GOAST: Genomics Optimization And Scalability Tool.

An Intel® Select Solution for Genomics Analytics Enabling Precision Medicine at the Population Scale.

Authors: Mileidy Giraldo, Ph.D.[a], Kevin Dean[a], Florian Merz. Ph.D.[a].
[a]Genomics R&D, HPC and AI, Lenovo

Powered by Intel®

Lenovo

Introduction

The development of Next-Generation Sequencing (NGS) technologies in the late 2000's led to a dramatic decrease in the cost of DNA sequencing. The advent of NGS coupled to the advancements in High Performance Computing (HPC) storage and computing technologies at the time created the perfect storm for a deluge of genomics data. This data deluge is propelling the birth of Precision Medicine, which aims to deliver individualized prevention, diagnosis, and treatment by leveraging knowledge from a person's specific genomic and environmental backgrounds.

Given the new affordability of NGS methods and the increased computing and storage capacities of the last decade, genomics can now be performed at the population level. Large national genomics initiatives such as the "UK Biobank", the "All of Us program" in the US, Singapore's "GenomeAsia", "Genomics Thailand", etc. are emerging all around the world. With goals of sequencing in the range of 500K to over 1M participants in a few years' time, these country-wide efforts aim to capture the genetic variation of their people to make Precision Medicine a reality.

The greatest challenges population genomics efforts face are scale and time. Population genomics requires scaling up in input data from exomes (the portions of a genome that code information for protein synthesis) to whole genomes, scaling up production levels (from a handful to tens of thousands of samples), and having to do so under very short time frames. Three out of the four stages in population sequencing take place in the HPC environment of a cluster or supercomputer, including genome assembly (assembling the DNA "letters" into words), variant analyses (comparing how a gene is "spelled" in different people), and downstream bioinformatics (e.g. measuring the effect of variations on function or disease). Therefore, scaling out population genomics productions in a timely fashion largely depends on the HPC technologies and the underlying acceleration they can offer.

Even with today's technologies, it is still taking genomics datacenters around the world typically 150-160 hrs. to process a single whole genome and 4-6 hrs. for an exome. In 2017, Intel partnered with the Broad Institute to search for technological innovations that would reduce genomics analytics times while retaining the scientific rigor and accuracy scientists have come to know and expect of the Broad's Genome Analysis Toolkit (GATK)[1]. Key outcomes of the Broad-Intel partnership include open-sourcing GATK, the Genomics Kernel Library (including AVX-512 optimizations of the Smith-Waterman and Pair-HMM algorithms), the GenomicsDB variant store for population analytics, and Intel-recommended hardware and workflow configurations. Since its release in 2017, Intel's reference architecture has afforded users of GATK's Germline Variant Calling Workflow accelerated performance by reducing runtimes to 10.8 hrs. and 25 min. for whole-genomes and exomes respectively.

Here, we describe how Lenovo validated and improved on Intel's original solution as we tested additional optimizations for increased genomics performance. Our search for a genomics solution prioritized identifying the best hardware, software and workflow combination for performance, cost, and usability. To that end, Lenovo's Genomics R&D team conducted a comprehensive systematic study of the performance of many parameters on 30+ tools in GATK against an extensive set of permutations of hardware building blocks, system tunings, data types, execution modes, and software implementations. As a result, we have become the first solution provider to validate the Intel® reference architecture successfully. Moreover, our benchmarks show that Lenovo's optimizations improve the performance of the original Intel® Select Solution for Genomics Analytics. GOAST, Lenovo's genomics reference architecture, provides a 27X to 40X performance improvement compared to typical runtimes at genomics datacenters around the world. As we will detail below, GOAST leverages non-specialty hardware, our system optimizations, and the Broad Institute's open-source software to deliver a solution that is affordable, scalable, and conducive to population-level analytics.

Methods

Input Datasets. The present study analyzed both Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) versions of the NA12878 genome—a well curated reference human genome representing a high-quality gold standard in the Genomics field[2].

Systematic study as a permutation test. A comprehensive study systematically evaluated the effect of hardware components, system tunings, and tool parameters on the performance of the Broad Institute's GATK Single-Sample Germline Variant Calling workflow. Our study evaluated a range of software versions from legacy releases through the latest versions released as of April 2019. The information listed in Table 1 represents the tool versions in our best-performing pipeline. The genomics workflow evaluated here involves a sequential pipeline calling 12 main software steps (Figure 1) that require 7 software suites executing 30+ tools. See Figure 1 for an overview of a typical NGS analysis from sample to variant output. Each tool, in turn, features a variable number of software parameters that can be tuned for performance. Our tests made no code-level changes to any of the tools in the workflow thus the scientific integrity of the software tools

Table 1: Software suites in GOAST's best-performing benchmarks

Software	Version	Release Date
BWA	0.7.17	2018-07-12
Cromwell	29	2017-08-16
GATK	4.1.2.0	2019-04-23
JAVA JDK	8u181	2018-07-07
Picard	2.20.0	2019-04-29
Python	3.6	2016-12-23
Samtools	1.9	2018-07-18

performance in two execution modes: latency vs. throughput. The latency runs evaluated one permutation of the workflow per node (*i.e.* all resources of 1 node assigned to executing only 1 WGS/WES workflow), whereas the throughput runs tested 2-100 jobs running concurrently on a single node.

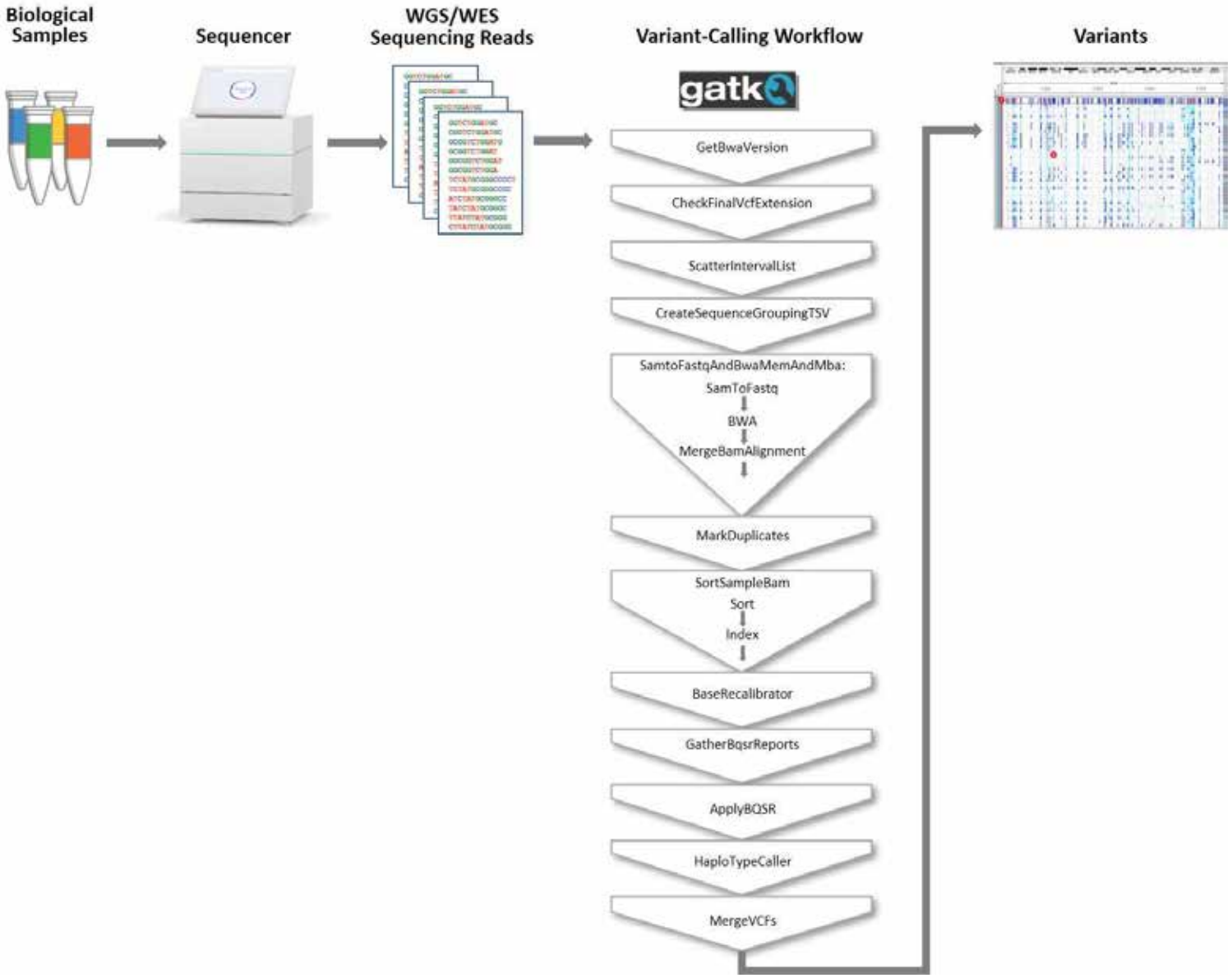

Figure 1: A typical Next-Generation Sequencing (NGS) workflow illustrating the sequential software calls required to perform WGS/WES variant-calling analyses. Biological samples (*i.e.* blood, saliva, etc.) processed experimentally are input into a sequencer, which in turn generates sequencing reads (fragments of DNA strings). The sequencing reads become the input for the variant calling workflows. The GATK Single-Sample Germline Variant Calling workflow shown here consists of 12 main software calls using 7 different software suites that in turn execute 30+ tools. The output of the genomics workflow feeds into variant analyses (comparing how a gene is "spelled" in different people), and downstream tertiary bioinformatics work (*e.g.* measuring effect of variations on function or disease).

Measuring performance. The present study measured performance as execution time for the entire workflow, from *.ubam input to the *.vcf output. Our search for a “hardware + software + system recipe” prioritized identifying the best solution for performance, cost, and usability.

Hardware evaluations. Two different Intel® Xeon® Scalable processors were used for all full workflow analyses; Intel® Xeon® Gold 6148 and Intel® Xeon® Platinum 8168 processors. An array of storage options was likewise tested, including local NVMe drives, local SSDs, local HDDs, and a few variants of Lenovo’s Distributed Storage Solution for IBM Spectrum Scale (DSS-G). Intel Select Solutions for Genomics Analytics take advantage of the high-performance capabilities of Intel® architecture, including Intel® Xeon® Scalable processors and Intel® SSD Data Center Family drives. Solutions incorporating the latest Intel® Xeon® Gold 6252 processors, Intel® Xeon® Gold 6226 processors, and Intel® Xeon® Platinum 8280 processors deliver the same performance or incremental performance gains as compared to similarly configured solutions based on previous-generation Intel® Xeon® Scalable processors.

Results

GOAST: A Two-pronged HPC-Architecture plus HPC-Scaler Solution

Determining the factors affecting the performance of genomics workflows is a complex problem. Genomics workflows string together a combination of 30+ single-threaded, non-distributed parallel, and distributed tools. Given the heterogeneity of the genomics tools and their corresponding profiling characteristics, the path to an optimal hardware recipe is not obvious. To date, most of the guidance in the research community regarding improving genomics HPC boiled down to speculations of the role of high memory, high storage speeds, and/or high core counts as possible indicators of performance.

Given the lack of consensus among genomics developers and the scientific computing staff that supports them, Lenovo set out to systematically evaluate as many of these dependencies as possible and the effect of permutations of hardware, software versions, tool parameters, execution modes, system tunings, and data types on the performance of GATK workflows. Lenovo’s search for a “hardware + software + system” recipe prioritized identifying the best solution for performance, cost, and usability. Such a comprehensive evaluation led to hundreds of simulations on many different hardware configurations spanning over a year. The study yielded two resources for deploying and scaling HPC for Genomics: GOAST Architecture and GOAST Scaler.

GOAST Architecture delivers a 27X to 40X speed-up in Genomics

As a result of Lenovo’s permutation tests of the hardware, software, and system factors affecting the performance of genomics workflows we identified an optimized architecture that can process 1 whole genome in 5.5 hrs. and 1 exome in 4 minutes with no specialty hardware (Table 2). With Lenovo’s GOAST Architecture, a data center can expect to process 4.5 genomes or 343 exomes per node per day; thus, gaining a 27X to 40X performance improvement over typical runtimes at genomics datacenters worldwide. GOAST Architecture leverages an optimized variant-calling workflow and a concise, simple, non-specialty hardware recipe to deliver an affordable solution with peak performance.

Table 2. GOAST Architectures: Two "Hardware + Software + System" Recipes*

Ingredient	1-node Configuration	4-node Configuration
Expected Annual Productivity	1.5K WGS/yr.*	6K WGS/yr.*
Compute Node		
Server Type	1x ThinkSystem SR630 Rack Server (1U)	4x ThinkSystem SD530 Dense Server (2U/4N)
Processor	2x Intel® Xeon® Platinum 8268 processors (24 cores, 2.7GHz) or Intel® Xeon® Gold 6248 processors (20 cores, 2.5GHz)	2x Intel Xeon Platinum 8268 processors (24 cores, 2.7GHz) or Intel Xeon Gold 6248 processors (20 cores, 2.5GHz)
Memory	12x ThinkSystem 32GB TruDDR4 2933MHz (2Rx4 1.2V) RDIMMs	12x ThinkSystem 32GB TruDDR4 2933MHz (2Rx4 1.2V) RDIMMs
Local Storage	4x 2.5" Intel® SSD D3-S4610 960GB Mainstream SATA 6Gb Hot Swap SSDs 1x 2.5" Intel® SSD D3-S4610 240GB Mainstream SATA 6Gb Hot Swap SSD (Boot) Up to 46TB	1x 2.5" Intel SSD D3-S4610 240GB Mainstream SATA 6Gb Hot Swap SSD (Boot) Customizable Local Storage Capacity: Up to 9TB
Data Network Adapter		1x ThinkSystem D2 10Gb 8 port EIOM Base T RJ45 (one per 4 nodes)
Host Adapter		1x Mellanox ConnectX-5 EDR IB/100GbE VPI Single-Port x16 PCIe 3.0 HCA
Host Network Adapter		Integrated 1 gigabit Ethernet (GbE)
Network Infrastructure		
Data Network		1x Mellanox SB7800 EDR IB Switch (36 port, 100 Gbps, 1U)
Management Network		1x ThinkSystem NE0152T RackSwitch (48 ports, 1 Gbps, 1U)
Management Node		
Server Type		1x ThinkSystem SR630 Rack Server (1U)
Processor		2x Intel® Xeon® Gold 6226 processors (12 cores, 2.7GHz)
Memory		12x ThinkSystem 8GB TruDDR4 2933MHz (1Rx8 1.2V) RDIMM
Local Storage		2x 2.5" Intel SSD D3-S4610 240GB Mainstream SATA 6Gb Hot Swap SSDs (Boot)
Host Adapter		1x Mellanox ConnectX-5 EDR IB VPI Dual-port x16 PCIe 3.0 HCA
Storage Infrastructure		
Solution Type		Lenovo Distributed Storage Solution for IBM Spectrum Scale, DSS-G201
Server Type		2x ThinkSystem SR650 Rack Servers (2U)
Processor		2x Intel® Xeon® Gold 6240 processors (18 cores, 2.6GHz)
Memory		12x ThinkSystem 32GB TruDDR4 2933MHz (2Rx4 1.2V) RDIMMs
Local Storage		2x 2.5" Intel SSD D3-S4610 240GB Mainstream SATA 6Gb Hot Swap SSDs (Boot)
Host Adapters		2x Mellanox ConnectX-5 EDR IB VPI Dual-port x16 PCIe 3.0 HCA
JBOD Enclosure		1x Lenovo Storage D1224 Disk Exp Enclosure (2U)
JBOD Storage Drives		24x Lenovo Storage 800GB 3DWD 2.5" SAS SSDs Customizable JBOD Storage Capacity: Up to 44TB
Software		
	• GATK, BWA, and GATK workflows optimized for Intel technologies • Optimized Cromwell workflow[3] • Intel Genomics Kernel Library (Intel GKL) with optimized routines for accelerating developer codes • Slurm job scheduler for running clustered analytics jobs • IBM Spectrum Scale, high-performance cluster file system	• GATK, BWA, and GATK workflows optimized for Intel technologies • Optimized Cromwell workflow[3] • Intel GKL with optimized routines for accelerating developer codes • Slurm job scheduler for running clustered analytics jobs • IBM Spectrum Scale, high-performance cluster file system
Firmware Optimizations		
	• Intel® Advanced Vector Extensions 512 (Intel® AVX-512) • Intel® Turbo Boost Technology – enabled • Intel® Hyper-Threading Technology (Intel® HT Technology) – enabled • Lenovo Scalable Infrastructure Release 19B Best Recipe	• Intel Advanced Vector Extensions 512 (Intel AVX-512) • Intel Turbo Boost Technology – enable • Intel Hyper-Threading Technology (Intel HT Technology) – enabled • Lenovo Scalable Infrastructure Release 19B Best Recipe

**Contact Lenovo to customize GOAST's configurations for other workload sizes.*

GOAST Scaler can Customize the Intel® Select Solution Architecture

Another byproduct of Lenovo's systematic genomics performance testing was the ability to generate a customizable architecture for genomics. While the reference architecture in Table 2 provides base configurations for small datacenters processing 1.5K and 6K WGS/year (1-node vs. 4-node configurations), we are aware that every genomics data center adopts a different mix of workloads, analyses workflows, has different active and archiving storage needs, a different mix of research types to support, and therefore needs a customized architecture tailored to their specific needs. Thus, we converted the lessons learned from our genomics benchmarking and systematic testing into formulas in GOAST Scaler—a tool for sizing and scaling HPC for Life Sciences workloads.

GOAST Scaler calculates the projected HPC usage for an expected workload; for example, it outputs the compute nodes, active, and archive storage needed to meet a workload quota (*e.g*. 50K genomes/yr.). GOAST Scaler can also be used to size the current production capabilities of an existing cluster: *e.g.*, to answer the question of how many genomes can I process with my current cluster? Or, how many genomes/yr. can this year's budget afford me?

Lenovo's Genomics experts work with your researchers to understand your initiative's expected workloads, growth plans, and data management policies. The Lenovo team uses this information to create HPC designs and usage projections to simulate how data will grow and populate the cluster over time. These exercises in HPC usage and projections are proving invaluable in workload management, budget planning, IT expenditure justification and allocation, and resource accountability.

Benefits of Lenovo's GOAST Solution

To Researchers and Developers

- **Acceleration.** Optimized workflows process genomes faster thus reducing execution times and decreasing time to scientific insight.
- **Budget planning.** The HPC usage projections by GOAST Scaler can help your funding agencies understand your expected IT expenditures and justify your proposed budget.
- **Wide acceptance in the scientific community.** Using open source software means that the resulting publications will not be subject to black-box algorithms problematic for peer-review or suspect in scientific accuracy.
- **Customizable for any Life Science workload.** Lenovo's experts are available to size HPC for any mix of workloads and are open to optimizing other bioinformatics workflows of interest.

To Funding Agencies and Scientific Administrators

- **Resource planning and funding allocation.** GOAST Scaler generates HPC usage projections that can help stakeholders plan resource and funding allocation.
- **Accountability**. The HPC usage designs by GOAST can help leadership hold researchers and developers accountable to budget plans and milestone deliverables.
- **Affordability.** GOAST Architecture delivers faster processing speeds with non-specialty hardware thus lowering costs and increasing purchasing potential.

To HPC and IT Departments supporting scientific computing

- **Assists HPC management.** The HPC usage designs by GOAST will simulate the growth of Life science-related data over time and provide alternative models for data flow, storage, and management across the cluster.
- **Scalable and flexible.** Lenovo's experts are available to engage your researchers and developers to understand your workloads, growth plans, and data management policies. GOAST Scaler can customize the solution in Table 2 for any workload size or Life Sciences application mix.

To OEM partners

- **Lenovo's strong scientific and benchmarking expertise in-house.** With a vertical team of experts and expanding partnerships, including genomics researchers, performance engineers, and software developers Lenovo can support your efforts by bringing technical depth to your interactions and by facilitating technical conversations with customers.
- **Simplifies Genomics Analytics.** A solution encompassing architecture and sizing capabilities makes it easier for integrators of HPC to break into the genomics space faster without having to hire vertical technical expertise.

Summary

The Intel-Broad Center for Genomic Data Engineering worked to optimize GATK on Intel architecture and technologies and to define a reference architecture for genomics analytics. The result was Intel Select Solutions for Genomics Analytics, developed by Intel and the Broad Institute and delivered by Intel solution providers like Lenovo. The original Intel solution demonstrated a 5X overall performance improvement running GATK. The Broad Institute certified the performance level and quality of results by Intel's original solution.

The work described here shows how Lenovo extended Intel's original solution and improved it further attaining 27X to 40X performance improvement on GATK Whole-Genome and Whole-Exome Sequencing, respectively. Lenovo's work in genomics resulted in two resources offered together as the GOAST solution. On the one hand, GOAST Architecture bundles a "hardware + software + system-tuning" recipe for Genomics HPC; on the other, GOAST Scaler calculates the HPC infrastructure required for scaling out Life Sciences workloads for any user-defined volumes.

Lenovo is leveraging the GOAST solution to help data centers around the world accelerate their workflows and plan their HPC resources more effectively as they embark on ever increasing workloads from cohort-level and population-level genomics projects. We have experts who can advise you on a complete, end-to-end deployment of population-level genomics; from workload planning, to cluster sizing, to accelerating secondary and tertiary NGS workflows. Lenovo's commitment to developing and adopting cutting-edge technological innovation is enabling the worldwide movement to sequence entire populations and bringing such initiatives closer to making Precision Medicine a reality.

To learn more about how Lenovo's GOAST solution can help you accelerate your genomics workloads and scale your HPC infrastructure:

Contact us
at mgiraldo@lenovo.com

or Visit
Intel Select Solutions: intel.com/selectsolutions
Intel Xeon Scalable Processors: intel.com/xeonscalable

Intel Select Solutions are supported by Intel® Builders: http://builders.intel.com.
Follow us on Twitter: #IntelBuilders

1. Van der Auwera GA et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. 2013. Curr Protoc Bioinformatics. 43:11.10.1-11.10.33.
2. Zook J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32, 246–51 (2014).
3. The optimized workflow will be available as a containerized workflow to run on Lenovo hardware. Optimal settings for genomics analyses are consequential and unique to the data type and execution mode of the analysis. Therefore, system tunings may require some customization.