Lenovo GOAST Bioinformatics Solution

Solution Brief

A Genomics and Bioinformatics Optimized System

Bioinformatics is the computational analysis of biological data powering research all the way from basic biology to medicine, to drug discovery, to agriculture, and more. The last two decades has seen an explosion of bioinformatics data which has been enabled by the advancements in computational resources. Yet, to date, even at cluster and supercomputer speeds, large-scale bioinformatics still faces long execution times on massive volumes of data, which delays “time to answer”.

The only high-performance solutions to mitigate these challenges are single-purpose boutique solutions requiring expensive specialty hardware and substantial licensing fees. These single-purpose genomics solutions require organizations to purchase multiple architectures to support other types of research in their datacenter. Even those organizations working in a single subfield (e.g. in genomics) find that the single-purpose boutique solutions are not enough.

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. Therefore, Lenovo developed Genomics Optimization and Scalability Tool (GOAST), a Genomics and Bioinformatics Optimized platform.

Extremely Fast Bioinformatics Analytics

Lenovo Genomics Optimization and Scalability Tool (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 (WGS) variant calling analytics from days to minutes. This process typically takes 40 hours in many data centers, but runs in just ~23.5 minutes in GOAST systems.

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

	GOAST Intel Base	**GOAST AMD Base**
Processor	2x Intel 8592+ CPUs (64 cores, 1.9GHz)	2x AMD 9654 CPUs (96 cores, 2.4 GHz)
Memory	1 TB RAM, 16x 64GB/5600MT/s RDIMMs	1.5 TB RAM, 24x 64GB/4800MHz DIMMs
Storage	Minimum 7TB SATA or NVMe SSD	Minimum 7TB SATA or NVMe SSD

Key features of Lenovo GOAST

The key features of Genomics Optimization and Scalability Tool (GOAST) include the following:

- **Extremely fast analytics**

  Bioinformatics optimized hardware runs sequential workflows faster: e.g. process 30x whole genomes at a throughput rate of up to 2.5 samples/hour and 50x whole exomes at 60 samples /hour

- **Increases lab productivity**

  Faster time to insight: e.g. up to ~22K whole genomes/node/year

- **Multi-purpose Bioinformatics use**

  Leverage BOSS's high-core, fast I/O, and high-memory specs to run any Bioinformatics or HPC tools or scripts

- **Cost effective**

  Up to 50% less than boutique solutions relying on GPUs or FPGAs without additional licensing fees. A single GOAST server can replace up to 50 standard nodes

- **Scalable**

  Deploy as a single-node appliance or as a cluster and grow linearly with flexibility

- **Easy to use**

  Simplified wrapper scripts for omics at the command line

GOAST v4.0 Capabilities

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

- Germline variant calling (WES and WGS)
- Germline joint calling (WES and WGS)
- Somatic short variant discovery (WES and WGS)
    - Tumor + Normal pair
    - Tumor Only

All needed dependencies come pre-installed in a Conda environment which 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, utilizing Conda to manage software installations as well as a complete rewrite of the backend GOAST Util Tool.

Increased 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 we tune our 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).

Table 2. Lab Productivity for Omics expected on a single GOAST system*

Expected Lab Productivity	**30x WGS Samples processed (n)**		**50x WES Samples processed (n)**	
	GOAST Intel Base	**GOAST AMD Base**	**GOAST Intel Base**	**GOAST AMD Base**
WGS/day/node	42.4	60.0	1,011	1,440
WGS/year/node	15,500	21,900	369,000	525,600

* 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.

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 the research questions, instead of spending valuable time tuning hardware, tweaking software versions, and optimizing workflows.

Multi-purpose Bioinformatics use

GOAST is a high-performance system for multi-purpose Bioinformatics use. The 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 necessary to run the Broad Institute's GATK Best Practices for Germline and Somatic SNP and Indel discovery. Lenovo GOAST also provides pre-configured 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 our team about pre-installing 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 pre-installed by GOAST 4.0

Software	Version
GATK	4.4.0.0
BWA	0.7.17
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% 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 since a single GOAST Plus 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. Scale linearly simply by adding compute and storage building blocks as needed.

Which GOAST Configuration is right for me?

The optimal GOAST configuration depends on your lab's throughput needs. Both GOAST Intel Base (42.4 WGS/day/node) and GOAST AMD Base (60.0 WGS/day/node) can support the output of an Illumina NovaSeq 6000 at full capacity, which produces 26 samples per day. The purpose of Lenovo GOAST is to help enable research though increasing efficiency and usability. Lenovo will work with you to find the best solution to support your bioinformatics analytics needs.

For more information see NovaSeq 6000 System Specifications:
https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html

Learn More

For more information, see the following resources:

- Genomics and Bioinformatics
- GOAST and Deep sequencing and De novo assembly at the University of Delhi
- Bringing personalized medicine to citizens
- How Multi-Scaled HPC-Enabled Genomics Will Help Save Your Life (nextplatform.com)

For questions reach out to Dana Alegre, M.S., Solutions Architect, Life Sciences, Lenovo HPC & AI.
dalegre@lenovo.com

Author

Dana Alegre is Lenovo's Solution Architect in the HPC Life Sciences vertical leading the Lenovo Genomics Optimization and Scalability Tool (GOAST) team. GOAST works to enable research through optimizing genomics workflows on Lenovo hardware. She has been asking and answering biological questions by analyzing next generation sequencing data, first in the Genomics Core at the Stowers Institute for Medical Research and at the Center for Quantitative Life Sciences at Oregon State University. Collaborating with dozens of groups conducting genomics and bioinformatics research over the years, has resulted in publications in Science and Nucleic Acids Research. Her professional accomplishments include developing a bioinformatics pipeline to support the Oregon Health Authority's efforts to identify and monitor variants in wastewater to help combat the COVID-19 pandemic.

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1888, was created or updated on May 1, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1888
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1888.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®

The following terms are trademarks of other companies:

Intel® is a trademark of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Other company, product, or service names may be trademarks or service marks of others.