

Solution Snapshot

Intel® Select Solutions for Genomics Analytics



Genomics Analytics Overview + Benefits

Intel® Select Solutions for Genomics Analytics add value to the alignment and variant calling step of genetic analyses. Intel has optimized performance in these steps by tuning Broad Institute's Genomics Analytics Toolkit (GATK) to take advantage of Intel® architecture.

Working together, Intel and the Broad Institute are pursuing three goals:

1. Develop an **optimized software stack** to analyze genomes faster and at greater scale.
2. Enable and qualify **turnkey configuration, setup, and deployment** of infrastructure to run genomics analysis.
3. **Take advantage of key technologies** to allow for ever-increasing datasets to be analyzed to deliver scientific and medical breakthroughs.

The Challenge

There is currently a race to find a cure for diseases such as cancer and COVID-19.

Setting up and scaling a genomics analytics cluster poses many challenges that can create bottlenecks for genomics researchers. For example, experts predict that in a few years, the field of genomics may generate up to 40 exabytes of data per year.¹ Processing this amount of data, with the scalability required in a fast-changing field, is a substantial compute problem.

Today, Researchers need:

- Ability to run a broad range of applications with diverse system requirements
- Ways to address increasing datasets and complexity
- Faster time to results
- Reduced total cost of ownership (TCO)

Use Cases:



Genomic Sequencing



Population Scaling



Cancer/Rare Disease Research



Agricultural Crop Research



Endangered Species Protection

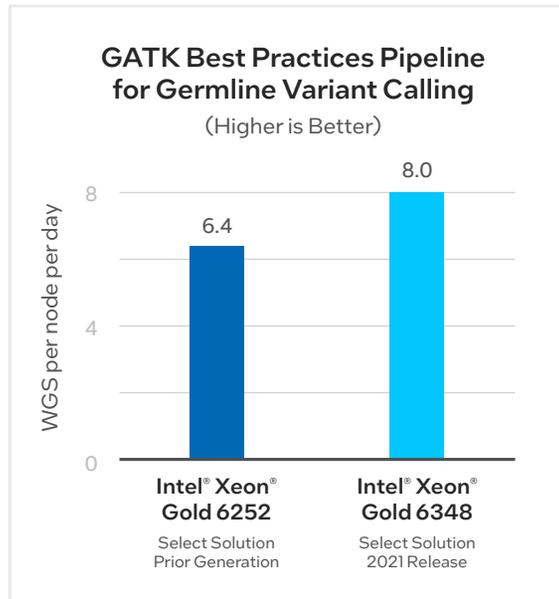


COVID-19 Pandemic Research

Intel® Select Solutions for Genomics Analytics are designed to scale from small (four-node minimum) to very large, clustered supercomputers. The customized systems can be quickly and dynamically configured to meet specific needs for organizations to scale as they grow their workloads. In addition, end-to-end optimized hardware and software solutions for genomic analysis.

- Optimized HPC cluster that can support multiple genomics pipelines as well as other HPC workloads
- Uses Intel® technology: 3rd Gen Intel® Xeon® processors and Intel® Solid State Drives (Intel® SSDs)
- WDL (workflow description language) scripts to allow users to replicate Genome Analysis Toolkit best practice pipelines and create their own pipelines quickly and easily

Proof Points



Pipeline for Germline Variant Calling benefits from the increased number of memory channels and support for 3200 MHz DRAM, along with higher core count (compared to 2nd Gen Intel® Xeon® Scalable processors).

Large-capacity Intel® SSD Data Center Family drives—either NAND-based or Intel® Optane™ SSDs—also contribute to the solution's high performance.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Why Intel for Genomic Analytics?

Intel® technologies integrated into Intel® Xeon® Scalable processors deliver further performance and reliability gains:

- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512):** Boosts performance for the most demanding computational workloads, with up to double the number of FLOPS per clock cycle, compared to previous-generation Intel® processors.
- **Intel® Cluster Checker:** Inspects more than 100 characteristics related to cluster health. Intel® Cluster Checker examines the system at node and cluster levels, enabling all components to work together to deliver optimal performance.
- **Cluster Management Software Stack:** Provides a software stack required to deploy and manage Linux HPC clusters. The stack includes provisioning and development tools, resource management, I/O clients, and scientific libraries. Resource management tools such as Bright Cluster Manager, Warewulf, and xCAT support the software stack.
- **Intel® Cluster Checker Runtimes:** Supplies key software runtime elements required on each cluster to ensure optimal performance paths for applications. Intel® runtime performance libraries, such as Intel® Math Kernel Library and Intel® MPI Library, are available in the Intel® oneAPI toolkit. This toolkit helps deliver excellent performance that is optimized for clusters based on Intel® architecture. They can also simplify the development and deployment of data-centric workloads across CPUs, GPUs, FPGAs, and other accelerators.

Want More Information?

What are Intel® Select Solutions for HPC?

Intel® Select Solutions for HPC removes the guesswork of selecting the right set of high-performance computing products, hardware configurations, and software.

Resources

[Intel® Select Solutions for HPC](#)

Read more on the Intel® Select Solutions Genomics Analytics Solution Brief

[Intel® Select Solutions for Genomics Analytics](#)

[3rd Generation Intel® Xeon® Scalable processors](#)

[Broad Institute of MIT and Harvard](#)

[Genome Analytics Toolkit](#)

1 National Center for Biotechnology Information, "Big Data: Astronomical or Genomic?" <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494865/>

3rd Generation Intel® Xeon® Scalable Processor Configuration: Test by Intel as of August 8, 2021. One front-end node and four compute nodes, all using Intel® Server Board

M50CYP25B-003. Front-end node configuration: 2x Intel® Xeon® Gold 6348 processor (28 cores, 2.90 GHz); Intel® Hyper-Threading Technology = ON, Intel® Turbo Boost Technology = ON, total memory 256 GB (16 slots/16 GB/3200 MHz); BIOS version: 22D08; BMC 2.66, SDR 0.31, CPLD 3p0; uCode: 0x0b000280; storage – boot drive 1x Intel® SSD P4610 1.6 TB (3D NAND PCIe 3.1 x4, 3D1, TLC); high-performance network: 1x Intel® Ethernet Converged Network Adapter X550-T2 (10 GbE), model X550T2. Compute node configuration: 2x Intel Xeon Gold 6348 processor (28 cores, 2.60 GHz); Intel Hyper-Threading Technology = ON, Intel Turbo Boost Technology = ON, total memory 512 GB (16 slots/32 GB/3200 MHz); BIOS version: 22D08; BMC 2.66, SDR 0.31, CPLD 3p0; uCode: 0x0b000280; storage – scratch drive: 1x Intel SSD P4610 1.6 TB (3D NAND PCIe 3.1 x4, 3D1, TLC); high-performance network: 1x Intel Ethernet Converged Network Adapter X550-T2 (10 GbE), model X550T2.

2nd Generation Intel® Xeon® Scalable Processor Configuration: Test by Intel as of November 14, 2019. One application node and four compute nodes. Application node configuration: 2x Intel® Xeon® Gold 6252 processor (24 cores, 2.10 GHz); 1x Intel® Server Board S2600WFT; 64 GB (4x16 GB DDR4 2933 MHz); 1x 960 GB Intel® SSD D3-S4510 Series (2.5 in SATA 6 Gb/s, 3D2,TLC); 1x 1.6 TB Intel® SSD DC P4610 Series (2.5 in PCIe 3.1 x4, 3D2, TLC); Microcode: 0x500002c, BIOS: SE5C620.86B.02.01.0009.092820190230; CentOS Linux Installation ISO (minimal or full) 7.7 build 1910; Intel® Cluster Runtimes 2019.4; Intel® Cluster Checker 2019.3.5; Intel® Select HPC Solution for RPM packages for EL7 2018.0; OpenHPC 1.3.8. 4x compute nodes configuration: 2x Intel® Xeon® Gold 6252 processor (24 cores, 2.10 GHz); 1x Intel Server Board S2600WFT; 384 GB (12x 32 GB DDR4 2933MHz); 1x 960 GB Intel SSD D3-S4510 Series (2.5 in SATA 6 Gb/s, 3D2, TLC); 1x 1.6 TB Intel SSD DC P4610 Series (2.5 in PCIe 3.1 x4, 3D2, TLC); Network devices: 1x Intel® C620 Series Chipset Ethernet Connection; Intel® Ethernet Adapter X722 onboard 10 GbE; Microcode: 0x500002c, BIOS: SE5C62 0.86B.02.01.0009.092820190230; 1x distributed 10 GB Lustre 2.10 ZFS system, 6 OST, 3 OSS, LNet Router with single 10 GB link for all I/O traffic clients to Lustre servers plus 1x 1.6 TB Intel SSD DC P4610 Series (2.5 in PCIe 3.1 x4, 3D2,TLC)