DELL Technologies

Ready Architecture

HPC Ready Architecture for Genomics with NVIDIA Clara Parabricks

Accelerating genomic data analysis with NVIDIA GPUs

Abstract

The HPC Ready Architecture for Genomics with NVIDIA Clara Parabricks is a modular, scale-out solution composed of NVIDIA Parabricks application software, a Dell EMC PowerEdge DSS 8440 Server with 16 NVIDIA T4 GPUs, and Dell EMC Isilon F800 network-attached storage. This document provides detail of the solution architecture and performance testing results for next generation sequencing (NGS).

November 2020

Revisions

Date	Description
November 2020	Initial Release

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email or provide your comments by completing our documentation survey.

Authors: Adnan Khaleel, Kihoon Yoon

Table of contents

Executive summary

Accelerated genomics analysis with NVIDIA Clara Parabricks

The HPC Ready Architecture for Genomics is a tested/validated solution that leverages Dell EMC PowerEdge servers with NVIDIA® T4 GPUs, NVIDIA Clara™ Parabricks® software, Dell EMC PowerSwitch networking, and Dell EMC Isilon storage. It combines IT resources required for various forms of genomic data analysis in a compact, scalable solution.

The Dell EMC Ready Architecture for Genomics with NVIDIA® Clara™ Parabricks uses a flexible and modular approach to High Performance Computing (HPC) system design that leverages individual building blocks. These integrated, tested and tuned solutions include the resources required for next-generation sequencing (NGS) secondary analysis while providing an optimal balance of compute density, energy efficiency and performance.

This solution is capable of processing 25 50X human genomes per day using all 16 GPUs. However, the design of the Dell EMC DSS 8440 server allows multiple secondary analyses in parallel. More importantly, achieving this daily output using NVIDIA T4 GPUs is ~5x less expensive than using a design that incorporates other GPUs.[1]

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

Intended audience

Researchers and scientists who are responsible for the analysis of NGS data and IT professionals who are responsible for providing a technical computing environment designed to support NGS applications are encouraged to read this paper.

Acknowledgments

Thank you Parabricks Inc. for providing the Parabricks application suite. We thank “Shop” Mallick, the David Reich Laboratory, Harvard Medical School, and the Simons Foundation for providing access to source data generated by the Simons Diversity Genome Project. We also would like to thank Glen Otero, VP of Scientific Computing, Translational Genomics Research Institute (Tgen), and Kihoon Yoon, Principal Engineer, Dell Technologies HPC & AI Innovation Lab for input and consultation.

[1] Based on Google Shopping cost of NVIDIA T4 GPU 16GB $2580 vs. NVIDIA V100 GPU 32GB $13,061, conducted December 29, 2020.

Introduction

DNA genomics analysis

DNA is the code of life. This molecule carries the genetic instructions for growth, development and reproduction of all living organisms. The building blocks of DNA are four chemical bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The order in which the bases occur is responsible for traits like eye color or drug sensitivity. DNA sequencing is the process of writing out the order of the bases for an organism of interest. The entire complement of DNA for an organism is a genome. After approximately ten years and over $2.7B USD, the first draft of the human genome sequence was published in April 2003 (NHGRI, 2019).

Next-generation sequencing (NGS) automates the rapid sequencing of DNA and can produce a human genome in approximately 24 hours. Consequently, NGS now plays an increasingly important role in clinical practice and public health. The information encoded in a person's genome is instrumental in assessing the response to diagnosis, treatment, and disease prevention strategies due to person-to-person variability (Suwinski, 2019). Identifying variants or differences for a genome is done by comparing an individual's genome to a DNA reference sequence. Also known as secondary analysis, this process for generating a list of variants can take minutes to days depending on the available software, computing and storage resources.

Keeping pace with NGS data generation while reducing secondary analysis time

Extending this approach to assess the genetic variability of patient populations requires operating the latest NGS instrumentation and computing resources at scale. For example, the latest Illumina® NovaSeq™ 6000 system can output approximately 5x more DNA bases than the previous generation of instrumentation (Illumina Inc., 2019). One Illumina NovaSeq system can produce between ~1.5 and 2.5TB of raw data per day, representing approximately 20 to 48 whole genome sequences per day.

Today, it is not uncommon for life science organizations to operate more than one NGS instrument and routinely process from 200 to over 1,000 samples per week. Ideally, an organization has enough computing and storage resources matched to the output capacity for a fleet of sequencing instruments such that the rate of secondary analysis keeps pace with the rate of raw NGS data generation. Otherwise, the organization risks experiencing an analysis backlog.

Working with NGS data

The desired product of whole-genome sequencing (WGS) is a list of variants or differences for a given sample when compared to a reference genome. Although motivations may be different, minimizing the time to generate this list of variants is a common goal shared by many healthcare and life science organizations. Research organizations competing for grants want to move into variant interpretation and analysis as soon as possible while avoiding costly false positives.

To recognize revenue, a DNA sequencing provider must return a list of variants to its customer per agreed-on timelines. While in a clinical setting, a diagnostic variant report is needed at a speed that impacts the care of a patient. To better understand how software, computing and storage technology choices impact the time to generate a variant list, it is worthwhile to review the three analysis phases of NGS data (Figure 2).

Note: There are many NGS applications. The patterns for working with NGS data are common across applications. For simplicity, this document focuses on WGS.

Primary analysis

The first step for processing NGS data is called primary analysis. This step is specific to the sequencing instrument and generates multiple FASTQ files containing sequencing reads. In the next step, known as secondary analysis, the FASTQ sequencing reads are mapped to a reference genome or a reference transcriptome. Additional processing identifies variants, or differences, between the sample of interest and a reference. The variants are annotated and interpreted in subsequent downstream steps. The secondary analysis time for a single sample ranges from hours to days, depending on data size, available computing resources, software and analytical workflow.

The most common output file format for this data as they arrive from the sequencer is FASTQ. The FASTQ format is ASCII text data containing the short sequences of DNA bases and associated quality score for each base. This short sequence data is unordered and unaligned and commonly referred to as “reads.” Depending on the type of sequencing instrument, instrument settings and application, FASTQ files can range from a small number of large (>120GB) files to an extremely large number of smaller files.

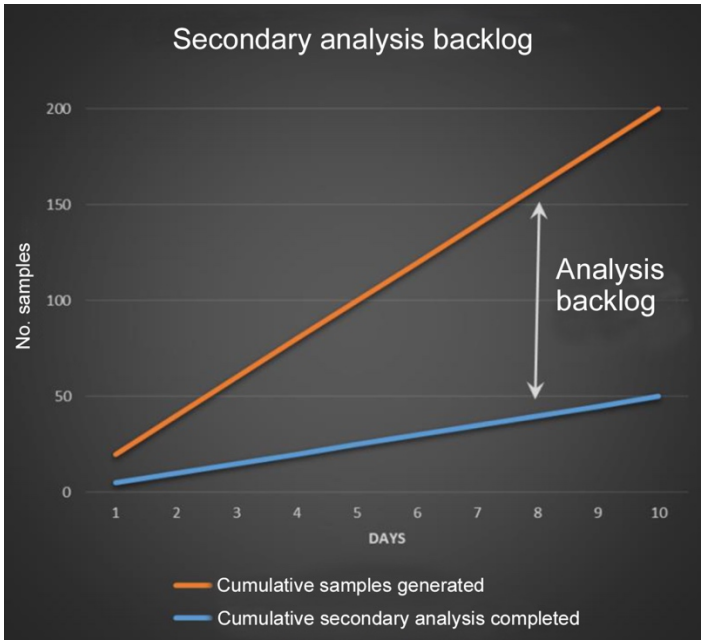

Figure 1: *Keeping pace with data generation*

Secondary analysis

During a secondary analysis, the raw reads contained in one or more FASTQ files are mapped and aligned to a reference genome. A binary alignment map (BAM) file is the output and represents the genome for the sample of interest. The genome of interest (BAM file) is passed to a variant calling step that identifies the significant variants between the genome of interest and a reference. The identified differences are written to a variant call file (VCF).

Secondary analysis is a computing and storage-intensive process, especially when processing hundreds to thousands of genomes. Many strategies exist to avoid secondary analysis bottlenecks. Until recently, the adoption of hardware acceleration using server accelerators remained low due to customized software requirements. Parabricks’ genomics software, which was acquired by NVIDIA in 2019, has pioneered a software stack performing various genomic analysis workflows with GPUs. We tested Parabricks software on a Dell EMC PowerEdge C4140 server with 4x NVIDIA V100 GPUs previously. We then tested a Dell EMC DSS 8440 with 16x NVIDIA T4 GPUs for accelerating secondary analysis while offering an attractive balance between price and performance. This document shares a new reference architecture and benchmark results for NVIDIA Clara Parabricks secondary analysis.

Note: Typically, short read segments range from 75 to 250+ bases depending on NGS application

---

Tertiary analysis

A tertiary analysis focuses on interpreting the variants for a given sample or a population of samples to understand their significance in the context of additional biological and clinical information.

Note: Tertiary analysis is beyond the scope of this paper.

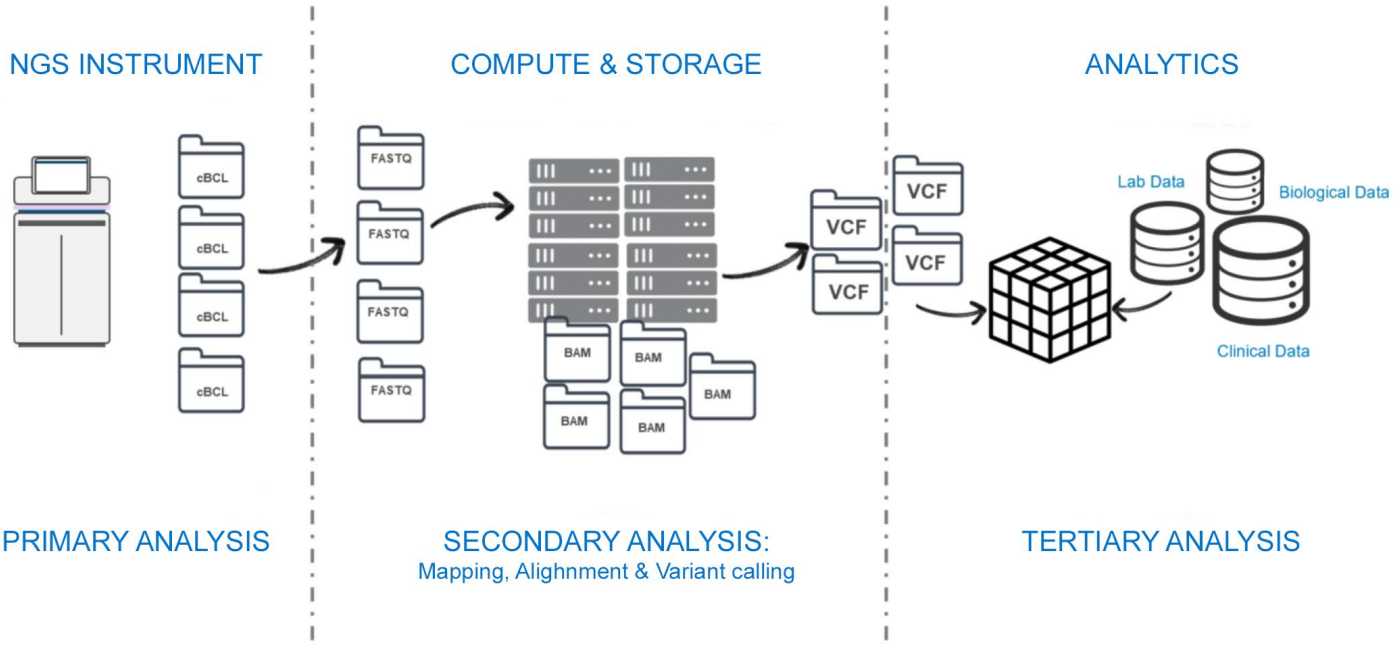

Figure 2: *Three phases of NGS analysis*

Reducing secondary analysis time to keep pace with NGS data generation

Due to the size of individual sample data and volume of samples, WGS secondary analysis is a compute and storage-intensive process. The most commonly used and cited methods for secondary analysis include the Burrows-Wheeler Alignment (BWA-Mem) (Li, 2009), and the Genome Analysis Tool Kit (GATK) (McKenna, 2010). Using the Broad GATK best practices workflow (pipeline) requires over 30 hours to process a 30X WGS (Goyal, 2017). Analyzing a few genomes per day is far from the ideal when a modern, high-throughput NGS instrument can generate unanalyzed, raw NGS data for 20 or more WGS per day.

It is important to consider critical variables that may impact the total secondary analysis (wall-clock) time when choosing technologies that enable secondary analysis of NGS data. These variables range from the type of NGS sequencing application, analysis software and strategies, output file types, application file access patterns, and number and type of available computing resources.

Sequence depth coverage

When planning time and resources to complete secondary analysis, it is essential to be aware of the sequencing depth of coverage for sample data as it will impact analysis time per sample. Coverage describes the average number of reads that align to, or cover, a known reference sequence. The coverage often determines if a variant exists with a certain degree of confidence at a specific genomic location. Coverage requirements vary by sequencing application. For example, 30X to 50X coverage is common for human WGS applications (Illumina, 2019). However, the analysis of cancer genomes may require sequencing to a depth of coverage higher than 100X to achieve the necessary sensitivity and specificity to detect rare variants (Griffith, 2015).

Coverage is also a measure of the amount of data per sample. As coverage increases, so does the amount of data per sample. For example, a 30X (coverage) WGS sample contains approximately 3x more data than a 10X WGS sample, which means secondary analysis time also increases.

Interplay between analysis and computing resources

Given many degrees of freedom between software and computing choices, it can be one of the most challenging and time-consuming tasks in minimizing secondary analysis time. Organizations with access to resources with deep computer science expertise may implement system-level optimizations achieving a 70% reduction in execution time (Kathiresan, 2017). Alternatively, it can be as simple as updating existing server technology yielding a 12% increase in daily output (Yoon, 2018).

To avoid limitations of hardware scalability, modern server accelerators such as GPUs, in combination with purpose-built software, can lead to significant reductions in secondary analysis times. For example, using PowerEdge C4140 servers with NVIDIA V100 GPUs, Dell Technologies engineers demonstrated, in collaboration with Parabricks, over 25X reduction in analysis time compared to a CPU-only solution (Dell Technologies, 2018).

Data placement

Like the interplay between software and computing resources, data storage solutions and their related file systems also offer opportunities to accelerate secondary analysis. It is worthwhile to inspect and understand the general file access patterns for the methods used in the secondary analysis. Some analysis applications used in secondary analysis, especially those like BWA used for sorting and alignment, can create many temporary files.

As a best practice, these temporary files should be placed on direct-attached storage (DAS) when feasible, instead of any network file storage. However, mounting a temporary or scratch directory on a shared storage resource is an acceptable approach, if it introduces opportunities for eliminating manual, time-consuming steps to stage large data sets next to computing resources. This approach also offers opportunities to minimize or prevent data loss.

Storage media types

Implementing secondary analysis workflows using shared storage resources often prompts groups to purchase more expensive flash storage with the anticipation that it will significantly reduce analysis time. However, the benefits from flash storage are highly dependent on the software application, available compute host memory, data set size, and application IOPS requirements. Only 50% of commonly used bioinformatics tools demonstrated 2X or more speed up from flash or solid-state drive (SSD) (Lee, 2016). Relative to other technical constraints using lower-cost hard disk drives (HDD) is an acceptable approach.

Simplifying choices

To simplify and streamline technology choices that lead to significantly reduced secondary analysis times while keeping pace with NGS data generation, Dell Technologies and Parabricks set out to identify a modular, easy-to-scale reference architecture.

A modular scale-out GPU solution architecture for NGS analysis

System architecture

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 infrastructure building blocks are infrastructure management, compute, networking and storage. 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.

Figure 3 illustrates the technical computing system used to evaluate the acceleration of NGS genomics analysis. It is comprised of a Dell EMC DSS8440 server populated with 16x NVIDIA T4 GPUs, a Dell EMC Isilon F800 scale-out NAS storage cluster, Dell EMC PowerSwitch networking, and NVIDIA Clara Parabricks software. Note that in a customer deployment, the number and type of server and storage nodes will vary and can be scaled independently to meet the requirements specific to an organization. The last section of this document will discuss a starting configuration that can be scaled-out as requirements change.

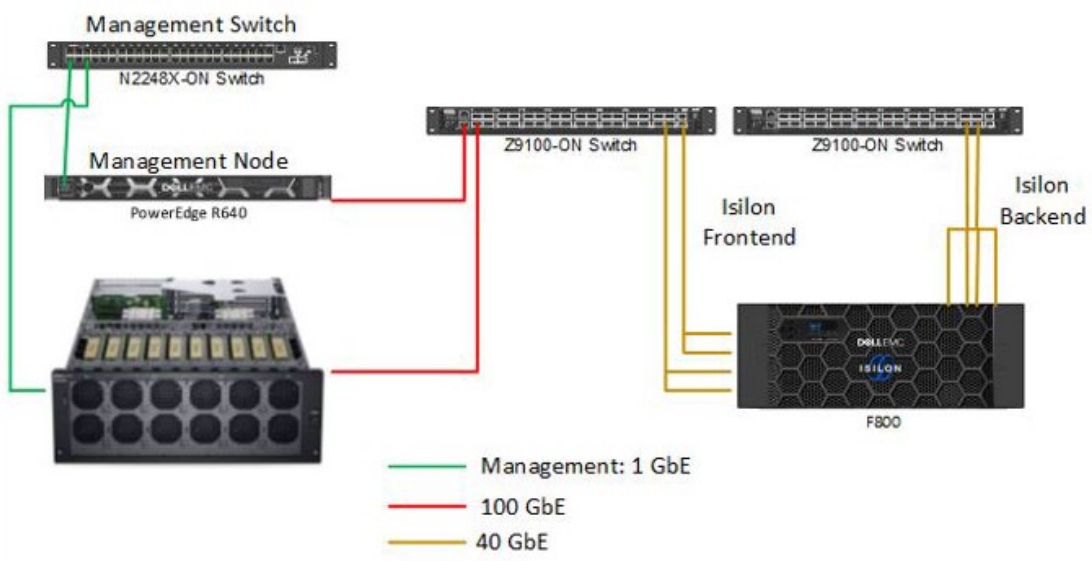

Figure 3: *HPC Ready Architecture for Genomics with NVIDIA Clara Parabricks showing the various modular scaling components*

Infrastructure nodes

Infrastructure nodes (also called log-in 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.

A recommended base configuration for infrastructure nodes is shown in Table 1.

Table 1: Infrastructure node recommended configuration

PowerEdge server	1x R640
CPU	2x Intel Xeon Gold 6132 at 2.6GHz 14 cores
Memory	192GB (24x 16GB) at 2666MT/s dual rank
Disk	8x 1.2TB SAS HDD
RAID controller	PERC H740P RAID Controller
Network adapter	Intel X550 10Gb Base-T, Intel X710 DP 10 Gb SFP+
Operating System (OS)	Red Hat® Enterprise Linux® (RHEL)
Power supply	2x 550W power supply units (PSUs)

Compute node

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

The Dell EMC DSS 8440, 2-socket 4U server can take up to 10x industry-leading NVIDIA V100S GPUs, up to 10x NVIDIA Quadro RTX™ GPUs, or up to 16x NVIDIA T4 GPUs, providing tremendous horsepower. This solution utilizes the DSS 8440 server with 16x NVIDIA T4 GPUs.

Table 2 lists the recommended options for the DSS 8440 server as configured in the HPC Ready Architecture for Genomics with NVIDIA Clara Parabricks.

Table 2: Compute node recommended configuration

PowerEdge server	1x DSS8440
CPU	2x Intel Xeon Gold 6248R 24 cores 3.0 GHz
Co-processors	16x NVIDIA T4 GPUs
Memory	24x 64GB at 2933MT/s dual rank
Disk	OS storage: 4x 480GB mixed use SATA SSD Optional: 2x 1.6TB mixed use NVMe
RAID controller	PERC H730P+ RAID Controller
Network adapter	Intel X550 10Gb Base-T, Intel X710 DP 10 Gb SFP+
Operating System (OS)	Red Hat Enterprise Linux (RHEL)
Power supply	2x 550W power supply units (PSUs)

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. 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 on 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 solely on local direct-attached storage (DAS). Typically, using DAS offers the best 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 or regenerate the data.

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 storage** may make the most sense, as the performance aspect of archival data tends to not impede HPC activities. This particular architecture utilizes the Dell EMC Isilon F800, which is all-flash NAS that scales easily and keeps management simple, no matter how large your data environment becomes. For best use of your resources, automated storage tiering lets you tier less frequently accessed data to other storage systems or to the cloud.

The Dell EMC Isilon F800 provides massive performance and capacity. It delivers up to 250,000 IOPS and up to 15GB/s aggregate throughput in a single chassis configuration and up to 15.75M IOPS and up to 945GB/s of aggregate throughput in a 252-node cluster. Each chassis houses 60 SSDs with a capacity choice of 1.6TB, 3.2TB, 3.84TB, 7.68TB or 15.36TB per drive. This allows you to scale raw storage capacity from 96TB to 924TB in a single 4U chassis and up to 58PB raw storage in a single cluster.

The Dell EMC Isilon F800 is part of the PowerScale series. PowerScale is the next evolution of OneFS the operating system powering the industry's leading scale-out NAS platform. The PowerScale family includes Dell EMC PowerScale platforms and the Dell EMC Isilon platforms configured with the PowerScale OneFS operating system. OneFS provides the intelligence behind the highly scalable, high-performance modular storage solution that can grow with your business.

A OneFS-powered cluster is composed of a flexible choice of data storage including all-flash, hybrid and archive nodes. These solutions provide the efficiency, flexibility, scalability, security and protection for you to store massive amounts of unstructured data within a cluster. The new PowerScale all-flash platforms co-exist seamlessly in the same cluster with your existing Isilon nodes to drive your traditional and modern applications.

Networking components

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 compute 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.

Two Dell EMC PowerSwitch Z9100-ON networking switches provided the interconnect between the compute node and the Isilon F800 storage cluster. 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. An additional switch, the Dell EMC PowerSwitch N2248X-ON, is used for management.

Software: NVIDIA Clara Parabricks application suite

Parabricks is a software suite for performing secondary analysis of NGS data. The suite provides access to many GPU accelerated mapping, alignment, post-processing and variant calling methods. Users can construct secondary analysis pipelines designed to deliver results at fast speeds and low cost. Parabricks analyzes whole human genomes in ~45 minutes, compared to ~30 hours using traditional CPU hardware for 30X WGS data.

The Parabricks software suite runs on a range of server GPUs available on-premises or in the cloud. It scales linearly with the number of GPU resources. Parabricks produces consistent results across different GPUs and generates the same results with each execution. The results are equivalent to Broad GATK best practices pipeline. The current version of Parabricks supports all versions of GATK through v4.0.4. Furthermore, Parabricks analysis pipelines are readily customizable, and new steps can be added effortlessly. Parabricks v3.0.0.05 was used for this architecture.

Services and support

The HPC Ready Architecture for Genomics with NVIDIA Clara Parabricks is available with optional services and support.

Performance evaluation and analysis

Methodology

To determine the recommended software and hardware configuration capable of keeping pace with the daily output of the latest NGS instrumentation, three test cases were evaluated. For each case, the observed wall-clock time was recorded for Parabricks analysis pipeline(s) using different resource configurations, data layouts and sample data.

NGS sample data sets

Data for benchmarking secondary analysis runtime consisted of three human WGS data sets, ERR091571, SRR3124837 and ERR194161, representing 10X, 30X and 50X sample coverage respectively. These data sets are available at the European Nucleotide Archive (ENA).

Parabricks secondary analysis pipeline

Analysis performed on NGS data is often described as a pipeline. A pipeline is simply a collection of methods or operations where the output of one operation becomes the input for the next operation. Four critical operations — mapping, alignment, pre-processing, and variant calling — make up most secondary analysis WGS pipelines.

Parabricks is a software suite for genomic analysis methods designed to take advantage of GPU acceleration. Many of the Parabricks methods are functionally equivalent to existing open-source methods. Parabricks operations are stitched together to create a secondary analysis pipeline best matched to the requirements for the sequencing application of interest such as germline and somatic analysis. Parabricks is available as either a Docker® or Singularity container and uses a variety of server GPU resources. Figure 4 highlights the Parabricks v3.0.0.05 application suite.

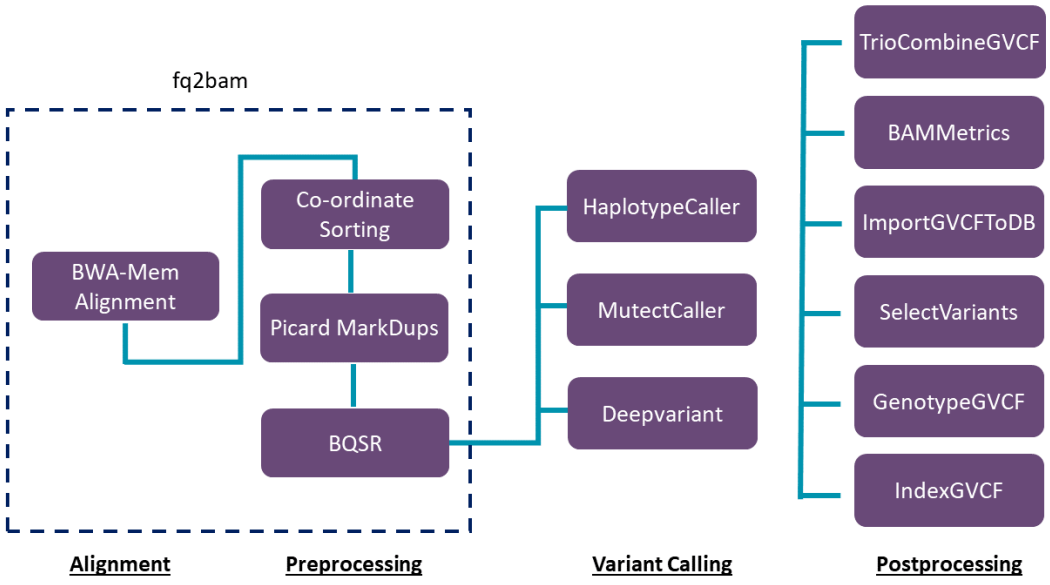

Figure 4: *Parabricks application suite*

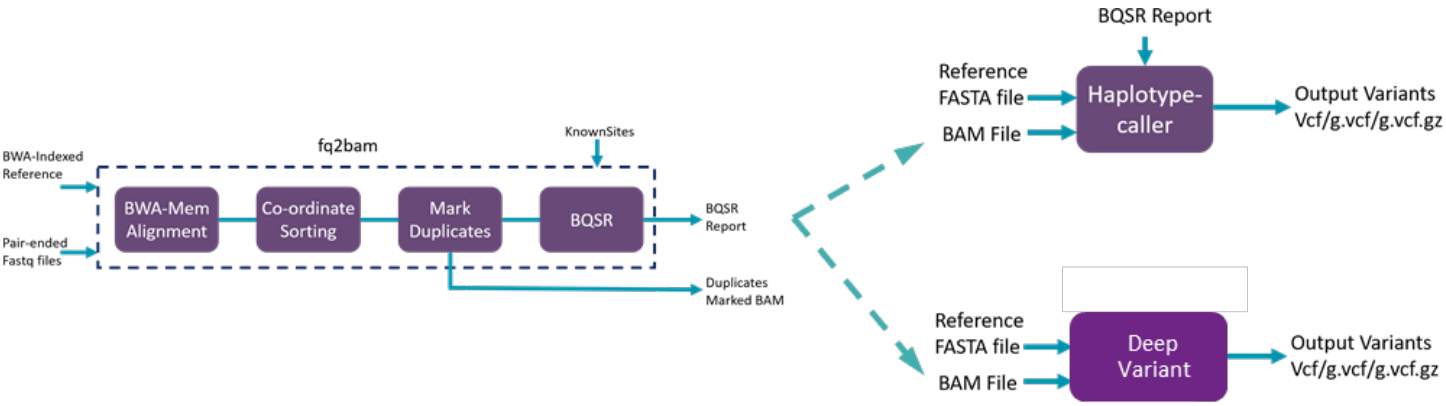

Figure 5: *Parabricks germline pipeline*

Why two variant callers?

Calling genetic variants present in an individual genome relies on billions of short, error-prone sequence reads. Despite more than a decade of effort and thousands of dedicated researchers, the hand-crafted and parameterized statistical models used for variant calling still produce thousands of errors and missed variants in each genome (Poplin, 2016). Many groups run consensus variant calling pipelines that use more than one variant calling method to minimize the likelihood of missing a variant.

DeepVariant, a variant calling method developed by Google®, applies a deep convolutional neural network and has been shown to outperform expert-driven statistical methods. However, calling variants for a 30X human genome and writing the variants out to a gVCF file takes approximately four hours and requires at least 1,024 compute cores. The Parabricks GPU accelerated version of DeepVariant executes in less than 20 minutes for a 30X genome. The fast analysis time makes it possible to use DeepVariant alone or in combination with other expert-driven methods while minimizing the potential of creating a secondary analysis backlog.

Performance evaluation of software

NVIDIA continues to introduce software improvements to NVIDIA Clara Parabricks. Figure 6 shows the runtime reduction between two versions of the Parabricks executing the germline pipeline using the Dell EMC PowerEdge C4140 server with four V100 GPUs test environment. Moving from v2.1.0 to v3.0.0 reduced the runtime by 42%.

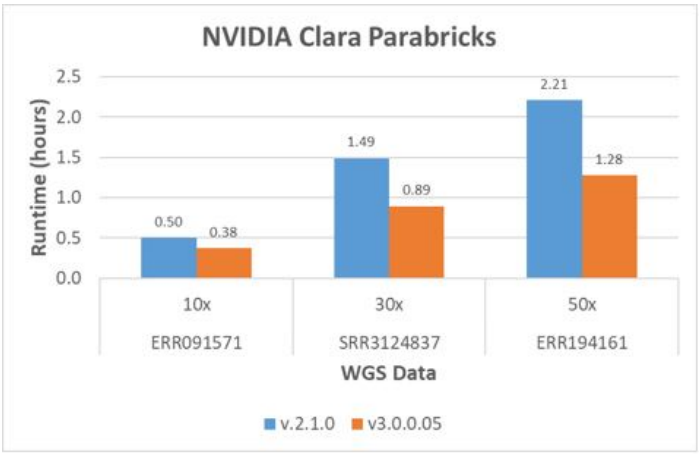

Figure 6: *Latest version of Parabricks germline variant calling pipeline runtime*

Performance of DSS 8440 with 16x NVIDIA T4s

The runtime for a NVIDIA Clara Parabricks secondary analysis using a single NVIDIA T4 GPU is approximately 30% slower than using one V100 GPU. However, two (2) T4 GPUs provide about 10% more TFLOPS than one (1) V100 GPU at approximately half the cost. The DSS 8440 provides up to 16x PCIe slots, which opens the possibility to design a T4 GPU based server that delivers similar runtime performance as a C4140 system with four V100 GPUs, but at a lower cost.

The Parabricks germline analysis was performed using a PowerEdge DSS 8440 with 16x T4 GPUs. For each WGS sample data set described earlier, the runtime was recorded using 1x, 2x, 4x, 8x and 16x T4 GPUs per secondary analysis. The results are plotted in Figure 7 through 9. Overall, the runtime does not scale linearly as the number of GPUs per analysis increases. The scaling pattern is similar to the amount of data per sample increases from 10X to 50X coverage.

Although not presented here, an earlier Dell EMC investigation of Parabricks runtime results using eight or more V100 GPUs per analysis did not scale as efficiently as the T4 GPUs. Additional testing demonstrated that six T4 GPUs generated runtime results nearly identical to four V100 GPUs.

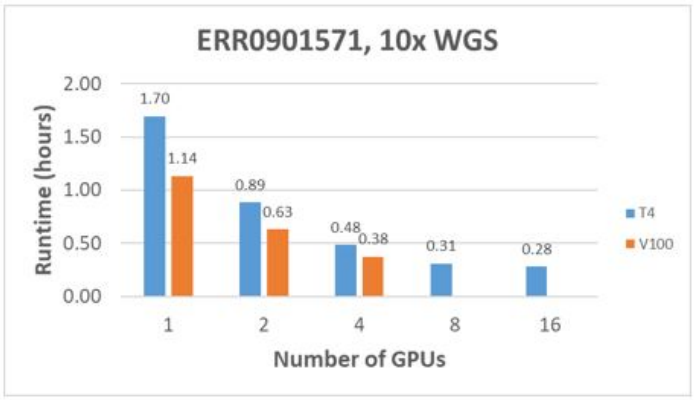

Figure 7: *Performance comparisons with 10X WGS*

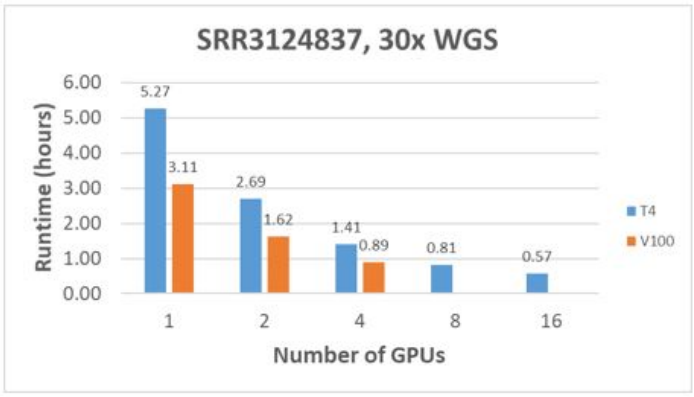

Figure 8: *Performance comparisons with 30X WGS*

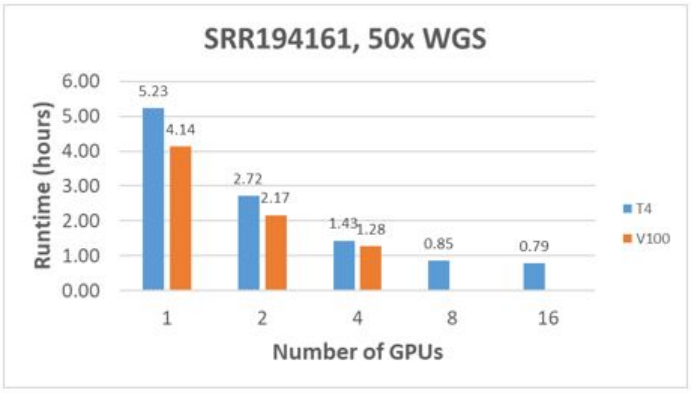

Figure 9: *Performance comparisons with 50X WGS*

Conclusion

NVIDIA GPUs provide tremendous boost to productivity on NGS analysis with NVIDIA Clara Parabricks

A Dell EMC DSS 8440 server with 16x T4 GPUs is capable of processing 30 50X human genomes per day. A similar daily analysis throughput using a traditional x86 CPU architecture requires ten PowerEdge C6420 compute nodes. The complete architecture is discussed in a previous Dell Technologies publication.

However, dedicating all 16x server T4 GPUs to process one sample offers little benefit as using 16 GPUs per analysis is at best 10% faster than using eight GPUs. The design of the DSS 8440 server allows multiple secondary analyses in parallel. By assigning eight NVIDIA T4 GPUs per sample, the daily analysis throughput increases to ~50 genomes per day. Using four GPUs per sample increases the analysis throughput to ~70 genomes per day. More importantly, this daily output using NVIDIA T4 GPUs is less than half the cost of using NVIDIA V100 GPUs.

In addition to speed, compatibility with other analysis tools is essential for the comparability of results. The Parabricks germline analysis results are nearly identical to the well-known BWA-GATK Haplotype caller analysis from prior testing. We wanted to also compare the Parabricks variant calling results to other toolsets like samtools/mpileup. These two completely different tools reach ~90% overall agreement for identified variants, and variations in many well-known genomic regions containing important genes agree more than 99%.

References (optional)

Birney, E. (2017). Genomics in healthcare: GA4GH looks to 2022. Retrieved from https://doi.org/10.1101/203554

Dell EMC. (2018). Dell EMC Isilon and NVIDIA DGX-1 Servers for Deep Learning. Retrieved from https://www.dellemc.com/resources/en-us/asset/whitepapers/products/storage/Dell_EMC_Isilon_and_NVIDIA_DGX_1_servers_for_deep_learning.pdf

Dell Technologies. (2018, October). High Performance Secondary Analysis of Genomic Data. Retrieved from https://www.dell.com/support/article/us/en/04/sln314233/high-performance-secondary-analysis-of-genomicdata?lang=en

Goyal, A. (2017). Ultra-Fast Next Generation Human Genome Sequencing Data. Retrieved from http://www.scirp.org/journal/paperinformation.aspx?paperid=74603

Griffith, M. (2015). Optimizing Cancer Genome Sequencing and Analysis. Cell Systems. Retrieved from https://doi.org/10.1016/j.cels.2015.08.015

Illumina. (2019, July 22). What is NGS Coverage? Retrieved from https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html

Illumina Inc. (2019, July 25). NovaSeq™ 6000 Sequencing System. Retrieved from https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-6000-spec-sheet-770-2016-025/novaseq-6000-spec-sheet-770-2016-025.pdf

Kathiresan, N. (2017). Accelerating Next Generation Sequencing Data Analysis With System Level Optimizations. Nature Scientific Reports. doi:10.1038/s41598-017-09089-1

Lee, S. (2016). Will solid-state drives accelerate your bioinformatics? In-depth profiling, performance analysis and beyond. Briefings in Bioinformatics. doi:10.1093/bib/bbv073

Li, H. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. doi:10.1093/bioinformatics/btp324

Mallick, S. (2016). The Simons Genome Diversity Project: 300 Genomes From 142 Diverse Populations. Nature. doi:10.1038/nature18964

McKenna, A. (2010). The Genome Analysis Toolkit: A MapReduce Framework For Analyzing Next Generation Sequence Data. Genome Research. doi:10.1101/gr.107524.110

NHGRI. (2019, July 25). Retrieved from https://www.genome.gov/human-genome-project/Completion-FAQ

Poplin, R. (2016). Creating A Universal SNP and Small Indel Variant Caller With Deep Neural Networks. Nature Biotechnology. doi: https://doi.org/10.1038/nbt.4235

Stephens. (2015). Big Data: Astronomical or Genomical? PLOS Biology. doi:10.1371/journal.pbio.1002195

Stephens, Z. D. (2015). Big Data: Astronomical or Genomical? PLOS Biology. Retrieved from https://doi.org/10.1371/journal.pbio.1002195

Suwinski, P. (2019). Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. Frontiers in Genetics. Retrieved from https://doi.org/10.3389/fgene.2019.00049

Yoon, K. (2018, March). Reference Architectures of Dell EMC Ready Bundle for HPC Life Sciences refresh with 14th generation servers. Retrieved from https://downloads.dell.com/manuals/allproducts/esuprt_software/esuprt_it_ops_datcentr_mgmt/high-computing-solution-resources_white-papers27_en-us.pdf

DELL Technologies