Parabricks™

SkyScale

Accelerated Genomic Analysis

Applying Massive Parallel Computing to Genomics Secondary Analysis

Modern genomics is characterized by rapid production of vast amounts of raw sequencing data (sequencing reads) using next-generation sequencing (NGS) and the equally massive computing requirements for conversion of that data into useful results.

That conversion from sequence reads to usable results is through public and private genome analysis toolkits, of which the most popular is the open-source Broad Institute Genome Analysis Toolkit (GATK) [1].

This white paper presents results from a new method developed by Parabricks run on a SkyScale Accelerated Cloud system with up to 16 NVIDIA Tesla V100 GPUs that greatly reduces that time (compared to CPU-based servers) to execute GATK4 Best Practices pipelines.

The Problem

The computational challenge of genomics analysis is immense and growing. The widely cited article "Big Data: Astronomical or Genomical?" by Zachary Stephens, et. al. [2], forecasts that by 2025, genomics will create 1 zetta bases/year (10^{21} = 1 trillion giga bases) and require ~2 trillion CPU hours for variant calling and an astronomical (genomical?) ~10,000 trillion CPU hours for whole genome alignment for a year's reads.

Narrowing that to an example that a lab might routinely perform as part of a population genetics study or a drug development, consider whole genome sequencing analysis of 1,000 individuals. Running Broad Institute GATK4 on computing equipment in a typical genomics lab will require 30-35 hours for one run on each genome; for 40X coverage, that is 1,200 to 1,400 hours for each individual (50+ days). The total time for 1,000 individuals will depend on the number of systems deployed at once, but for most labs time for the study will be measured in months if practical at all.

Solution, Part 1—Parallelism in the Cloud

The algorithms used in parts of a full genomic analysis in general, and in GATK4 in particular, are amenable to being parallelized—individual streams of an entire data set can be processed on many processors in parallel. Initial efforts operated on multi-core CPUs in parallel. This helps—but with the data generated by NGS, is not sufficient.

The next logical step is to run many clustered (communicating) machines, each with one or more multicore processors, at the same time. But building large

Figure 1. NVIDIA Tesla V100 Accelerator from One Stop Systems / SkyScale Accelerated Cloud. See "Solution Part 2" below.

clusters of computers is expensive and beyond the ability of most university and small to mid-size companies to build and operate.

A partial solution is cloud computing. The National Institute of Standards and Technology (NIST) defines cloud computing as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources … that can be rapidly provisioned and released with minimal management” [3].

The use of cloud computing for genomics analysis is becoming increasingly well-established. In 2015, Philip Groth and Gerhard Reuter provided an introduction in “Analysis of Genomic Data in a Cloud Computing Environment” [4]. Three years later with an article in *Nature*, Ben Langmead and Abhinav Nellore authored a comprehensive review in the article “Cloud computing for genomic data analysis and collaboration” [5]. This article has several useful features:

- Provides examples of large genomics projects and resources.
- Details a second major benefit of genomics in the cloud: facilitating collaboration worldwide, including a number of large example projects.
- Summarizes genomics data types, giving for each the number of bases, bytes required, and core hours to analyze. This last type is whole genome sequencing of human DNA from the TOPMed project: ~18,000 human samples with 30X coverage requiring ~30,000 hours to analyze 100 samples (Table 2 in the article).
- Defines key cloud attributes and advantages for genomics research, and lists 24 cloud providers including those specializing in genomics (Table 3).
- Address the critical issues of privacy, security, and regulation, noting that “while the NIH initially disallowed analysis of protected data on commercial clouds, this policy was reversed in 2015” (217).

Fully endorsing the cloud trend, with GATK4 the Broad Institute has partnered with multiple cloud providers, including Alibaba, Amazon, Google, and Microsoft for cloud services. This enables users to commission and run GATK4 on large clusters of processors in the cloud. However,

- It is often expensive.
- The massive parallelism described below has limited availability in the cloud.
- Services are usually provided through virtual machines using multi-tenant hardware (issues described below).
- Even with cloud providers’ attention to security, an organization may find security insufficient.

Solution Part 2—Massive Parallelism, GPUs, Parabricks, and SkyScale

Modern GPU-based hardware can provide computing nodes with tens of thousands of cores. Generally a GPU has a single instruction, multiple data (SIMD) architecture. The implementation described in this whitepaper is based on the ***Tesla V100 Accelerator from NVIDIA*** (see Figure 1). The Tesla V100 GPU goes beyond SMID to implement SIMT—Single Instruction, Multiple Thread. SIMT is unique to NVIDIA. It allows each thread in a “warp” of up to 32 parallel threads to access its own registers and follow divergent control flow paths (which ordinary SIMD cannot do).

A second key feature of the Tesla V100 is its support for mixed precision. As shown in Figure 2, each Tesla V100 GPU has 5,120 32-bit floating point cores, 2,560 64-bit floating point cores, and 640 “tensor” cores that support FP32, FP16, and 8 and 4-bit integer operations.

The algorithms used in secondary analysis can be parallelized to take advantage of dense GPU architecture. But doing so is challenging.

Parabricks’ novel contribution is that it has adapted GATK4 Best Practices workflows to execute on the NVIDIA Tesla V100 hardware, including optimizing for its SIMT and mixed precision architecture. When executed on the SkyScale Accelerated Cloud maximum configuration with 16 V100 GPUs, execution speed for secondary analysis is increased by a factor of 40+ and a two-day run is reduced to an hour compared to a typical CPU-based server.

See Figure 2 for configuration and Figures 3 and 4 for detailed results. The Tesla V100 is described in [6], with details on the Tensor Core in [7].

* Parabricks current release uses both FP64 and FP32 cores, with Tensor Cores being added.

SkyScale and HPC in the Cloud

The high performance computing platform (HPC) used by Parabricks and incorporating the Tesla V100 Accelerators was developed by One Stop Systems and is provided in the cloud by its subsidiary, SkyScale. Configurations with 4, 8, and 16 GPUs are available, either on site from One Stop Systems or in the cloud from SkyScale.

Figure 2 shows the main elements of the most powerful of SkyScale's Accelerated Cloud Platforms, with details for a single NVIDIA Tesla V100 GPU module (see Figure 1) and for the SkyScale 16-GPU configuration.

Metric	Tesla V100	16-GPU Node
NVIDIA Tesla V100 Module with Volta GPU	1	16
Volta Streaming Multiprocessors (SM) / GPU	80	1,280
FP32 Cores / SM	64	1,024
FP32 Cores / GPU	**5,120**	**81,920**
FP64 Cores / SM	32	512
FP64 Cores / GPU	**2,560**	**40,960**
Tensor Cores / SM	8	128
Tensor Cores / GPU	**640**	**10,240**
Peak FP32 Teraflops	**15.7 Tflops**	**224 Tflops**
Peak FP64 Teraflops	**7.8 Tflops**	**112 Tflops**
Peak Tensor Teraflops	**125 Tflops**	**1,792 Tflops**
Memory Interface	4096-bit HMB2[1]	
Memory Size	16 GB	256 GB
Control processor (CPU)	Dual Intel Xeon (3.2 GHz)	
System Memory	512 GB	
System Storage	1.6 TB NVMe SSD[2]	

1. HMB2 = High Bandwidth Memory 2nd Generation
2. NVMe = Non-Volatile Memory Express Solid State Storage

Figure 2. SkyScale NVIDIA Volta V10016xP 16-GPU Accelerated Cloud Platform specification.

As shown, the 16-GPU node has a total of 81,920 32-bit cores capable of 224 teraflops per second, 40,960 64-bit cores at 112 TFLOPS, and 640 Tensor cores at nearly 1.8 peak petaflops.

With multiple GPUs, high-bandwidth efficient GPU-GPU and GPU-CPU interconnect is critical to performance for some problems. The SkyScale 16-GPU system uses PCIe Gen3 16 lane non-blocking GPU interconnect for high bandwidth GPU to GPU communication.

For even more demanding applications requiring more than a single 16-GPU node, SkyScale can interconnect a cluster of nodes using InfiniBand and Remote Direct Memory Access (RDMA) Ethernet technology.

Results

The results of using the Parabricks version of GTK4 on SkyScale with 16 NVIDIA Tesla V100 modules are shown below. NVIDIA Tesla V100 processors are also available on Amazon Web Services Cloud, albeit with lower performance compared to SkyScale at the same number of GPUs and no option for 16 GPUs; together the two factors give SkyScale more than 2:1 greater throughput.

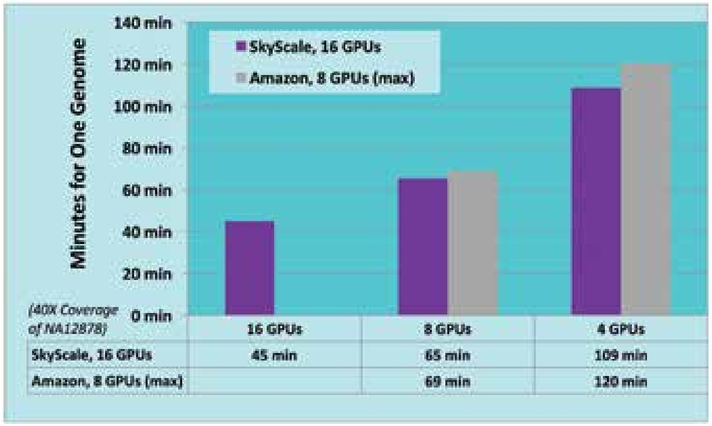

Figure 3. Minutes to analyze a single genome as a function of how many GPUs are used.

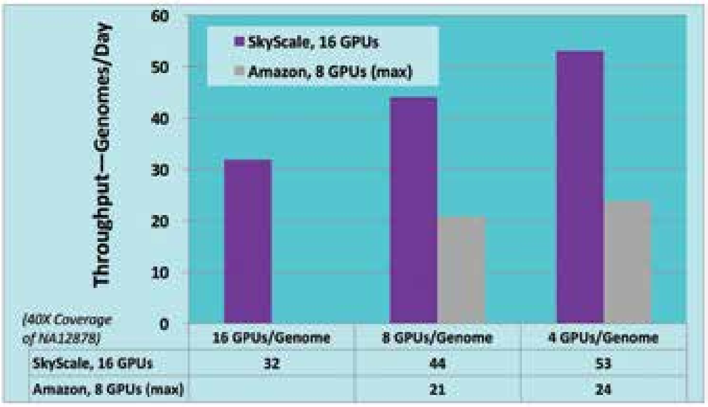

Figure 4.Throughput in genomes per 24 hours as a function of the number of GPUs used.

The figures also reveal an additional feature of the Parabricks implementation: the SkyScale 16-GPU node can be configured to use the 16 GPUs to analyze a single genome, to use eight GPUs on each of two genomes, and to use four GPUs on each of four genomes. Using fewer GPUs on a single genome increases the total runtime, but less than linearly. For example, Figure 3 shows that a single genome analysis using all 16 GPUs requires 45 minutes for one genome; on four GPUs time increases to 109 minutes, but since four analyses are done in that time, the effective time is one-quarter that or 27 minutes per genome. The same effect is seen in Figure 4, going from a throughput of 32 genomes per day to 53 per day by running four analyses at once.

SkyScale High Performance Computing in the Cloud

The advantages of the cloud for high performance computing are clear: scalability (add/subtract resource easily and immediately), elasticity (scalability plus pay-as-you-go), easy collaboration worldwide, and no need to buy equipment and hire personnel to run it.

But there are also disadvantages: high performance computing resource of the type described here have limited availability at most commercial cloud providers, security may be a concern with most commercial clouds implemented through virtual machines on multi-tenant hardware, which may also be subject to excessive latency and inconsistent performance (doubled runtimes have been reported), ability to commission specific hardware and install software may be limited, use of specific software features on a commercial cloud may lead to lock-in, and the quality of support from large vendors may vary.

These downsides are eliminated or reduced with Parabricks GATK4 running on the SkyScale Accelerated Cloud Platform access in a “bare metal” environment.

With SkyScale, the customer “rents” remote computing resources that are purchased, managed, and provisioned by SkyScale. There is no virtualized environment and no concern for multiple tenants on the same hardware. Depending on the needs of the application, data may reside permanently on storage equipment at SkyScale or may be loaded over the network to begin execution and unloaded when complete.

Because SkyScale is not subject to the impact of one customer on another, it offers direct access to the hardware, ability to make operating system and software changes, and minimal changes to customer workflow.

Bare-metal cloud service providers are generally highly focused on the needs of their particular customers and the exact resources and configurations those customers require, and can offer high levels of support consistent with that customer focus.

For security, SkyScale’s deploys enterprise-grade intrusion prevention, detection, and recovery systems and monitors them 24x7, and its datacenters have manned security 24x7, with biometric identity verification and HD camera coverage.

Finally, SkyScale requires no complex setup and no challenging configuration across multiple locations—log in and go—by the hour, week, month, or year.

Try Parabricks and SkyScale Now at No Cost

To run Parabricks adaption of the Broad Institute GATK4 Best Practices workflows with your data on a SkyScale Accelerated Cloud or onsite, and benefit from the flexibility, customizability, performance, and affordability of 16 NVIDIA Tesla V100 GPUs, with 81,920 cores delivering 224 teraflops of 32-bit floating-point and 40,960 cores at 112 teraflops of 64-bit floating point now, and 10,240 Tensor cores at ***1.8 petaflops*** in the next Parabricks release, contact info@parabricks.com or http://www.skyscale.com/contact/.

[1] Broad Institute. Genome Analysis Toolkit (GATK). GATK version 4.0 was released Jan 9, 2018. https://software.broadinstitute.org/gatk/.
[2] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLOS Biology. Public Library of Science (PLoS); 2015;13: e1002195. doi:10.1371/journal.pbio.1002195.
[3] Mell, P. M. & Grance, T. SP 800–145. The NIST definition of cloud computing. National Institute of Standards and Technology. http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf (2011).
[4] Groth P, Reuter G, Thieme S. Analysis of Genomic Data in a Cloud Computing Environment. Big Data Analytics in Bioinformatics and Healthcare. IGI Global; pp. 186–214. doi:10.4018/978-1-4666-6611-5.ch009.
[5] Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nature Reviews Genetics. Springer Nature; 2018;19: 208–219. doi:10.1038/nrg.2017.113.
[6] NVIDIA. NVIDIA Tesla V100 GPU Architecture. WP-08608-001_v1.1; 2017: https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.
[7] NVIDIA. NVIDIA Tensor Cores: https://www.nvidia.com/en-us/data-center/tensorcore/.