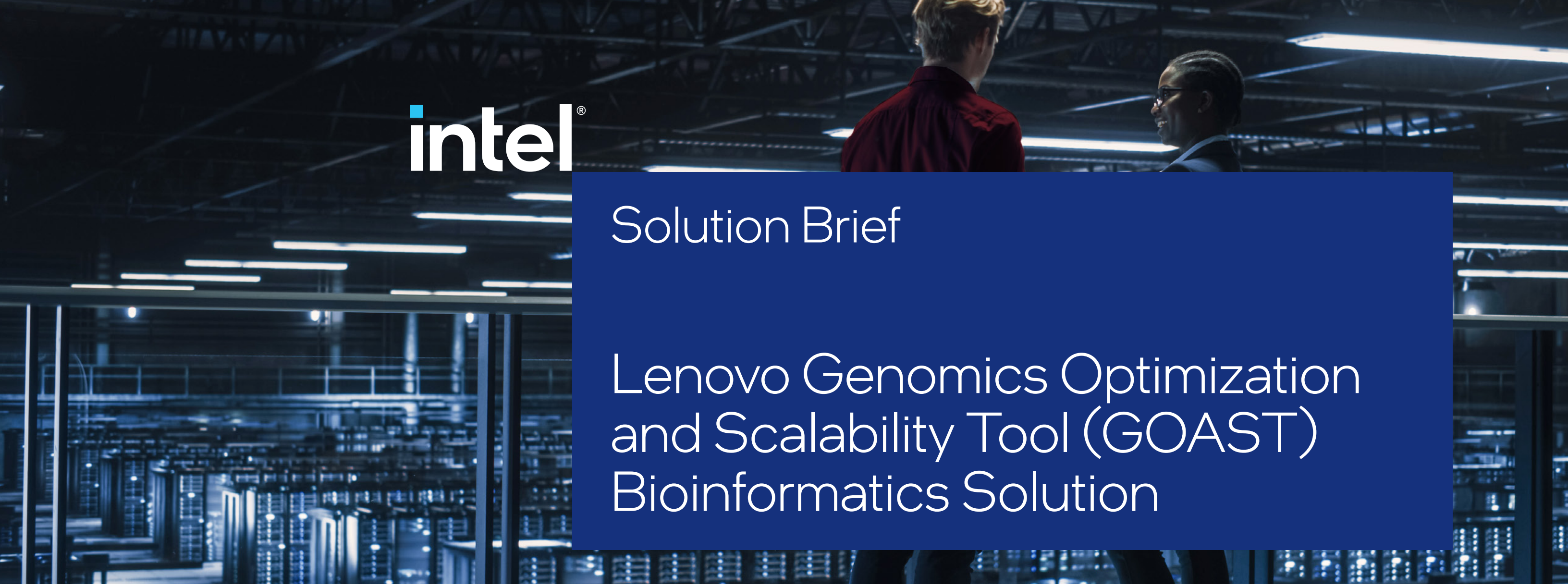


Solution Brief

Lenovo Genomics Optimization and Scalability Tool (GOAST) Bioinformatics Solution

GOAST v4.0 optimized hardware accelerates execution speeds to boost lab productivity

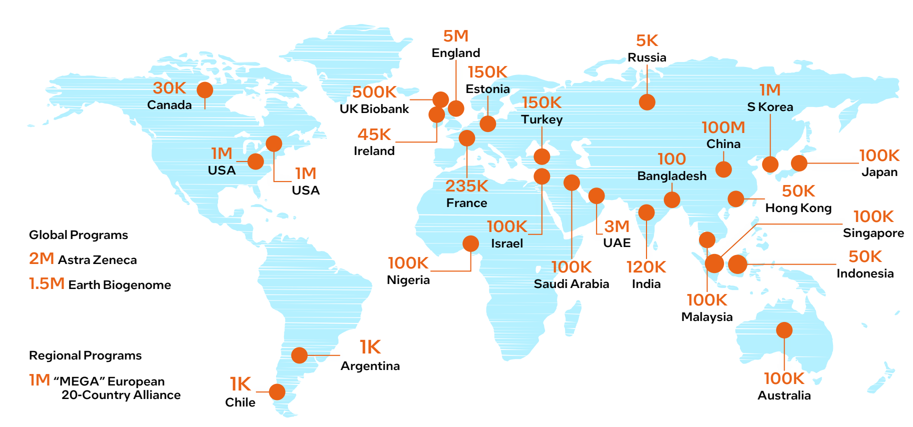
Highlights

Lenovo GOAST v4.0 powered by 5th Gen Intel® Xeon® Scalable processors offers:

- a leap in performance over previous generations, speeding sequential workflow processing for faster time to insight
- preconfigured, optimized workflows to free researchers' time for research
- GPU-like performance at CPU-level prices, up to 50 percent less than boutique solutions relying on GPUs or FPGAs without additional licensing fees
- flexibility to deploy as a single-node appliance or a cluster and grow linearly with ease
- Snakemake as the workflow manager

Introduction

Bioinformatics powers genomics research from basic biology to precision medicine, drug discovery, agriculture, and more. While the field has existed for over 50 years, the past two decades have seen an explosion of bioinformatics data. Much of this data comes from large national genomics initiatives such as the "Genomics England/UK Biobank," the "All of Us program" in the US, and Singapore's "GenomeAsia."



Source: Frost & Sullivan, "Global Precision Market Growth Opportunities, Forecast to 2025," Originally published in January 2017 and updated by Intel with internal market data

Figure 1. Large Scale Sequencing is already underway!

Propelled by the increasing affordability of Next-Generation Sequencing (NGS) technologies and advancements in high-performance computing (HPC) storage and computing technologies, these initiatives aim to capture the genetic variation of each nation's people to spur advances in prevention, diagnosis, and treatment for individuals and populations. They're making significant progress. At the UK Biobank, the whole genomes of 500,000 people recently went live to help researchers search their genetic code for links to disease. Analysis of this data has led to a variety of insights, including the realization that 4.8 percent of the UK population has elevated liver iron levels, a key risk factor for significant liver disease.[1] Such discoveries point the way to more accurate diagnosis, less invasive testing, and quicker treatment.

The greatest challenges facing population genomics efforts—and therefore downstream bioinformatics—have long been scale and time. Population genomics requires scaling up input data from exomes 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. Two of three stages in population sequencing take place in the HPC environment of a cluster or supercomputer, including variant analyses (comparing how a gene is "spelled" in different people), and downstream bioinformatics (for example, 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.

To address these challenges, in 2017, Intel began an ongoing partnership with the Broad Institute to search for technological innovations that would reduce genomics analytics times while retaining the scientific rigor and accuracy scientists had come to know and expect of the Broad's Genome Analysis Toolkit (GATK).[2] 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 Genomics DB 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 for whole genomes and exomes respectively. Intel continues to work with the Broad Institute, helping to optimize the Broad Institute's pipelines and GATK with Intel libraries, including the Intel Genomics Kernel Library.

Despite these advances, today—even at cluster and supercomputer speeds—large-scale bioinformatics still faces long execution times on massive volumes of data, which delays "time to answer." Currently, the only high-performance solutions to mitigate these challenges are single-purpose boutique solutions that require expensive specialty hardware and substantial licensing fees. These single-purpose solutions require organizations to purchase multiple architectures to support other types of research in their data centers. Even organizations working in a single subfield (for example, in genomics) find that the single-purpose solutions frequently do not offer sufficient computational resources for them to accomplish their goals. For example, those performing secondary genomics analytics find that they also need computational resources to gather, select, manage, transform, and describe their data, in both primary and tertiary downstream analyses. General-purpose data centers worldwide feel this need even more acutely since the omics (genomics, transcriptomics, proteomics) are only a fraction of the users they must serve.

To meet the diverse needs of these organizations, Lenovo partnered with Intel in 2019 to develop the GOAST system, a multi-purpose Genomics and Bioinformatics Optimized platform, and has continued to evolve GOAST ever since. Now, Lenovo has released new GOAST v4.0 with major updates to accelerate bioinformatics performance further.

GOAST Capabilities

Extremely Fast Bioinformatics Analytics

Lenovo GOAST is a multi-purpose system specifically engineered to meet the demands of bioinformatics workloads. GOAST leverages an architecture of carefully selected hardware tuned to accelerate bioinformatics performance. Lenovo GOAST's high-core, fast I/O, and high-memory specs (Table 1) excel at running the massively parallel applications and sequential workflows common in bioinformatics, including multi-omics (genomics, transcriptomics, proteomics) applications. GOAST accelerates mapping and whole-genome sequencing variant calling analytics from days to hours—a process that today in many data centers around the world takes more than 24 hours runs at a throughput rate of 1.77 samples per hour or 16 samples in 9 hours.

GOAST v4.0 Capabilities

Lenovo GOAST v4.0 is a recipe that comes with several preconfigured and optimized workflows:

- Germline variant calling
  - Whole-Exome Sequencing (WES)
  - Whole-Genome Sequencing (WGS)
- Somatic variant calling
  - Tumor + Normal Pair
  - Tumor Only

All needed dependencies come pre-installed in a Conda environment that is easily replicated for additional users. Workflows are submitted using the command line GOAST Util Tool, which enables users to submit complex GATK workflows with a single command. This tool automatically allocates resources to be used most efficiently based on the number of samples submitted and the available computational resources. Users may also monitor progress, abort, and restart jobs, and manage temporary files.

GOAST v4.0 comes with several major updates including additional workflows, Snakemake as the workflow manager, using Conda to manage software installations, and a complete rewrite of the backend GOAST Util Tool.

Table 1. GOAST v4.0 reference architectures for bioinformatics (fully customizable)

	GOAST v4.0 Intel	GOAST v4.0 Intel
Processor	▪ 2x Intel 8480+ CPUs (56 cores, 2.0 GHz, 350W), Lenovo ThinkSystem SR650 V3 Server	▪ 2x Intel 8592+ CPUs (64 cores, 1.9 GHz, 350W), Lenovo ThinkSystem SR650 V3 Server
Memory	▪ 1 TB RAM, 16x 64GB 4800 MT/s RDIMM	▪ 1 TB RAM, 16x 64GB 5600 MT/s RDIMM
Storage	▪ Min. 7 TB local SSD or NVMe	▪ Min. 7 TB local SSD or NVMe

GOAST Benefits

Increases Lab Productivity

Accelerated execution speeds mean you get to process more samples, find answers faster, and generate breakthroughs that much sooner. GOAST outperforms any other competing CPU-based (and even the FPGA- and GPU-based) systems because Lenovo tunes GOAST systems to meet the requirements of bioinformatics pipelines running in-node workloads rather than those assumed in traditional HPC workloads. The result is the ability to run software pipelines in higher throughputs. Higher throughput capacity means batches of samples analyzed in less time. (Table 2). Lenovo has performed the heavy lifting of optimizing workflows as well as ensuring software updates work seamlessly together. This gives researchers the opportunity to focus on research questions, instead of spending valuable time tuning hardware, tweaking software versions, and optimizing workflows.

Table 2. Lab productivity for omics expected on a single GOAST system*

Expected Lab Productivity	50x WES samples** GOAST Intel		30x WGS samples** GOAST Intel	
	4th Gen Intel® Xeon® Processor	5th Gen Intel® Xeon® Processor	4th Gen Intel® Xeon® Processor	5th Gen Intel® Xeon® Processor
Samples/ Node/Day	826	1011	36	42.4
Samples/ Node/Year	~301K	~369K	~13.1K	~15.5K

* Performance is based on the processing of this NA12878 sample which was sequenced on the NovaSeq 6000. All processing was performed on local NVMe drives. Performance may vary based on hardware setup and coverage of sample.

** WES - 64 samples were used and processed. WGS - 16 samples were used and processed.

Multi-purpose Bioinformatics Use

The high-performance GOAST system comes preloaded with omics tools to get you up and running on day one, or it can be fully customized with the bioinformatics tools of your choice.

- For multi-omics analytics: Lenovo pre-installs the tools and other dependencies in Table 3 that are necessary to run the Broad Institute's GATK Best Practices for Germline and Somatic SNP and Indel discovery. Lenovo GOAST also provides preconfigured scripts to allow you to run (submit, monitor, manage) samples on the Germline workflow and Somatic workflow optimally on Lenovo hardware with the GOAST Util Tool.
- For other bioinformatics: Install any tools of your choice on GOAST systems or talk to the Lenovo team about preinstalling your software pipeline of choice. GOAST nodes are configured to support a wide range of bioinformatics workflows.

Table 3. Genomics analytics software and other dependencies preinstalled by GOAST v4.0

Software	Version
GATK	4.4.0.0
BWA-MEM2	2.2.1
Samtools	1.17
Picard Tools	3.0.0
OpenJDK	17.0.3
Snakemake*	7.32.3
SLURM (Optional)*	23.02.4
OS (Recommended)	Rocky Linux 9.2

*GOAST systems currently use Snakemake as the workflow manager and job scheduler on a single-node setup and Slurm as the job scheduling system on a multi-node setup.

Cost Effective

GOAST leverages an optimized CPU-based architecture; thus, it requires no FPGAs or GPUs of any kind for acceleration. Users should expect GPU-like performance for the optimized workflows – at CPU-level prices, or 50 percent lower than boutique solutions relying on FPGAs or GPUs, and no licensing fees. The Lenovo Bioinformatics R&D group continually tests new bioinformatics pipelines and releases to its customers hardware-tuned versions of standardized workflows such as the Broad Institute's GATK Best Practices at no cost. In addition, GOAST solutions can reduce investments needed to support large-scale projects because a single server can replace up to 40 standard nodes, reducing hardware, maintenance costs, and other expenses, including power consumption and cooling.

Scalable

The performance of Lenovo GOAST scales linearly from single-node appliance to cluster implementation to serve the needs of labs of all sizes, from small research groups to commercial labs, and to national population-level projects. This includes transitioning from WES to WGS, undertaking a new project with greater scope and complexity, and expanding both data and users. Organizations can scale linearly simply by adding compute and storage building blocks as needed.

Methods

Methodology used for testing:

Lenovo's GOAST contains pre-tuned GATK scripts to maximize the performance of variant calling on specially tuned Lenovo hardware. The workflow manager used here is Snakemake, and Slurm is the recommended job scheduler for a cluster setup.

The pipelines for Germline Variant Calling and Somatic Variant Calling support both whole genome and whole exome samples and follow the GATK Best Practices pipeline. The preprocessing phase performs alignment with BWAmem, continues with MarkDuplicates to remove various duplication artifacts, and finishes with Base Quality Score Recalibration (BQSR) to ensure that accurate quality scores are being passed to the variant caller. GOAST v4.0 provides the option to use BWA mem2 for up to a 30 percent increase in performance and to visualize FASTQC reports with MultiQC (optional).

The Germline Variant Calling pipeline uses the standard GATK workflow utilizing Haplotypecaller with an option to emit VCF or GVCF files as output. Joint Calling uses the option to choose between the GenomicsDB or CombineVariants methods, depending on data types and cohort sizes. Variants from an individual sample or from a jointly called cohort of samples may be filtered with the hard-filtering workflow.

The Somatic Variant Calling pipeline call variants with GATK's Mutect2 calculates and filters contaminants. This supports matched Tumor and Normal sample pairs (recommended) in addition to Tumor Only samples. Optional annotation step of Funcotator.

Running GOAST showing 50X coverage for WES, 30X coverage for WGS:

The data provided here are based on single-node testing. These results are the average of three iterations. See Figure 1 for an overview of a typical NGS analysis from sample to variant output.

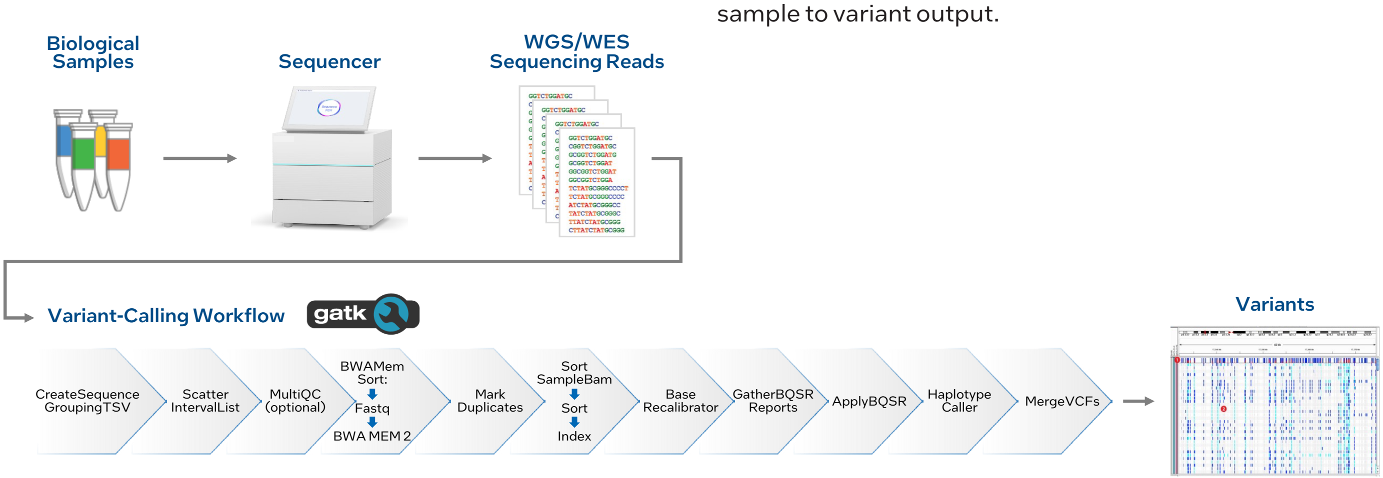


Figure 2. A typical Next-Generation Sequencing (NGS) workflow illustrating the sequential software calls required to perform WGS/WES variant-calling analyses. Biological samples (i.e., blood, saliva, etc.) processed experimentally are input into a sequencer, which in turn generates sequencing reads (fragments 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).

The number of samples (exome) processed per day on the Intel platform, starting from a 2nd Gen Intel CPU to a 5th Gen Intel CPU with the updated GOAST platform, is shown in Figure 2.

a. **Measuring performance.** The present study measured performance as execution time for the entire workflow, starting from the input file, called a FASTQ file, to the *.vcf output file. Lenovo's search for a "hardware + software + system recipe" prioritized identifying the best solution for performance, cost, and usability. Lenovo also provided a throughput measure in the form of genomes (or exomes) per compute node per day.

b. **Hardware evaluations.** Two different Intel Xeon Scalable processors were used for all full workflow analyses: Intel Xeon **Platinum 8480+** and Intel Xeon Platinum **8592+** processors. Solutions incorporating the latest Intel Xeon **Platinum 8480+** processors and Intel Xeon Platinum **8592+** processors deliver significant performance gains as compared to similarly configured solutions based on previous-generation Intel Xeon Scalable processors.

Results

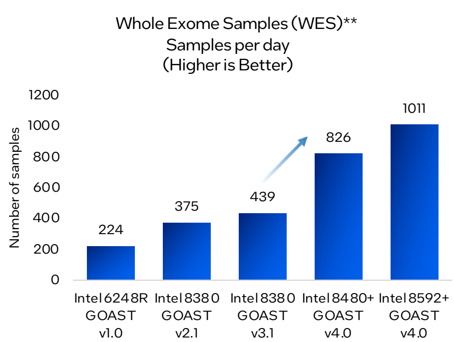
GOAST: A Two-pronged HPC-Architecture Plus HPC-Scaler Solution

Determining the factors affecting the performance of genomics workflows is a complex problem. Genomics workflows string together a combination of 28 java applications, 24 of which are single-threaded, and 4 of which are multi-threaded. 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 has 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.

As Figure 2 illustrates, GOAST architecture has delivered up to a 4.5X speed-up for whole exomes (WES) and up to a 6X increase for all genomes (WGS) since the introduction of GOAST v1.0.

As a result of permutation tests of the hardware, software, and system factors affecting the performance of genomics workflows, Lenovo identified an optimized architecture that can process over 1,000 WES samples per node per day and 42 WGS samples per node per day. GOAST architecture leverages an optimized variant-calling workflow and a concise, simple, non-specialty hardware recipe to deliver an affordable solution with peak performance.



From v2.1 to v3.1, Lenovo began using fastq files as input instead of ubams.

From v3.1 to v4.0, they began using a new input file with RTA3 quality scores instead of RTA2.

These changes meant that the files were smaller, and performance improved due to improved I/O issues.

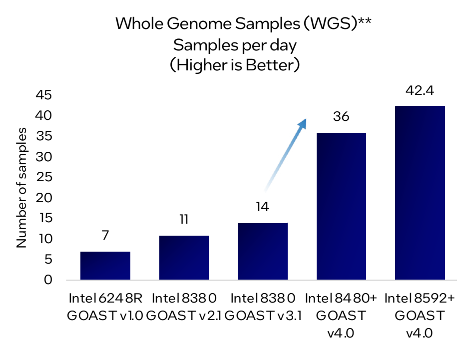


Figure 3. Results for multigenerational comparison data.

Similarly, GOAST v4.0, running on 5th Gen Intel Xeon processors, enables processing of approximately 14 percent more tumor normal samples per day than was possible with 4th Gen Intel Xeon processors (Figure 3).

** WES - 64 samples were used and processed. WGS - 16 samples were used and processed.

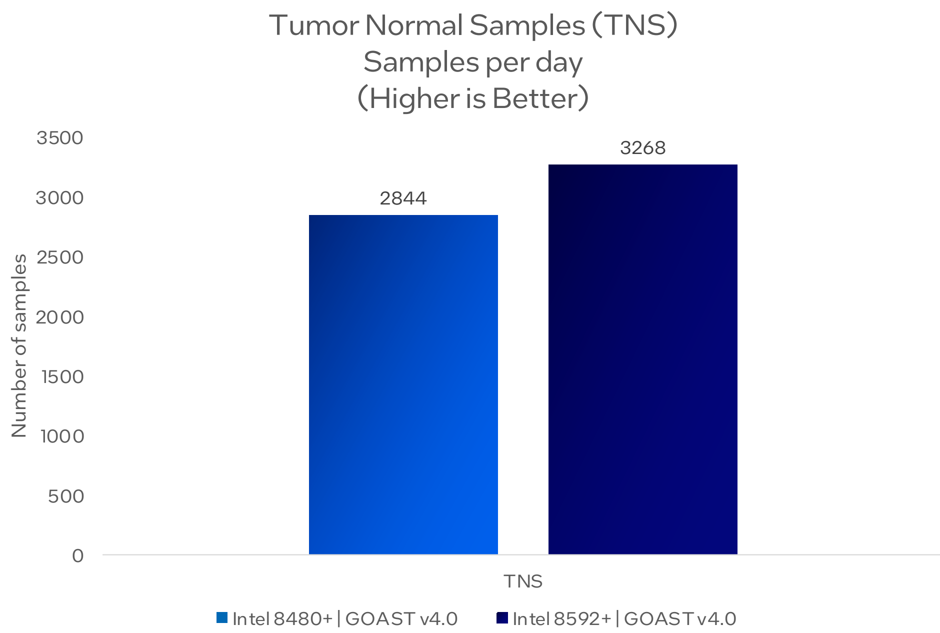


Figure 4. Gen-over-gen increase in tumor normal samples processed daily. For WES, 64 samples were used and processed. For WGS, 16 samples were used and processed.

Together, these results highlight how the optimized performance of GOAST v4.0 can ease the data analysis bottleneck and decrease time to discovery.

Summary

Through five iterations of GOAST, Lenovo and Intel have worked closely together to meet the computational needs of customers' bioinformatics workloads. This collaboration has yielded steady improvements in performance with each new iteration. With GOAST v4.0, the latest update from 4th Gen to 5th Gen Intel Xeon processors is, again, accelerating execution speeds, allowing for the processing of 22 percent more WES samples, 17 percent more WGS samples, and 14 percent more tumor normal samples every day.

Improvements are ongoing, as GOAST remains in active development with Lenovo adapting for changes of prevalent datatypes, ensuring that software updates work seamlessly together, and adding workflows as customers request them. The most recent additions include Somatic Variant Calling and Joint Calling.

Preconfigured, optimized hardware and workflows streamline processing for reduced time to insight. Equally consequential, the performance of GOAST scales linearly, which makes it easy for organizations of all sizes to take on projects with greater scope and complexity. Moving forward, Lenovo and Intel plan to continue their collaboration on GOAST, keeping the focus on finding new ways to enable research through increased efficiency and usability.

Accelerated by Intel

To deliver the best experience possible, Lenovo and Intel have optimized this solution to leverage Intel capabilities like processor accelerators not available in other systems. Accelerated by Intel means enhanced performance to help you achieve new innovations and insight that can give your company an edge.

accelerated by intel

For More Information, Visit These Sites

- **HPC Solutions | High-Performance Computing | Lenovo US**
- Powering groundbreaking plant genomics research | Lenovo US
- Bringing personalized medicine to citizens across Saudi, Arabia | Lenovo US
- How Multi-Scaled HPC-Enabled Genomics Will Help Save Your Life | nextplatform.com
- Unlocking new insights into serious diseases | Lenovo US

[1] McKay A, Wilman HR, Dennis A, Kelly M, Gyngell ML, Neubauer S, et al. (2018) Measurement of liver iron by magnetic resonance imaging in the UK Biobank population. PLoS ONE 13(12): e0209340. https://doi.org/10.1371/journal.pone.0209340

[2] 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.

Configurations

1-node, 2x Intel® Xeon® Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS ESE118F-2.22, microcode 0x2b0004b1, 2x Ethernet Controller E810-XXV for SFP, 4x I350 Gigabit Network Connection, 2x BCM57508 NetXtreme-E 10Gb/25Gb/40Gb/50Gb/100Gb/200Gb Ethernet, 2x 894.3G Micron_7450_MTFDKBA960TFR, 1x 7T SAMSUNG MZQL27T6HBLA-00A07, Red Hat Enterprise Linux 9.2 (Plow), 5.14.0-284.30.1.el9_2.x86_64, WORKLOAD+VERSION, COMPILER, LIBRARIES, OTHER_SW, score=?UNITS. Test by Lenovo as of 10/27/23.

1-node, 2x Intel® Xeon® Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS ESE122N-3.10, microcode 0x21000161, 2x Ethernet Controller E810-XXV for SFP, 4x I350 Gigabit Network Connection, 2x BCM57508 NetXtreme-E 10Gb/25Gb/40Gb/50Gb/100Gb/200Gb Ethernet, 2x 894.3G Micron_7450_MTFDKBA960TFR, 1x 7T SAMSUNG MZQL27T6HBLA-00A07, Red Hat Enterprise Linux 9.2 (Plow), 5.14.0-284.30.1.el9_2.x86_64, WORKLOAD+VERSION, COMPILER, LIBRARIES, OTHER_SW, score=?UNITS. Test by Lenovo as of 10/31/23.

1-node, 2x Intel® Xeon® 8380,40 cores (2.4GHz, 270W), SR630 V2, 512GB RAM, 16x 32GB/3200MHz DIMMs, 7TB SSD

1-node, 2x Intel® Xeon® 8380, 40 cores(2.30GHz, 270 W), 1024GB RAM, 16x 32GB/3200MHz DIMMs, 7TB SSD

1-node, 2x Intel® Xeon® 6248R CPUs, 24 cores (3.0GHz, 205W), SD650, 384GB RAM, 12x 32GB/2933MHz DIMMs, 1.92TB, 1x 2.5" SAS SSD

Genomics Glossary

Term/Acronym	Definition
Exome	An exome is the sequence of all the exons in a genome, reflecting the protein-coding portion of a genome. In humans, the exome is about 1.5 percent of the genome.
Genome	The genome is the entire set of DNA instructions found in a cell. In humans, the genome consists of 23 pairs of chromosomes located in the cell's nucleus, as well as a small chromosome in the cell's mitochondria. A genome contains all the information needed for an individual to develop and function.
Variant	A DNA sequence observed at a site on a single chromosome. Note that the DNA sequence in a reference sequence at that site is also a variant. A ref-matching variant is the variant found in the reference sequence. A ref-mismatching variant is any other variant.
Tumor only (Comparison Mode)	Compares a tumor-normal pair of sequences to find variants called in the tumor but not in the normal.