Accelerate Genomics Discovery in AWS with NVIDIA GPUs

aws partner network

Advanced Technology Partner

Storage Competency

The life science industry has experienced explosive growth in data generation, with conservative estimates predicting 40 exabytes of human genome data alone generated by 2025. If put to good use, this data has the potential to revolutionize precision medicine along with the drug discovery and development and genomics industry. Before meaningful insights can be gleaned from all of the data at a reasonable pace, the complexity of compute processing and storage need to be addressed.

Modern sequencing workflows produce a large amount of data- with high throughput sequencing platforms generating multiple terabytes per day and many labs run their multiple instruments 24/7. Genomic analysis requires very good, large, and small file sequential and random-access performance to support preprocessing, alignments, variant calling, joint genotyping, variant processing, and quality checking as part of the workflow. This is especially important for workflows that are accelerated by GPUs. Amazon Web Services (AWS) provides the opportunity to bring together on-demand GPU accelerated compute with genomic analytics software, plus critical storage infrastructure from WekaIO to solve the most demanding genomics research and clinical needs.

WekaFS AND NVIDIA UNRIVALED GENOMIC ANALYTICS PERFORMANCE

The combination of WekaIO's filesystem (WekaFS™) and NVIDIA Clara™ Parabricks Pipelines offer superior sequence analysis performance, enabling researchers to immediately begin running pipelines and obtain mission-critical results much faster than CPU-only solutions. Clara Parabricks Pipelines is a software suite for genomic analysis and delivers major improvements in throughput time for common analytical tasks in genomics, including germline, somatic and RNA workflows. The core of the Clara Parabricks Pipelines software starts with FASTQ data and the user can create their own analysis workflows using the individual tools within the Pipelines suite to ultimately generate a variant call file (VCF) or a genomic variant call file (GCVF).

Running on best-of-class NVIDIA® GPUs in AWS, the CUDA-accelerated Clara Parabricks Pipelines software suite replicates market leading and open-source CPU-based sequencing tools, while significantly reducing computation time between 20 to 60x compared to CPU-only workflows. This enables researchers to analyze whole genomes in as little as 40 minutes. The Clara Parabricks Pipelines suite includes several tools to compare results with other analysis tools, easily allowing researchers to verify that the analysis results are nearly identical to, to CPU generated instances.

Keeping the GPU compute cluster fed with data is critical to overall sequencing pipeline performance and efficiency. Storage based on WekaFS pushes the limits of what's possible. Not only can WekaFS be deployed on-premises, it also supports hybrid and cloud deployments and integrates seamlessly with AWS. WekaFS solves storage performance problems for the most challenging genomic sequencing environments.

UP TO 72 TIMES FASTER
Dramatically reduces time required to analyze a genome pipeline compared to CPU-only solution

HYBRID CLOUD BURSTING
Build filesystems directly from S3 data buckets in AWS, or snapshot from on premises to the cloud

CLOUD SCALE & CONVENIENCE
Go from no hardware to sequencing genomes on a new cluster in the AWS cloud in as little as 30 minutes

WORLD-CLASS FILESYSTEM PERFORMANCE

Built to solve the most demanding sequencing storage challenges, WekaFS is the world's fastest shared parallel file system delivering unmatched performance for genome sequencing at ANY scale. WekaFS delivers the highest-bandwidth, lowest-latency performance to any InfiniBand or Ethernet-enabled GPU-based compute cluster. Combined with Clara Parabricks Pipelines, WekaFS pushes common genomic analysis performance to 72x faster compared to a CPU-only cluster (Figure 1).

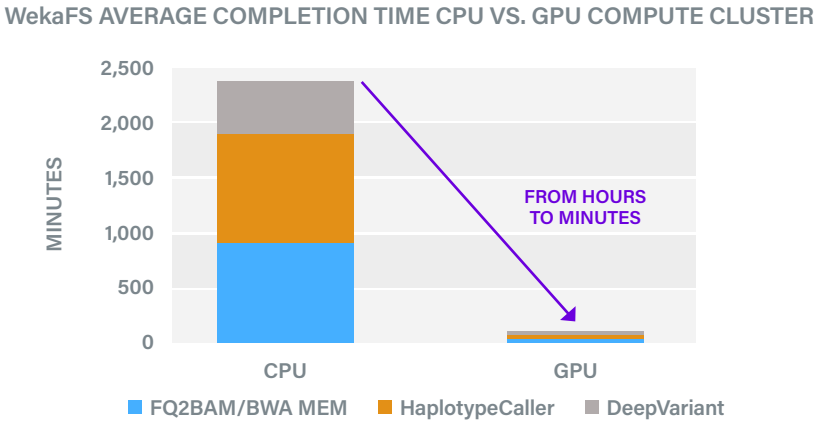

Figure 1 – WekaFS Delivers 72x better performance than a CPU only cluster

WekaFS is distributed, scale-out and POSIX compliant, and built on a modern architecture using ultra-low-latency NVMe-over-Fabrics (NVMe-oF) plus massively distributed metadata. Running on standard x86 server platforms with local NVMe SSDs, WekaFS avoids wasting precious resources waiting for data ingest. With Weka's integrated tiering to S3 object storage system, storage capacity is near limitless and addresses the exploding genomics data requirements. Valuable data is protected using patented data protection and distribution algorithms which allow the storage system to sustain up to four simultaneous node or SSD failures.

HYBRID CLOUD BURSTING

Whether working on-premise or in the cloud, WekaFS makes it easy to transition datasets between locations with just a few clicks. By connecting an S3 bucket in AWS to an on-premise filesystem, researchers can experiment locally to prove functionality prior to moving a dataset in the cloud. WekaFS makes it possible to build filesystems directly from S3 data buckets stored in AWS, removing the hassle of downloading, parsing, and filtering datasets.

CLOUD CONVENIENCE AND SCALE STREAMLINE DEPLOYMENTS

With the flexibility of AWS, researchers can go from having no hardware, to sequencing genomes on a GPU-enabled cluster in the cloud in as little as 30 minutes. With GPU clients running EC2 instances with pre-configured Amazon Machine Images (AMIs), additional compute power can easily be added to increase overall productivity. Once all results have been collected, a snapshot to an S3 bucket can be done to save cost. Figure 2 depicts a production deployment of Clara Parabricks Pipelines with WekaFS for the storage layer, on premises or in AWS.

Baylor College of Medicine

The Aiden Lab cluster required a new solution to improve application performance and facilitate the deployment of a high-performance filesystem in a cloud computing environment.

We required a solution that could support the team and their research related to Genome Architecture and felt that [legacy parallel filesystems] could not keep up with our workload.

***David Weisz,** Lead Scientific Programmer, Aiden Lab at Baylor College of Medicine*

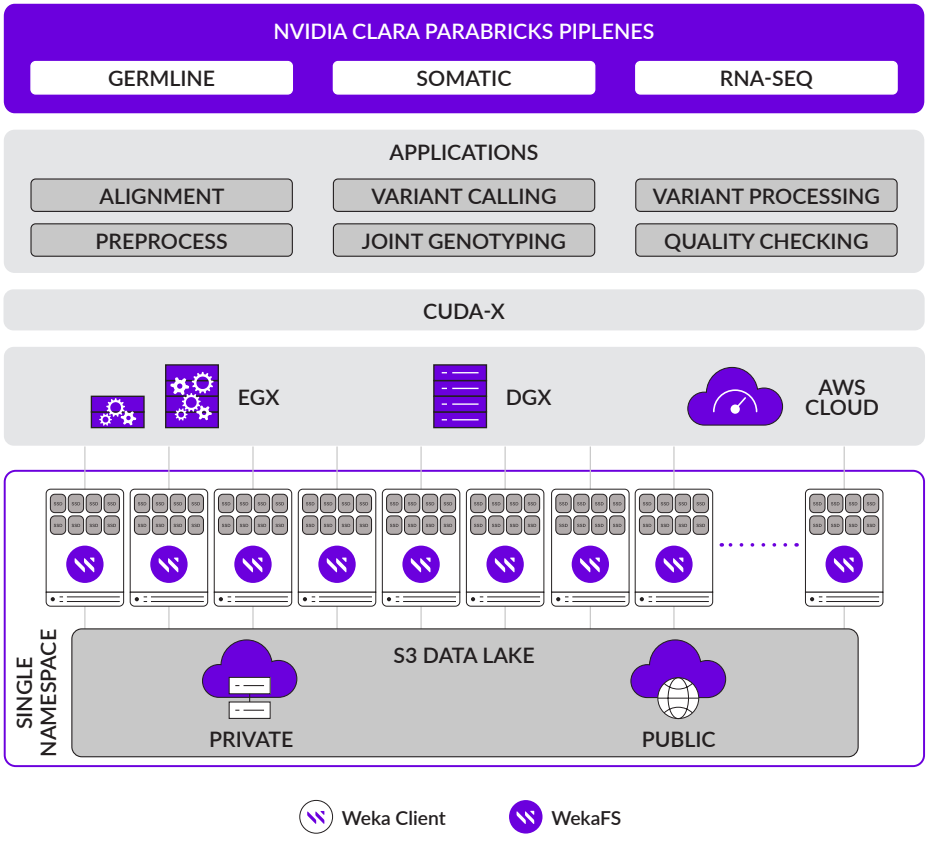

Figure 2 – WekaFS presents a performance storage layer to the GPU compute cluster

Weka Innovation Network

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io

W24SB202011