Accelerated throughput of next-generation sequencing workflows

Advancing precision medicine with Parabricks software on the HPE Apollo 6500 Gen10 system and NVIDIA GPUs

Contents

Executive summary .......................................................................................................................... 2
Introduction .................................................................................................................................. 2
Parabricks accelerates GATK Best Practices pipeline .................................................................. 2
HPE Apollo 6500 Gen10 system—Optimized for genomics processing ........................................... 3
Performance and best practices ........................................................................................................ 4
   Speedup over CPU ....................................................................................................................... 4
Accuracy ...................................................................................................................................... 5
Minimizing turnaround time versus maximizing throughput .......................................................... 6
Total system cost reduction ............................................................................................................ 6
Additional performance advantages ............................................................................................... 7
The HPE advantage for genomics .................................................................................................. 7
Resources ...................................................................................................................................... 7
Executive summary

Next-generation sequencing (NGS) technologies have revolutionized the field of genomics and fueled the growth of biotechnology, health care, pharmaceutical, and life sciences organizations worldwide. This rapid growth in production capabilities, however, requires significant compute and storage capacities to meet the increased demand for processing. In this white paper, we demonstrate an order of magnitude improvement in the throughput performance of whole genome sequencing (WGS) workflows. We achieve this impressive speedup by using Parabricks™ GPU-accelerated implementation of the Broad Institute’s Genome Analysis Toolkit (GATK) Best Practices pipeline on an HPE Apollo 6500 Gen10 system with NVIDIA® GPUs.

Introduction

Genomic sequencing has flourished rapidly over the last decade to the point where a whole human genome can be sequenced in less than a day for about $1000.¹ The plummeting cost of DNA sequencing, however, has resulted in an explosion of genetic data volume from NGS methods that are challenging businesses to adapt to new production processes, changing technologies, and rising global competition.

From a computational standpoint, NGS workflows require high-performance processing and storage infrastructures to achieve maximum productivity and gain deeper insight from their most data-intensive workflows. Figure 1 shows a typical NGS workflow, starting 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.

The sequencing machines, in particular, generate large volumes of data per whole genome. A 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 using CPU solutions.² This center stage can, therefore, become a computational bottleneck when processing thousands of genomes, or when a patient is waiting critically in a clinical setting.

![Figure 1. Schematic of NGS workflow](image)

This white paper examines and elucidates the performance, scalability, and affordability of the HPE Apollo 6500 Gen10 system for tackling one of the greatest computational bottlenecks in genomic processing today.

Parabricks accelerates GATK Best Practices pipeline

Parabricks is a genomic analysis company whose flagship software (GenomeBricks™) accelerates the processing of the standard GATK Best Practices pipeline from the Broad Institute. It delivers superior GPU performance that is faster than the same CPU-only algorithms, along with statistically equivalent results. Figure 2 shows the standard GATK Best Practices pipeline that is implemented by the Parabricks software, which consists of three sequential stages of secondary analysis. This analyzes and converts the FASTQ formatted sequence files from an NGS sequencer into a VCF-formatted file containing the computed sequence variants that are used for subsequent tertiary analysis.

1 DNA Sequencing Costs
2 Data is based on the single-node performance of a 30X coverage workflow using the GATK Best Practices pipeline. The workflow is on an HPE ProLiant XL170r Gen10 server with 2 x Intel® Xeon® Gold 6152 Processors (30.25M Cache, 2.10 GHz) and 384 GB memory running RHEL 7.4.
Figure 2. Computational stages of the Parabricks GenomeBricks workflow

Features of the Parabricks software include:

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

**HPE Apollo 6500 Gen10 system—Optimized for genomics processing**

With up to eight high-performance GPUs per server, including NVIDIA Tesla V100, P100, and P40, the HPE Apollo 6500 Gen10 system provides superior performance per dollar for GPU-intensive workloads—delivering up to 125 TFLOPs single-precision compute. Purpose-built for accelerated computing, this platform features both PCIe and NVIDIA NVLink™ GPU interconnects, providing the flexibility to suit a wide variety of requirements. High-bandwidth and low-latency networking adapters (up to four high-speed Ethernet, Intel® Omni-Path Architecture, InfiniBand EDR, and future InfiniBand HDR per server) are tightly coupled with the GPU accelerators. This allows the system to leverage the network bandwidth for the most demanding genomics workflows.

In addition to the HPE Apollo 6500 Gen10 system, Hewlett Packard Enterprise provides biotechnology, health care, pharmaceutical, and life sciences institutions a unique portfolio of high-value capabilities and resources to support their growing production needs.

---

2 Theoretical peak performance.
Performance and best practices

Speedup over CPU

The total execution time of the GATK Best Practices pipeline can be dramatically reduced by an average of 36X by using the Parabricks software on an HPE Apollo 6500 Gen10 system with NVIDIA GPUs. This means that a WGS workflow that normally takes 1.5 days on a CPU-only node can be processed on a GPU-enabled node in less than one hour.4

Figure 3 shows 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 June 2018. For each of the five workflows, the performance scales near linearly with the number of GPUs used and the speedup of GenomeBricks with 8 GPUs (in blue) ranges from 33X to 39X. Figure 4 shows the average GPU-enabled performance improvement for each number of GPUs, which ranges from an average speed of 7X with only 1 GPU to 36X with 8 GPUs.

Figure 3. Speedups of Parabricks GenomeBricks pipeline over CPU equivalent pipeline

---

4 Calculation is based on the single-node performance of an equivalent CPU-only implementation of the Parabricks GenomeBricks workflow. The CPU-only workflow is on an HPE ProLiant XL270d Gen10 server with 2 x Intel Xeon Gold 6152 Processors (30.25M Cache, 2.10 GHz) and 768 GB memory running Ubuntu 16.04.
Accuracy

In addition to impressive compute performance, the GPU-enabled Parabricks software produces intermediate output files from the GATK4 BaseRecalibrator and GATK4 ApplyBQSR tools that are fully equivalent to the corresponding CPU-generated outputs. Furthermore, the final VCF results from the GATK4 HaplotypeCaller tool are 99.99% accurate in sensitivity and precision for both the single nucleotide polymorphism (SNP) and insertion-deletion (INDEL) results.\textsuperscript{5} It should be noted that the baseline variant caller in GATK4 is nondeterministic and can generate slightly different results depending upon certain run-time parameters, such as number of threads, so the differences are consistent with these variations in GATK4 execution.

\textsuperscript{5} Comparison is based on the single-node results of an equivalent CPU-only implementation of the Parabricks GenomeBricks workflow. The CPU-only workflow is on an HPE ProLiant XL270d Gen10 server with 2 x Intel Xeon Gold 6152 Processors (30.25M Cache, 2.10 GHz) and 768 GB memory running Ubuntu 16.04.
Minimizing turnaround time versus maximizing throughput

If the desired goal for genomics processing is to minimize the individual time to solution for a single workflow, then the best way to run each workflow is to use the maximum number of GPUs like the 8-GPU workflow jobs in Figure 3. Running each workflow in this manner minimizes the turnaround time of each individual job by maximizing the number of GPUs used.

However, if the goal is to maximize the number of workflows processed in a given time, then the best way to run each workflow is to run it simultaneously with other workflows on the same node. But use a fewer number of GPUs. Running each workflow in this manner maximizes the overall throughput of multiple jobs by minimizing any parallel overheads and inefficiencies associated with running a multi-GPU job.

Figure 6 demonstrates four different ways to run the GATK pipeline on an 8-GPU compute node and the resulting throughput rate in genomes/day/node. The workflow used is the previously used Sample 1 (26X coverage) test case.

The first way is the traditional method of running one job with all 8 GPUs (1 x 8 GPUs), and the resulting throughput is 42.0 genomes/day/node. The second way is to run two jobs using 4 GPUs each (2 x 4 GPUs), and the resulting throughput increases by 1.2X to 53.6 genomes/day/node. Further, increasing the number of workflows to four jobs using 2 GPUs each (4 x 2 GPUs) and eight jobs using 1 GPU each (8 x 1 GPU) continues to improve throughput (relative to the 1 x 8 GPU case). It improves by 1.3X to over 56 genomes/day/node. A slight optimum throughput is achieved with the 4 x 2 GPU case most likely due to a small contention for resources with eight concurrent jobs.

Figure 6. Optimizing throughput by running multiple jobs on a node

Total system cost reduction

The dramatic decrease in the solution time of the GATK pipeline can significantly drop the total system cost for processing genomic workflows by 90% in server and infrastructure acquisition costs. The reduced number of servers required, combined with lower power consumption and networking overhead, can result in a total cost savings of 9X–10X.

---

6 Comparison is based on the single-node results of an equivalent CPU-only implementation of the Parabricks GenomeBricks workflow. The CPU-only workflow is on an HPE ProLiant XL270d Gen10 server with 2 x Intel Xeon Gold 6152 Processors (30.25M Cache, 2.10 GHz) and 768 GB memory running Ubuntu 16.04.

7 Cost savings are based on the HPE US. list prices for reference servers as of September 4, 2018. Calculation is based on the single-node performance of an equivalent CPU-only implementation of the Parabricks GenomeBricks workflow. The CPU-only workflow is on an HPE ProLiant XL170r Gen10 server with 2 x Intel Xeon Gold 6152 Processors (30.25M Cache, 2.10 GHz) and 384 GB memory running RHEL 7.4.
Additional performance advantages

In addition to GPUs and GPU-enabled software, the HPE Apollo 6500 Gen10 system leverages the latest Intel Skylake processors, which can provide a 20% performance increase over a comparable Broadwell processor.\(^8\) Furthermore, using local SSD storage for the inputs and reference databases can improve performance by an additional few percent when compared to a remote HDD file system.

The HPE advantage for genomics

As a market leader in life sciences solutions, Hewlett Packard Enterprise is partnering with NVIDIA and independent software vendors (ISVs) in the genomics industry to help institutions maximize their productivity and return on investment. This partnership is anchored on the HPE Apollo 6500 Gen10 system powered by NVIDIA GPUs.

HPE and NVIDIA have extensive experience working with industry-leading pharmaceutical companies and independent software providers to deliver solutions that offer high performance, flexible capacity, secured availability, and simplified management. With a broad choice of standards-based solutions, proven in many of the most demanding research and production environments in the world, HPE ensures businesses find the optimal solution to fit their needs and budget.

Resources

hpe.com/servers/apollo6500
parabricks.com

Learn more at

\(^8\) Calculation is based on the single-node performance of the Parabricks GenomeBricks workflow. The Broadwell-based system is an HPE ProLiant XL270d Gen9 server with 2 x Intel Xeon E5-2697 v4 Processors (45M Cache, 2.30 GHz), 8 x NVIDIA Tesla V100-PCIe-16GB GPUs, and 512 GB memory running Ubuntu 16.04. The Skylake-based system is an HPE ProLiant XL270d Gen10 server with 2 x Intel Xeon Gold 6152 Processors (30.25M Cache, 2.10 GHz), 8 x NVIDIA Tesla V100-SXM2-16GB GPUs, and 768 GB memory running Ubuntu 16.04.