DELL Technologies

HPC Ready Architecture

Dell EMC HPC Ready Architecture for Genomics

Abstract

The Dell EMC HPC Ready Architecture for Genomics uses a flexible and modular approach to HPC system design where individual building blocks can be combined to build HPC systems that are optimized for specific workloads and use cases. These tested and tuned solutions leverage Dell EMC servers, networking and storage, with services available from consulting to financing. They include the hardware resources required for various forms of genomic data analysis while providing an optimal balance of compute density, energy efficiency and performance.

The HPC Ready Architecture for Genomics is designed to speed time to production, improve performance with purpose-built solutions, and scale more easily with modular building blocks for capacity and performance. The white paper provides configuration guidance based on next generation sequencing (NGS) and de novo assembly applications and workloads.

In this document, the Dell Technologies Engineering team outlines the system design and performance benchmarking results. The solution can process 56 whole human genomes at 50X coverage can be processed in roughly 54 hours on the medium sized configuration, which consists of eight compute nodes inside 2x Dell EMC PowerEdge C6420s coupled with Dell EMC Ready Solutions for HPC BeeGFS High Capacity Storage.

July 2020

Revisions

Date	Description
July 2020	Initial Release

Authors

- Kihoon Yoon, Senior Principal Systems Development Engineer, HPC Solutions, Dell Technologies
- Adnan Khaleel, Global HPC Sales Strategy, Dell Technologies
- Michael McManus, Principle Engineer and Sr. Life Sciences Solution Architect, Intel

Table of contents

Introduction

The HPC Ready Architecture for Genomics is a tested and tuned solution that leverages Dell EMC servers, software, networking and storage. It includes resources required for various forms of genomic data analysis while providing an optimal balance of compute density, energy efficiency and performance.

The Dell EMC HPC Ready Architecture for Genomics uses a flexible and modular approach to HPC system design where individual building blocks can be combined to build HPC systems that are optimized for specific workloads and use cases.

The HPC Ready Architecture for Genomics is built on Dell EMC PowerEdge servers with second generation Intel® Xeon® Scalable processors. This is a boon to HPC applications, especially in the field of genomics, which requires a flexible architecture to accommodate various system requirements. In addition to offering more cores, the solution offers support for Intel® Optane™ memory and storage technology, faster DRAM (DDR4-2933 in 1 DPC configuration), and more DRAM configurations (1TB, 2TB and 4TB).

This document illustrates how the updated architecture with new Intel Xeon processors behave on two different genomics workloads — variant calling and de novo assembly — especially with the new Dell EMC Ready Solution for HPC BeeGFS® Storage.

It begins with an overview of the architecture of the HPC Ready Architecture for Genomics and a description of the building blocks. It then goes on to describe the system configuration, software and application versions. Finally, the performance benchmarks for genomics sequencing data analysis such as variant calling and de novo assembly are presented and analyzed.

Please visit delltechnologies.com/hpc for an overview of Dell Technologies HPC solutions. Detailed reference architectures, performance testing results and guidance are available from the HPC & AI Innovation Lab engineering team at http://www.hpcatdell.com.

Solution overview

The HPC Ready Architecture for Genomics is designed using preconfigured building blocks. This building block architecture allows HPC systems to be optimally designed for specific end-user requirements while still making use of standardized, domain-specific system recommendations.

The available building blocks are infrastructure management, storage, networking and compute. Configuration recommendations are provided for each of the building blocks, which are designed to deliver good performance for typical applications and workloads within the genomics domain. The overall solution is designed to be flexible and scalable. The focus of this solution is for genomics processing using GATK/BWA running on PowerEdge C6420 server nodes, with the added benefit of supporting de novo assembly running on PowerEdge R740xd server nodes. Using a combination of both, the solution is presented here as three sizing options:

- **Small** – 4x compute nodes that are capable of processing 360 genomes/month at 50 with an optional 1x compute node for de novo assembly
- **Medium** – 8x compute nodes that are capable of processing 720 genomes/month at 50 with an optional 1x compute node for de novo assembly
- **Large** – 12x compute nodes that are capable of processing 1080 genomes/month at 50 with an optional 1x compute node for de novo assembly

For the sake of simplicity, the Small, Medium and Large configurations are all designed to fit within one standard 42U rack. The storage has been configured to meet the performance and monthly storage needs of each configuration and users may opt for higher density hard disk drives (HDDs) based on their capacity needs. Additional *variant calling* and *de novo assembly* configurations are possible, and more nodes of each can be added based on your specific needs, but please refer to a Dell Technologies HPC specialist for configuring custom options.

With this flexible building block approach, appropriately-sized HPC clusters can be designed based on individual workloads and requirements. Figure 1 shows three examples of HPC clusters designed using the HPC Ready Architecture for Genomics scalable design architecture.

Clinical genomic studies are generally focused on the identification of genetic *variants* from DNA sequencing data, where variants are defined as single nucleotide variants (SNVs), small insertions and deletions (indels), and structural variants (SVs). The primary computational challenge in DNA sequencing data analysis is identifying and differentiating “true variants” from “noise” for a given sample, a task referred to as *variant calling*. Genomic variant discovery may appear to be a straightforward problem that consists of mapping reads to a reference sequence and at every position, counting the mismatches and construing the genome variant types. However, multiple sources of error in the sequence make this process much more complex than it is at first glance. There are numerous mathematical algorithms for variant calling with various levels of performance, but in general, they tend to be compute-bound.

The Dell EMC PowerEdge C6420 has been designed with such compute-intensive work loads in mind, offering industry-leading performance and compute density and in a very convenient form factor. The PowerEdge C6420 forms the basis of the scalable compute node used for variant calling in this solution.

In addition, the solution includes an optional compute node that is optimized for de novo sequence assemblers, which assemble short nucleotide sequences into longer ones without the use of a reference genome. This assembly technique is becoming increasingly popular in bioinformatic studies to assemble genomes or transcriptomes.

Typical de novo assembly applications such as SOAPdenovo2 and SPAdes are memory-intensive and require server platforms that can support very large DRAM memory configurations for ultra-deep sequencing data analysis, and consequently such systems can be expensive. However, new memory technologies — such as Intel Optane — provide comparable performance to all-RAM configurations at a lower cost.

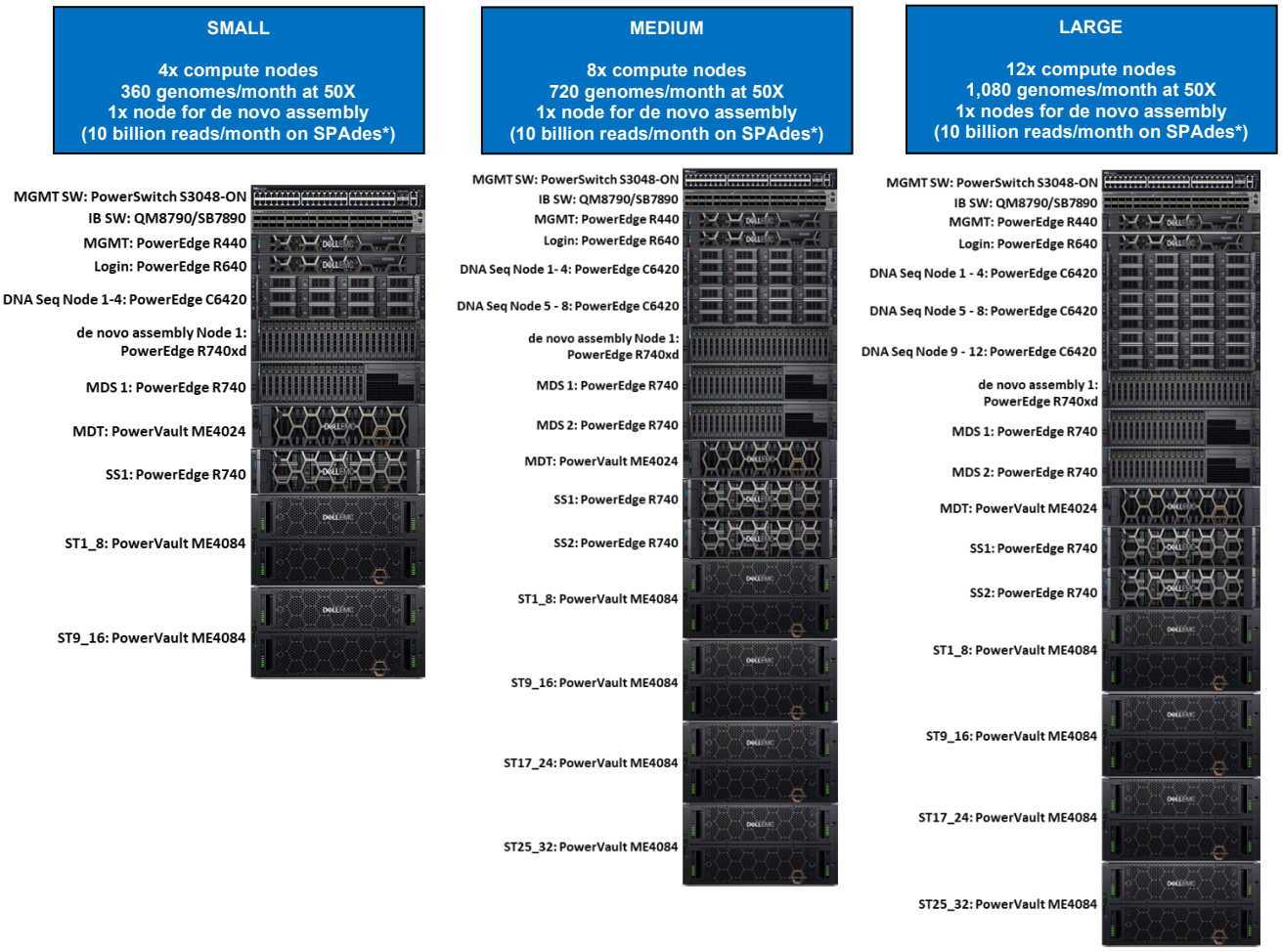
A typical server uses two kinds of devices to process data: RAM to cache data temporarily, and hard disk drives (HDDs) or solid-state drives (SSDs) to store data indefinitely. HDDs and SSDs are relatively inexpensive but slower for the system to access, while RAM is often more expensive but quite fast. In this way, RAM and storage each make up for the other's shortcomings. But, as the amount of data use explodes, organizations are beginning to see the limitations of the traditional storage/RAM pairing.

According to Intel[1], the act of data passing back and forth between storage and RAM negatively impacts latency and bandwidth — consequences that can carry over to customers and end users. Intel Optane DC persistent memory is an innovative memory technology that delivers a unique combination of affordable large capacity and support for data persistence. It acts as a hybrid of the two traditional media, potentially reducing the latency and bandwidth penalties of moving data around separate server components. It is one of the key technologies enabling a more cost-effective solution for HPC problems that demand large memory, such as genomics processing.

The Dell EMC PoweEdge R740xd offers 24 DIMM slots and is an ideal platform for large memory footprint applications like de novo assembly. In this solution, 12 DIMM slots are populated with 32GB DRAM modules, with the remaining 12 slots fitted with 128GB Intel Optane DC 128GB persistent memory module (1,536GB), for a total of 1.92TB.

[1] To learn more about Intel Optane DC persistent memory, visit https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html.

Below are rack diagrams of small, medium and large configurations of the HPC Ready Architecture for Genomics. The solution is comprised of modular, scalable building blocks.



* Based on running SPAdes using the ERR318658 dataset.

Figure 1. *HPC Ready Architecture for Genomics capable of processing 600, 1,200 and 1,080 genomes/month and de novo assembly.*

Infrastructure nodes

Infrastructure server nodes (also called “master” or “login” servers) are used to administer the system and provide user access. They provide services that are critical to the overall HPC system. For small clusters, a single physical server can provide the necessary system management functions. Infrastructure nodes can also be used to provide storage services via network file system (NFS); in which case, they must be configured with additional disk drives or an external storage array.

One infrastructure node is required to deploy and manage the HPC system. If high-availability (HA) management functionality is required, two infrastructure nodes are necessary. Ready Solutions for HPC BeeGFS Storage requires one infrastructure node for administration and management.

A recommended base configuration for infrastructure nodes is:

Dell EMC PowerEdge R440 Server

- CPU: 2x Intel Xeon Gold 6230 at 2.1GHz 20 cores
- Memory: 192GB (12x 16GB at 2666MHz)
- Disk: 10x 1.92TB SSD SATA read intensive 6Gbps 512e 2.5 inch hot-plug S4510 drive in RAID 10
- BIOS system profile: Performance optimized
- Logical processor: Enabled
- Virtualization technology: Disabled
- Operating system (OS): Red Hat® Enterprise Linux® (RHEL) 7.6
- Power supplies: 2x 550W power supply units (PSUs)

The Dell EMC PowerEdge R440 Server is well-suited to be an infrastructure node. Typical HPC clusters will only use a few infrastructure nodes, meaning density is not a priority while manageability is important. The Intel Xeon Gold 6230 processor, with 20 cores per socket, is a basic recommendation for this role. If the infrastructure node will be used for CPU-intensive tasks, such as compiling software or processing data, then a more capable processor may be appropriate. The server’s 192GB of memory provided by 12x 16GB DIMMs provides sufficient memory capacity, with minimal cost per gigabyte, while also providing good memory bandwidth. These servers are not expected to perform much I/O, so a single mixed-use SATA SSD should be enough for the operating system.

Login nodes are optional and one login server per 30 to 100 users is recommended. A typical configuration for a login node is given below.

Dell EMC PowerEdge R640 for login node and CIFS gateway (optional)

- CPU: 2x Intel Xeon Gold 6230 at 2.1GHz 20 cores
- Memory: 192GB (12x 16GB at 2666MHz)
- Disk: 10x 1.92TB SSD SATA read intensive 6Gbps 512e 2.5 inch hot-plug S4510 drive in RAID 10
- BIOS system profile: Performance optimized
- Logical processor: Enabled
- Virtualization technology: Disabled
- OS: RHEL 7.6
- Power supplies: 2x 750W PSU

For most systems, Mellanox® HDR100 or EDR InfiniBand® is likely to be the data interconnect of choice, to provide a high-throughput, low-latency fabric for node-to-node communications or access to a Dell EMC Ready Solution for HPC BeeGFS Storage. Although other parallel file systems like Lustre can be use in place of BeeGFS®, this document only explores the use of BeeGFS. In addition, a high performance NAS storage system like PowerScale can be used for long term persistent data storage.

Compute building blocks

Compute building blocks (CBB) provide the computational resources, in this case for variant calling and de novo assembly respectively. The best configuration for these servers depends on the specific mix of applications and types of computations being performed by each HPC system. In this case, we provide building blocks for both DNA sequencing and de novo assembly.

The modular nature of the HPC Ready Architecture for Genomics allows for constructing systems with CBB for both DNA sequencing and for de novo assembly. Table 2 lists the recommended options for these servers. The configuration can then be designed based on the specific system and requirements of each workload. Relevant criteria to consider when making these selections are discussed in the performance section of this document. The recommended configuration options for the DNA sequencing building block are provided in Table 1.

Servers	• Dell EMC PowerEdge C6400 enclosure with 4x Dell EMC PowerEdge C6420 Servers
Processors	Choice of: • Dual Intel Xeon Gold 6242 @ 2.8GHz (16 cores) • Dual Intel Xeon Gold 6248 @ 2.5GHz (20 cores) • Dual Intel Xeon Gold 6252 @ 2.1GHz (24 cores)
Memory Options	Choice of: • 192GB (24x 8GB 2933 MT/s DIMMs) • 384GB (24x 16GB 2933 MT/s DIMMs) • 768GB (24x 32GB 2933 MT/s DIMMs)
Storage Options	• PERC H330, H730P or H740P RAID controller • 2x 750GB Intel Optane DC P4800X and Intel Memory Drive Technology (IMDT) • 4x 480GB 12Gbps mixed-use SAS SSDs
iDRAC	• iDRAC9 Express (for the servers)
Power Supplies	• 2x 2000W PSU (for the enclosure)
Networking	• Mellanox® ConnectX®-6 HDR100 or Mellanox ConnectX®-5 EDR InfiniBand adapter

Table 1. *Recommended configuration for the DNA sequencing building block*

The recommended configuration options for the de novo assembly are provided in Table 2.

Platforms	• Dell EMC PowerEdge R740xd Servers
Processors	• Dual Intel Xeon Gold 6248R @ 3.0GHz (24 cores)
Memory Options	• 12x 16GB 2666 MT/s DIMMs (384GB) • 12x Intel Optane DC 128GB persistent memory module (1,536GB)
Storage Options	• PERC H330, H730P or H740P RAID controller • 20x 480GB mixed use SAS SSDs
iDRAC	• iDRAC9 Express
Power Supplies	• 2x 2000W PSU
Networking	• Mellanox ConnectX®-6 HDR100 or Mellanox ConnectX®-5 EDR InfiniBand adapter

Table 2. *Recommended configurations for the de novo assembly building block*

Storage components

Dell Technologies offers a wide range of HPC storage solutions. For a general overview of the entire HPC solution portfolio, please visit delltechnologies.com/hpc, Ready Solutions for HPC Storage. There are typically three tiers of storage for HPC which differ in terms of size, performance and persistence. These are: scratch storage, operational storage and archival storage.

Scratch storage tends to persist for the duration of a single simulation. It may be used to hold temporary data which is unable to reside in the compute system's main memory due to insufficient physical memory capacity. HPC applications may be considered "I/O bound" if access to storage impedes the progress of the simulation. For these HPC workloads, the most typical and cost-effective solution is to provide enough direct-attached local storage in the compute nodes. For situations where the application may require a shared file system across the compute cluster, a high-performance shared file system may be better suited than relying on local direct-attached storage. Typically, using direct-attached local storage offers overall price/performance and is considered best practice for most genomics workloads. For this reason, local storage is included in the recommended configurations with appropriate performance and capacity for a wide range of production workloads. If anticipated workload requirements exceed the performance and capacity provided by the recommended local storage configurations, care should be taken to size scratch storage appropriately based on the workload.

Operational storage is typically defined as storage used to maintain results and other data — such as home directories — over the duration of a project, such that the data may be accessed daily for an extended period. Typically, this data consists of simulation input and results files, which may be transferred from the scratch storage, typically in a sequential manner, or from users analyzing the data, often remotely. Since this data may persist for an extended period, some or all of it may be backed up at a regular interval, where the interval chosen is based on the balance of the cost to either archive the data or regenerate it if need be. Archival data is assumed to be persistent for a very long term, and data integrity is considered critical. For many modest HPC systems, use of the existing enterprise archival data storage may make the most sense, as the performance aspect of archival data tends not to impede HPC activities.

The rack system layouts in Figure 1 consist only of the scratch storage needed during computation. Please visit delltechnologies.com/hpc for references to operational storage that are compatible with our solutions. A variety of options are available based on user needs for capacity and performance. Smaller storage capacities may be satisfied with the Dell EMC Ready Solution for HPC NFS Storage. Larger enterprise needs can be met with the Dell EMC PowerScale product family.

Dell EMC Ready Solution for HPC BeeGFS Storage
The Dell EMC Ready Solution for HPC BeeGFS Storage is a fully supported parallel file system designed with the needs of high-throughput computing applications in mind. This high capacity solution is designed with high availability tenets so there is no single point of failure in the cluster. In this solution, if a node becomes unavailable or inoperative, the services fail over from that node to another without interruption to cluster clients, and the faulty node can be removed from the cluster to prevent data corruption. The Dell EMC Ready Solution for HPC BeeGFS Storage uses a flexible approach where individual building blocks can be combined with offer various configurations to meet specific workloads and use cases and to provide for future expansion.

BeeGFS is an open-source parallel cluster file system software that can be downloaded from www.beegfs.io. The software also includes enterprise features such as high availability (HA), quota enforcement and access control lists. BeeGFS distributes user data across multiple storage nodes. There is parallelism of access as it maps data across many servers and drives and provides a global namespace, a directory tree, that all nodes can see. It is easy to deploy BeeGFS and integrate it with existing systems. The BeeGFS server components are user space daemons. The client is a native kernel module that does not require any patches to the kernel itself.

BeeGFS components can be installed and updated without even rebooting the machine; therefore, clients and servers can be added to existing systems without downtime. BeeGFS is a highly scalable file system. By increasing the number of servers and drives, the performance and capacity can be increased to the required level from small clusters up to enterprise-class systems with thousands of nodes. As BeeGFS is software-defined storage that decouples the storage software from its hardware, it offers flexibility and choice. In BeeGFS, the roles and hardware are not tightly integrated. The BeeGFS clients and servers can even run on the same machine. BeeGFS supports a wide range of Linux distributions, Red Hat® Enterprise Linux® (RHEL), CentOS® and SUSE®. BeeGFS is designed to be independent of the local file system used. The local storage can be formatted with any of the standard Linux file systems — xfs or ext4.

The BeeGFS architecture consists of four main services:

- Management service: Each BeeGFS file system or namespace has only one management service. The management service must be set up first because all other services must register with the management service.
- Metadata service: This is a scale-out service, which means that there can be many metadata services in a BeeGFS file system. However, each metadata service has exactly one metadata target to store metadata. On the metadata target, BeeGFS creates one metadata file per user created file. BeeGFS metadata is distributed on a per-directory basis. The metadata service provides the data striping information to the clients and is not involved in the data access between file open/close.
- Storage service: Stores user data. A BeeGFS file system can be made of multiple storage servers where each storage service can manage multiple storage targets. On those targets, the striped user data is stored in chunk files for parallel access from the client.
- Client service: A kernel module.

The management, metadata and storage services are user space processes. Figure 2 illustrates the general architecture of the BeeGFS file system.

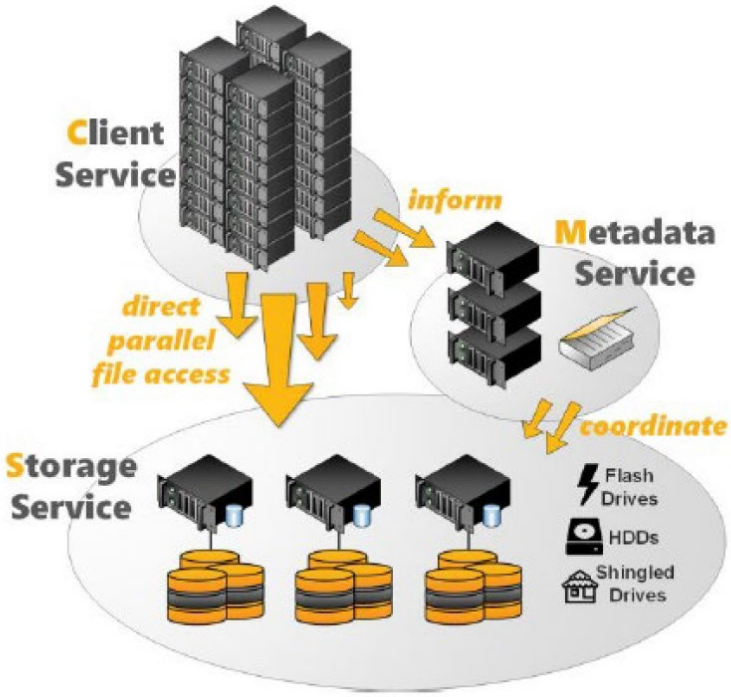


Figure 2. *BeeGFS architecture overview (source: www.beegfs.io)*

Ready Solutions for HPC BeeGFS Storage architecture

The Dell EMC Ready Solution for HPC BeeGFS Storage consists of a management server, a pair of metadata servers, a pair of storage servers and the associated storage arrays. The solution provides storage that uses a single namespace that is easily accessed by the cluster's compute nodes. The following figure shows the solution reference architecture with these primary components:

- Management server
- Metadata server (MDS) pair with Dell EMC PowerVault ME4024 as back-end storage
- Storage server pair with Dell EMC PowerVault ME4084 as back-end storage

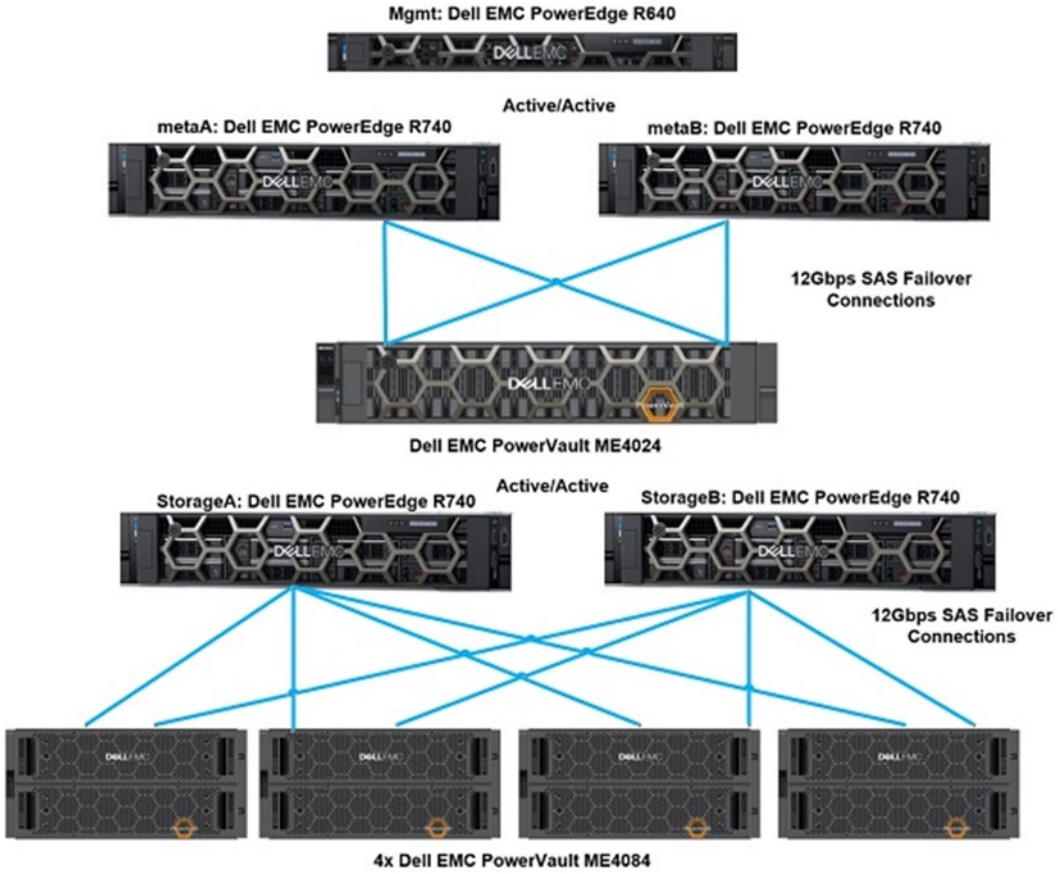


Figure 3. *Ready Solution for HPC BeeGFS Storage*

In Figure 3, the management server running the BeeGFS monitoring daemon is a Dell EMC PowerEdge R640. The two metadata servers (MDS) are PowerEdge R740 servers in an active-active high availability (HA) configuration. The MDS pair is connected to the 2U, PowerVault ME4024 array by 12Gb/s SAS links. The PowerVault ME4024 storage array hosts the MetaData Targets (MDTs). Another pair of PowerEdge R740 servers, also in active-active HA configuration, are used as storage servers (SS). This SS pair is connected to four fully populated PowerVault ME4084 storage arrays using 12Gb/s SAS links. The ME4084 arrays support a choice of 4TB, 8TB, 10TB or 12TB NL SAS 7.2K RPM HDDs and host the storage targets (STs) for the BeeGFS file system. This solution uses Mellanox InfiniBand HDR100 for the data network. The clients and servers are connected to the 1U Mellanox Quantum HDR Edge Switch QM8790, which supports up to 80 ports of HDR100 by using HDR splitter cables.

The recommended hardware and software configuration options for the Dell Technologies Ready Solution for HPC BeeGFS Storage as used with the RA for Genomics are given in Table 3.

Management server	• 1x Dell EMC PowerEdge R640
MDS	• 2x Dell EMC PowerEdge R740
Storage servers	• 2x Dell EMC PowerEdge R740
Processors	• Management server: Dual Intel Xeon Gold 5218 • MDS and SS servers: Dual Intel Xeon Gold 6230
Memory	• Management server: 12x 8 GB 2666 MT/s DDR4 RDIMMs • MDS and SS servers: 12x 32 GB 2933 MT/s DDR4 RDIMMs
Local disks and RAID controller	• Management server: PERC H740P Integrated RAID, 8GB NV cache, 6x 300GB 15K SAS hard drives (HDDs) configured in RAID10 • MDS and SS servers: PERC H330+ Integrated RAID, 2x 300GB 15K SAS HDDs configured in RAID1 for OS
InfiniBand HCA	• Mellanox ConnectX-6 HDR100 InfiniBand adapter
External storage controllers	• On each MDS: 2x Dell 12 Gb/s SAS HBAs • On each SS: 4x Dell 12 Gb/s SAS HBAs
Object storage enclosures	• 4x Dell EMC PowerVault ME4084 fully populated with a total of 336 drives
Metadata storage enclosure	• 1x Dell EMC PowerVault ME4024 with 24 SSDs
RAID controllers	• Duplex RAID controllers in the ME4084 and ME4024 enclosures
HDDs	• On each ME4084 Enclosure: 84 x 8 TB 3.5 in. 7.2 K RPM NL SAS3 • ME4024 Enclosure: 24 x 960 GB SAS3 SSDs
Operating system	• CentOS Linux release 8.1.1911 (Core)
Kernel version	• 4.18.0-147.5.1.el8_1.x86_64
Mellanox OFED version	• 4.7-3.2.9.0
BeeGFS file system version	• 7.2 (beta2)

Table 3. *Recommended hardware configuration options for the Ready Solution for HPC BeeGFS Storage*

System networks

Most HPC systems are configured with two networks: an administration network and a high-speed/low-latency switched fabric. The administration network is typically Gigabit Ethernet that connects to the onboard LOM/NDC of every server in the cluster. This network is used for provisioning, management and administration. On the CBB servers, this network will also be used for intelligent platform management interface (IPMI) hardware management. For infrastructure and storage servers, the iDRAC Enterprise ports may be connected to this network for out-of-band (OOB) server management. The management network typically uses the Dell EMC PowerSwitch S3048-ON Ethernet switch. If there is more than one switch in the system, they should be stacked with 10Gb Ethernet cables.

Mellanox QM8790 (HDR) or Mellanox SB7890 (EDR) InfiniBand Switch
A high-speed/low-latency fabric is recommended for clusters with more than four servers. The current recommendation is for HDR100 or EDR InfiniBand fabric. The fabric will typically be assembled using Mellanox QM8790 40-port HDR100 or SB7890 36-port EDR InfiniBand switches. This 40-port/36-port non-blocking InfiniBand HDR100/EDR 100Gb/s switch system provides the highest performing fabric solution in a 1U form factor by delivering up to 7.2Tb/s of non-blocking bandwidth with 90 nanosecond port-to-port latency.

Dell EMC PowerSwitch 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 infrastructure nodes and for reporting data from client to the infrastructure node. Dell EMC PowerSwitch S3048-ON is recommended for the management network.

Bright Cluster Manager
Bright Computing® provides comprehensive software solutions for deploying and managing HPC clusters, big data clusters and OpenStack in the data center and in the cloud. Bright Cluster Manager® is the recommended cluster management software to install and monitor the HPC system. 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 user interface monitors every single node and reports detection of any software or hardware events.

Services and support
The HPC Ready Architecture for Genomics is available with optional services and support.

Software components
In addition to the hardware and cluster management software, BioBuilds was installed on this system. BioBuilds is a well maintained, versioned and continuously growing collection of open-source bio-informatics tools from L7 Informatics. They are built and optimized for a variety of platforms and environments. BioBuilds solves most software challenges faced in the life sciences domain.

Performance evaluation and analysis

Variant calling 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
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, a new workflow with three phases was introduced:

- 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 analysis

Here we tested phase 1, phase 2A and phase 3 for a germline variant calling pipeline. The details of commands in the benchmark are in APPENDIX A. Genome Reference Consortium Human build 37 (GRCh37) and were used as a reference genome sequence. The 50X whole human genome sequencing data from the Illumina platinum genomes project named ERR194161_1.fastq.gz and ERR194161_2.fastq.gz were used for a baseline test.

While it's ideal to use non-identical sequence data for each run, it is extremely difficult to collect non-identical sequence data having more than 50X 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 time, the run time would be slightly longer in real cases. Despite these built-in inaccuracies, this variant analysis performance test can provide valuable insights when estimating the level of resources required for an identical or similar analysis pipeline with a defined workload.

Total run time is the elapsed wall-clock 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 4.

The test data was chosen from one of Illumina's Platinum Genomes. ERR194161 was processed with Illumina HiSeq 2000 submitted by Illumina and can be obtained from EMBL-EBI. The DNA identifier for this individual is NA12878. The description of the data from the linked website shows that this sample has a >30x depth of coverage, and it reaches to ~53x.

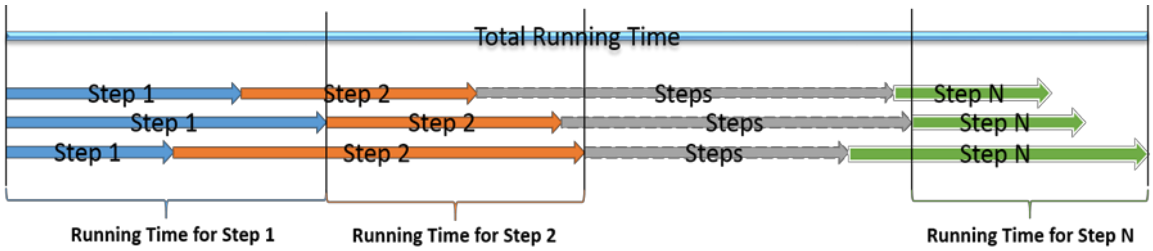


***Figure 4**. Running time measurement method*

Multiple sample/multiple nodes performance

A typical way of running NGS pipeline is to process multiple samples on a compute node and use multiple compute nodes to maximize the throughput. The number of compute nodes used for the tests was eight C6420 compute nodes, and the number of samples per node was seven samples. Hence, up-to 56 samples are processed concurrently to estimate the maximum number of genomes per day without a job failure. Eight Dell EMC PowerEdge C6420 nodes were used for the tests with seven samples per node. Seven cores and 30GB of memory are allocated for each sample. Up to 320 samples were processed concurrently to estimate the maximum number of genomes per day without a job failure. More information on BWA-GATK pipeline, used on this test, can be obtained from the Broad Institute website.

As shown in Figure 1, single C6420 compute node can process 3.69 of 50x whole human genomes per day when seven samples are processed together. For each sample, five cores and 20 GB memory are allocated.

A single PowerEdge C6420 server as configured in Table 4 can process 3.69 50X whole human genomes per day when seven samples are processed concurrently as shown in Figure 5. For each sample, 5 cores and 20

GB memory are allocated. Fifty-Six 50X whole human genomes can be processed with 8x Dell EMC PowerEdge C6420 compute nodes in ~54 hours. In other words, the performance of the test configuration (8-nodes) summarizes as 25.11 genomes per day for whole human genome with 50x depth of coverage.

As the data size of WGS has been growing constantly. The current average size of WGS is about 55x. This is 5x larger than a typical WGS 4 years ago when we started to benchmark BWA-GATK pipeline. The increasing data size does not strain storage side capacity since most applications in the pipeline are also bounded by CPU clock speed. Hence, the pipeline runs longer with larger data size rather than generating heavier IOs.

However, more temporary files are generated during the process due to the larger data needs to be parallelized, and this increased number of temporary files opened at the same time exhausts the open file limit in a Linux operating system. One of the applications silently fails to complete by hitting the limit of the number of open files. A simple solution is to increase the limit to >150K.

The results in Figure 5 shows that the throughput tests did not hit the maximum capacity of the system. Since there was not any sign of significant slowdown by adding more samples, it must be possible to process more than seven samples if compute nodes are setup with larger memory.

CPU	• 2x Xeon® Gold 6248 20 cores 2.5 GHz (Cascade Lake)
RAM	• 12x 16GB at 2933 MTps
OS	• Red Hat Enterprise Linux Server release 7.4 (Maipo)
Interconnect	• Mellanox EDR InfiniBand
BIOS System Profile	• Performance Optimized
Logical Processor	• Disabled
Virtualization Technology	• Disabled
BWA	• 0.7.15-r1140
Sambamba	• 0.7.0
Samtools	• 1.6
GATK	• 3.6-0-g89b7209

Table 4. *Configuration of Dell EMC PowerEdge C6420 as used in the BWA-GATK benchmarks*

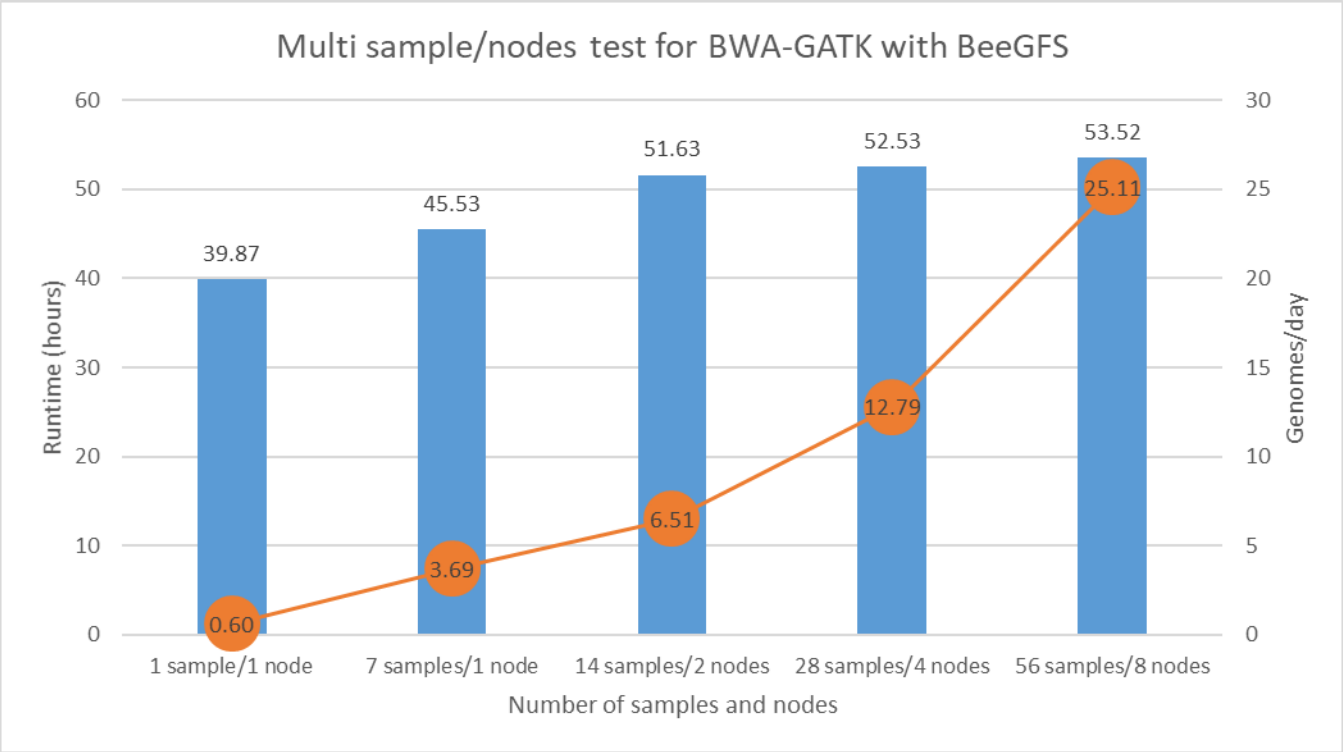


Figure 5. *Throughput tests with up to 8 Dell EMC PowerEdge C6420 Servers with BeeGFS*

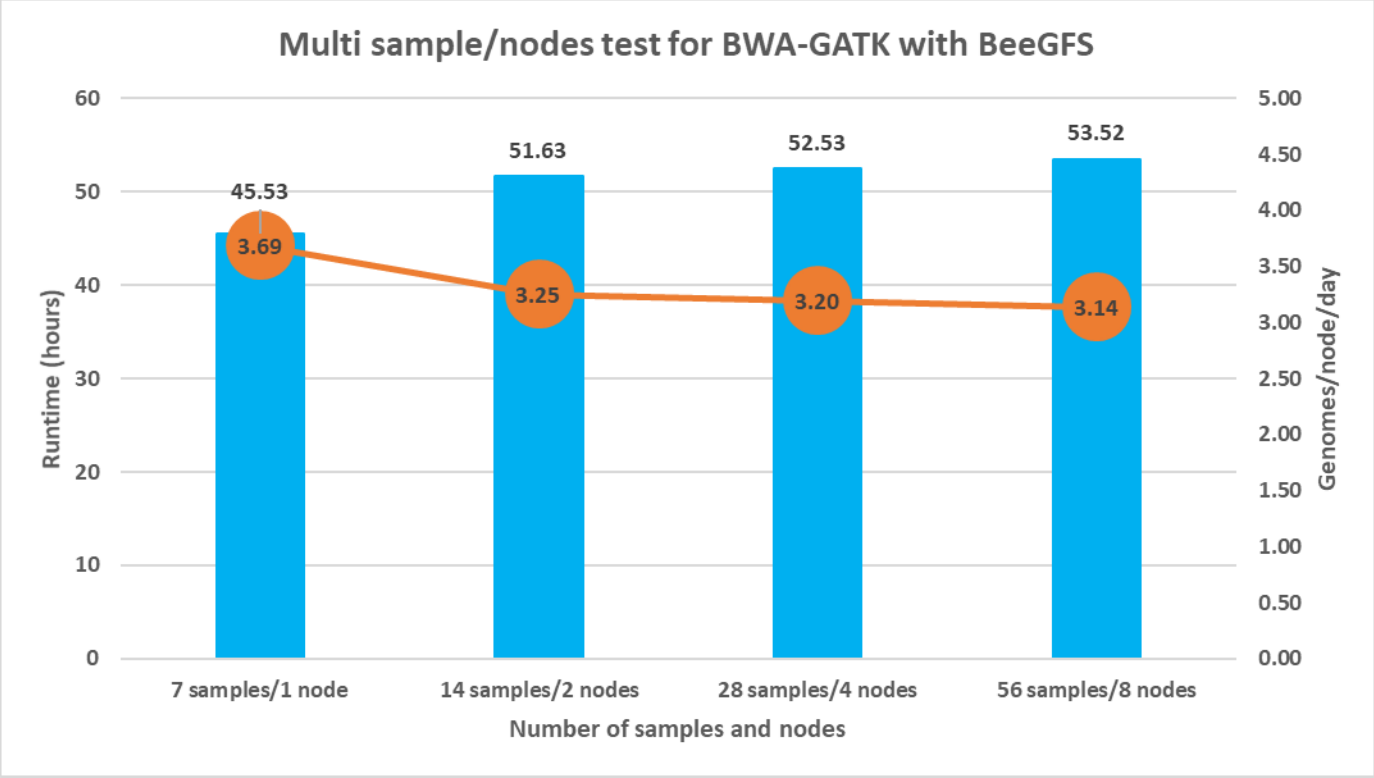


Figure 5a. *Throughput tests with up to 8 Dell EMC PowerEdge C6420 Servers with BeeGFS. Performance results have been restated in genomes per node per day*

SPAdes assembler test

SPAdes is a relatively new application and reported for some improvement on the Euler-Velvet-SC assembler (2011) and SOAPdenovo. SPAdes is also based on the de Bruijn graph algorithm like most of the assemblers targeting NGS data. De Bruijin graph-based assemblers would be more appropriate for larger datasets having short reads in the hundred-millions.

As shown in Figure 6, greedy-extension and overlap-layout-consensus (OLC) approaches were used in the very early next-gen assemblers. Greedy-extension's heuristic is that the highest scoring alignment takes on another read with the highest score. However, this approach is vulnerable to imperfect overlaps and multiple matches among the reads and leads to an incomplete or arrested assembly. The OLC approach works better for long reads such as Sanger or other technology generating more than 100bp due to minimum overlap threshold (454, Ion Torrent, PacBio and so on). De Bruijin graph-based assemblers are more suitable for short read sequencing technologies such as Illumina. The approach breaks the sequencing reads into successive k-mers, and the graph maps the k-mers. Each k-mer forms a node, and edges are drawn between each k-mer in a read.

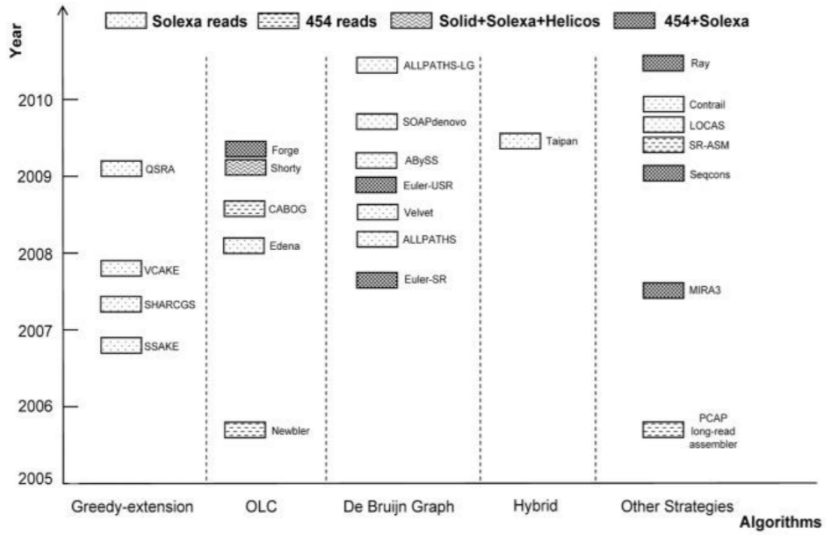


Figure 6: *Overview of de novo short reads assemblers*[2]

SPAdes is based on de Bruijn graph for both single-cell and multicell data. It improves on the recently released Euler Velvet Single Cell (Euler + Velvet-SC) assembler (specialized for single-cell data) and on popular assemblers, Velvet and SoapDeNovo (for multicell data).

The data used for the tests is a paired-end read, ERR318658, which can be downloaded from the European Nucleotide Archive (ENA). The read generated from blood sample as a control identifies somatic alterations in the primary and metastatic colorectal tumors. This data contains 3.2 billion reads with the read length of 101 nucleotides.

[2] Source: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056720/*

In the benchmark comparison presented here, SPAdes runs three sets of de Bruijn graphs with 21-mer, 33-mer and 55-mer, consecutively. Hyperthreading is enabled since IMDT is optimized for use with hyperthreading turned on. The number of cores tested here – 28, 46 and 92 – are shown in Figure 7.

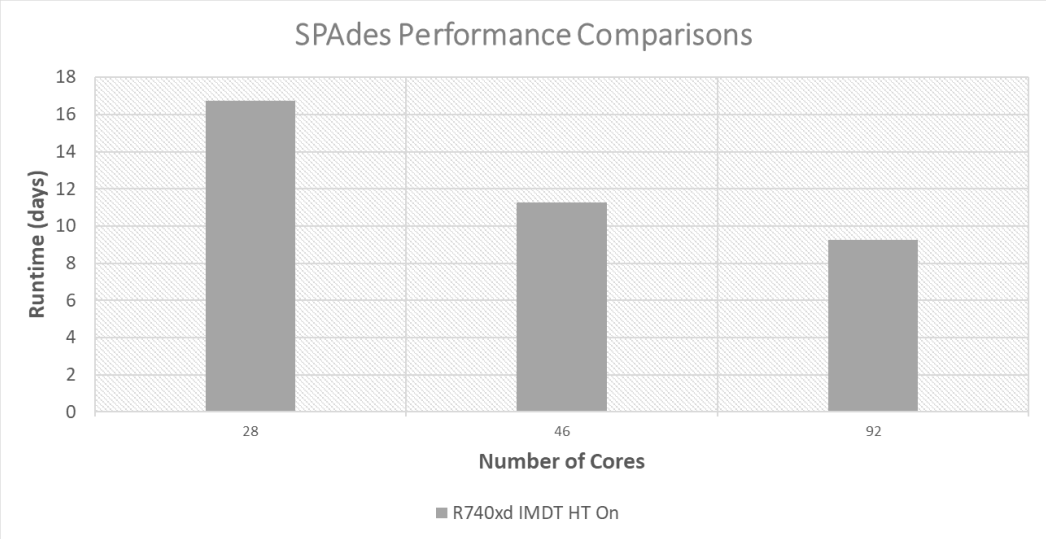


Figure 7. *Runtime in days for SPAdes tests using ERR318658 dataset on R740xd with IMDT*

Based on performance runs on a Dell EMC PowerEdge R740xd with IMDT, we can achieve around 10 billion reads per month using the ERR318658 dataset.

Conclusion

The Dell EMC HPC Ready Architecture for Genomics uses a building block approach with deployment of an HPC system optimized for specific computation requirements. The design addresses computation, storage, networking and software requirements and provides a solution that is easy to install, configure and manage with installation services and support readily available. The performance benchmarking bears out the solution design, demonstrating system performance with genomics software.