdge FC430 with IB FDR

Molecular dynamics software performance

Over the past decade, GPUs became popular in scientific computing because of their great ability to exploit a high degree of parallelism. NVIDIA has a handful of life sciences applications optimized to run on their general purpose GPUs. Unfortunately, these

GPUs can only be programmed with CUDA, OpenACC and the OpenCL framework. Most of the life sciences community is not familiar with these frameworks, and so few biologists or bioinformaticians can make efficient use of GPU architectures. However, GPUs have been making inroads into the molecular dynamics and electron microscopy fields. These fields require heavy computational work to simulate biomolecular structures or their interactions and reconstruct 3D images from millions of 2D images generated from an electron microscope.

We selected three different molecular dynamics applications to run tests on a PowerEdge C4130 with four K80s. The applications are Amber 16, HOOMD-blue and NAMD [14, 15, 16].

The HPC focused Tesla series K80 GPU provides 8.74/2.91 TFLOPs (single/double precision) compute capacity, which is 31%-75% more than the K40, the previous Tesla card [17].

A benchmark suite is available from the Amber and HOOMD-blue web sites, and we used these suites with minor adjustments for the benchmark. The command line and a configuration file used to test NAMD performance is listed in APPENDIX A. All the tests are repeated on local storage and Lustre.

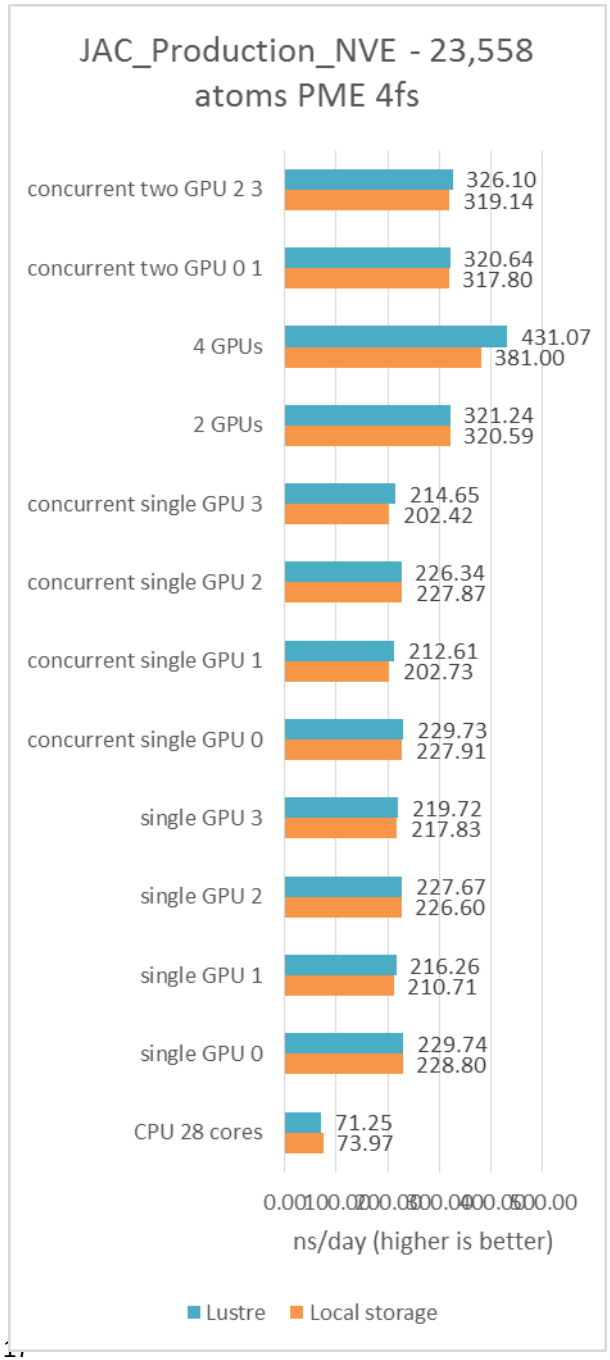
Molecular dynamics application test configuration

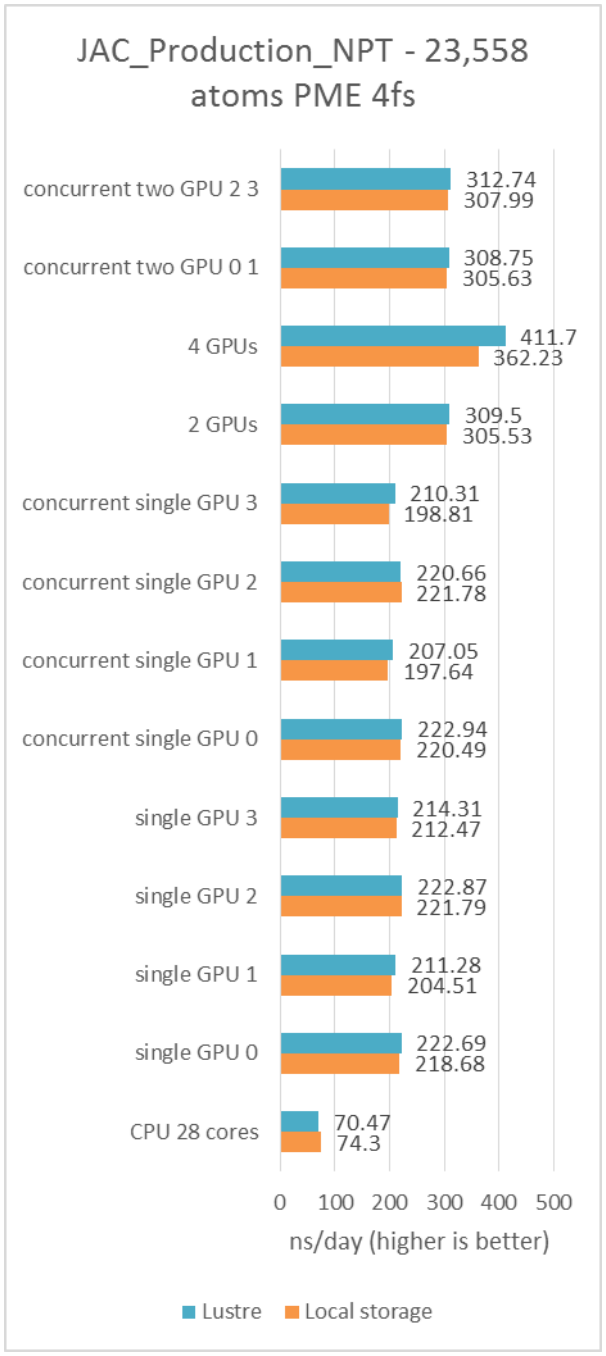
A single PowerEdge C4130 with Intel® Xeon® Dual E5-2690 v3 with 128 GB DDR4 2133MHz and four K80 in C configuration [17]. Local storage and Lustre connected through Intel® OPA are tested.

Amber benchmark suite

This suite includes the Joint Amber-Charmm (JAC) benchmark considering dihydrofolate reductage (DHFR) in an explicit water bath with cubic periodic boundary conditions. The major assumptions are that the DHFR molecule presents in water without surface effect and its movement assumed to follow microanonical (NVE) ensemble which assumes constant amount of substance (N), volume (V), and energy (E). Hence, the sum of kinetic (KE) and potential energy (PE) is conserved, in other words, Temperature (T) and Pressure (P) are unregulated. JAC benchmark repeats simulations with Isothermal–isobaric (NPT) ensemble that assumes N, P and T are conserved. It corresponds most closely to laboratory conditions with a flask open to ambient temperature and pressure. Beside these settings, Particle mesh Ewald (PME) is the choice of algorithm to calculate electrostatic forces in molecular dynamics simulations. Other biomolecules simulated in this benchmark suite are Factor IX (one of the serine proteases of the coagulation system), cellulose and Satellite Tobacco Mosaic Virus (STMV). Here, we report the results from DHFR and STMV data.

The test results from Amber benchmark suite cases including three simulations, two JAC production simulations with DHFR and one STMV production simulation are illustrated In Figure 8. The “CPU 28 cores” label refers to Amber simulations run strictly on CPUs. In this case, all 28 cores are used. The “single” GPU 0, 1, 2 and 3 labels refer to tests performed on individual GPUs separately, whereas the “concurrent” GPU 0, 1, 2 and 3 tests were run simultaneously. The “2 GPUs” and “4 GPUs” labels indicate that multiple GPUs were used to solve a single job. Finally, “concurrent two GPU 0 1” and “concurrent two GPU 2 3” are where the tests ran simultaneously. Each test set has two performance measurements from different storage settings, local versus Lustre storage. Overall, GPUs can take advantage of Lustre storage, and Lustre helps significantly when multiple GPUs are used for a single job. 4 GPU tests across different simulations show more than 10% speed gain.





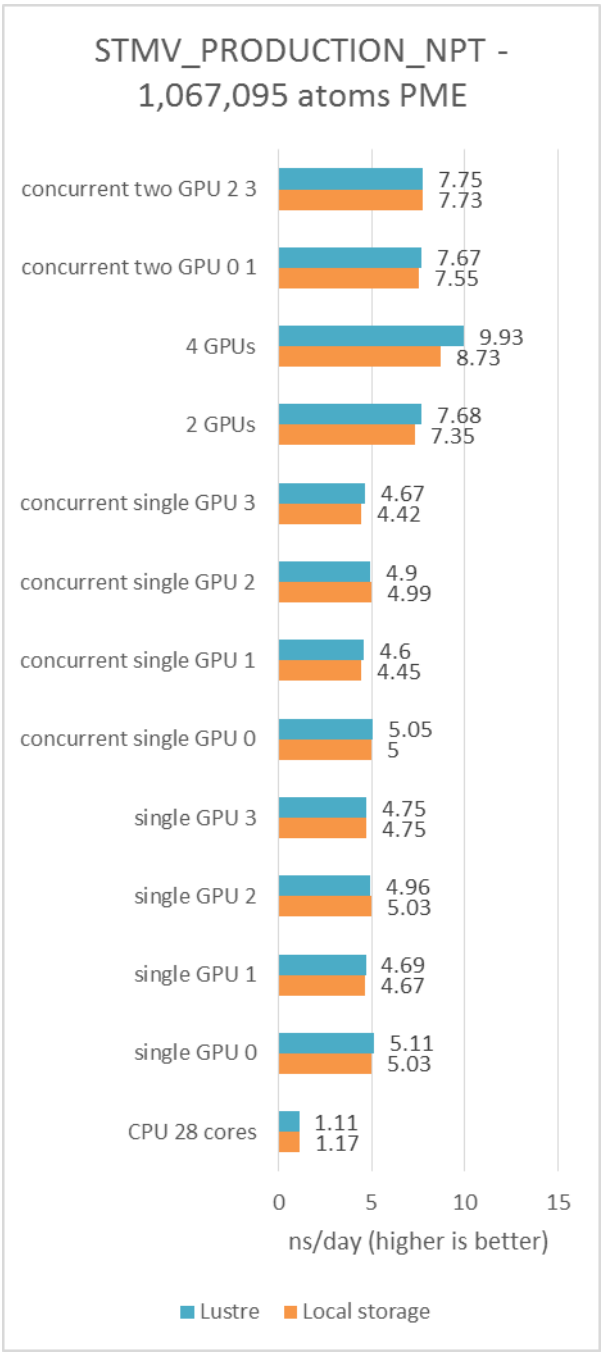


Figure 8 Benchmark results from Amber

HOOMD-blue benchmark suite

HOOMD-blue is a general-purpose particle simulation toolkit. It scales from a single CPU core to thousands of GPUs.

Performance results in Figure 9 are reported in hours to complete ten million Monte Carlo sweeps, where one sweep is N trial moves on 1,048,576 particles for 10e6 steps. HOOMD-blue shows almost linear scaling behavior on both CPUs and GPUs. Unlike the results from Amber, HOOMD-blue does not take an advantage of Lustre for this hexagon benchmark. All other benchmarks such as liquid, microsphere, quasicrystal and triblock copolymer benchmarks show identical scaling behavior.

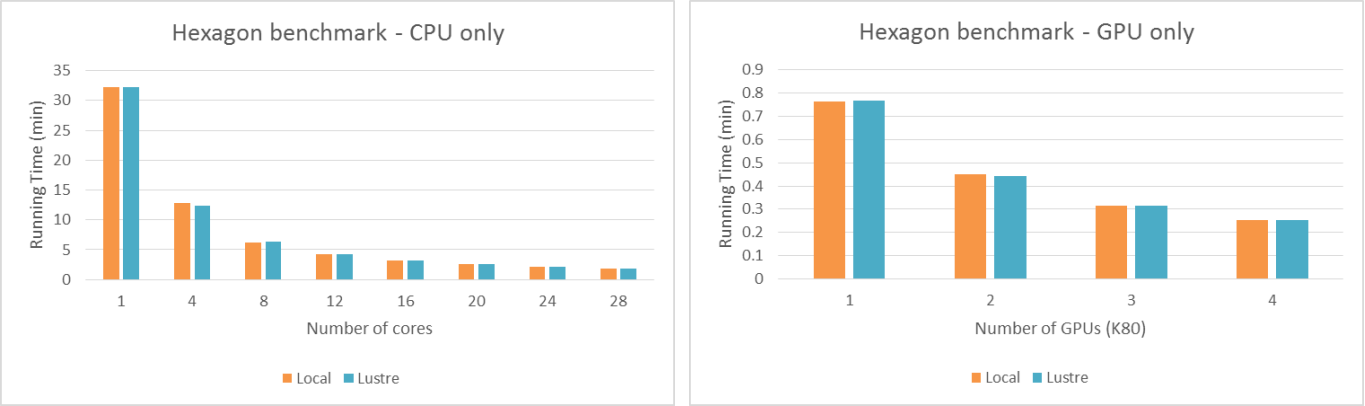


Figure 9: Hexagon benchmark results

NAMD benchmark on Apoa1, F1atpase and STMV

NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. The benchmark APOA1 (92224 atoms) is a high density lipoprotein found in plasma, which helps extraction of cholesterol from tissues to liver. F1-ATPase (327506 atoms) is responsible for the synthesis of the molecule adenosine triphosphate. STMV is a small, icosahedral virus, which worsens the symptoms of infections by tobacco mosaic virus. STMV is a large benchmark case with 1,066,628 atoms.

NAMD scales well over number of cores, number of GPUs and the size of data, as shown in Figure 10, Figure 11 and Figure 12. Like HOOMD-blue, this simulation tool does take advantage of Lustre storage. The performance from 28 CPU cores with 4 GPUs for all three simulations becomes worse when Lustre is used. This can be explained by NAMD's strong dependency on CPUs.

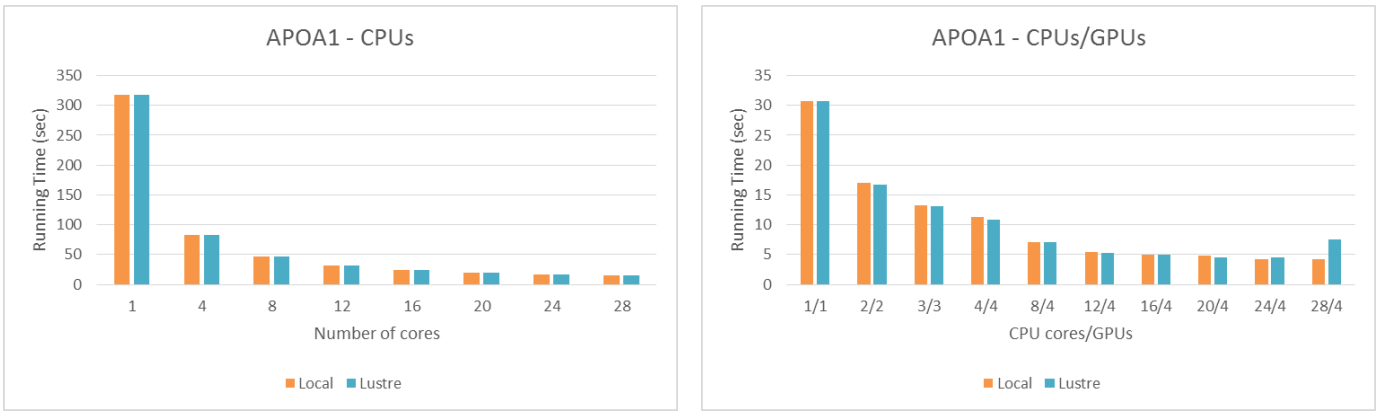
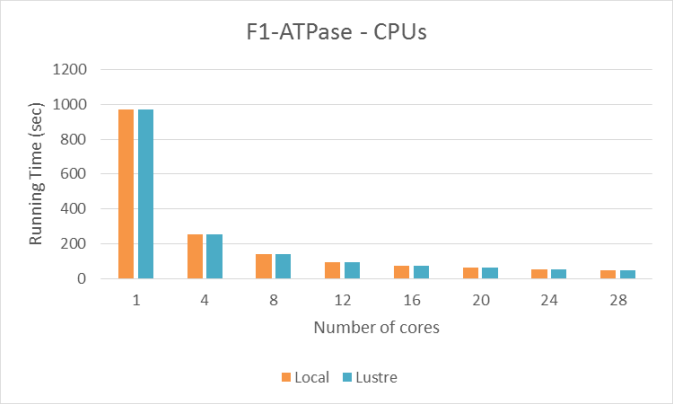


Figure 10: APOA1 benchmark results



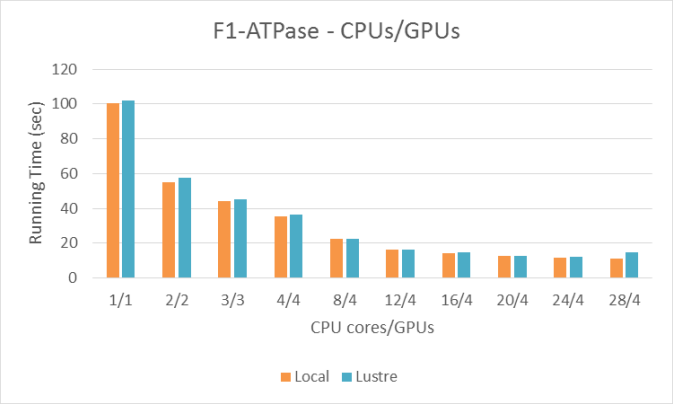
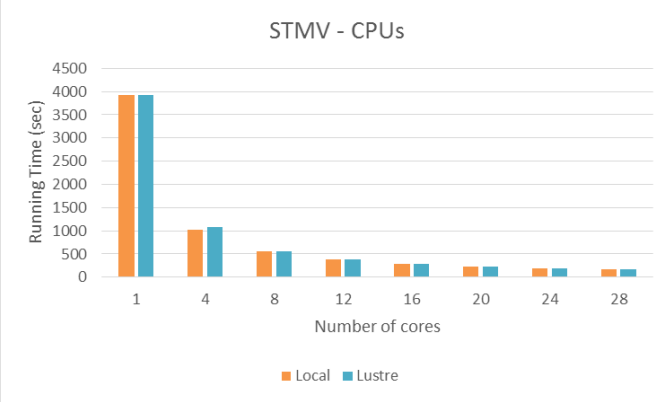


Figure 11: F1-ATPase benchmark results



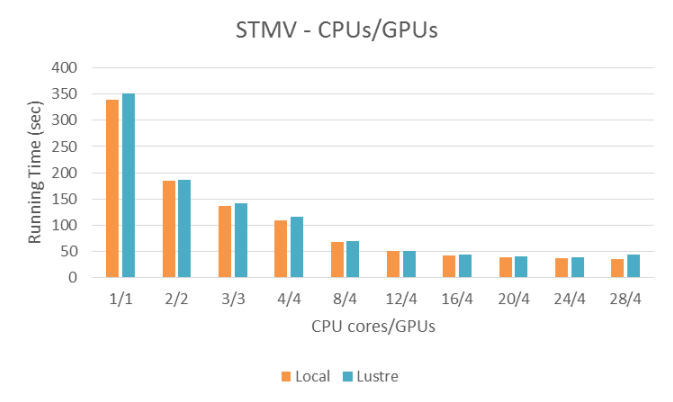


Figure 12: STMV benchmark results

CONCLUSION

This white paper is focused on testing working solutions for diverse life sciences applications. Upon successful iteration of Dell EMC's HPC Solution for genomics v2.0, we incorporated a molecular dynamics simulation solution into the flexible architecture in addition to improving the performance of the genomics data analysis platform.

Four K80 GPU unit in PowerEdge C4130 adds significant molecular dynamics performance to the solution, as it is able to simulate 1 million atoms in 10e6 steps in less than a minute. We showed the Dell EMC HPC System for Life Sciences improved variant analysis pipeline performance by at least 17% compared to the previous version, and we believe actual throughput of this solution will be higher, since only 72% of total cores were used to make a fair comparison.

Overall, the Dell EMC HPC System for Life Sciences is a cost-effective solution flexible enough to be used for many diverse life sciences applications.

APPENDIX A: Benchmark commands

BWA scaling test command

bwa mem -M -t [number of cores] -v 1 [reference] [read fastq 1] [read fastq 1] > [sam output file]

BWA-GATK commands

Phase 1. Pre-processing

Step 1. Aligning and sorting

bwa mem -c 250 -M -t [number of threads] -R '@RG\tID:noID\tPL:illumine\tLB:noLB\tSM:bar' [reference chromosome] [read fastq 1] [read fastq 2] | samtools view -bu - | sambamba sort -t [number of threads] -m 30G --tmpdir [path/to/temp] -o [sorted bam output] /dev/stdin

Step 2. Mark and remove duplicates

sambamba markdup -t [number of threads] --remove-duplicates --tmpdir=[path/to/temp] [input: sorted bam output] [output: bam without duplicates]

Step 3. Generate realigning targets

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -nt [number of threads] -R [reference chromosome] -o [target list file] -I [bam without duplicates] -known [reference vcf file]

Step 4. Realigning around InDel

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T IndelRealigner -R [reference chromosome] -I [bam without duplicates] -targetIntervals [target list file] -known [reference vcf file] -o [realigned bam]

Step 5. Base recalibration

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct [number of threads] -l INFO -R [reference chromosome] -I [realigned bam] -known [reference vcf file] -o [recalibrated data table]

Step 6. Print recalibrated reads - Optional

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T PrintReads -nct [number of threads] -R [reference chromosome] -I [realigned bam] -BQSR [recalibrated data table] -o [recalibrated bam]

Step 7. After base recalibration - Optional

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct [number of threads] -l INFO -R [reference chromosome] -I [recalibrated bam] -known [reference vcf file] -o [post recalibrated data table]

Step 8. Analyze covariates - Optional

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T AnalyzeCovariates -R [reference chromosome] -before [recalibrated data table] -after [post recalibrated data table] -plots [recalibration report pdf] -csv [recalibration report csv]

Phase 2. Variant discovery – Calling germline variants

Step 1. Haplotype caller

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T HaplotypeCaller -nct [number of threads] -R [reference chromosome] -ERC GVCF -BQSR [recalibrated data table] -L [reference vcf file] -I [recalibrated bam] -o [gvcf output]

Step 2. GenotypeGVCFs

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -nt [number of threads] -R [reference chromosome] -V [gvcf output] -o [raw vcf]

Phase 3. Preliminary analyses

Step 1. Variant recalibration

java -d64 -Xms512m -Xmx2g -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R [reference chromosome] --input [raw vcf] -an QD -an DP -an FS -an ReadPosRankSum -U LENIENT_VCF_PROCESSING --mode SNP --recal_file [raw vcf recalibration] --tranches_file [raw vcf tranches]

Step 2. Apply recalibration

java -d64 -Xms512m -Xmx2g -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R [reference chromosome] -input [raw vcf] -o [recalibrated filtered vcf] --ts_filter_level 99.97 --tranches_file [raw vcf tranches] --recal_file [raw vcf recalibration] --mode SNP -U LENIENT_VCF_PROCESSING

NAMD commands

namd2 +p [number of threads] [run configuration file]

Example of a configuration file

```
cellBasisVector1     108.8612 0.0 0.0
cellBasisVector2     0.0 108.8612 0.0
cellBasisVector3     0.0 0.0 77.758
cellOrigin           0.0 0.0 0.0

coordinates          apoa1.pdb
temperature          300
seed                 74269

switching            on
switchdist           10
cutoff               12
pairlistdist         13.5
margin               0
stepspercycle        20

PME                  on
PMEGridSizeX         108
PMEGridSizeY         108
PMEGridSizeZ         80

structure            apoa1.psf
parameters           par_all22_prot_lipid.xplor
parameters           par_all22_popc.xplor
exclude              scaled1-4
1-4scaling           1.0
```

```
timestep             1.0
fullElectFrequency   4

numsteps             500
outputtiming         100
outputEnergies       100

outputname           [output file]
```

APPENDIX B: Test data descriptions

Sequence data for BWA scaling tests

Table 6: Data sets for BWA scaling tests

Sample	File Size (GB)	Read Length	# of Million Fragments
SRR1060762	0.21	76	1.987
SRR593165	1.59	102	10.083
SRR786500	3.92	51	39.251
ERR754356	11.58	102	69.225
ERR754362	15.51	102	83.950
SRR1299474	11.27	52	122.253
ERR754364	29.61	102	167.366
SRR1299472	20.24	52	207.848

Sequence data for BWA-GATK pipeline tests

Table 7: Whole genome sequencing data

Species	Fastq files	Coverage Depth	Number of Raw Reads	Total Nucleotides
Homo sapiens	ERR091571	10x	422,875,838	42,710,459,638
Homo sapiens	ERR194146	50x	1,626,361,156	171,588,070,386
Bos taurus	SRR1805809	12x	163,405,319	32,681,063,800
Oryza sativa indica	SRR3098100	30x	48,379,772	12,191,702,544
Oryza sativa japonica	SRR1450198	132x	248,384,796	49,676,959,200
Sus scrofa	SRR1056427	8x	123,273,020	24,901,150,040
Sus scrofa	SRR1178925	14x	206,940,772	41,802,035,944
Zea mays	SRR1575496	11x	180,961,086	36,192,217,200

APPENDIX C: Running time details of BWA-GATK pipeline

Table 8: Detailed running time measure for BWA-GATK pipelines

Steps	Human 10x		Human 50x		Cow 12x		Rice - indica 30x		Rice - japonica 132x		Pig 8x		Pig 14x		Corn 11x	
	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR	C6320 OPA	FC430 FDR
Aligning & Sorting	3.23	3.93	10.51	15.79	4.42	5.77	0.99	1.18	4.58	9.50	2.39	3.04	3.73	6.16	8.59	11.16
Mark/Remove Duplicates	0.66	0.66	3.28	2.62	0.65	0.73	0.11	0.12	1.02	1.27	0.27	0.27	0.91	0.79	0.73	0.72
Generate Realigning Targets	0.17	0.29	0.42	1.08	0.20	1.57	0.03	0.05	0.12	0.22	0.11	0.27	0.14	2.17	0.07	0.26
Realign around InDel	2.25	2.50	7.65	8.90	2.46	3.15	1.07	1.25	5.62	7.37	1.67	1.83	2.51	5.02	2.76	3.18
Base Recalibration	0.99	1.82	4.12	6.80	1.19	1.96	0.30	0.36	1.46	3.16	0.70	1.01	1.08	3.74	0.79	1.91
HaplotypeCaller	3.74	4.52	8.75	10.28	8.10	9.33	1.55	1.77	5.70	8.95	8.01	14.65	9.11	16.02	8.87	16.72
GenotypeGVCFs	0.02	0.03	0.02	0.03	0.05	0.05	0.00	0.01	0.01	1.12	0.06	0.06	0.04	0.26	0.03	0.04
Variant Recalibration	0.52	0.67	0.21	0.37	0.46	0.86	0.04	0.04	0.01	0.92	0.47	0.56	0.46	0.86	0.53	0.46
Apply Recalibration	0.02	0.04	0.01	0.04	0.03	0.06	0.00	0.01	0.00	0.03	0.05	0.08	0.03	0.06	0.02	0.05
Total Run Time	11.61	14.50	34.97	45.90	17.55	23.50	4.09	4.78	18.53	32.50	13.72	21.80	18.01	35.08	22.40	34.50
Number of Genomes per Day	165	133	55	42	109	82	469	402	104	89	140	88	107	55	86	56

REFERENCES

[1] "Dell HPC System for Life Sciences White Paper," [Online]. Available: http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20441607.

[2] "Bioinformatics," [Online]. Available: https://en.wikipedia.org/wiki/Bioinformatics.

[3] "Supercomputer models one second of human brain activity," [Online]. Available: http://www.telegraph.co.uk/technology/10567942/Supercomputer-models-one-second-of-human-brain-activity.html.

[4] "Dell 13th Generation PowerEdge Server Resources," [Online]. Available: http://en.community.dell.com/techcenter/extras/w/wiki/7520.dell-13th-generation-poweredge-server-resources.

[5] "PowerEdge C6320," [Online]. Available: http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/Dell-PowerEdge-C6320-Spec-Sheet.pdf.

[6] "PowerEdge C4130," [Online]. Available: http://www.dell.com/us/business/p/poweredge-c4130/pd.

[7] "Dell HPC NFS Storage Solution - High Availability Configuration White Paper (NSS 7.0-HA)," [Online]. Available: http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20442675.

[8] "Dell HPC Lustre Storage solution with Intel Omni-Path," [Online]. Available: http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20442903.

[9] "Dell HPC Omni-Path Fabric White Paper," [Online]. Available: http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20442683.

[10] "Bright Cluster Manager," [Online]. Available: http://www.brightcomputing.com/News-Bright-Computing-Collaborates-with-Dell-for-Genomic-Analysis.php.

[11] "Biobuilds by Lab7," [Online]. Available: https://biobuilds.org/downloads/.

[12] "BWA," [Online]. Available: http://bio-bwa.sourceforge.net/.

[13] "Illumina plantinum genomes project," [Online]. Available: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR091/ERR091571.

[14] "Amber," [Online]. Available: http://ambermd.org/.

[15] "HOOMD-blue," [Online]. Available: http://glotzerlab.engin.umich.edu/hoomd-blue/.

[16] "NAMD - Scalable Molecular Dynamics," [Online]. Available: http://www.ks.uiuc.edu/Research/namd/.

[17] "Accelerating HPC applications using K80 GPUs," [Online]. Available: http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2015/11/12/accelerating-hpc-applications-using-k80-gpus.

[18] "Gromacs," [Online]. Available: http://www.gromacs.org/.