Genomics Optimized Servers Accelerate COVID-19 Discovery with High-Throughput Analytics

Technical White Paper

1 October 2020

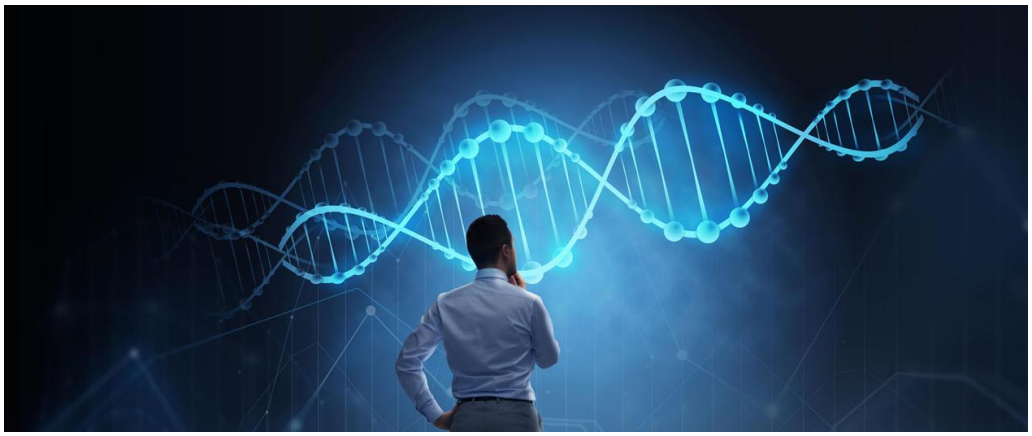
Mileidy Giraldo, Ph.D.,
Global Lead, Life Sciences,
HPC and AI
Lenovo Data Center Group

Lenovo.com/systems

Powered by Intel®

As of today, Johns Hopkins University reports over 31 million confirmed cases of COVID-19 and close to 1 million deaths in 188 countries around the world[1]. The effects of the COVID-19 pandemic are still permeating every sector of society and changing how we live, learn, work, and interact with others. With COVID-19 statistics still climbing, understanding the novel coronavirus and developing an effective vaccine are front and center in the minds of biomedical researchers and healthcare practitioners worldwide.

This article reviews how High-Performance Computing (HPC) clusters optimized for high-throughput analytics are crucial to the success of COVID-19 efforts. Here, we also show our latest results processing a Human Whole Genome (WGS) in as little as 48 minutes with Lenovo's Genomics Optimization and Scalability Tool (GOAST)—Lenovo's CPU-based architecture for Genomics Analytics. GOAST leverages specially tuned hardware and pre-configured pipelines to accelerate the GATK software suite from the Broad Institute[3], thus enabling genomics analytics at speed-ups up to 188X compared to typical[a] GATK environments. GOAST is improving lab productivity by allowing researchers to process more genomes concurrently—this means higher throughputs and shorter execution times. Processing more genomes means getting answers faster and HPC-driven discoveries saving more lives at a fraction of the cost of alternatives.



WHAT TYPE OF INSIGHTS ARE COVID-19 RESEARCHERS INTERESTED IN?

There are many unknowns regarding the mechanisms of infection of SARS-CoV-2 (the virus causing COVID-19) and the host's immune response it elicits. Therefore, researchers are tackling this unknown from multiple angles. Vaccine design, improving testing/diagnostic kits, identifying determinants of susceptibility, assessing virulence, tracking virus origin, tracking virus spread, identifying antiviral drug targets, and designing protective measures are some of the highest-priority efforts in the biomedical and pharmaceutical communities today. What do all these efforts have in common? "OMICS" analytics (*i.e.,* Genomics, Transcriptomics, etc.) is one of the first steps undertaken to derive scientific insight.

Genomics Analytics focusing on finding genetic differences. A single respiratory droplet from an infected person can contain millions to billions of viral particles showing minute differences in their DNA with respect to the viral population in the next person. Those differences could be responsible for the severity of infection (for example, some people manifest the infection as a mild cough, while others progress to pneumonia), transmissibility (how easily the virus jumps to another person), or susceptibility to treatment (how effective medication would be at attenuating or eradicating the infection). Often, researchers sequence the genomes of infected individuals with differing symptoms to identify whether the severity of infection can be explained by host genetics (Fig. 1). Perhaps, the mild cases present in people with innate protective factors. In the case of COVID-19, genetic protective factors are still unaccounted for, but identifying them is of extreme interest since they would shed light on possible treatments and other protective measures. A different research approach might instead sequence the viruses infecting the group with a mild case vs. the viruses in the pneumonia group to study whether the severity of infection is associated with carrying specific variants of the virus (Fig. 1).

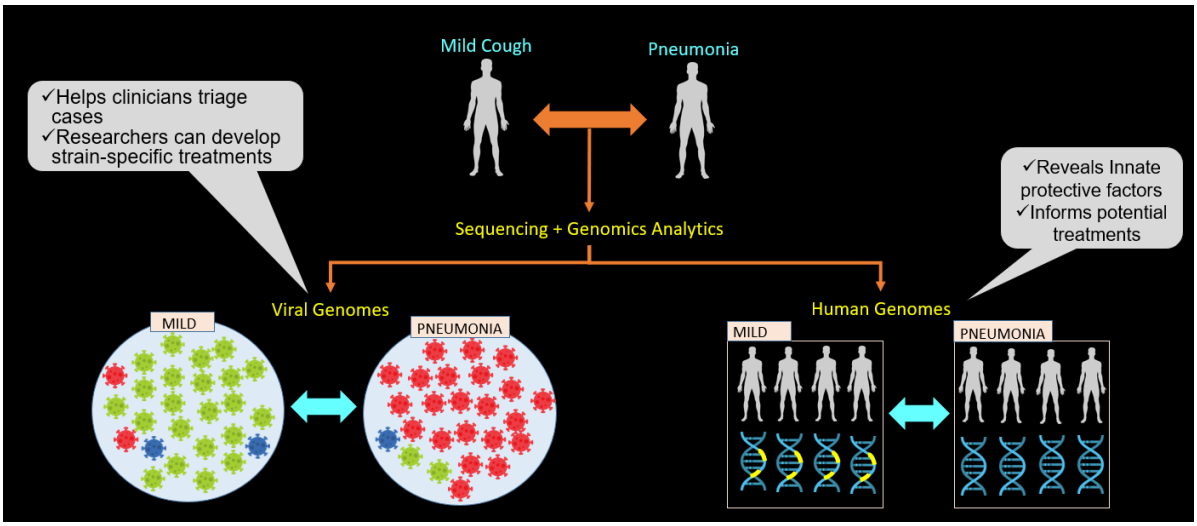


Figure 1: Genomics analytics unveiling genetic differences. An example of how biomedical researchers might use genomics analytics to understand how severity of SARS-CoV-2 infection manifests differently in different individuals. To understand why COVID-19 manifests as a mild cough in some people, while leading to pneumonia in others, some projects may analyze the genomes of the viruses while others may analyze the genomes of the patients in the two groups. Should the difference in severity result from differences in the dominant strain, it would help clinicians and hospitals triage cases and assign strain-specific treatments. Should the key to severity be on the patient's genomes instead, it would open a path to retro-engineer a treatment that would mimic the same protective mechanism that those with mild cases have naturally.

When genetic similarities inform vaccine targets. Other times, researchers focus on the genetic similarities among viruses, rather than on the differences. Take vaccine design, for example (Fig. 2). Often, virologists try to identify conserved regions among viruses: *i.e.*, areas of the viral genome that remain unchanged over time. Let me explain. If we think of genes as sentences relaying instructions, then conserved gene regions are words in the instructions with identical spelling in all circulating viruses. We have all heard that viruses mutate over time; in other words, they change the way their DNA "spells" biological instructions in some areas of their genome. Rapid DNA mutation is the reason why designing an effective treatment or vaccine against viruses is so difficult. A vaccine against a changing target rapidly loses effectiveness. When researchers identify unchanging regions in the viral genomes it gives them potential vaccine targets (Fig. 2). Through immunological methods we can train the human immune system to learn that those target conserved regions are foreign and must be attacked; hence, eliciting an immune response to keep you healthy. Those are the basics of how vaccines work.

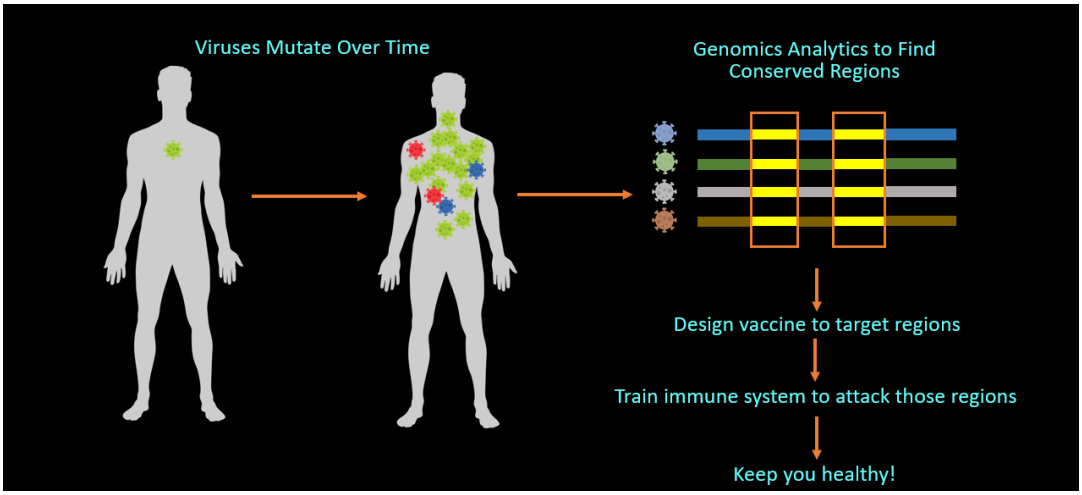


Figure 2: Genomics analytics to find genetic similarities in vaccine design. A diagram showing a typical vaccine design process leveraging genomics analytics to find conserved regions on the viral genomes. The conserved regions reveal essential areas of the genome from which the virus cannot mutate away thus indicating potential vaccine targets. In turn, the vaccine targets are used to train the immune system to recognize as foreign and to mount an immune response next time the vaccinated person encounters the pathogen.

HOW CAN HPC ENABLE COVID-19 DISCOVERIES?

Enabling Life Sciences research, and specifically empowering COVID-19 discoveries, requires HPC environments that support high-throughput volumes, accelerate speeds, optimize infrastructure, and ensure the protection of data privacy.

Supporting High-Throughput Volumes. The efficacy of COVID-19 vaccine and treatment development will depend largely on researchers' ability to assemble and analyze many genomes more quickly and at a larger scale. Since we can neither sequence every virus in every patient, nor can we sequence the genome of every infected human, scientists rely on sampling. The bigger the sample size, the more we will learn about the viruses circulating in the population and the more effective our treatments, vaccines, and other protective measures will be. Thus, increasing the throughput capacity of the analytics on cohort-level or population-level genomics is crucial to the success of COVID-19 research efforts.

Accelerating Genomics Analytics. The outputs of genomics analytics (also known as secondary analyses) in turn become the input of tertiary analyses, which also require a robust HPC environment: molecular modelling, 3D-structural analyses, biostatistics, etc. Together, the secondary and tertiary analyses guide drug discovery, vaccine design, identify treatments, and can eventually deliver the promise of precision medicine. Unfortunately, running genomics workflows creates a bottleneck in the discovery process: e.g., a typical environment running the GATK[3] workflow on a 30X-coverage whole genome takes 60 to 150 hours. Deploying genomics in standard HPC environments is not scalable for sizeable efforts. Large-scale genomics projects require an HPC infrastructure delivering accelerated speeds and scaling modularly to handle increasing datasets and population-level sample sizes.

Optimizing infrastructure while increasing usability. Setting up and managing a large-scale Next-Generation Sequencing (NGS) operation faces a steep learning curve. Genomics analytics require sequential workflows of over 30 different bioinformatics tools strung together in non-traditional computing set-ups difficult to optimize. Traditional HPC typically distributes one tool across many nodes; genomics instead runs many tools in one node. This is problematic because the various genomics tools have different profiling characteristics resulting in long runs and underutilized hardware. Most organizations lack teams combining HPC + Genomics

expertise working together to optimize infrastructure specifically for genomics workloads. Researchers are focused on their science, while the HPC support teams are spread too thin to begin the arduous task of creating DIY alternatives. Multidisciplinary teams systematically testing the effect of hardware and software components on performance is the key to addressing the bottlenecks in genomics analytics.

Securing Data Privacy. Any research project analyzing human data must employ robust mechanisms to protect the privacy of patient data. The sensitive nature of human data in COVID-19 or any other biomedical project mining patient information requires a platform that ensures appropriate user authentication and controls data access appropriately.

FINDING THE BEST ARCHITECTURE FOR GENOMICS ANALYTICS

For the past two years, the supercomputers at Lenovo's benchmarking datacenter have been hard at work systematically testing hundreds of permutations of the hardware building blocks and software parameters used when running the 30-plus-tool genomics workflows. Lenovo's genomics R&D group has been searching for the best "hardware + system" recipe to support high-throughput volumes, deliver accelerated execution speeds, and increase usability in an environment securing data privacy. In 2019, we published a genomics analytics solution delivering speed-ups of up to 40X: Lenovo GOAST originally processed a 30X-coverage whole genome (WGS) in 5.5 hrs. and a 50X whole exome (WES) in 6 minutes[2]. Such an achievement afforded us the first certificate for Intel® Select Solution for Genomics Analytics awarded to any solution provider.

After the onset of the COVID-19 pandemic at the beginning of 2020, our team increased in-house optimizations even further to ensure we could do our part in the fight against the pandemic. Our goal has been to focus on optimizing the technology and increasing its usability so that our collaborators and customers can instead focus on moving science forward.

Here, we report new results from a comprehensive study systematically evaluating the effect of hardware building blocks, workflow set-up strategies, and tool parameters on the performance of the Broad Institute's GATK Single-Sample Germline Variant Calling workflow[3]. This study analyzed both WGS and WES versions of the NA12878 genome—a well curated reference human genome representing a high-quality gold standard in the Genomics field[4]. In terms of hardware building blocks, we tested many options of processors, storage, memory, and server types. The CPUs in our tests ranged from 20 to 28 cores. Our storage profiling evaluated HDDs, SATA SSDs, SAS SSDs, and NVMes available locally and over the network. The nodes featured memory capacities ranging from 192GB to 12TB of RAM. Our tests also evaluated 2-socket, 4-socket, and 8-socket servers. For software optimizations we decided to stay away

from any code-level changes to the tools and instead focused on tool parameter optimizations, thus leaving the tools scientists know and trust untouched. We instead evaluated software parameters that tuned hardware resources, such as memory allocation, thread count, etc. We also investigated factors related to how the workflow is set up. Our study also measured the effects of progressively increasing the number of threads, shards, and concurrent in-node jobs.

Our study was quite comprehensive, requiring hundreds of runs to systematically measure the effect of each variable in different permutations of architectures and workflow set-ups. The high dimensionality of the problem, in addition to the extensive list of options evaluated for each variable make it a monumental benchmarking effort seldom or never undertaken by the HPC teams supporting genomics research given the resources and time it requires.

LENOVO'S GENOMICS OPTIMIZED SERVERS: HPC INFRASTRUCTURE OPTIMIZED TO ACCELERATE HIGH-THROUGHPUT GENOMICS ANALYTICS

GOAST options to balance cost vs. extreme performance. We based our genomics optimizations on extensive systematic testing of the factors affecting performance at our in-house Genomics R&D lab. As a result of our R&D work, we have developed two configurations to accelerate variant calling workflows: GOAST Base and GOAST Plus, optimized for cost vs. extreme performance, respectively. GOAST Base, optimized for cost, can process one WGS in 3.27 hours, thus analyzing 7.3 WGS/node/day or up to ~2700 samples/node/year[b] (Fig. 3). The GOAST Plus configuration, which is our extreme performance option, will analyze one genome in 48-53 minutes, which means 27-30 whole-genomes per node per day or up to ~11K per node per year[c] (Fig. 3).

		Lenovo GOAST Base	Lenovo GOAST Plus
			
Specifications	**Server Type**	**1x ThinkSystem SR630**	**1x ThinkSystem SR950**
	Processor	2x Intel® Xeon® Gold 6248R (24 cores, 205W, 3.0 GHz)	8x Intel Xeon Platinum 8280 (28 cores, 205W, 2.7GHz)
	Memory	384GB RAM, 12x ThinkSystem 32 GB, TruDDR4 2933 MHz (2Rx4 1.2V RDIMM)	1.5TB RAM, 48x ThinkSystem 32GB TruDDR4 2933MHz (2Rx4 1.2V) RDIMM
	Local storage	ThinkSystem 2.5" PM1643a 1.92TB Entry SAS 12Gb Hot Swap SSD	4 x 3.2TB U.2 NVMe PCIe 3.0 x 4 Hot Swap SSD
Performance Metrics	Per sample (Time/WGS)	3.3 hr.	48-53 min.
	Per day (WGS/node/day)	7.3	30
	Per year (WGS/node/year)	2,700	10,950

Figure 3. Technical specifications and performance metrics for GOAST architectures. GOAST comes in two configurations: GOAST Base and GOAST Plus. Both configurations are modular units that can be used as a single appliance or scaled out onto a cluster design. The figure lists specifications for both architectures as well as their corresponding performance metrics. Local storage is optional when sufficient network storage of at least SSD-level or above is available.

The impact of GOAST innovation is staggering. Typical environments run GATK in 60 to 150 hours per whole-genome (Fig. 4). The original Intel® Select Solution for genomics analytics, BIGstack, achieved a 5X-14X speed-up. By leveraging our technology and expertise, Lenovo improved the Intel® Select Solution for Genomics Analytics to deliver a 18X-45X improvement with GOAST Base and a 75X-188X speed-up with GOAST Plus compared to standard environments (Fig. 4).

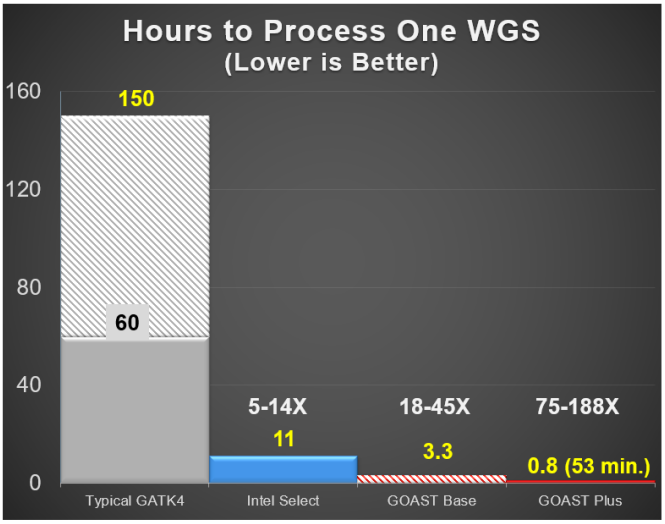


Figure 4. Performance of the GATK Workflow with different architectures. The bar graph compares the performance of the GATK Germline Variant Calling workflow on the NA12878 30X-deep WGS dataset in typical HPC environments vs. three Intel® Select Solutions for Genomics Analytics. Lenovo GOAST processes one WGS in either 3.3 hr./node with GOAST Base or as little as 48-53 min./node when running GOAST Plus.

GOAST performance is validated and repeatable. To test the repeatability of our results, we ran the same architecture permutation test at least five times. See Table 1 for repeatability tests on GOAST Base and GOAST Plus. Some variation in execution time is always expected due to the programming inherent in the underlying GATK tools. Nonetheless, it is reassuring that GOAST results deviate only 1-2% away from the average performance for WGS datasets and 1-3% for WES. Considering the entire deviation, from shortest to longest run, results never exceeded 4.6% from the performance metrics reported in this work.

Table 1
Repeatability of GOAST performance when processing WGS and WES datasets

	WGS		WES	
Performance	**GOAST Base**	**GOAST Plus**	**GOAST Base**	**GOAST Plus**
Average run (time/sample/node)	3.27 hr./WGS	0.89 hr./WGS	6.42 min./WES	1.6 min./WES
Shortest run (% diff. from average run)	-2.0%	-1.5%	-1.8%	-2.0%
Longest run (% diff. from average run)	+1.4%	+1.3%	+1.3%	2.5%
Repeatability margin	3.4%	2.8%	3.2%	4.6%

Increasing throughput capacity dramatically improves genomics execution time. Figure 5 shows the results of workflow scaling studies for GATK when analyzing up to 256 concurrent WES samples or batches of up to 32 WGS in a single node. The blue bars in Fig. 5A show the stark contrast between running in latency mode (i.e. one sample per node) at 30.66 min./WES vs. running 32 samples concurrently and reaching speeds of 6.42 min./WES. The same benefit is seen for WGS for GOAST Base (Fig. 5B). While one WGS/node runs in 5.6 hrs., a 16-sample batch achieves 3.1 hrs./sample. The real benefit of GOAST Plus is evident in the throughput runs. The latency exome runs in GOAST Plus (Fig. 5B) perform similarly to the latency runs in GOAST Base: 30.54 vs. 30.66 min./WES. In contrast, the best throughput exome runs in GOAST Plus show execution times as low as 1.6 min./WES (Fig. 5C). GOAST Plus is particularly useful for throughout WGS runs (Fig. 5D) reaching 0.88 hr./WGS (i.e. 53 min. to process one whole genome or 48 min. when more memory is available to the node). These results also indicate that genomics performance does not improve linearly or indefinitely with increasing concurrent samples; instead, performance plateaus after reaching a local minimum and at different values for the different datasets and architectures.

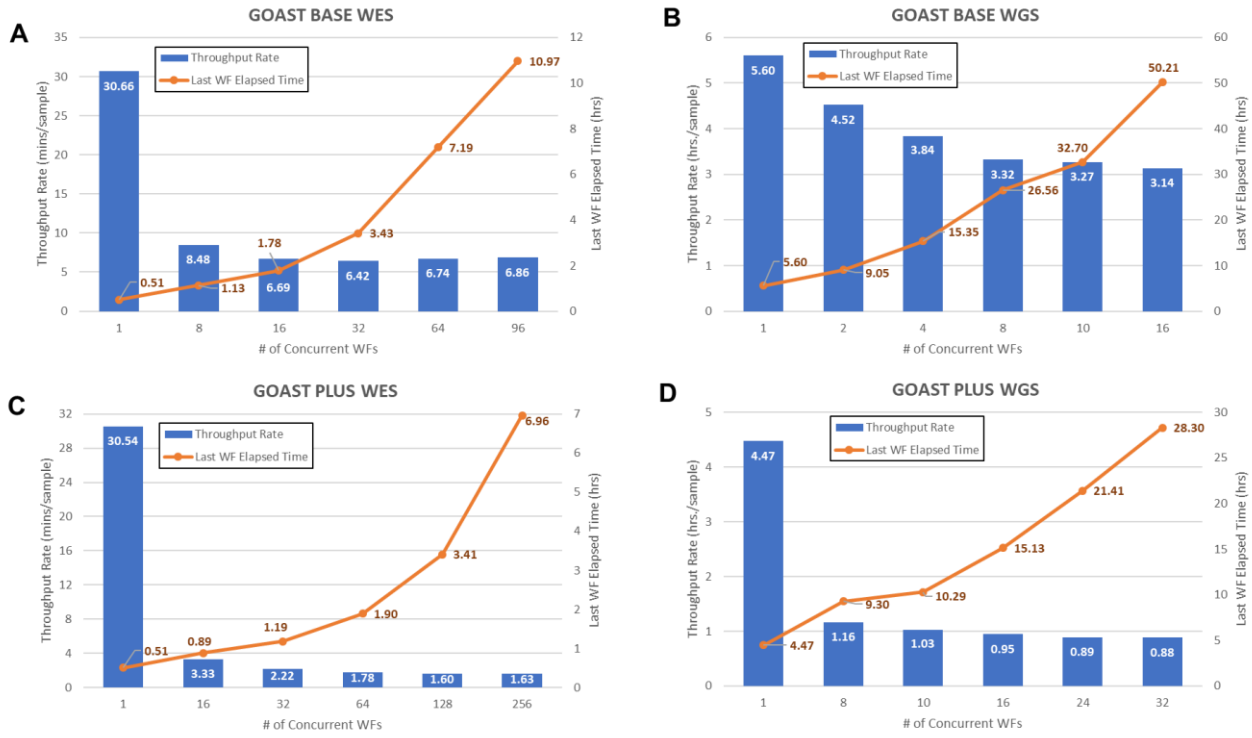


Figure 5. Workflow scaling study of the GATK workflow when running multiple concurrent WES and WGS samples on a single node. The X-axis shows the number of samples running on a single node (i.e., the batch size) and their corresponding performance or throughput rate as time per sample on the left vertical axis (blue bar). The graphs also show the performance of each batch of samples (orange trend) on the right vertical axis as the execution time for the last workflow (WF) in the batch. The results for GOAST Base running WES (A) and WGS (C) and for GOAST Plus running WES (B) and WGS (D) show the effect of latency (one sample) vs. throughput (multiple samples) execution in genomics analytics.

High-core processors and SSD-level storage or above are the best indicators of genomics performance. By far, the best indicators of performance in genomics workflows are high-core count and fast storage at SSD-level or above. In our permutation benchmarking tests, access to higher core processors always yielded considerable performance boosts particularly in the throughput tests. In terms of I/O usage, architectures with at least SSD-level storage always outperformed HDD latency runs by as much as 30-50%. NVMe storage did not offer any performance advantage over SSD storage locally or over the network for latency runs, but NVMe showed an 8-10% performance boost in the throughput runs presumably because of the increased I/O. See Fig. 6 and Fig. 7 for profiling snapshots of throughput WES and WGS runs capturing patterns of CPU, storage and memory usage by the two GOAST architectures.

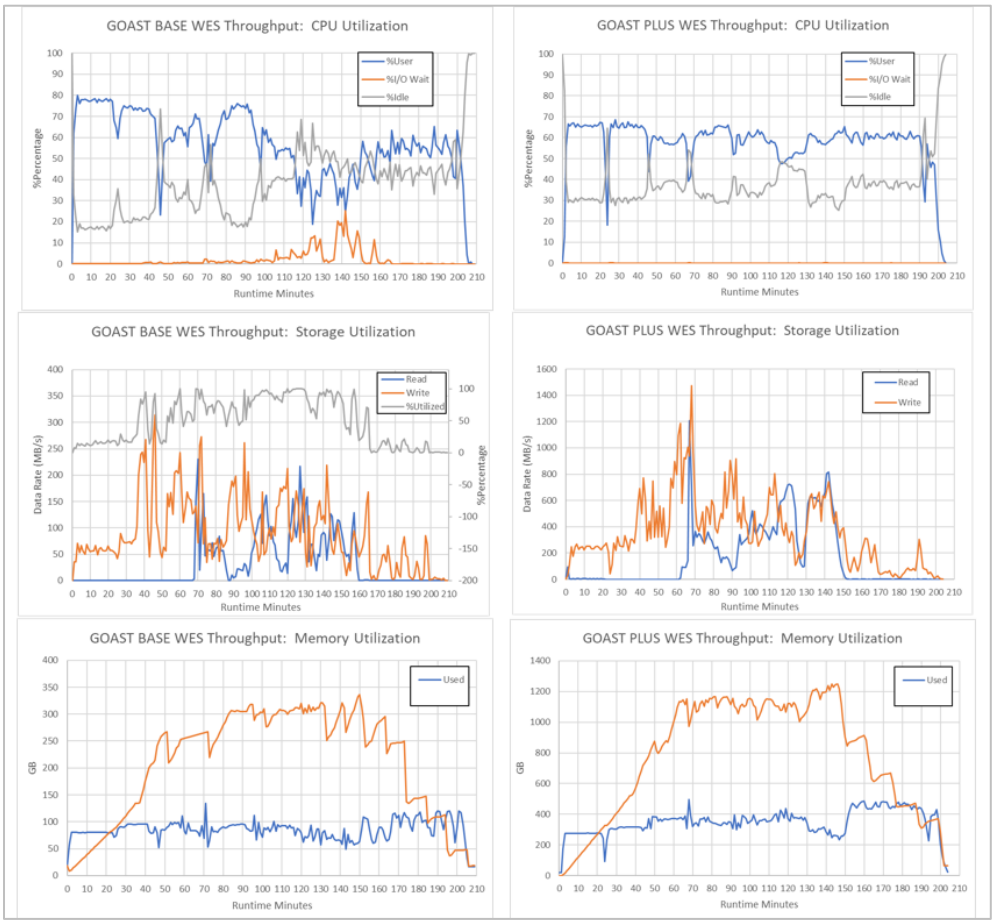


Figure 6. HPC profiling for WES datasets. CPU utilization, I/O, and Memory usage averaged per minute.

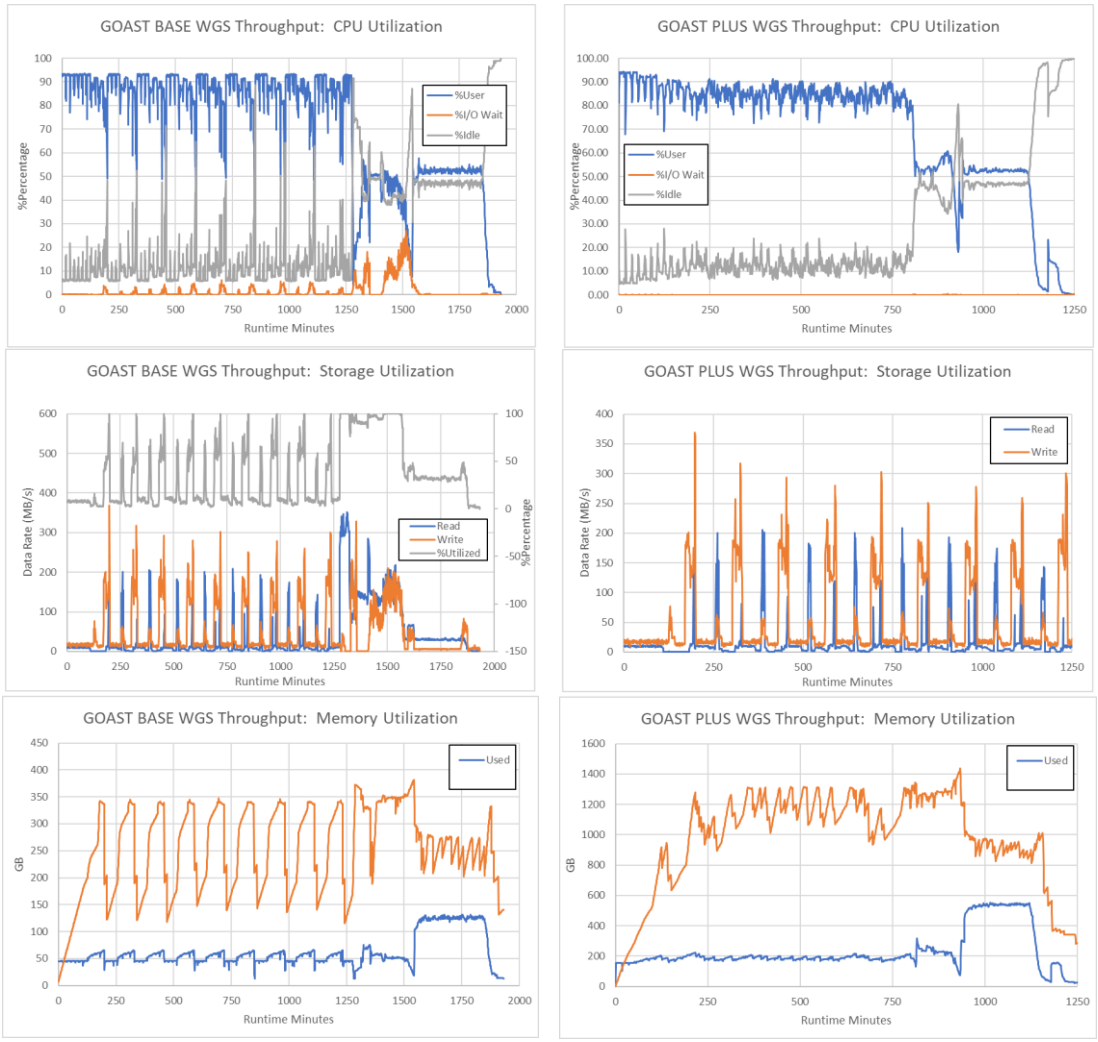


Figure 7. HPC profiling for WGS datasets. CPU utilization, I/O, and Memory usage averaged per minute.

Run multiple life sciences workloads with a single optimized GOAST architecture. Preliminary results from our in-house testing also show that the same GOAST architectures considerably speed up transcriptomics workflows (manuscript in progress). Our findings suggest that bioinformatics sequential pipelines with mixed HPC profiling characteristics (Fig. 6-7) benefit from the optimizations of this work. In addition to genomics and transcriptomics, GOAST architectures can be sized to support proteomics, molecular dynamics, and cryoEM workloads.

SUMMARY

GOAST is the result of two years of systematic benchmarking in a large-scale permutation study of the factors affecting the performance of genomics analytics. Developing GOAST required hundreds of runs to systematically measure the effect of different permutations of architectures and workflow set-ups on performance. The result of this work is an easy-to-use, pre-configured, pre-installed hardware + software bundle delivering GPU-level speeds but at CPU-level costs. Lenovo GOAST delivers execution speeds unprecedented in CPU-based technologies and only possible before with costly specialty hardware.

GOAST is decreasing time to scientific insights. Before GOAST, it took datacenters 60-150 hrs. to process a single human genome. Genomics workflows on GOAST run in as little as 48-53 min (a 188X speed-up). The same GOAST architecture appears to also optimize transcriptomics and it can be sized to support proteomics, molecular dynamics, and cryo-EM workloads thus GOAST provides a versatile and affordable solution for OMICS and other HPC-heavy Life Sciences workloads.

GOAST supports high-throughput volumes. GOAST architectures process up to 30 whole-genomes per node per day and up to 11K per node per year. The two GOAST architectures shown here can easily support OMICS analytics efforts as small as a single research lab with a single sequencer all the way to the population-level efforts sequencing hundreds of thousands of genomes per year. The Lenovo Genomics team understands the factors affecting biomedical analytics on HPC and can customize the GOAST configurations to fit the performance level or budget constraint of any organization.

GOAST is lowering the barriers to usability. We have tremendously increased usability of the genomics workflows by creating pre-configured pipelines leveraging discoveries from our systematic testing. We have created easy-to-run scripts to set up the tools, libraries, dependencies and to set-up tool parameters, which should lower the barriers of usability. We also pre-tune the relevant hardware building blocks to yield the performance boosts reported here.

GOAST secures data privacy. By deploying GOAST on premise and on hybrid environments we circumvent many security issues plaguing the public cloud. Our optimizations allow on-premise and hybrid efforts to increase productivity and customization so that they can deploy HPC in an environment where they have the ultimate control.

GOAST is democratizing knowledge by making HPC accessible. GOAST is both high-performing and affordable requiring no specialty hardware accelerators. GOAST is making high-throughput, high-performance analytics accessible to organizations that cannot afford the more expensive GPU-based solutions. Our solution relies on optimized CPUs leveraging the open-

source software scientists already know and trust to deliver execution speeds only matched by solutions based on “GPUs + proprietary software”.

How is GOAST accelerating the path to COVID-19 discovery? The efficacy of COVID-19 vaccine and treatment development will depend largely on researchers’ ability to assemble and analyze many genomes/transcriptomes more quickly and at a larger scale. Genomics and transcriptomics are often the first steps in the highest-priority efforts around COVID-19 research (vaccine design, improving diagnostic kits, assessing virulence, identifying determinants of susceptibility, tracking virus origin, identifying drug targets, etc.). Therefore, by accelerating OMICS and making the optimized hardware more accessible, GOAST is helping accelerate the path to COVID-19 discoveries and enabling many more to walk it.

Lenovo is leveraging the GOAST Genomics Optimized Servers to help datacenters around the world accelerate their workflows, increase their throughput, and plan their HPC resources more effectively. 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 of sequencing ever larger samples and is empowering scientists on the frontlines of COVID-19 research to accelerate their path to discovery. The tech revolution of accessible innovation is here, be it for basic research, infectious disease, or precision medicine. The GOAST way is the smarter way of decoding genomics at scale with a validated, pre-configured genomics analytics solution built on reliable, high-performance Lenovo ThinkSystem.

References

1. COVID-19 Dashboard. Johns Hopkins University Center for Systems Science and Engineering. https://coronavirus.jhu.edu/map.html
2. Giraldo, M., Dean, K., Merz, F. GOAST: Genomics Optimization And Scalability Tool—An Intel® Select Solution for Genomics Analytics Enabling Precision Medicine at the Population Scale. https://www.lenovo.com/us/en/resources/data-center-solutions/solution-brief-documents/goast-genomics-optimization-and-scalability-tool/
3. 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.
4. 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).

Quantifiable Benefits with Substantiation

a. **Smarter genomics analytics**
Quant: "GOAST leverages specially tuned hardware and pre-configured pipelines to accelerate the GATK software suite from the Broad Institute[3], thus enabling genomics analytics at speed-ups up to 188X compared to typical GATK environments."

Substantiation: Lenovo internal calculation. **Calculation performed by Lenovo employee Mileidy Giraldo, Ph.D. September 2020.**
Typical GATK environments, both on-prem and in the cloud, take 60 to 150 hours to process a single whole genome. GOAST Plus reduces GATK execution time down to 48-53 minutes. Therefore,

- **Formula for GOAST Plus speedup** = Standard_speed_in_hours / (GOAST_Plus_speed_in_min/ min_in_1_hr)
- **Calculation of GOAST Plus speedup** = 150 hr./ (48 min./60 min.)
- **Result of GOAST Plus speedup** = 188X speedup*

b. **Lenovo GOAST Base Quantifiable Benefits**
Quant: "GOAST Base, optimized for cost, can process one WGS in 3.27 hours, thus analyzing 7.3 WGS/node/day or up to ~2,700 samples/node/year"

Substantiation: Lenovo internal calculation. Calculation performed by Lenovo employee Mileidy Giraldo Ph.D. September 2020.

- Samples per node per year = daily node production * 365
- Samples per node per year = (24 hr./3.3 hr.) * 365
- Samples per node per year = 7.3 * 365
- **Samples per node per year = ~2700 WGS/node/year**

Quant: GOAST Base delivers <u>18X-45X improvement</u>

Substantiation: Lenovo internal calculation. Calculation performed by Lenovo employee Mileidy Giraldo Ph.D. September 2020.

Fastest Standard GATK environments processes one WGS in 60 hours.

- **GOAST Base Speed-up against <u>fastest</u> standard environment = 60/3.3 = 18.18**

Slowest Standard GATK environments processes one WGS in 150 hours.

- **GOAST Base Speed-up against <u>slowest</u> standard environment = 150/3.3 = 45.45**

c. **Lenovo GOAST Plus Quantifiable Benefits**
Quant: “The GOAST Plus configuration, which is our extreme performance option, will analyze one genome in 48-53 minutes, which means 27-30 whole-genomes per node per day or up to ~11K per node per year”

Substantiation: Lenovo internal calculation. Calculation performed by Lenovo employee Mileidy Giraldo Ph.D. September 2020.

- Samples per node per year = daily node production * 365
- Samples per node per year = (24 hr./(48 min./60 min.)) * 365
- Samples per node per year = **30** * 365
- **Samples per node per year = 10950 WGS/node/year**