Hewlett Packard Enterprise | NVIDIA

Technical white paper

Check if the document is available in the language of your choice.

# AI-ENABLED BREAKTHROUGH GENOMICS

Simple, scalable, and optimized HPE and NVIDIA solutions for better tests, drugs, therapies, and personalized healthcare

## CONTENTS



## EXECUTIVE SUMMARY

Next-generation sequencing (NGS) technologies have revolutionized the field of genomics and fueled the growth of biotechnology, healthcare, pharmaceutical, and life sciences organizations worldwide. This rapid growth in production capabilities, however, requires significant compute and storage capabilities to meet the increased demand for processing. In this paper, we demonstrate a 76x improvement in the throughput performance of whole genome sequencing (WGS) workflows compared to a traditional CPU-only compute platform.

This impressive speedup is delivered by using the NVIDIA Clara Parabricks Pipelines GPU-accelerated implementation of the Broad Institute's Genome Analysis Toolkit (GATK) best practices on the flagship NVIDIA®-certified HPE Apollo 6500 Gen10 Plus system.

In addition to this flagship HPC and AI offering, HPE also provides unique capabilities across compute, storage, interconnect, software, services, and consumption models for an end-to-end solution on-premises, hybrid, or as a service. These solutions help simplify system and data management, reduce costs and complexity, and scale to deliver excellent performance for genomics clients in academic/genome research institutions, pharmaceutical/biotech companies, healthcare research organizations, and more.

The ultimate beneficiary is the patient with better health outcomes and quality of life with better drugs, vaccines, therapies, and personalized healthcare.

## THE TRANSFORMATIVE IMPACT OF GENOMICS

Prior to COVID-19, the healthcare and life sciences industry was already at a crossroads with patent expirations, reimbursement pressures, and a quest for new ways of delivering better health outcomes through selective tests, personalized drugs, and therapies. Targeted drugs, tests, and vaccines offer significant benefits to quickly combat health emergencies such as a global pandemic and improve patient outcomes and lifestyles.

The three foundational domains (Figure 1) of genomic medicine include:

- **NGS:** Processing and reducing the amount of raw data (often in tens of terabytes) into a workable format (variant call format [VCF]) so that it can ultimately be used in drug/vaccine discovery and medicine.
- **Translational medicine:** Discovering early and establishing the relationship between genotypes to understand the influence individual DNA variances have on medical outcomes.
- **Precision or personalized medicine:** Tailoring disease prevention and treatment by considering differences in people's genes, environments, and lifestyles.

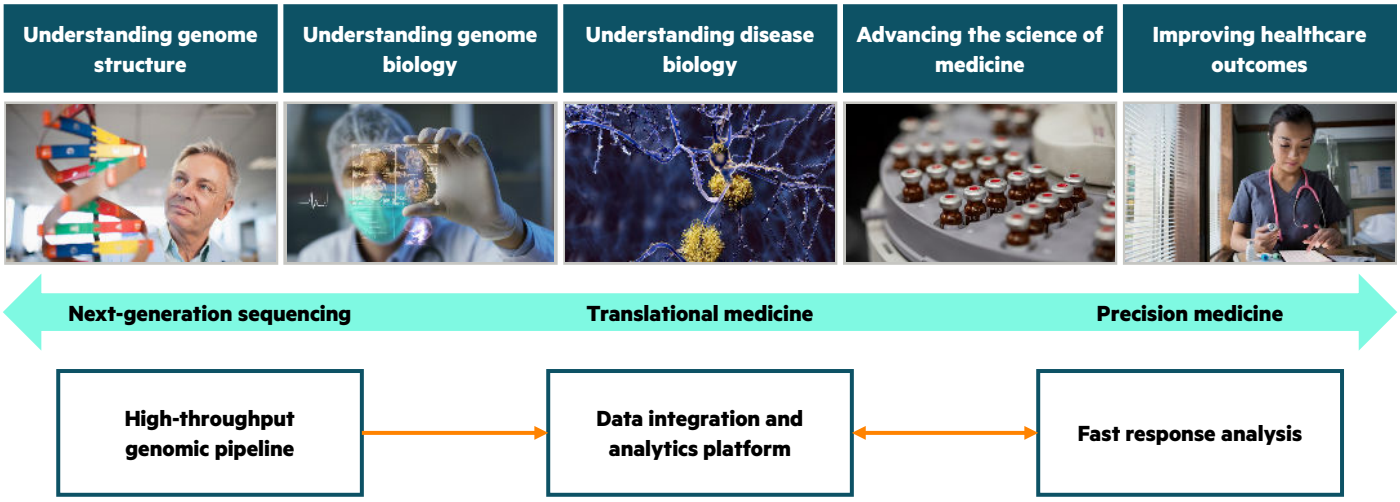



**FIGURE 1.** From NGS to translational and precision medicine

While the rewards of genomics are real and remarkable—more so now than ever before, simple, scalable, and optimized HPC and AI solutions are critical for the continued advancement of these radical innovations.

## SIMPLE, SCALABLE, AND OPTIMIZED HPC AND AI SOLUTIONS CRITICAL TO MAXIMIZE VALUE

It took 13 years and $2.7 billion to sequence one human genome in 2003.[1] Now with the growing use (17%/year[2]) of NGS, the cost of human genome sequencing is plummeting more rapidly than Moore's law to $1000/genome,[3] and it only takes a few hours. Consequently, human genome data analysis requirements are growing at an unprecedented scale.

By 2025, data volumes are expected to be 1 zettabase/year[4] requiring about 40 exabytes of storage. To understand just how much data that is, all videos on YouTube™ will require 1–2 exabytes of storage and all of Twitter will only need .001 to .017 exabytes.[5] Likewise, computing requirements are also escalating with organizations expected to perform a large number of whole genome alignments (WGA).

HPC and AI solutions are needed to overcome these large-scale challenges and continue to propel breakthroughs in genomics. The integration of these NGS breakthroughs with other related chemical imaging, and clinical data analysis is enabling remarkable advances in translational and precision medicine (Figure 1) for longer and healthier lives.

The ROI from HPC and AI for life sciences can be in the hundreds of percent.[6] However, organizations need reliable and secure HPC and AI solutions that scale and perform consistently to generate time-critical insights and maximize the ROI.

Together, HPE and NVIDIA provide these accelerated genomic solutions for academic/genome research institutions, pharmaceutical/biotech companies, and healthcare research organizations. Central to these solutions is the NVIDIA Clara Parabricks Pipelines with NGS software optimized for the HPC and AI platform from HPE.

## HPC INFRASTRUCTURE IS CRITICAL TO SPEED UP NGS

NGS is crucial to get a better insight into the root cause of disease, identify new biomarkers for specific diseases, investigate new drug candidates, and personalize treatments based on a patient's genotype.

Figure 2 shows a typical NGS workflow such as the GATK best practices pipeline. It starts from the unknown DNA fragments on the left to the base-calling sequencers and computational analyses in the middle, to the subsequent annotation and interpretation to the right.


**FIGURE 2.** Schematic of an end-to-end NGS process

Sequencing machines generate large volumes of data per whole genome. One whole human genome at 30x coverage can require several hundred gigabytes of storage during the alignment and variant calling stages, and it can take more than 30 hours to process this data on CPUs.[7] Hence, these middle stages can become computational bottlenecks when processing thousands of genomes, or when a patient is waiting critically in a clinical setting for test results.

Many life sciences organizations operate several NGS instruments concurrently and routinely processing hundreds of samples per week. To keep pace with the output speed of these sequencers, these organizations must quickly process the rapidly growing quantities of genomic and other types of complex life science data, and seamlessly integrate this data on a common HPC and AI platform (Figure 3).

[1] "The Cost of Sequencing a Human Genome," National Human Genome Research Institute, December 2020
[2] Global Next-generation Sequencing Market (2020 to 2025)—Industry Trends, Share, Size, Growth, Opportunity and Forecast—ResearchAndMarkets.com | Business Wire
[3] genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost
[4] "The World Will Store 200 Zettabytes Of Data By 2025," Cybercrime Magazine, June 2020
[5] Medical Breakthrough Driving 10x Returns | InvestorPlace
[6] "The Business Value of Leading-Edge High Performance Computing," Hyperion Research, 2019
[7] "Modern Workloads in Pharma and Life Sciences," Lynn Orlando, WEKA, March 2021

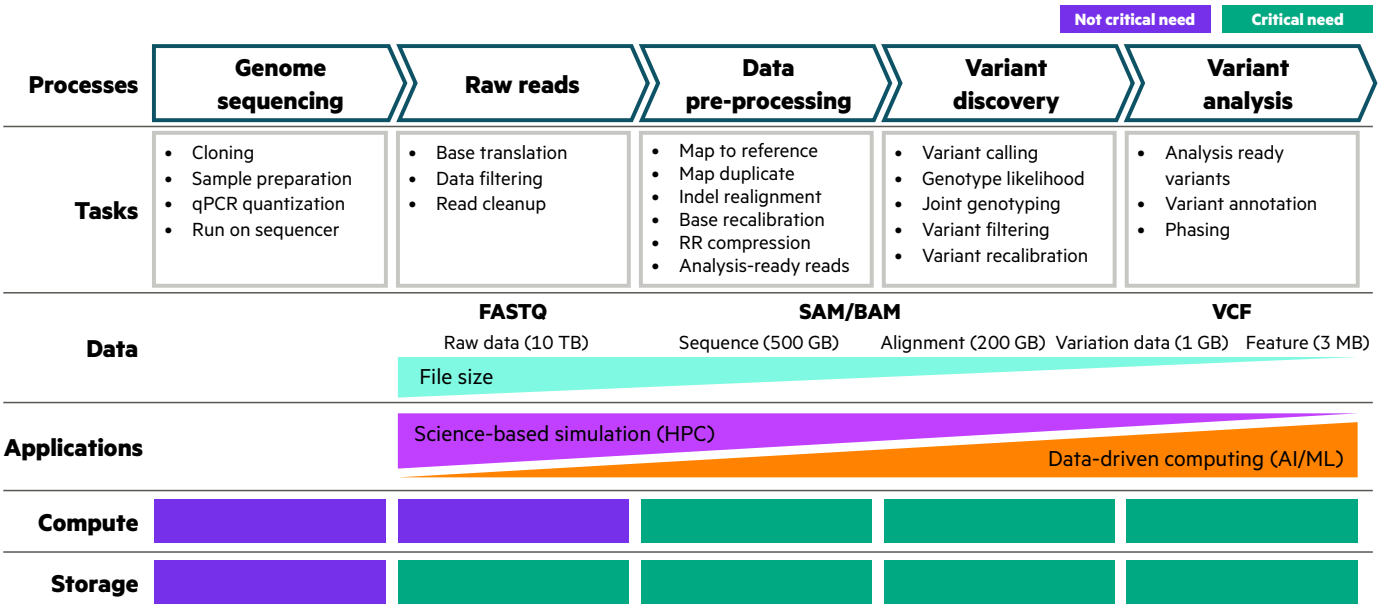



**FIGURE 3.** An HPC and AI platform for NGS

Here is a brief description of the key layers of the HPC and AI platform for genomics:

**Processes and tasks:** Several computational tasks underpin the end-to-end NGS processes in the top layer. Raw sequencing reads are mapped and compared with a known reference genome. The sequence alignment information is stored and compressed. The differences (variants) between the reference and sequence genome are identified. Often this data is infused with analytics capability to increase its value to downstream users, including academic/healthcare researchers and the bio-pharmaceutical industry.

**Data and applications:** The first NGS step is specific to the sequencing instrument and generates raw reads (multiple FASTQ files). These are ASCII text data containing the unordered and unaligned short sequences of DNA bases with the quality score for each base. FASTQ files can range from a small number of big (over several GB) files to a massive number of smaller files. Cumulatively, depending on the desired accuracy and the number of samples, this could total to tens of terabytes/day and to multiple petabytes annually.

Next, this data is transformed into sequence alignment map (SAM) or binary alignment map (BAM) files. These files are much smaller in size—BAM files are compressed. Then these files are compared to a reference genome and the differences (variants) are recorded in even smaller text files in VCF. Finally, specific features are extracted and stored.

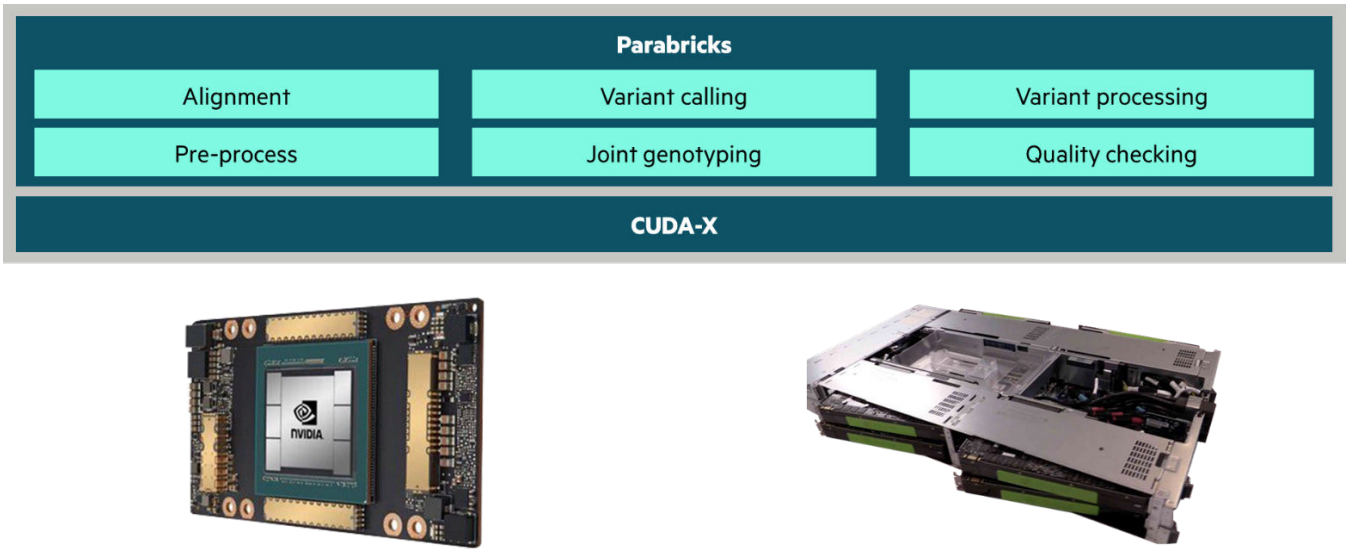
All the applications are typically compute- and data-intensive. In the early stages of the NGS pipeline, they are based on traditional HPC bioinformatics algorithms, while the latter stages use analytics, AI, and machine learning (ML) methods. For example, DeepVariant from Google™, a variant calling method, applies a deep convolutional neural network and often outperforms expert-driven statistical methods in terms of accuracy and sensitivity.

**Compute/Storage:** To accelerate the NGS pipeline, it is critical to optimize every NGS step to the HPC infrastructure. Moreover, the data must be secure and private. While highly parallel computing approaches can accelerate NGS significantly, load imbalances caused by varying reads can result in processing delays. Also, storage choices need to carefully consider tradeoffs in performance, size, persistence, and costs across the NGS data lifecycle.

NVIDIA Clara Parabricks Pipelines, a core turnkey software component of this HPC and AI platform, significantly accelerates genomic data analyses on-premises or in the cloud.

## NVIDIA CLARA PARABRICKS PIPELINES FOR BREAKTHROUGH NGS PERFORMANCE

NVIDIA Clara Parabricks Pipelines (Figure 4) accelerates best practices for somatic and germline variant calling pipelines, such as the GATK, DeepVariant, and other popularly used genomic pipelines. It delivers GPU performance that's 2x faster than the same CPU-only algorithms, along with statistically equivalent results.[8] It is fully configurable, allowing users to choose which steps, parameter settings, and versions of the pipeline to run to build powerful compute tools for genomics.




Source: NVIDIA

**FIGURE 4.** NVIDIA Clara Parabricks Pipelines—a GPU-accelerated genomics toolkit

Key features of the software include:

- **Turnkey solution:** Runs on standard CPU and GPU nodes available and requires no additional setup by the user.
- **On-premises and cloud agnostic:** Can run on both the cloud and local servers such as the HPE Apollo 6500 Gen10 Plus system.
- **Fully deterministic and reproducible:** Any configuration of the software on any platform with any number of resources, generates the exact same results in every run.
- **Equivalent results:** Its pipeline generates equivalent results to the same algorithms used in the baseline GATK4 best practices pipeline.
- **Support for all tool versions:** The accelerated software supports multiple versions of BWA, Picard, and GATK, as well as all future versions of these tools.
- **Visualization:** Generates several key real-time visualizations while performing secondary analysis that can improve the user's understanding of the data.
- **Single-node implementation:** The entire pipeline is run using one compute node and does not incur any overhead of distributing data, orchestrating workflows, or reducing accuracy.

While NVIDIA Clara Parabricks Pipelines significantly accelerates NGS, clients also require other components of the HPC and AI platform across compute, storage, interconnect, software, and services for an end-to-end solution. HPE delivers these solutions on-premises, hybrid, or as a service to help simplify system and data management, reduce costs and complexity, and scale to deliver excellent performance for genomics clients in academic/genome research institutions, pharmaceutical/biotech companies, healthcare research organizations, and more.

[8] "NVIDIA Clara Parabricks Pipelines v3.5 Accelerates Google DeepVariant v1.0," NVIDIA Developer Blog, February 2021

## TAILORED HPC AND AI SOLUTIONS

The HPE and NVIDIA certified system is designed to overcome NGS challenges for clients in:

**Academic/genome research institutions** involved in high-throughput genome research. These institutions typically have a large, central genomics facility and are engaged in research—cancer, neurological diseases, microbial genomics, agrigenomics, and other areas.

They are typically interested in a strategic co-development partnership. Other key requirements include low cost and high performance of the sequencing equipment, capability to integrate easily with genomics software, scaling effectively to manage petabytes of data with security and privacy, on-premises, and especially on the cloud.

**Pharmaceutical/biotech companies** involved in drug/vaccine discovery, molecular modeling, and genomic medicine. Key focus areas include cancer, autoimmune disorders, gene therapy, immunotherapies, neurodegenerative diseases, microbial genetics, and others.

Key requirements include low cost and high performance of the genome sequencing equipment, capability to integrate genomic workflows with legacy infrastructure and data, and availability of on-premises and private cloud HPC infrastructure with data security and privacy.

**Healthcare research organizations** advancing precision medicine and modeling infectious diseases. Key focus areas include diagnostics and treatment of cancer, neurological disease, viral and bacterial disease, translational medicine, and more.

These organizations typically focus on individual patients to provide unique tailored therapies based on the individual's personal and genomic features that are also linked to other evidence-based personalized treatments.

Key customer requirements include low-cost and high-performance sequencing equipment, integration of genomic workflows with electronic health records (EHR) solutions and legacy infrastructure with minimal investment in new on-premises IT infrastructure, and a strong focus on cloud-based HPC solutions.

**The HPE and NVIDIA solution stack** (Figure 5) provides clients excellent choice, flexibility, and performance. It has a broad range of platforms (several servers, storage, interconnects, software, and services) and a fully optimized version of the NVIDIA Clara Parabricks Pipelines software. HPE and NVIDIA also have application engineers with deep genomics skills who collaborate with clients to develop and tune their unique NGS pipelines on HPE or NVIDIA platforms, often with the earliest releases of new software versions.

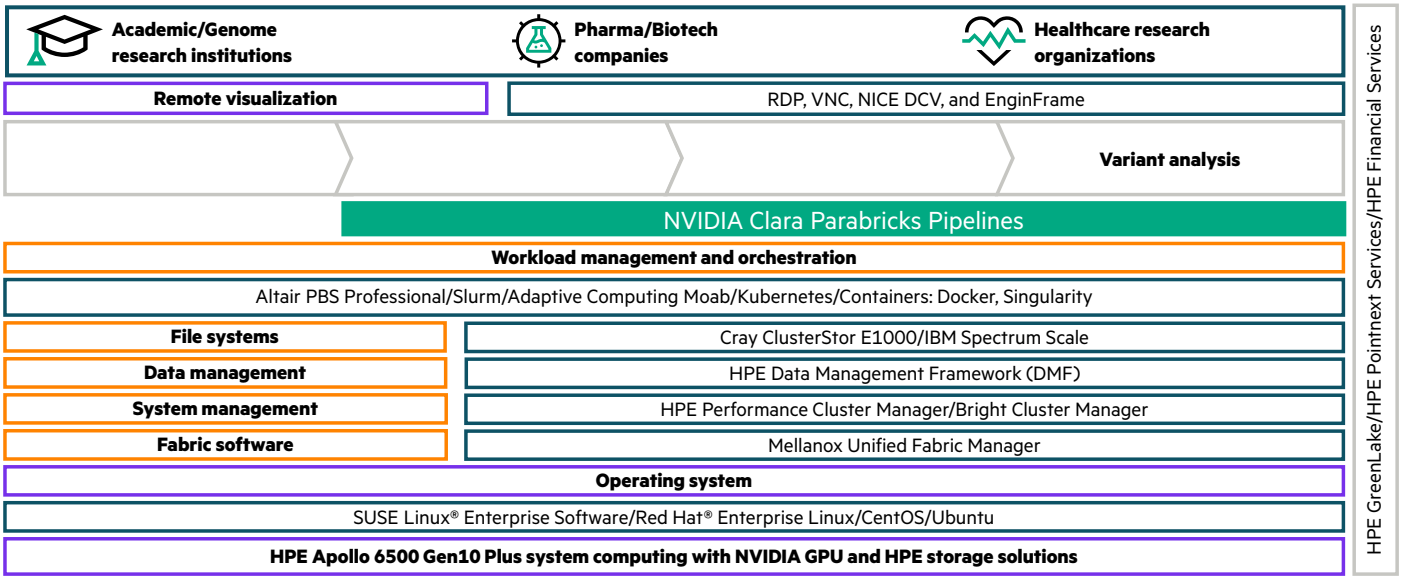



**FIGURE 5.** The HPE and NVIDIA solution stack for NGS

Clients can precisely tailor their NGS solutions to fit their unique requirements with a wide and flexible range of additional solution components.

## PUTTING IT ALL TOGETHER WITH THE NVIDIA-CERTIFIED HPE SYSTEM INFRASTRUCTURE

The key additional components of the HPE with NVIDIA NGS solution stack (Figure 5) are discussed in the following sections starting with the bottom infrastructure layer:

**HPE Apollo 6500 Gen10 Plus system** (Figure 6) is built for the exascale era and optimized for complex genomic workloads using NVIDIA HGX and NVIDIA A100 Tensor Core GPUs with NVLink. This purpose-built platform provides enhanced performance with GPUs, fast GPU interconnects, high-bandwidth fabric, and configurable GPU topology, providing solid reliability, availability, and serviceability (RAS). It can be configured with single or dual processor options for a better balance of processor cores, memory, and I/O. System flexibility can be improved with support for 4, 8, 10, or 16 GPUs and a broad selection of operating systems and options all within a customized design to reduce costs, improve reliability, and provide leading serviceability. Salient components and features include:

- HPE ProLiant XL675d server or HPE ProLiant XL645d server
- NVIDIA HGX and NVIDIA A100 Tensor Core GPUs with NVIDIA NVLink and NVSwitch
- Supports PCIe Gen4, InfiniBand, and Ethernet
- Enterprise RAS with HPE iLO 5, easy access modular design, and N+N power supplies
- Enterprise OS supported: Windows, VMware®, SUSE, Red Hat, and Ubuntu


Source: NVIDIA

**FIGURE 6.** HPE Apollo 6500 Gen10 Plus system, NVIDIA HGX, and NVIDIA A100 Tensor Core GPU

**HPC storage solutions from HPE** (Figure 7) span the whole storage hierarchy to accelerate time-to-insight while managing and protecting valuable data in a client's parallel file systems in a cost-effective way. Parallel file systems deliver aggregate speeds that exceed the architectural limitations of network attached storage (NAS), scale-out NAS, distributed file systems, or object storage. Two parallel file systems products with enterprise-grade support from HPE include:

- **Cray ClusterStor E1000**, which comes with the open-source Lustre PFS. This engineered parallel storage system is purpose-built, high-performance storage controllers for extreme speed and scalability. It is for organizations that do not require enterprise-grade functionality but value extreme price/performance and scale. Key features include:
  - Up to 80 GB per second from just 24 SSDs in just two rack units
  - Up to 3.3 GB per second per SSD data transfer to the compute nodes
  - Support for the Data-on-MDT (DoM) feature, which improves small file I/O by placing small files on the NVMe flash-based metadata targets (MDT)
  - Connected with 200 Gbps HPE Slingshot or InfiniBand EDR/HDR or 100/200GbE
  - Benefits of Lustre file system includes no software license per TB or storage drive
  - Additional premium support with in-house Lustre R&D team from HPE

- **HPE Parallel File System storage** embeds IBM Spectrum Scale, a general parallel file system (GPFS) for the enterprise. This is a software-defined storage solution built on cost-effective HPE ProLiant DL rack servers and offers a broad set of enterprise storage features—enterprise IT-grade data availability (backup and disaster recovery), data accessibility (NFS, SMB, HDFS, object), and data compliance (audit log, industry certifications). Key features include:
  - Combination of the leading parallel file system in the enterprise, along with the leading x86 rack servers in the enterprise (HPE ProLiant DL)
  - Starts as low as 27 TB in just four rack units and scales to more than 25 PB in a single file system
  - Connected with InfiniBand EDR/HDR or 100/200GbE
  - No separate software license per TB or storage drive
  - Operational support services for the full product—both hardware and software—from HPE Pointnext Services

**Cray ClusterStor E1000 storage system**

**HPE Parallel File System Storage**
(in 4 node starter configuration)

**FIGURE 7.** Two HPE parallel file system storage offerings with enterprise-grade support

**HPE GreenLake** is a market-leading IT-as-a-service offering for genomics. It offers easy and affordable access to dedicated, powerful computing and analytics capabilities, helping clients make faster decisions, and reduce time to discovery. It avoids overprovisioning costs with elastic capacity ready for growth or unpredictable spikes. HPE GreenLake combines the simplicity, agility, and economics of public cloud with the security and performance benefits of on-premises IT.

This consumption-based IT model helps clients accelerate time to value, align IT economics with business priorities, simplify IT operations, and gain better control. HPE GreenLake brings a consumption-based HPC model on-premises—or in a colocation—that delivers superior flexibility, scalability, and control.

With HPC and AI as a service, clients can design their own genomics infrastructure solution using industry-leading HPE technologies or can standardize their service with pre-sized configurations that are self-service and managed for them. A built-in technology refresh feature in an HPE GreenLake engagement allows clients to benefit from the latest technology available in the market so they can stay competitive. HPE can also buy out—and recycle—existing infrastructure to help meet sustainability targets.

HPE Performance Cluster Manager is a flexible, easy-to-use system management solution offering system administrators all the tools they need to turn even the most complex hardware into easily manageable systems capable of accommodating a growing variety of workloads:

- Accelerated system set-up
  - Deploy system over bare metal in minutes rather than hours or days
  - Nodes are provisioned in parallel for maximum system performance
  - During setup, hardware elements are automatically discovered and configured
- Fine-grained centralized monitoring and management of the cluster
- Image management features
  - Provision a wide variety of applications on any number of nodes to keep up with users' demands
  - Repurpose nodes on the fly to help minimize wait times
  - Update software and perform other maintenance tasks on idle nodes without interrupting running jobs on the rest of the cluster
- Advanced power management features allow smart management of power resources for better data center economics.

**HPE Data Management Framework optimizes** storage resource utilization and data accessibility by introducing a hierarchical, tiered storage management architecture. Data is moved between tiers based on service-level requirements defined by the administrator. For example, frequently accessed data can be placed on flash drives in a high-performance tier, while data accessed less often can be stored on hard drives in a capacity tier and data to be archived can be sent off to tape storage.

**HPE Pointnext Services** offers a spectrum of services for genomics—from services like application tuning to more integrated advisory service offerings such as project management, on-site consulting, technical account management, and solution architecture consulting.

Combined with the HPE HPC Cluster Management Solution, HPE Pointnext Services provide skilled consultants to assist clients with installation, configuration, and understanding the management of the entire HPE cluster environment.

Several HPE financing/sourcing options help clients purchase HPC systems and upgrade them frequently. The full genomics infrastructure stack includes:

- Classic purchase, where the client owns it runs it.
- HPE Financial Services (HPEFS), which finances the stack and the client runs it.
- HPE GreenLake is based on pay for usage, where the client subscribes and pays for what they use, and they run it.
- HPC as a service is where the client subscribes and pays for what they use, and HPE runs it.

Together, HPE and NVIDIA are accelerating the performance and impact of NGS.

## DELIVERING BREAKTHROUGH PERFORMANCE FOR NGS

The total runtime of the GATK best practices pipeline can be dramatically reduced to 59x[9] by using the NVIDIA Clara Parabricks Pipelines software on an HPE Apollo 6500 Gen10 Plus system with NVIDIA GPUs. This means that a WGS workflow that normally takes about three days on a CPU-only node can be processed on a GPU-enabled node in less than one and a half hours.

Figure 8 depicts the relative GPU-enabled speedup over the equivalent CPU-only implementation for five different WGS workflows, which are all human genome samples with WGS coverage levels ranging from 26x to 43x. HPE carried out the benchmarks in August 2021. For each of the five workflows, the performance scales near linearly with the number of GPUs used and the throughput rate of NVIDIA Clara Parabricks Pipelines with 8 GPUs (in orange) ranges from 45 to 80 genomes per day.

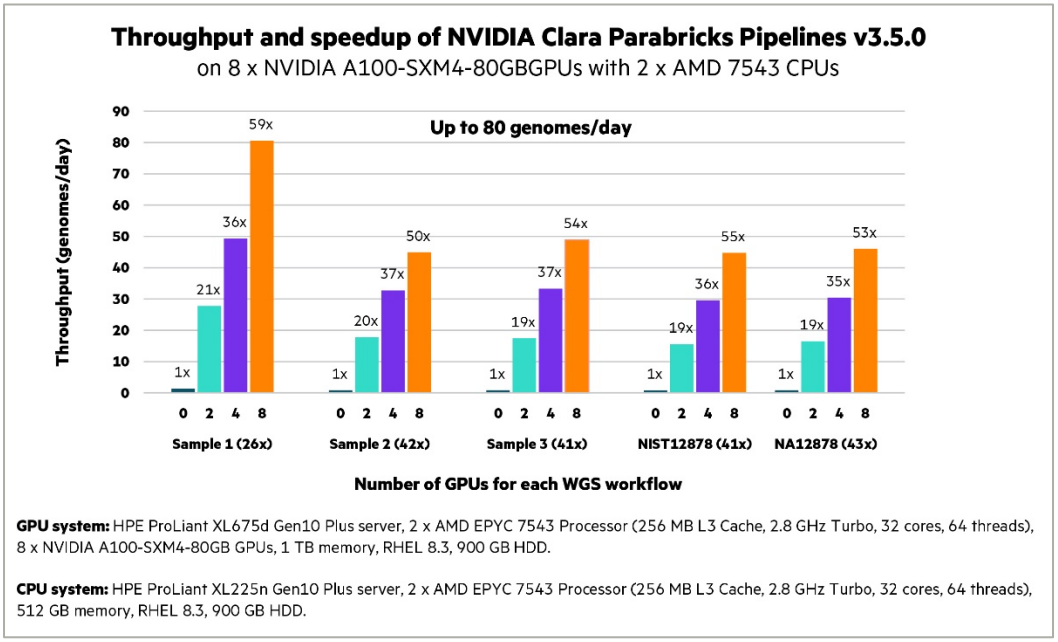



**FIGURE 8.** Throughput and speedup over CPU of five WGS workflows on a single 8 GPU node

[9] Based on HPE internal testing, August 2021.

Figure 9 shows the throughput and scaling of a 4 x NVIDIA A100-SXM4-40GB GPU configuration for the five WGS workflows. This graph demonstrates the excellent scaling of the HPE Apollo 6500 Gen10 Plus server and the ability to process up to 54 genomes/day using 4 NVIDIA A100 GPUs.[10]

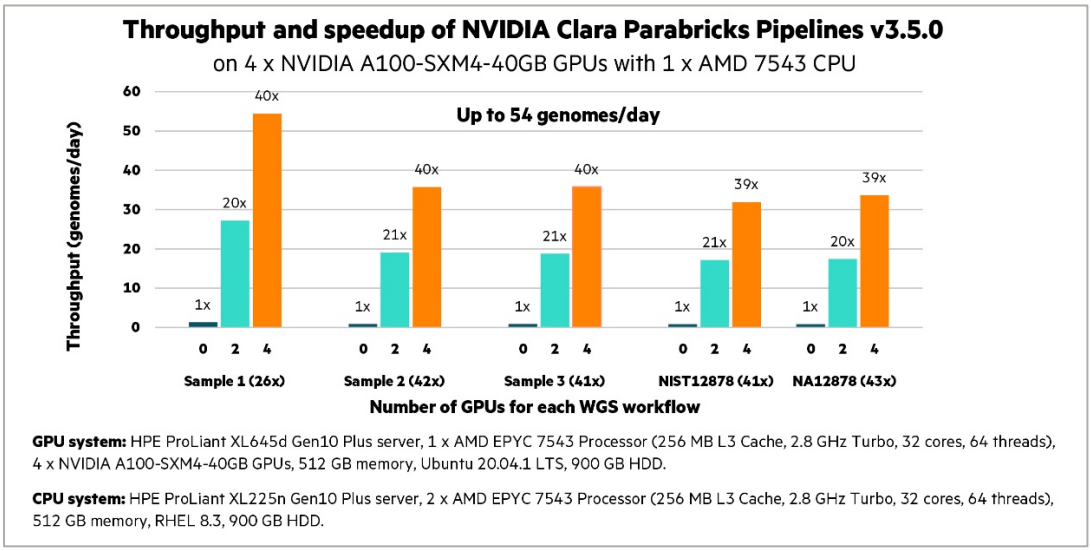



**FIGURE 9.** Throughput and speedup over CPU of five WGS workflows on a single 4 GPU node with NVIDIA A100-SXM4-40GB GPUs

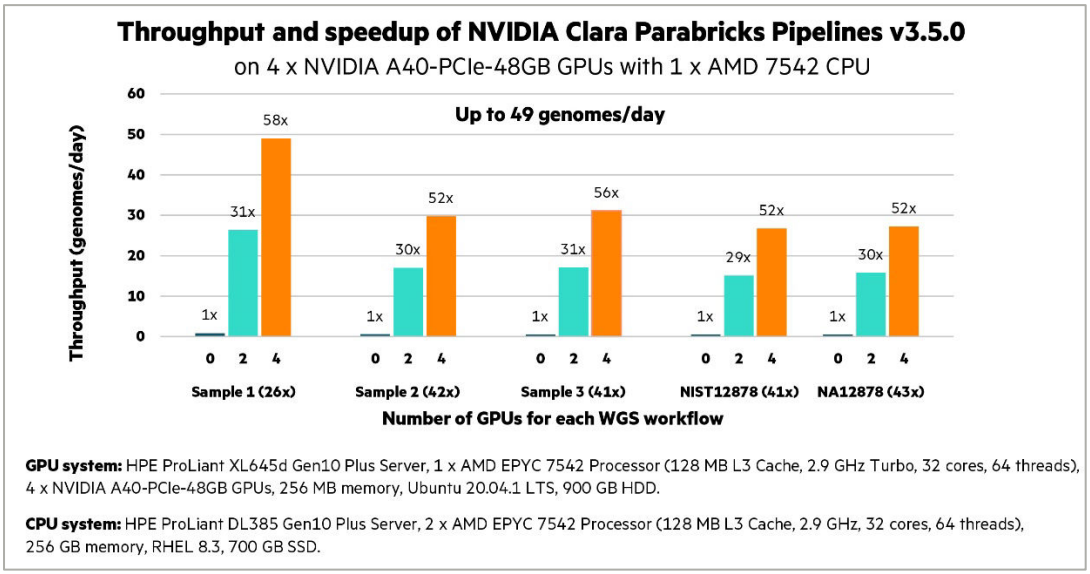



**FIGURE 10.** Throughput and speedup over CPU of five WGS workflows on a single 4 GPU node with NVIDIA A40-PCIe-48GB GPUs

Figure 10 shows the throughput and scaling of a 4 x NVIDIA A40 GPU system for the five WGS workflows, demonstrating comparable performance up to 49 genomes per day.[11]

In addition to the impressive performance gains, the GPU-enabled HPE with NVIDIA Clara Parabricks Pipelines solution also delivers 99.999%[12] accurate results compared to the CPU-only version. Worldwide, several clients are leveraging these impressive capabilities to overcome challenges with deploying genomics workloads across their organization.

[10, 11] Based on HPE internal testing, August 2021.
[12] NVIDIA Clara Parabricks Product Sheet

## THE HPE ADVANTAGE

As more life sciences and healthcare organizations implement and scale genomics solutions, they need a reliable partner with deep HPC, AI, and genomics expertise. As a market leader in life sciences solutions, HPE is partnering with NVIDIA to deliver breakthrough performance for the NVIDIA Clara Parabricks Pipelines next-generation sequencing (NGS) genomics software. This deep and sustained collaboration is anchored on the HPE Apollo 6500 Gen10 Plus system powered by NVIDIA GPUs.

In addition, HPE delivers a comprehensive portfolio of secure, enterprise-grade, high-performance systems, software, high-value services, and the top-tier ecosystem of partners to help clients in academic/genome research institutions, pharmaceutical/biotechnology companies, and research medical centers deploy and scale NGS solutions.

These solutions have been proven in many of the most demanding research and production environments in the world and deliver excellent productivity and return on investment. Most importantly, patients benefit from better health outcomes and quality of life with targeted drugs, vaccines, therapies, and personalized healthcare.

## LEARN MORE AT

hpe.com/us/en/compute/hpc/apollo-systems.html

nvidia.com/en-us/clara/genomics/





a00115076ENW, October 2021, Rev. 2

Hewlett Packard Enterprise