DELL EMC READY BUNDLE FOR HPC LIFE SCIENCES

Refresh with 14th Generation servers

ABSTRACT

Dell EMC's flexible HPC architecture for Life Sciences has been through a dramatic improvement with new Intel® Xeon® Scalable Processors. Dell EMC Ready Bundle for HPC Life Sciences equipped with better 14G servers, faster CPUs, and more memory bring a much higher performance in terms of throughput compared to the previous generation especially in genomic data processing.

March 2018

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Since Dell EMC announced Dell EMC HPC solution for Life Science in September 2016, the current Dell EMC Ready Bundle for HPC Life Sciences can process 485 genomes per day[i] with 64x C6420s and Dell EMC Isilon F800 in our benchmarking. This is roughly two-fold improvement from Dell EMC HPC System for Life Science v.1.1 due to the introduction of Dell EMC's14th generation servers which include the latest Intel® Xeon® Scalable processors (code name: Skylake), updated server portfolio, improved memory, and storage subsystem performance (1).

This whitepaper describes the architectural changes and updates to the follow-on of Dell EMC HPC System for Life Science v1.1. It explains new features, demonstrates the benefits, and shows the improved performance.

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

Although the successful completion of the Human Genome Project was announced on April 14, 2003 after a 13-year-long endeavor and numerous exciting breakthroughs in technology and medicine, there's still a lot of work ahead for understanding and using the human genome. As an iterative advance in sequencing technology has accumulated, these cutting-edge technologies allow us to look at many different perspectives or levels of genetic organization such as whole genome sequencing, copy number variations, chromosomal structural variations, chromosomal methylation, global gene expression profiling, differentially expressed genes, and so on. However, despite the scores of data we generated, we still face the challenges of understanding the basic biology behind the human genome and the mechanisms of human diseases. There will not be a better time than now to evaluate if we have overlooked the current approach, "analyze large numbers of diverse samples with the highest resolution possible" because analyzing a large number of sequencing data alone has been unsuccessful to identify key genes/variants in many common diseases where majority of genetic variations seem to be involved randomly. This is the main reason that data integration becomes more important than before. We have no doubt that genome research could help fight human disease; however, combining proteomics and clinical data with genomics data will increase the chances of winning the fight. As the life science community is gradually moving toward to data integration, **E**xtract, **T**ransform, **L**oad (ETL) process will be a new burden to an IT department (2).

Dell EMC Ready Bundle for HPC Life Sciences has been evolving to cover various needs for Life Science data analysis from a system only for **N**ext **G**eneration **S**equencing (NGS) data processing to a system that can be used for Molecular Dynamics Simulations (MDS), Cryo-EM data analysis and *De Novo* assembly in addition to DNA sequencing data processing. Further, we are planning to cover more applications from other areas of Life Sciences and re-design the system suitable for data integration. In this project, we tested an additional two Dell EMC Isilon storages as the importance of finding suitable storages for variable purposes grows. This is an additional step to build a HPC system as a tool to integrate all the cellular data, biochemistry, genomics, proteomics and biophysics into a single frame of work and to be ready for the high-demanding era of ETL process.

SOLUTION OVERVIEW

Especially, HPC in Life Sciences requires a flexible architecture to accommodate various system requirements. Dell EMC Ready Bundle for HPC Life Sciences was created to meet this need. It is a pre-integrated, tested, tuned, and leverages the most relevant of Dell EMC's high-performance computing line of products and best-in-class partner products (3). It encompasses all of 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 Ready Bundle for HPC Life Sciences provides high flexibility. The platform is available in three variants which are determined by the cluster interconnect selected for the storages. In the current version, the following options are available:

- PowerEdge™ C6420 compute subsystem with Intel® Omni-Path (OPA) fabric or Mellanox InfiniBand® (IB) EDR fabric
  - Storage choices:
    - Dell EMC Ready Bundle for HPC Lustre Storage as a performance scratch space
    - Dell EMC Ready Bundle for HPC NFS Storage as a home directory space

- PowerEdge C6420 compute subsystem with 10/40GbE fabric
  - Storage choices:
    - Either Dell EMC Isilon F800 or Dell EMC Isilon H600 as a performance scratch space
    - Dell EMC Ready Bundle for HPC NFS Storage as a home directory space

- Add-on compute nodes:
  - Dell EMC PowerEdge R940
    - This server covers large memory applications such as De Novo Assembly
  - Dell EMC PowerEdge C4130
    - A server for accelerators like NVIDIA GPUs

In addition to the compute, network, and storage options, there are several other components that perform different functions in the Dell EMC Ready Bundle for HPC Life Sciences. These include CIFS gateway, fat node, acceleration node and other management components. Each of these components is described in detail in the subsequent section.

The solutions are nearly identical for Intel® OPA and IB EDR versions except for a few changes in the switching infrastructure and network adapter. The solution ships in a deep and wide 48U rack enclosure, which helps to make PDU mounting and cable

management easier. Figure 1 shows the components of two fully loaded racks using 64x Dell EMC PowerEdge C6420 rack server chassis as a compute subsystem, Dell EMC PowerEdge R940 as a fat node, Dell EMC PowerEdge C4130 as an accelerator node, Dell EMC Ready Bundle for HPC NFS Storage, Dell EMC Ready Bundle for HPC Lustre Storage and Intel® OPA as the cluster's high speed interconnect.

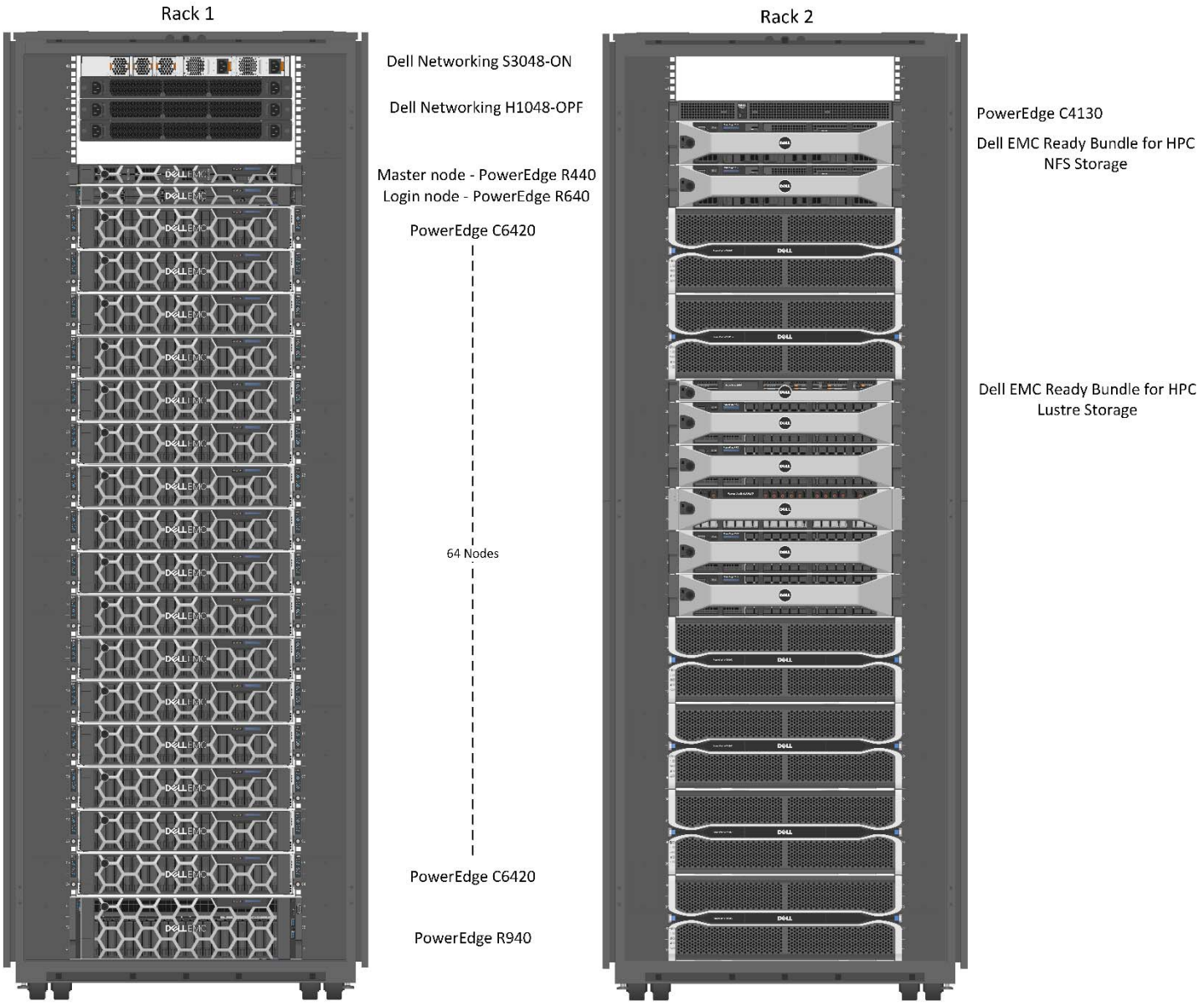

Figure 1 Dell EMC Ready Bundle for HPC Life Sciences with Intel® OPA fabric

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 node, 1U form factor Dell EMC PowerEdge R440 is recommended. The master node is responsible for managing the compute nodes and optimizing the overall compute capacity. The login node (Dell EMC PowerEdge R640 is recommended) 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 can be provided for master and login nodes. An example solution is illustrated in Figure 1, and the high-speed interconnect is configured to 2:1 blocking fat tree topology with Intel® OPA fabric.

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 careful execution to achieve the best performance; however, for Life Sciences applications, heterogeneous clusters make perfect sense to handle completely independent workloads such as DNA-Seq, De Novo assembly or Molecular Dynamics Simulations. These workloads require quite different hardware components. Hence, we recommend Dell EMC PowerEdge C6420 as a compute node to handle NGS data processing due to its density, a wide choice of CPUs, and high maximum memory capacity. Dell EMC PowerEdge R940 is an optional node with 6TB of RDIMM /LRDIMM memory and is recommended for customers who need to run applications requiring large memory such as De Novo assembly. Accelerators are used to speed up computationally intensive applications, such as molecular dynamics simulation applications. We tested configurations G and K for this solution.

The compute and management infrastructure consists of the following components.

- Compute
  - Dell EMC PowerEdge C6400 enclosure with 4x C6420 servers
    - 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.
    - It 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.
    - It also 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 with up to four NVIDIA® Tesla™ P100 or V100 GPUs
    - It provides supercomputing agility and performance in an ultra-dense platform purpose-built for scale-out HPC workloads. 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® E5-2690 v4 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 R940
    - The Dell EMC PowerEdge R940 is a 4-socket, 3U platform, equipped with the Intel® Xeon® Platinum 8168 2.7GHz (24 cores per socket – 96 cores in server or 28 cores per socket – 112 cores in server with Intel® Xeon® Platinum 8180(M) Processor) and is dubbed “the fat node,” because of its 6 TB of memory capacity.
    - This server is an optional component that can be added to the solution for *De Novo* assembly, visualization and/or large statistical analysis.

- Management
  - Dell EMC PowerEdge R440 for master node
    - It is used by Bright Cluster Manager® to provision, manage and monitor the cluster.
    - Optionally, a high availability (HA) configuration is also available by adding an additional server and set them in an active-passive HA state.
  - Dell EMC PowerEdge R640 for login node and CIFS gateway (*optional*)
    - These components are optional.
    - Typically, a master node can manage user logins without a problem. However, it is wise to separate login/job scheduling from a master node when the number of users grows larger. The login node can also be configured in HA mode by adding an additional server.
    - This optional CIFS gateway can help data transferring from NGS sequencers to a shared storage.

Storage Components

The storage infrastructure consists of the following components:

- Dell EMC Ready Bundle for HPC NFS Storage (NSS7.0-HA) (4)
- Dell EMC Ready Bundle for HPC Lustre Storage (4)
- Dell EMC Isilon Scale-out NAS Product Family; Dell EMC Isilon F800 All-flash (5)
  or Dell EMC Isilon Hybrid Scale-out NAS H600 (5)

Dell EMC Ready Bundle for HPC NFS Storage (NSS7.0-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 Dell EMC PowerVault™ storage arrays, Red Hat HA software stack. 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 Intel® OPA, IB or 10GbE. 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.

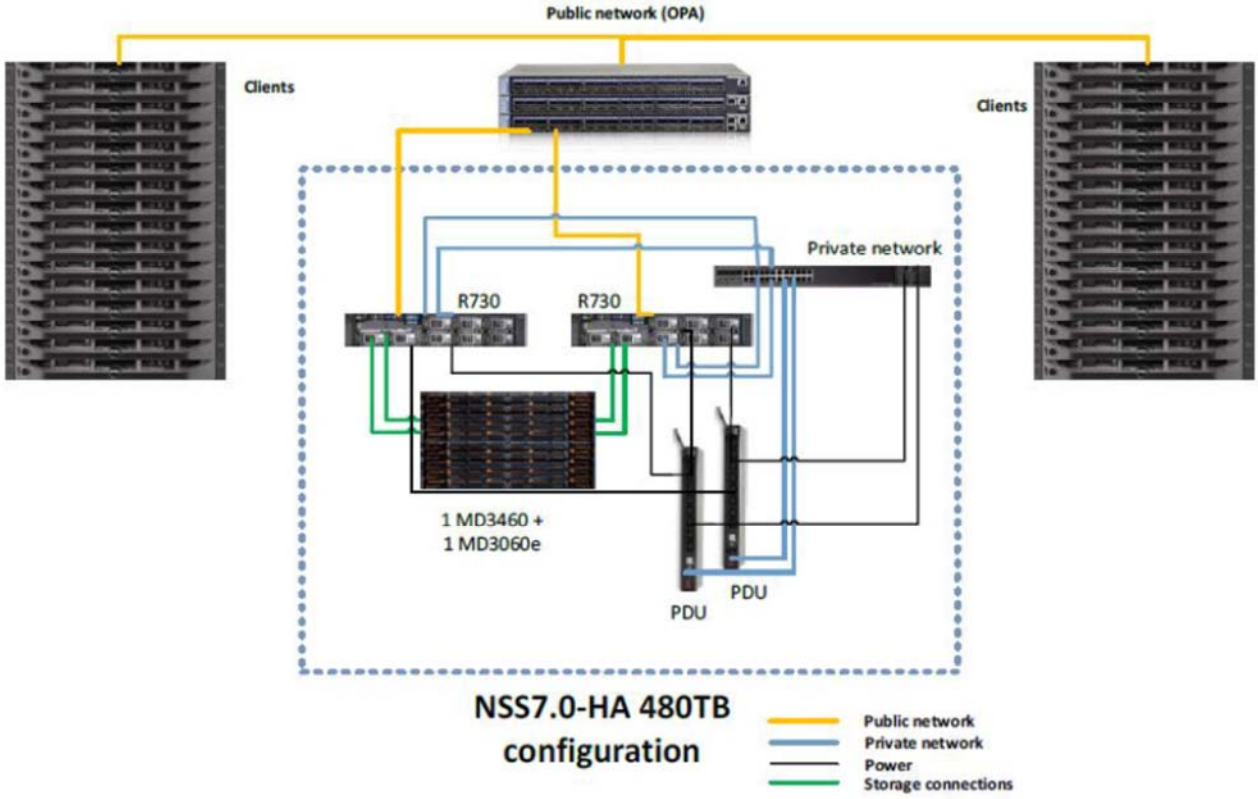

Figure 2 NSS 7.0-HA configuration

Dell EMC Ready Bundle for HPC Lustre Storage

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. Intel® Enterprise Edition (EE) for Lustre® software v.3.0. 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 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.

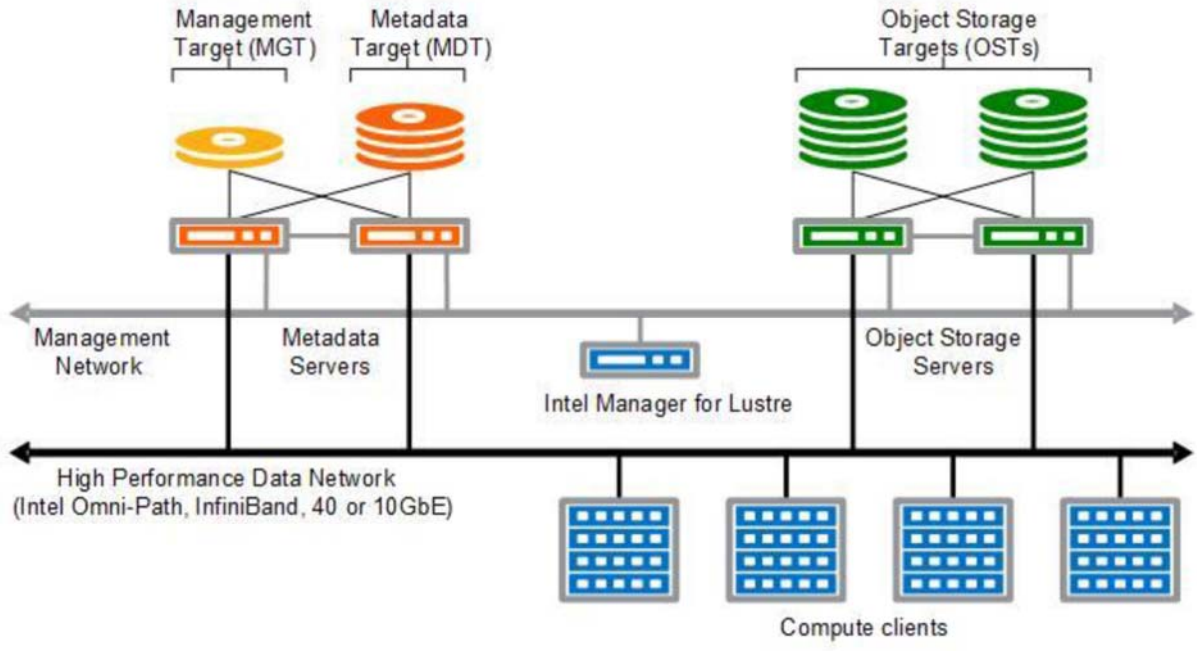

Figure 3 Lustre-based storage solution components

Dell EMC Isilon F800 All-flash and Dell EMC Isilon Hybrid Scale-out NSA H600

A single Isilon storage cluster can host multiple node types to maximize deployment flexibility. Node types range from the Isilon F (All Flash) to H (Hybrid), and A (Archive) nodes. Each provides a different optimization point for capacity, performance, and cost. Automated processes can be established that automatically migrate data from higher-performance, higher-cost nodes to more cost-effective storage. Nodes can be added "on the fly," with no disruption of user services. Additional nodes result in increased performance (including network interconnect), capacity and resiliency.

The Dell EMC Isilon OneFS operating system powers all Dell EMC Isilon scale-out NAS storage solutions. OneFS also supports additional services for performance, security, and protection:

- **SmartConnect** is a software module that optimizes performance and availability by enabling intelligent client connection load balancing and failover support. Through a single host name, SmartConnect enables client connection load balancing and dynamic NFS failover and failback of client connections across storage nodes to provide optimal utilization of the cluster resources.
- **SmartPools** provides rule based movement of data through tiers within an Isilon cluster. Institutions can set up rules keeping the higher performing nodes available for immediate access to data for computational needs and NL and HD series used for all other data. It does all this while keeping data within the same namespace, which can be especially useful in a large shared research environment.
- **SmartFail** and **Auto Balance** ensure that data is protected across the entire cluster. There is no data loss in the event of any failure and no rebuild time necessary. This contrasts favorably with other file systems such as Lustre or GPFS as they have significant rebuild times and procedures in the event of failure with no guarantee of 100% data recovery.
- **SmartQuotas** help control and limit data growth. Evolving data acquisition and analysis modalities coupled with significant movement and turnover of users can lead to significant consumption of space. Institutions without a comprehensive data management plan or practice can rely on SmartQuotas to better manage growth.

Through utilization of common network protocols such as CIFS/SMB, NFS, HDFS, and HTTP, Isilon can be accessed from any number of machines by any number of users leveraging existing authentication services.

Formerly known as "Project Nitro," the highly dense, bladed-node architecture of F800 provides four nodes within a single 4U chassis with capacity options ranging from 92TB to 924TB per chassis. Available in several configurations, F800 units can be combined into a single cluster that provides up to 92.4PB of capacity and over 1.5TB/s of aggregate performance. It is designed for a wide range of next-generation applications and unstructured workloads that require extreme NAS performance including, 4K streaming of data, genomic sequencing, electronic design automation, and near real-time analytics. It can also be deployed as a new cluster, or can

seamlessly integrate with existing Isilon clusters to accelerate the performance of an enterprise data lake and lower the overall total cost of ownership (TCO) of a multi-tiered all-flash and high capacity SATA solution. Powered by the OneFS operating system, this new offering from Dell EMC is designed to provide the essential capabilities that enterprises require to modernize IT: extreme performance, massive scalability, operational flexibility, increased efficiency, and enterprise-grade data protection and security.

Similarly, H600 also uses bladed-node architecture and has capacity options from 72TB to144TB per chassis. It is equipped with 120 SAS drives, SSD caching, and built-in multiprotocol capabilities.

Network Components

Dell Networking H1048-OPF

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.

Figure 4 Dell Networking H1048-OPF

Mellanox SB7700

This 36-port Non-blocking Managed InfiniBand EDR 100Gb/s Switch System provides the highest performing fabric solution in a 1U form factor by delerivering up to 7.2Tb/s of non-blocking bandwidth with 90ns port-to-port latency.

Figure 5 Mellanox SB7700

Dell Networking Z9100-ON

The Dell EMC Networking Z9100-ON is a 10/25/40/50/100GbE fixed switch purpose-built for applications in high-performance data center and computing environments. 1RU high-density 10/25/40/50/100GbE fixed switch with choice of up to 32 ports of 100GbE (QSFP28), 64 ports of 50GbE (QSFP+), 32 ports of 40GbE (QSFP+), 128 ports of 25GbE (QSFP+) or 128+2 ports of 10GbE (using breakout cable).

Figure 6 Dell Networking Z9100-ON

Dell Networking S3048-ON

Management traffic typically communicates with the Baseboard Management Controller (BMC) on the compute nodes using IPMI. The management network is used to push images or packages to the compute nodes from the master nodes and for reporting data from client to the master node. Dell EMC Networking S3048-ON is recommended for management network.

Figure 7 Dell Networking S3048-ON

Interconnect

Figure 8 describes how the network components are configured for different storages. InfiniBand EDR connection uses Mellanox SB7700 switch, and the cable connections are identical to Dell Networking H1048-OPF. Either Intel® OPA or IB EDR network is configured as 2:1 blocking fat tree topology.

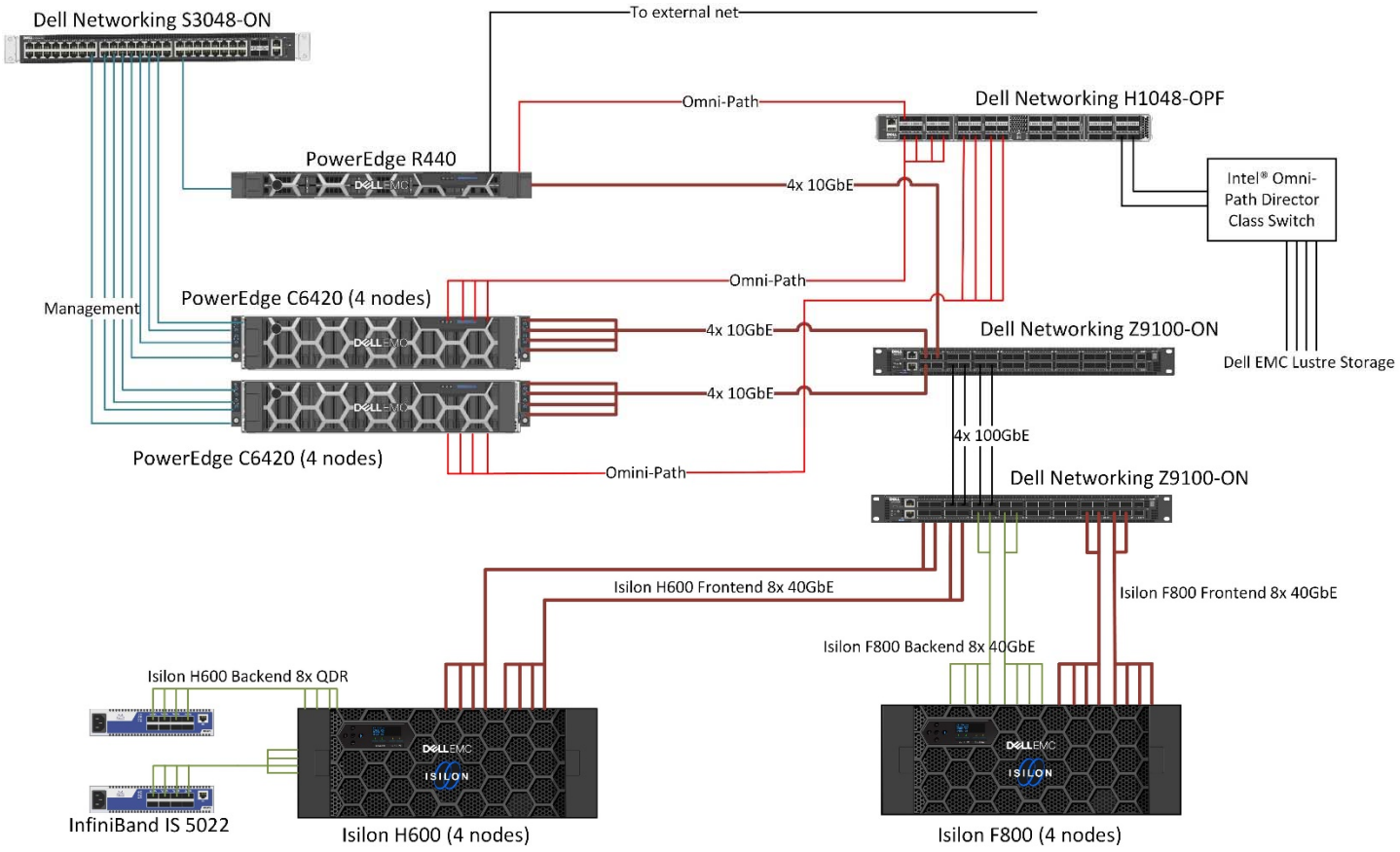

Figure 8 Illustration of Networking Topology

Software Components

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

- Bright Cluster Manager®
- BioBuilds

Bright Cluster Manager

Bright Computing is a commercial software that provides comprehensive software solutions for deploying and managing HPC clusters, big data clusters and OpenStack in the data center and in the cloud (6). Bright cluster Manager can be used to deploy complete clusters over bare metal and manage them effectively. Once the cluster is up and running, the graphical user interface monitors every single node and reports if it detects any software or hardware events.

BioBuilds

BioBuilds is a well maintained, versioned, and continuously growing collection of open-source bio-informatics tools from Lab7 (7). They are prebuilt and optimized for a variety of platforms and environments. BioBuilds solves most 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 of the dependencies the software has as well. BioBuilds includes the software and its supporting dependencies for ease of deployment.
- Using BioBuilds among all of the collaborators can ensure reproducibility since everyone is running the same version of the software. In short, it is a turnkey application package.

PERFORMANCE EVALUATION AND ANALYSIS

ALIGNER SCALABILITY

This is a base-line test 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. There are more than 90 short read alignment programs available as you can see in Figure 9 (8). Traditional sequence alignment tools like BLAST are not suited for NGS. To extract meaningful information from NGS data, one needs to align millions and millions of mostly short sequences. BLAST and similar tools are way to slow for the vast amount of data produced by modern sequencing machines.

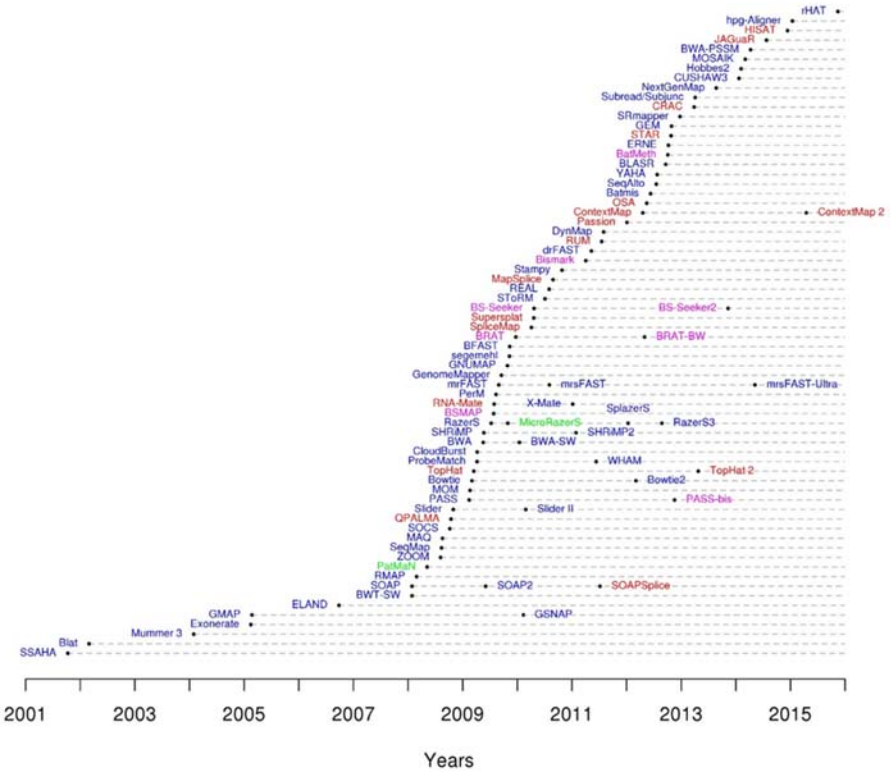

Figure 9 Aligners currently available

Hence, one might want to know what the best alignment software is; however, it is hard to answer to the question since the answer can be drawn from many different conditions. Even if we could compare all different alignment software, there will not be any conclusion which alignment tool is the best. It really depends on your goals and the specific use case like the reason we choose to test Burrows-Wheeler Aligner (BWA) is because it is a part of the popular variant calling workflow with Genome Analysis Toolkit (GATK) (9) (10).

Nonetheless, one of many aligners, BWA, scales stably for different numbers of cores and various NGS data size with Dell EMC PowerEdge C6420. A single PowerEdge C6420 server is used to generate baseline performance metrics and ascertain the optimum number of cores for running BWA for these scaling tests.

Figure 10 shows the run times of BWA on various sequence data sizes ranging from 2 to 208 million fragments (MF) and different number of threads. Although the results show speed-up due to increasing core count in general, the optimum number of cores for BWA is in between 12 - 20.

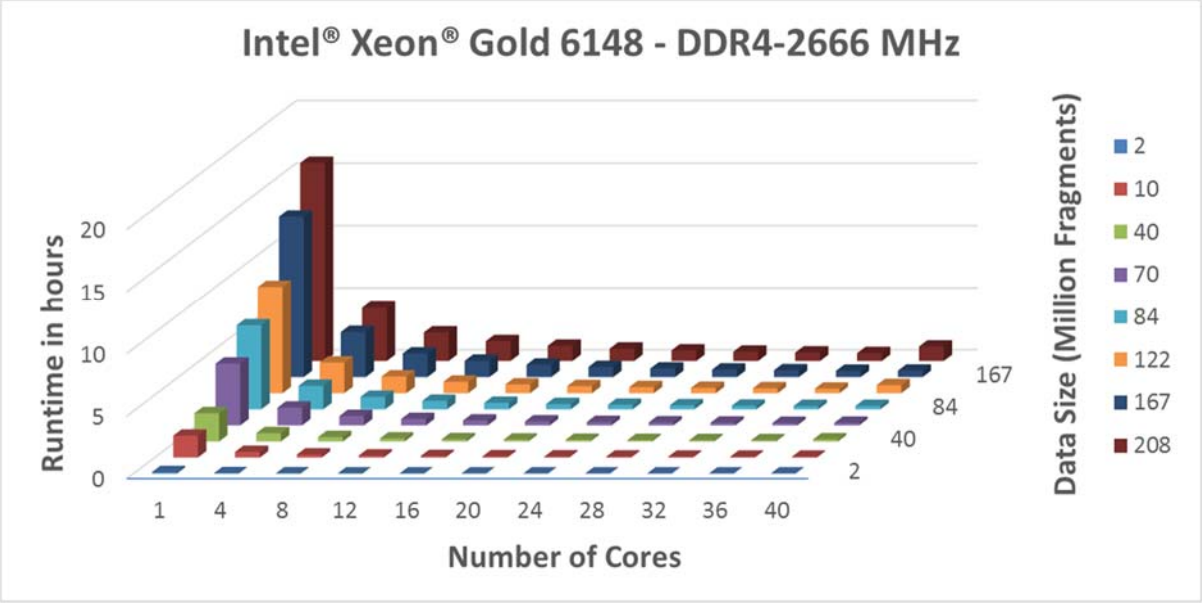

Figure 10 Scaling behavior of BWA

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 3.6, 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 30x whole human genome sequencing data from the Illumina platinum genomes project, named ERR091571_1.fastq.gz and ERR091571_2.fastq.gz were used for a baseline test (11).

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 of the simultaneous runs will read/write roughly at the same time, 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 much resources are required for an identical or even similar analysis pipeline with a defined workload.

The throughput of Dell HPC Solution 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 11.

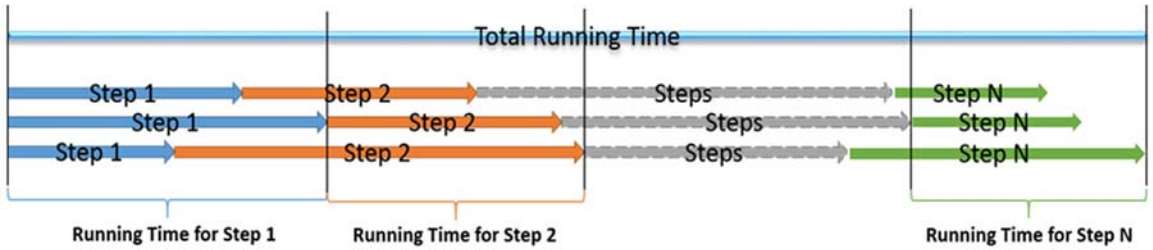

Figure 11 Running time measurement method

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 12, the throughputs in total number of genomes per day for all tests with various numbers of 30x whole genome sequencing data are summarized. The tests performed here are designed to demonstrate performance at the server level, not for comparisons on individual components. At the same time, the tests were also designed to estimate the sizing information of Dell EMC Isilon F800/H600 and Dell EMC Lustre Storage. The data points in Figure 12 are calculated based on the total number of samples (X axis in the figure) that were processed concurrently. The number of genomes per day metric is obtained from total running time taken to process the total number of samples in a test. The smoothed curves are generated by using a polynomial spline with the piecewise polynomial degree of 3 generating B-spline basis matrix. The details of BWA-GATK pipeline information can be obtained from the Broad Institute web site (10).

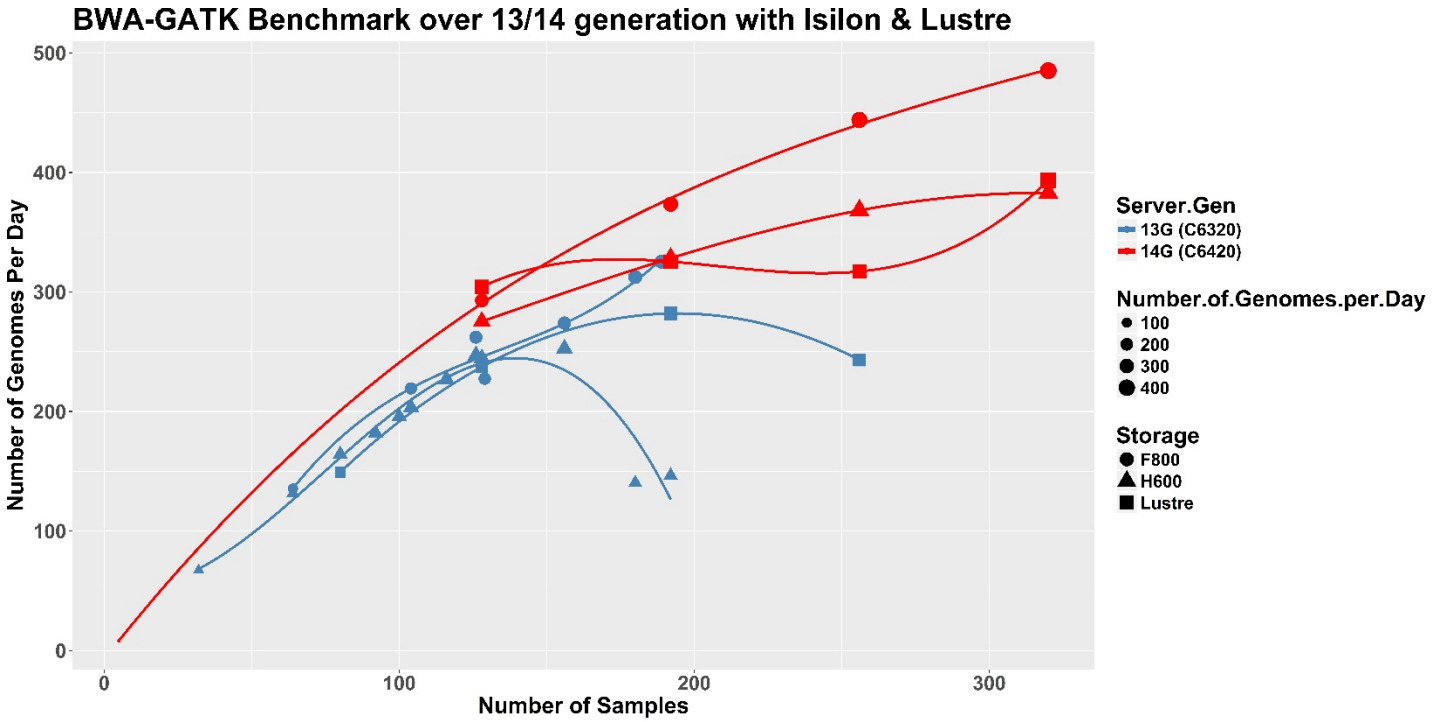

Figure 12 Performances in 13/14 generation servers with Isilon and Lustre

The number of compute nodes used for the tests are 64x C6420s and 63x C6320s (64x C6320s for testing H600). The number of samples per node was increased to get the desired total number of samples processed concurrently. For C6320 (13G), 3 samples per node was the maximum number of samples each node can process. 64, 104, and 126 test results for 13G system (blue) were with 2 samples per node while 129, 156, 180, 189 and 192 sample test results were obtained from 3 samples per node. For C6420 (14G), the tests were performed with maximum 5 samples per node. The plot for 14G was generated by processing 1, 2, 3, 4, and 5 samples per node. The number of samples per node is limited by the amount of memory in a system. 128 GB and 192 GB of RAM were used in 13G and 14G system, respectively. C6420s show a better scaling behavior than C6320s. 13G server with Broadwell CPUs seems to be more sensitive to the number of samples loaded onto system as shown from the results of 126 vs 129 sample tests on all the storages tested in this study.

Dell EMC PowerEdge C6420 has at least a 12% performance gain compared to the previous generation. Each C6420 compute node with 192 GB RAM can process about seven 30x whole human genomes per day. This number could be increased if the C6420 compute node is configured with more memory. In addition to the improvement on the 14G server side, four Isilon F800 nodes in a 4U chassis can support 64x C6420s and 320 30x whole human genomes concurrently.

MOLECULAR DYNAMICS SIMULATION SOFTWARE PERFORMANCE

Over the past decade, GPUs has become 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 and 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 two different molecular dynamics applications to run tests on a PowerEdge C4130 with P100 and V100. The applications are Amber and LAMMPS (12) (13).

Molecular dynamics application test configuration

The Dell EMC PowerEdge C4130 with Intel® Xeon® Dual E5-2690 v4 with 256 GB DDR4 2400MHz and P100 and V100 in G and K configurations (14) (15). Unfortunately, we were not able to complete integration tests with Dell EMC Ready Bundle for HPC Life Sciences with various interconnect settings. However, we believe the integration tests will not pose any problem since molecular dynamics applications are not bounded by either storage I/O or inter-communication bandwidth.

The NVIDIA® Tesla® V100 accelerator is one of the most advanced accelerators available in the market right now and was launched within one year of the P100 release. In fact, Dell EMC is the first in the industry to integrate Tesla V100 and bring it to market. As was the case with the P100, V100 supports two form factors: V100-PCIe and the mezzanine version V100-SXM2. The Dell EMC PowerEdge C4130 server supports both types of V100 and P100 GPU cards.

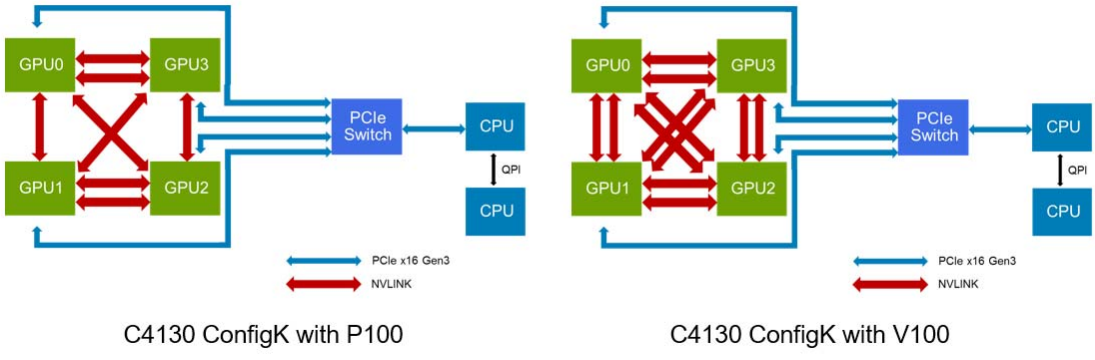

Figure 13 V100 and P100 Topologies on C4130 configuration K

Amber benchmark suite

This suite includes the Joint Amber-Charmm (JAC) benchmark considering dihydrofolate reductase (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 microcanonical (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.

Figure 14

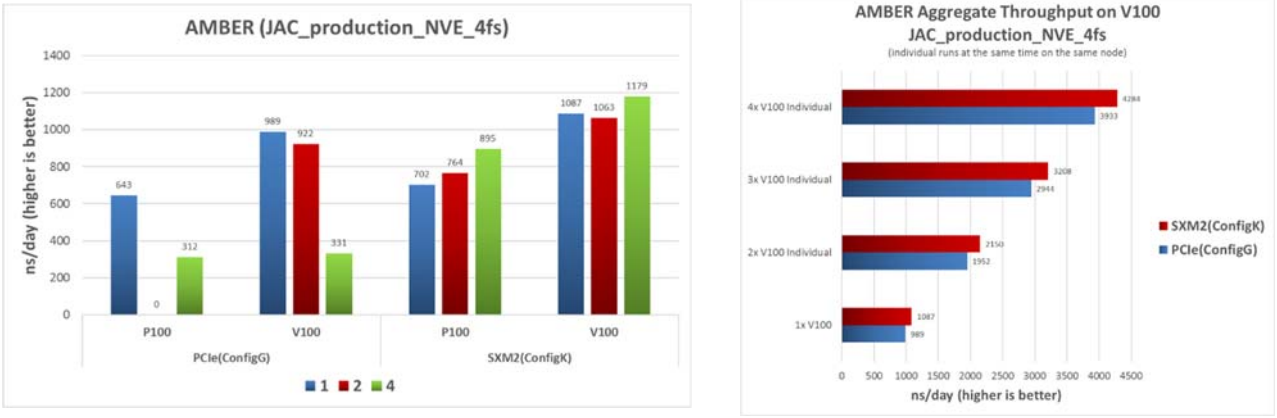

Figure 14 AMBER JAC Benchmark

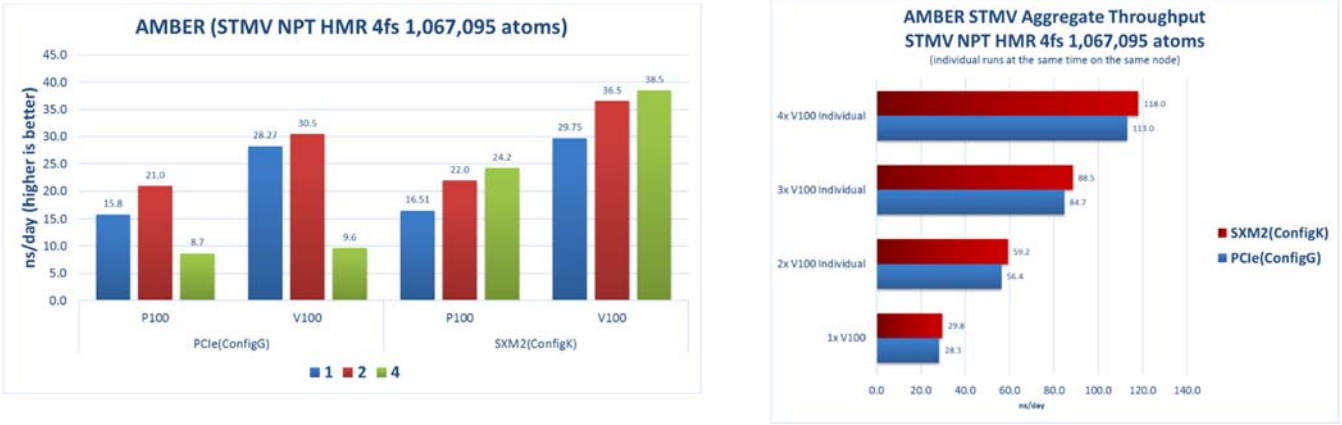

Figure 15 AMBER STMV benchmark

Figure 14 and Figure 15 illustrate AMBER's results with DHFR and STMV dataset. On SXM2 system (Config K), AMBER scales weakly with 2 and 4 GPUs. Even though the scaling is not strong, V100 has noticeable improvement than P100, giving ~78% increase in single card runs, and 1x V100 is actually 23% faster than 4x P100. On the PCIe (Config G) side, one and two cards perform similar to SXM2; however, four cards' results dropped sharply. This is because PCIe (Config G) only supports Peer-to-Peer access between GPU0/1 and GPU2/3 and not among all four GPUs. Since AMBER has redesigned the way data transfers among GPUs to address the PCIe bottleneck, it relies heavily on Peer-to-Peer access for performance with multiple GPU cards. Hence a fast, direct interconnect like NVLink between all GPUs in SXM2 (Config K) is vital for AMBER multiple GPU performance. To compensate for a single job's weak scaling on multiple GPUs, there is another use case promoted by AMBER developers, which is running multiple jobs in the same node concurrently but where each job uses only 1 or 2 GPUs. Figure 5 shows the results of 1-4 individual jobs on one C4130 with V100s and the numbers indicate that those individual jobs have little impact on each other. This is because AMBER is designed to run pretty much entirely on the GPUs and has very low dependency on the CPU. The aggregate throughput of multiple individual jobs scales linearly in this case. Without any card to card communication, the 5% better performance on SXM2 is contributed by its higher clock speed.

LAMMPS

Large-scale **A**tomic/**M**olecular **M**assively **P**arallel **S**imulator (LAMMPS) is a classical molecular dynamics code and has potentials for solid-state materials (metals, semiconductors) and soft matter (biomolecules, polymers) and coarse-grained or mesoscopic systems. It can be used to model atoms or, more generically, as a parallel particle simulator at the atomic, meso, or continuum scale. It runs on single processors or in parallel using message-passing techniques and a spatial-decomposition of the simulation domain. Many of its models have versions that provide accelerated performance on CPUs, GPUs, and Intel Xeon Phis. The code is designed to be easy to modify or extend with new functionality.

We performed a 3D Lennard-Jones melt simulation package which comes with LAMMPS. It simulates the movement of molecules confined in a squared box. LAMMPS has built-in functions for placing molecules inside the box and moving them according to Newton's laws. This is arguably one of the simplest models capable of reproducing the complete thermodynamic behavior of classical fluids. The details of the simulation set-up can be found in Appendix B.

Figure 16 shows LAMMPS performance on both configurations G and K. The testing dataset is Lennard-Jones liquid dataset, which contains 512,000 atoms, and LAMMPS compiled with the kokkos package. V100 is 71% and 81% faster on Config G and Config K respectively. Comparing V100-SXM2 (Config K) and V100-PCIe (Config G), the former is 5% faster due to NVLINK and higher CUDA core frequency.

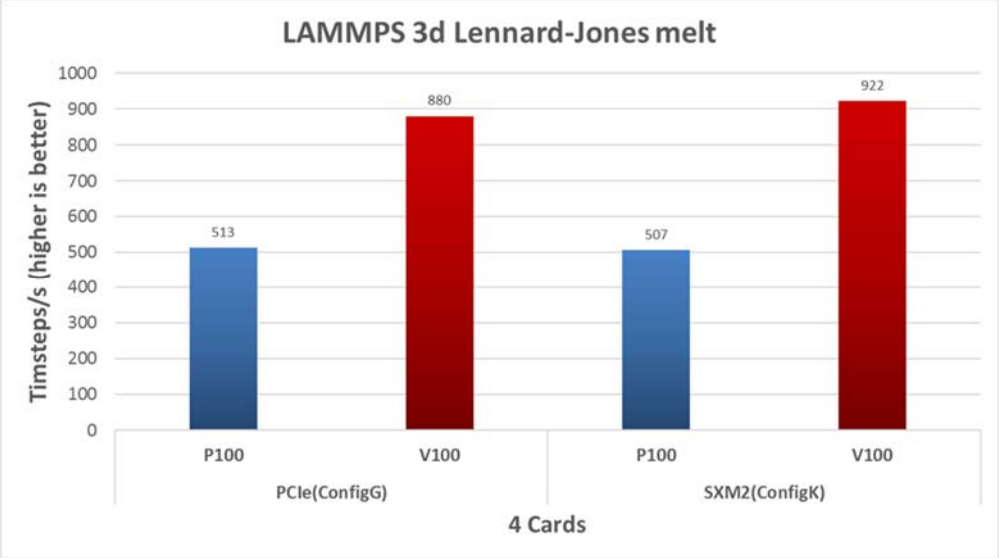

Figure 16 LAMMPS 3D Lennard-Jones melt simulation

The C4130 server with NVIDIA® Tesla® V100™ GPUs demonstrates exceptional performance for HPC applications that require faster computational speed and highest data throughput. Applications like AMBER and LAMMPS were boosted with C4130 configuration K, owing to P2P access, higher bandwidth of NVLink and higher CUDA core clock speed. Overall, a PowerEdge C4130 with Tesla V100 GPUs performs 1.54x to 1.8x faster than a C4130 with P100 for AMBER and LAMMPS.

CRYO-EM PERFORMANCE

The purpose of this study was to validate the optimized Relion (for **RE**gularised **LI**kelihood **O**ptimizatio**N**) on Dell EMC PowerEdge C6420s with Skylake CPUs. Relion was developed from the Scheres lab at MRC Laboratory of Molecular Biology. It uses an empirical Bayesian approach to refine multiple 3D images or 2D class averages for the data generated from CryoElectron Microscopy (Cryo-EM). The impressive performance gain from Intel®'s efforts in the collaboration of Relion development team reduced the performance gap between CPUs and GPUs. The CPU/GPU performance comparison results are not shown here; however, the performance gap becomes single digit fold between Skylake CPU systems and Broadwell CPU/Tesla P100 GPU systems.

Essentially, Cryo-EM is a type of **T**ransmission **E**lectron **M**icroscopy (TEM) for imaging frozen-hydrated specimens at cryogenic temperatures. Specimens remain in their native state without the need for dyes or fixatives, allowing the study of fine cellular structures, viruses and protein complexes at molecular resolution. A rapid vitrification at cryogenic temperature is the key step to avoid water molecule crystallization and forming amorphous solid that does almost no damage to the sample structure. Regular electron microscopy requires samples to be prepared in complex ways, and the sample preparations make hard to retaining the original molecular structures. Cryo-EM is not perfect like X-ray crystallography; however, it has quickly gained the popularity in the research community due to the simple sample preparation steps and flexibility of the sample size, complexity, and non-rigid structure. As the resolution revolution in Cryo-EM progresses due to the 40+ years of dedicated work from the structural biology community, we now can yield accurate, detailed 3D models of intricate biological structures at the sub-cellular and molecular scales.

The tests were performed on 8 nodes of Dell PowerEdge C6420s which is a part of Dell EMC Ready Bundle for HPC Life Sciences.

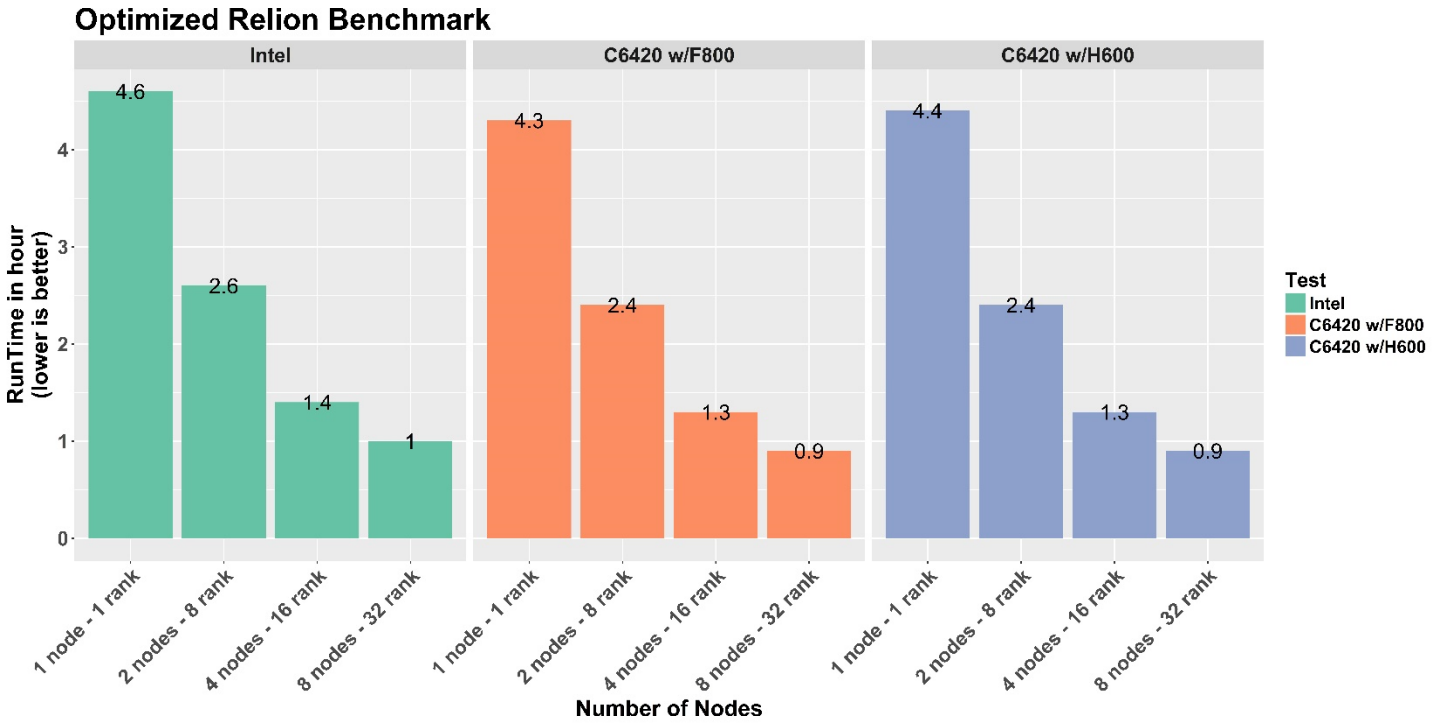

Figure 17 Optimized Relion Benchmark

Dell EMC PowerEdge C6420 shows that it is an ideal compute platform for the Optimized Relion. It scales well over various number of compute nodes with Plasmodium ribosome data. In the future study, we plan to use a larger protein data and more compute nodes to accomplish more comprehensive scaling tests.

CONCLUSION

Overall, 14th generation servers with Skylake and larger/faster memory size (due to higher number of memory channels compare with Broadwell) show a better throughput on BWA-GATK pipeline. The throughput for this type of work improved from four 30x genomes per day per C6320 to seven 30x genomes per day per C6420. Also, we verified that Dell EMC Isilon storage, F800 and H600, can be used for high-performance scratch storage while they provide all the conveniences of easy maintenance, scaling, and various supported file systems. In addition to that, we observed a better performance on Cryo-EM data process with Intel®'s optimized Relion codes. Unfortunately, we could not test Dell EMC PowerEdge C4140 with NVIDIA® Tesla™ V100; however, the V100 with PowerEdge C4130 still shows impressive performance compared to P100. We believe the current version of Dell EMC Ready Bundle for HPC Life Sciences is ready for data centric high-performance computing.

APPENDIX A

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

APPENDIX B

```
# 3d Lennard-Jones melt

variable      N string off          # Newton Setting
variable      w equal 10            # Warmup Timesteps
variable      t equal 7900          # Main Run Timesteps
variable      m equal 1             # Main Run Timestep Multiplier
variable      n equal 0             # Use NUMA Mapping for Multi-Node
variable      p equal 0             # Use Power Measurement

variable      x equal 4
variable      y equal 2
variable      z equal 2

variable      xx equal 20*$x
variable      yy equal 20*$y
variable      zz equal 20*$z
variable      rr equal floor($t*$m)

newton        $N
if "$n > 0"    then "processors * * * grid numa"

units         lj
atom_style    atomic

lattice       fcc 0.8442
region        box block 0 ${xx} 0 ${yy} 0 ${zz}
create_box    1 box
create_atoms  1 box
mass          1 1.0

velocity      all create 1.44 87287 loop geom

pair_style    lj/cut 2.5
pair_coeff    1 1 1.0 1.0 2.5

neighbor      0.3 bin
neigh_modify  delay 0 every 20 check no

fix           1 all nve
thermo        1000

if "$p > 0"    then "run_style verlet/power"

if "$w > 0"    then "run $w"
run           ${rr}
```

REFERENCES

1. Blueprint for High Performance Computing. *Dell TechCenter.* [Online] http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20443473.

2. ETL: The Silent Killer of Big Data Projects. *insideBIGDATA.* [Online] https://insidebigdata.com/2015/07/23/etl-the-silent-killer-of-big-data-projects/.

3. Dell EMC PowerEdge Servers. [Online] https://www.dellemc.com/en-us/servers/index.htm.

4. Dell EMC Ready Bundles for HPC Storage. [Online] https://si.cdn.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/HPC-Storage.pdf.

5. Dell EMC Isilon F800 AND H600 I/O Performance . [Online] https://www.emc.com/collateral/white-papers/f800-h600-performance-wp.pdf.

6. Bright Cluster Manager. [Online] http://www.brightcomputing.com/products.

7. BioBuilds. [Online] https://www.lab7.io/.

8. What is the best NGS alignment software? [Online] https://www.ecseq.com/support/ngs/what-is-the-best-ngs-alignment-software.

9. Burrows-Wheeler Aligner. [Online] http://bio-bwa.sourceforge.net/.

10. Genome Analysis Toolkit. [Online] https://software.broadinstitute.org/gatk/.

11. European Nucleotide Archive: ERR091571. [Online] https://www.ebi.ac.uk/ena/data/view/ERR091571.

12. Amber. [Online] http://ambermd.org/.

13. LAMMPS. [Online] http://lammps.sandia.gov/.

14. Application Performance on P100-PCIe GPUs. [Online] http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2017/08/17/application-performance-on-p100-pcie-gpus.

15. HPC Applications Performance on V100. [Online] http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2017/09/29/hpc-applications-performance-on-v100.

[i] The tested sequencing data's depth of coverage is 30x when we assume that the sequence reads are generated from germ cells which are haploids.