White Paper

NetApp ONTAP AI and Parabricks for Genome Sequencing Workflows

Parabricks Software with ONTAP AFF A800 Systems and NVIDIA GPUs

Karthikeyan Nagalingam, NetApp
Mehrzad Samadi and Ankit Sethia, Parabricks
April 2019 | WP-7295

In partnership with

Parabricks

Abstract

This document describes the performance of whole genome sequencing workflows using Parabricks GPU-accelerated genomic analysis of Broad Institute's Genome Analysis Toolkit pipeline on a NetApp® ONTAP® AI proven architecture, powered by NVIDIA DGX supercomputers and NetApp cloud-connected storage. The ONTAP AI system consists of a NetApp AFF A800 all-flash storage system and NVIDIA GPUs.

NetApp®

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

DNA analysis is transforming the treatment of many diseases, including cancer, Alzheimer's disease, and monogenic disorders in infants, by customizing patients' treatment based on their genetic characteristics. A patient's genetic information extracted by using DNA analysis can help to determine the most suitable method for curing the disease. This personalized approach to medicine, which promises to increase human longevity and quality of life, is beginning to evolve for a gamut of diseases. However, next-generation sequencing (NGS) is restricted to using limited DNA data in hospitals and labs because of the time and cost to process the whole genome sequencing (WGS) data and the large storage resource that it requires.

Research has shown that limited DNA data is not sufficient, and data from WGS is a crucial requirement for accurate and detailed analysis of the underlying disease. This limitation has led to hospitals using WGS data in their diagnostic processes by using techniques that carry prohibitive costs and time for rare cases. WGS has not been widely adopted due to the time consumed by the data processing. Because the data per patient can be 300GB to 1TB, processing can take several days. The number of patients who could benefit from analysis of their DNA data is expected to be about one billion by 2026. To serve this large number of patients will require a major breakthrough in WGS processing times and storage capacity.

2 Parabricks Overview

Parabricks provides high-performance GPU computing and deep-learning technologies that are tailored for NGS analysis. The accelerated software is a drop-in replacement of industry-standard tools that does not sacrifice output accuracy or configurability. Parabricks provides 30 to 50 times faster secondary analysis of FASTQ files coming out of sequencer to variant call files (VCFs) for tertiary analysis compared with conventional CPU-based approaches.

3 ONTAP AI Overview

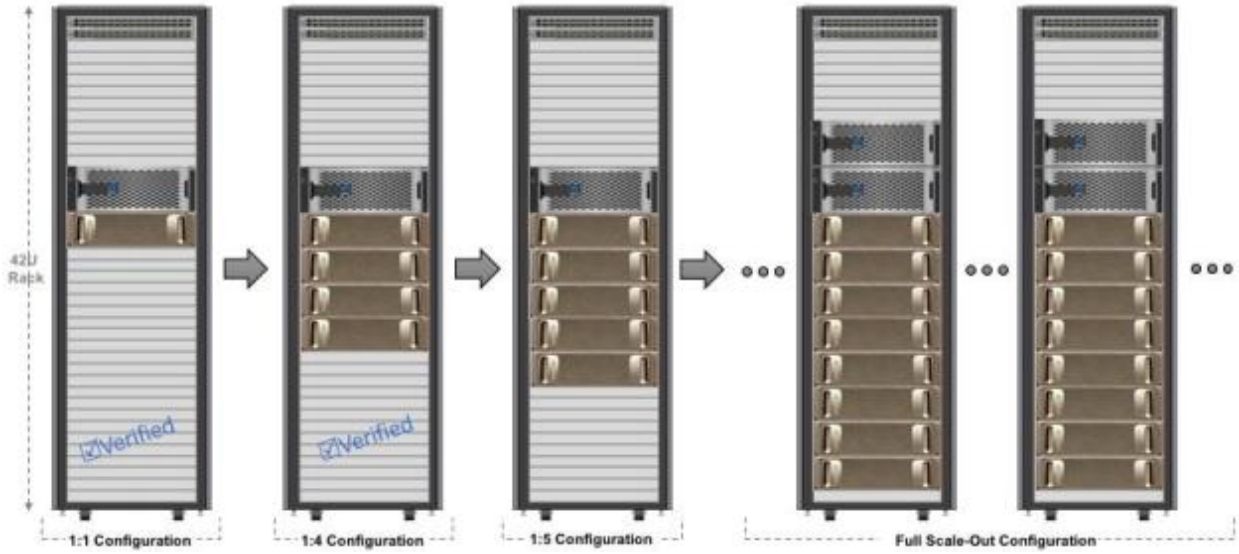
NetApp ONTAP AI proven architecture, powered by NVIDIA DGX supercomputers and NetApp cloud connected storage, was developed and verified by NetApp and NVIDIA. It offers organizations a prescriptive architecture that provides the following benefits:

- Eliminates design complexities
- Permits the independent scaling of compute and storage
- Can start small and scale seamlessly
- Offers a range of storage options for various performance and cost points

ONTAP AI integrates NVIDIA DGX-1 servers with NVIDIA Tesla V100 GPUs and a NetApp AFF A800 system with state-of-the-art networking. ONTAP AI simplifies artificial intelligence deployments by eliminating design complexity and guesswork, enabling enterprises to start small and grow nondisruptively while intelligently managing data from edge to core to cloud and back.

Figure 1 shows the scalability of the ONTAP AI solution. The AFF A800 system has been verified with four DGX-1 servers and has demonstrated sufficient performance headroom to support five or more DGX1 servers without affecting storage throughput or latency. By adding more network switches and storage controller pairs to the ONTAP cluster, the solution can scale to multiple racks to deliver high throughput and accelerate training and inferencing. This approach offers the flexibility of altering the ratio of compute to storage independently according to the size of the data lake, the deep learning (DL) models used, and the required performance metrics.

Figure 1) ONTAP AI solution rack-scale architecture.

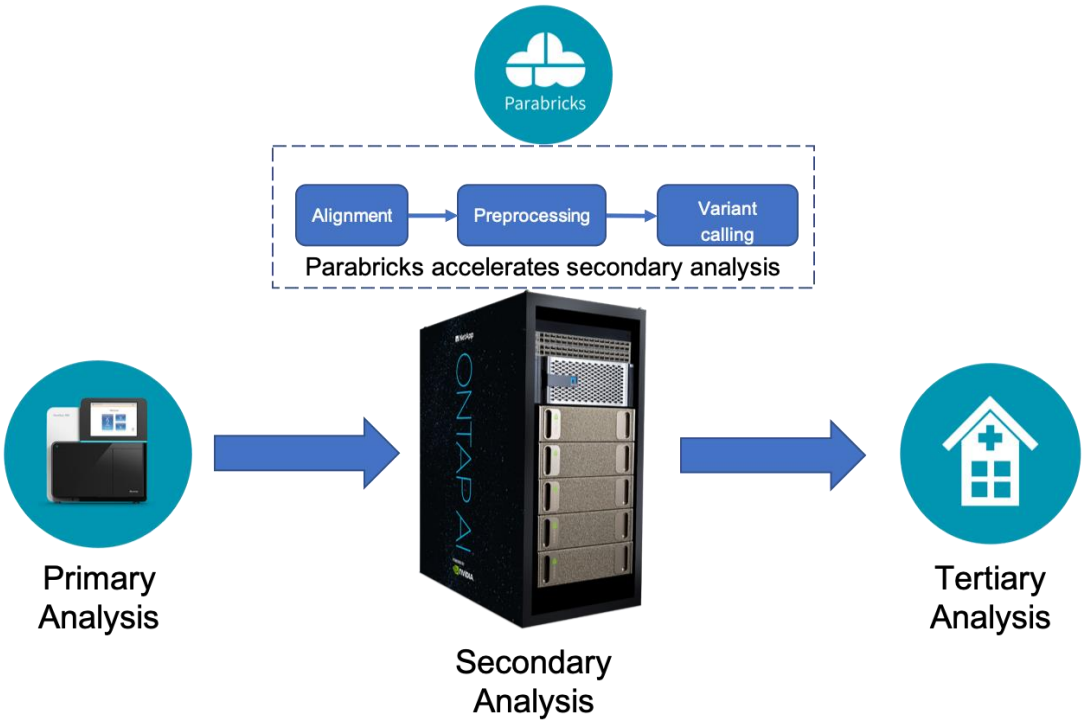


The number of DGX-1 servers and AFF systems that can be placed in a rack depends on the power and cooling specifications of the rack in use. Final placement of the systems is subject to computational fluid dynamics analysis, airflow management, and data center design.

4 ONTAP AI with Parabricks Solution

Figure 2 illustrates ONTAP AI with Parabricks primary, secondary, and tertiary analysis.

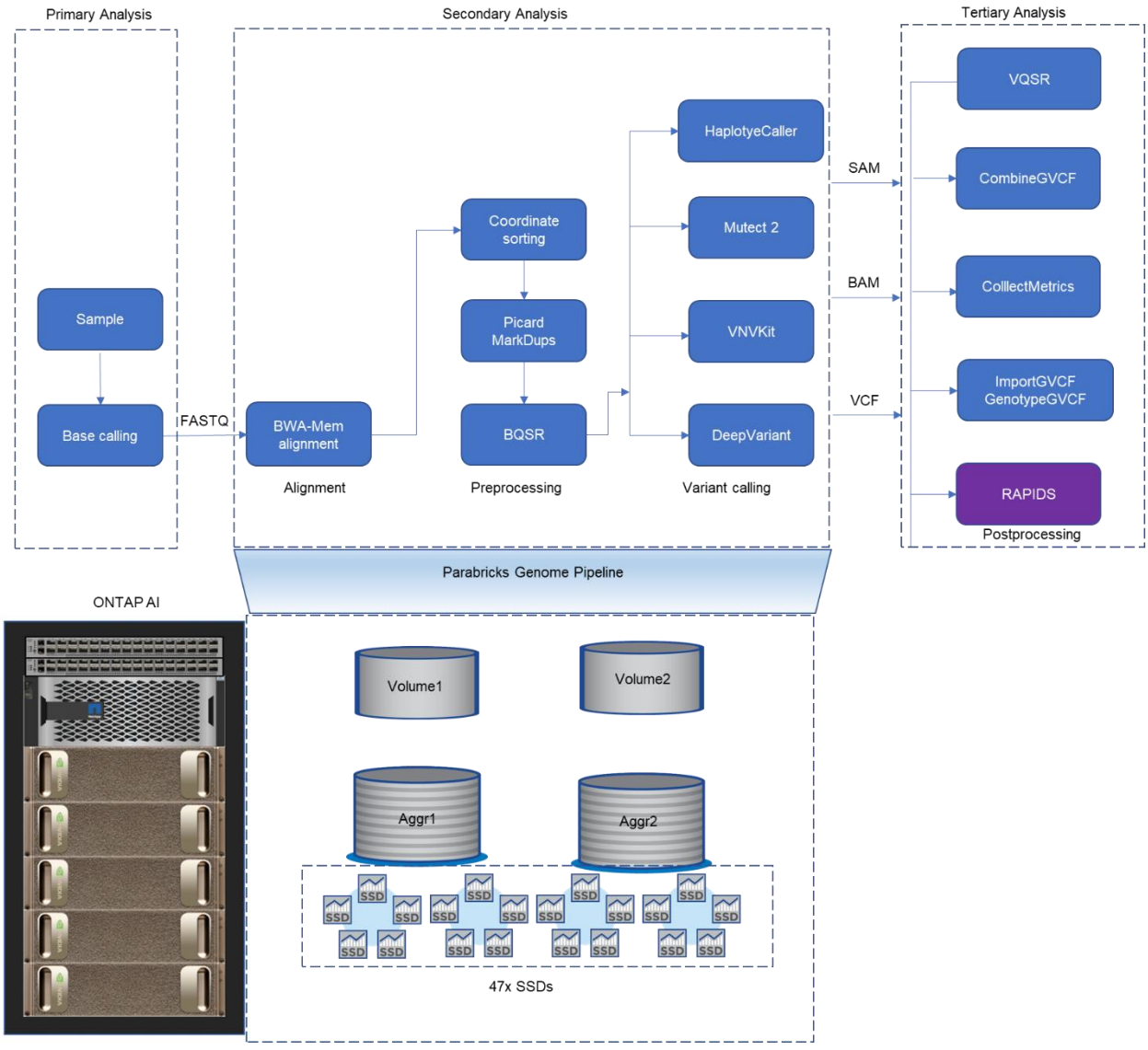
Figure 2) ONTAP AI with Parabricks solution.



NetApp completed genome analysis performance validation by using Parabricks software in ONTAP AI with a single NVIDIA DGX-1 box. Genome analysis is customarily divided into three stages: primary, secondary, and tertiary analysis. The primary analysis includes taking the human sample, cleansing the sample, and providing the fastq format of data by base calling operation. The secondary analysis includes aligning, preprocessing, and variant calling. The tertiary analysis includes postprocessing, where the NVIDIA RAPIDS platform now plays a major role.

Figure 3 is a detailed image of the genome pipeline. Parabricks accelerates secondary analysis by using GPUs for computing both on the premises and in the cloud. Parabricks runs an entire analysis by using a single DGX-1 node, which significantly reduces the computing cost. Parabricks uses the same algorithms as the Broad Institute's Genome Analysis Toolkit (GATK) pipelines to perform the genome analysis, but it provides 30 to 50 times faster results by using 8 GPUs versus 32 vCPUs.

Figure 3) Genome pipeline with ONTAP AI and Parabricks.



5 ONTAP AI and Parabricks Performance Validation

This section describes the solution validation with Parabricks software in ONTAP AI with five data samples and characterizes their results in terms of speed, accuracy, and throughput.

5.1 Comparative Pipeline Performance: GPU Versus CPU

The total execution time of the GATK best practices pipeline can be dramatically reduced by as much as 50 times by using the Parabricks software on ONTAP AI, which is composed of NVIDIA DGX-1 and an ONTAP AFF A800 storage system. The WGS workflow that normally takes 1.5 days on a CPU-only node can be processed on a GPU-enabled node in less than an hour.

Figure 4 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. NetApp performed the validation in February 2019. For each of the five workflows, the performance scales linearly with the number of GPUs used, and the speedup of Parabricks with eight GPUs ranges from 48x to 54x. Figure 5 shows the average GPU-enabled performance improvement for each number of GPUs, which ranges from an average speed of 10X with only 1 GPU to 50X with 8 GPUs.

Figure 4) Speedup of Parabricks pipeline over CPU-equivalent pipeline.

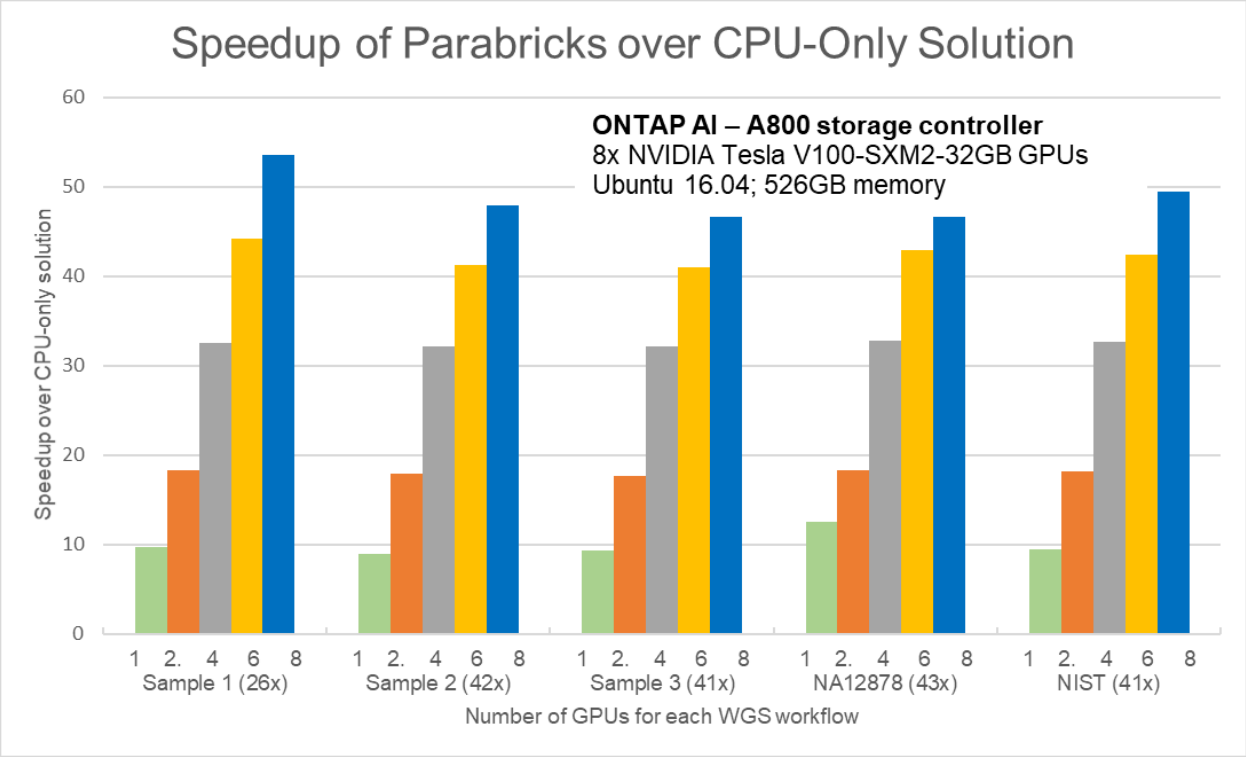
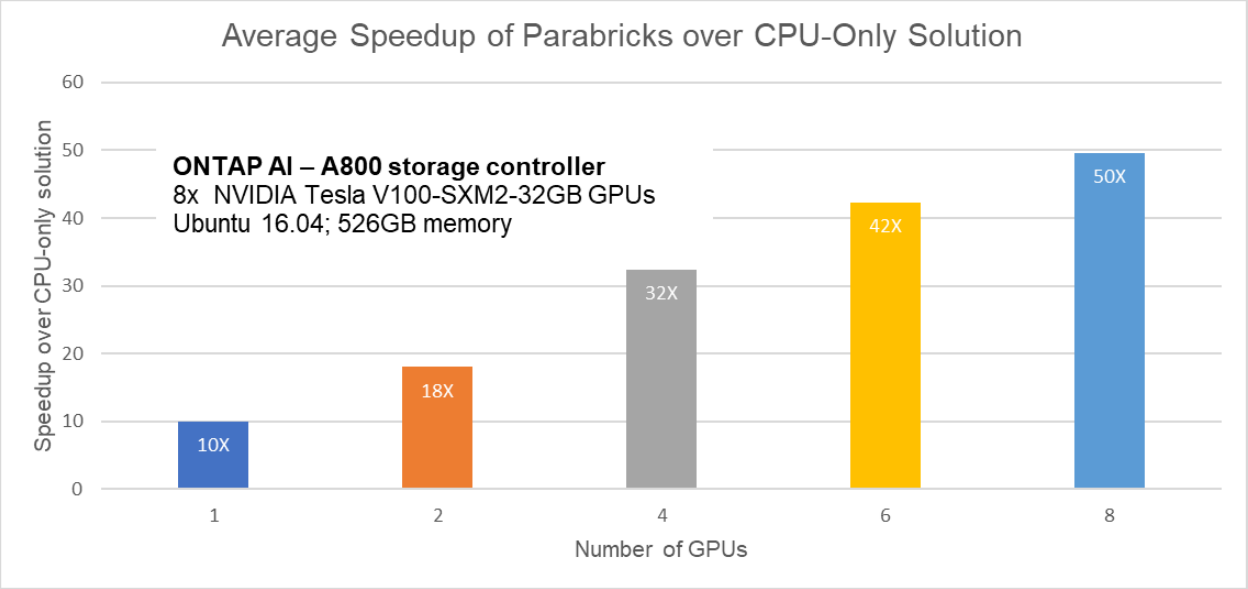


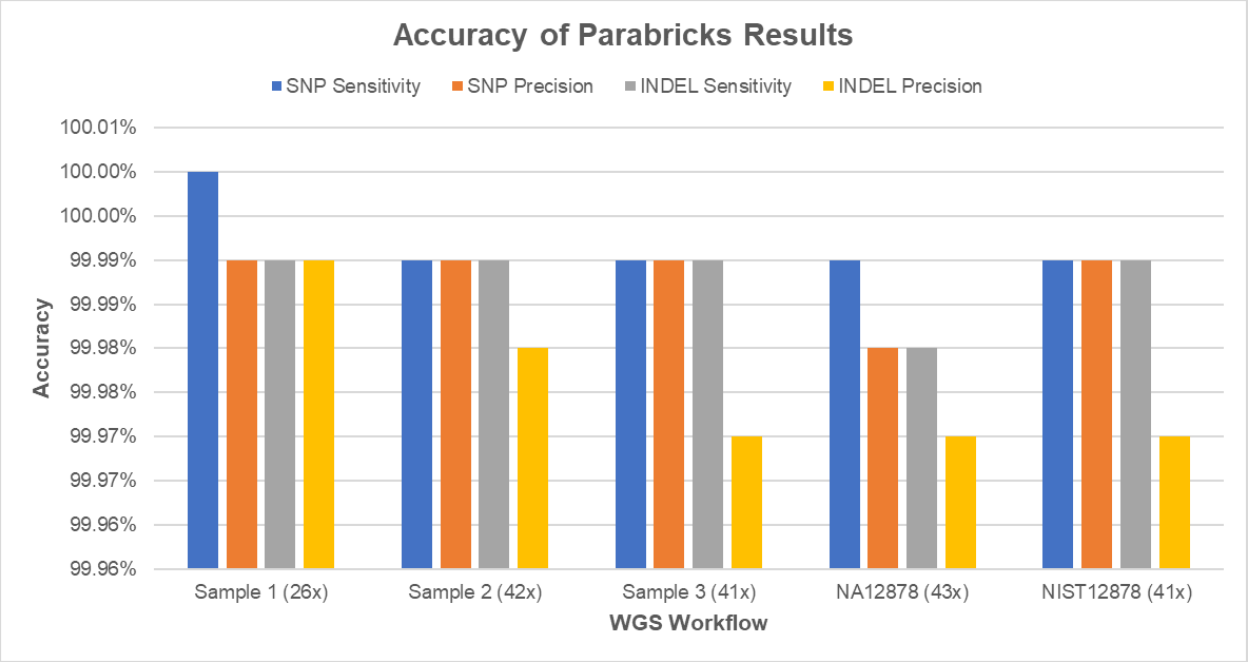
Figure 5) Average speedup of Parabricks pipeline over CPU-equivalent pipeline.



5.2 Accuracy

While achieving faster performance, the GPU-enabled Parabricks software produces intermediate output files from the GATK4 BaseRecalibrator and GATK4 ApplyBQSR tools, which 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. It should be noted that the baseline variant caller in GATK4 is nondeterministic and can generate slightly different results depending on certain run-time parameters, such as number of threads, so the differences are consistent with these variations in GATK4 execution.

Figure 6) Accuracy of Parabricks results compared to CPU-equivalent results.



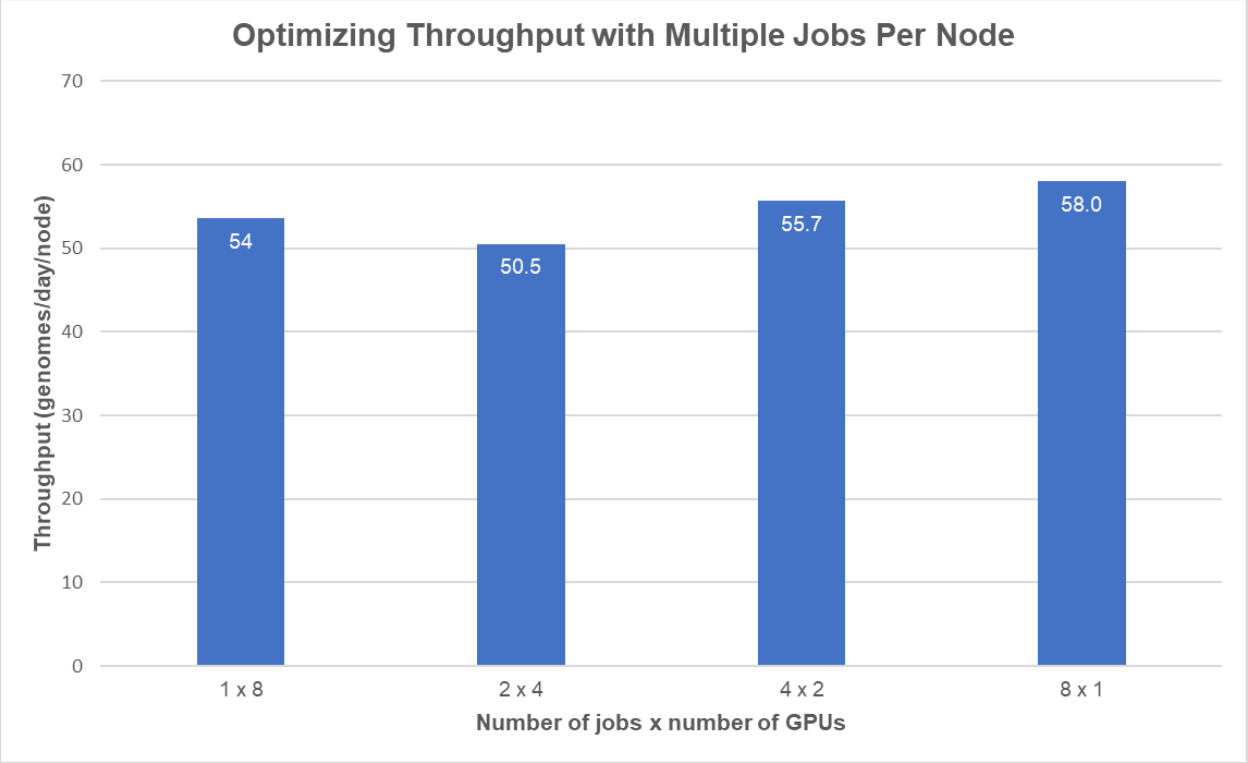
5.3 Throughput and Minimizing Time

This section describes four ways to run the GATK pipeline on an eight-GPU compute node and the resulting throughput rate in genomes per day, per node using Sample 1 (26 X) test case.

- **First job.** One worker and eight GPUs (1 x 8 GPUs); the resulting throughput is 54.0 genomes per day, per node.
- **Second job.** Two workers and four GPUs per worker; the resulting throughput is 50.5 genomes per day, per node.
- **Third job.** Four workers and two GPUs (4 x 2 GPUs) per worker; the resulting throughput is 55.7 genomes per day, per node.
- **Fourth job.** Eight workers with one GPU (8 x 1 GPU) per worker; the resulting throughput is 58 genomes per day, per node.

Figure 7 illustrates throughput optimization by running multiple jobs on a node.

Figure 7) Optimizing throughput by running multiple jobs on a node.



5.4 ONTAP AI for Genomics

ONTAP AI is a NetApp appliance that integrates NVIDIA DGX-1 systems. NetApp has partnered with NVIDIA and independent software vendors in the genomics industry to provide improved return on investment in their productivity. NetApp and NVIDIA have done multiple deployments in healthcare industries and independent software providers for better performance in scale-up and scale-out deployments. An AFF A800 storage system in ONTAP AI runs with ONTAP data management software, which has provided enterprise management features for more than 25 years. This management functionality helps genome analytics in terms of backup, restore, disaster recovery, duplicating genome analysis, quality of service, and eliminating duplicate copies by using NetApp storage efficiency. With NetApp ONTAP AI, you can run workloads of thousands of genomes, a capability that is not achievable in local disk GPU nodes.

Where to Find Additional Information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NVA-1121-DEPLOY: NetApp ONTAP AI, Powered by NVIDIA
  https://www.netapp.com/us/media/nva-1121-deploy.pdf
- ONTAP 9 Documentation Center
  http://docs.netapp.com/ontap-9/index.jsp
- NetApp Product Documentation
  https://www.netapp.com/us/documentation/index.aspx
- Parabricks website
  https://www.parabricks.com/

Version History

Version	Date	Document Version History
Version 1.1	April 2019	Minor updates.
Version 1.0	March 2019	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Trademark Information

WP-7295-0419