Solution Brief

November 2021

Speed Time to Insight with Intel® Select Solutions for Genomics Analytics

Quickly and easily build scalable genomics analytics clusters using a verified design powered by the latest technical innovations from Intel and the Broad Institute

intel. SELECT SOLUTIONS

Executive Summary

Genomics analytics is at the forefront of modern medicine. Examples include identifying variants of the SARS-CoV2 virus to confront the COVID-19 pandemic and designing targeted treatment programs for cancer, cardiovascular issues, and other diseases based on a patient's genetic information. The global genomics market is expected to more than double in the next four years—from USD 22.7 billion in 2020 to USD 54.4 billion in 2025.[1]

But processing genomic data requires monumental compute resources as well as advanced software applications. And constructing a genomics analytics cluster to optimize performance can be laborious. In a field where uncovering insights can literally affect the lives of people around the world, time is precious.

Intel® Select Solutions enable organizations to deploy high-performance compute infrastructure quickly and efficiently to achieve reliable, security-enabled, and workload-optimized performance on a balanced platform. Customers spend less time, effort, and expense evaluating hardware and software options. Intel Select Solutions also help customers simplify design choices by bundling hardware and software while guaranteeing that the system meets or exceeds minimum performance metrics.

Intel Select Solutions for Genomics Analytics provide optimized performance, scale, and ease of deployment for genomics analytics. This validated reference design is an end-to-end hardware and software package developed in collaboration with the Broad Institute of MIT and Harvard. The latest release of Intel Select Solutions for Genomics Analytics enables users to run the Broad Institute's GATK Best Practices Pipeline for Germline Variant Calling to process up to eight Whole Genome Sequencing (WGS) samples per node per day. This high-throughput performance—a 25 percent increase compared to the performance of the previous generation of this solution—is made possible by the 3rd Generation Intel® Xeon® Scalable processor platform.[2]

Intel® Select Solutions for Genomics Analytics

Intel Select Solutions power genomics analytics with 3rd Gen Intel® Xeon® Scalable processors, Intel® SSDs, Intel® Ethernet Network Adapters, and other Intel technologies.

- LATEST GATK RELEASE
- VALIDATED DESIGN STREAMLINES DEPLOYMENT
- HIGH-PERFORMANCE INTEL TECHNOLOGY
- DELIVER PERFORMANCE OPTIMIZED TO A SPECIFIC THRESHOLD

Business Challenge

Setting up and scaling a genomics analytics cluster pose many challenges that can create bottlenecks for genomics researchers. For example, the infrastructure must be able to store and process massive amounts of data—a typical genomics analytics input dataset is about 150 GB[3] (assuming 30X coverage), and higher coverage samples can be hundreds of GBs in size. Experts predict that in a few years, the field of genomics may generate up to 40 exabytes of data per year.[4] Processing this amount of data, with the scalability required in a fast-changing field, is a substantial compute problem.

Time is also a factor. It takes time and effort to design a genomics analytics cluster. Considerations include determining the best hardware and software components to deploy and integrating all the components, so they work well together. Pipelines that are critical for diagnosis and treatment can take weeks to run, so optimizing settings for the best performance is important. Every delay can have negative implications for fast diagnosis and treatment selection.

Solution Value

Intel® Select Solutions are ISV certified, OEM validated, and verified by Intel. Benefits of adopting an Intel Select Solution include the following:

- **Workload-optimized.** Jointly developed with ISVs, Intel Select Solutions are delivered as recipes that integrate the latest hardware and software across top workloads.
- **Premium configurations.** They are benchmarked to deliver optimized performance, price/performance, and security.
- **Verified by Intel.** They are delivered by OEM or solution providers who have met or exceeded a minimum performance level.

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. The solution also includes workflow description language (WDL) scripts tuned for deploying on local HPC infrastructure. These scripts allow users to replicate GATK Best Practices pipelines quickly and easily and to create their own pipelines.

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

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

Solution Benefits

- End-to-end optimized hardware and software solution for genomics analysis
- Optimized, scalable computing cluster that can support multiple genomic pipelines as well as other HPC workloads
- Predictable performance of up to eight Whole Genome Sequencing samples per node per day[2]

Solution Architecture Highlights

Intel's reference design for genomics clusters (see Figure 1) includes all the necessary components: compute, storage, memory, and network fabric in a balanced configuration that helps eliminate performance bottlenecks. Intel Select Solutions for Genomics Analytics take advantage of the high-performance capabilities of 3rd Gen Intel® Xeon® Scalable processors. In particular, the 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.

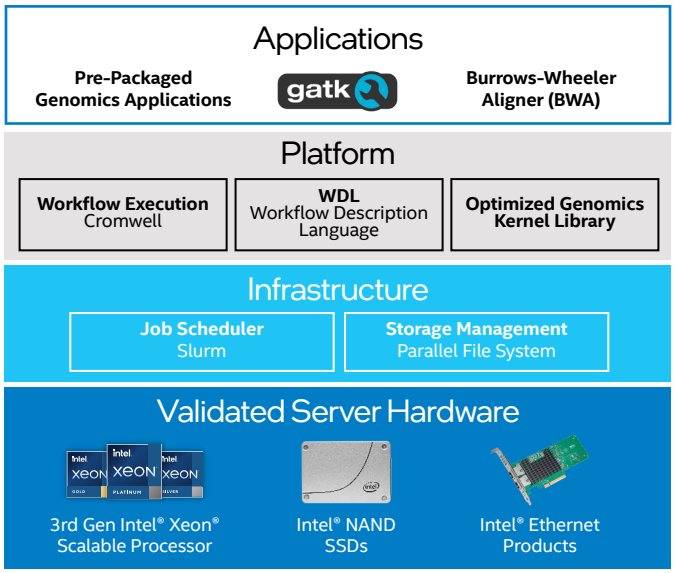

Figure 1. The balanced design of Intel® Select Solutions for Genomics Analytics helps eliminate performance bottlenecks.

What Are Intel® Select Solutions?

Predefined, workload-optimized solutions designed to minimize the challenges of infrastructure evaluation and deployment.

These solutions are validated by OEM/ODMs, certified by ISVs, and verified by Intel.

All Intel Select Solutions are a tailored combination of Intel® data center compute, memory, storage, and network technologies that deliver predictable, trusted, and compelling performance. Each solution offers assurance that the workload will work as expected, if not better, which can save individual businesses from investing the resources that might otherwise be used to evaluate, select, and purchase the hardware components to gain that assurance themselves.

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, and organizations can scale as they grow their workloads. In addition, Intel Select Solutions for Genomics Analytics include tools to discover, compose, and monitor resources with powerful, modern API-based software, along with a line-by-line installation guide.

In addition to the Intel hardware foundation used for Intel Select Solutions for Genomics Analytics, the following 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.[5]
- **Intel® oneAPI Runtimes.** Supplies key software runtime elements that are required on each cluster to help 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, and also simplifies development and deployment of data-centric workloads across CPUs, GPUs, FPGAs, and other accelerators.
- **Intel® Cluster Checker.** Inspects more than 100 characteristics related to cluster health. Intel Cluster Checker examines the system at both the node level and cluster level, making sure all components 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.

A Closer Look at the Broad Institute–Intel Collaboration

Intel and the Broad Institute of MIT and Harvard have been collaborating for several years to scale researchers' ability to analyze massive amounts of genomic data from diverse sources worldwide. Through this collaboration, researchers and software engineers are building, optimizing, and widely sharing new tools and infrastructure that can help scientists integrate and process genomic data.

We are working with Broad Institute in three areas:

- Overcome the challenge of diverse genomic datasets by optimizing Broad's GATK Best Practices hardware recommendations for genomic workloads for on-premises, public cloud, and hybrid cloud use cases.
- Simplify and accelerate the execution of genomics analytics by optimizing genomics software tools such as GATK and Cromwell on industry-standard Intel architecture-based platforms.
- Empower users such as healthcare providers, pharmaceutical companies, and academic research organizations to collaborate on workflow execution models across complex and distributed datasets. Achieving this goal will enable highly secure processing of data across organizations, which can stimulate research and discovery, drug development, clinical trial recruitment, and clinical decision-making across the entire research and discovery ecosystem.

The collaboration between Intel and Broad combines two powerful forces in the advancement of HPC and application-specific analysis. These dedicated teams are delivering solutions to the challenges of disease discovery and treatment.

Results

Customers who haven't recently refreshed their genomics analytics hardware and software can experience substantial improvement in throughput by upgrading to the latest Intel Select Solutions for Genomics Analytics. In this validated reference design, Intel has demonstrated a throughput of up to eight Whole Genome Sequencing (WGS) samples per node per day using 3rd Gen Intel Xeon Scalable processors and NAND-based Intel® SSDs. This is a 25 percent performance increase compared to the previous-generation solution,[6] and double the performance of the first generation of the solution (see Figure 2).

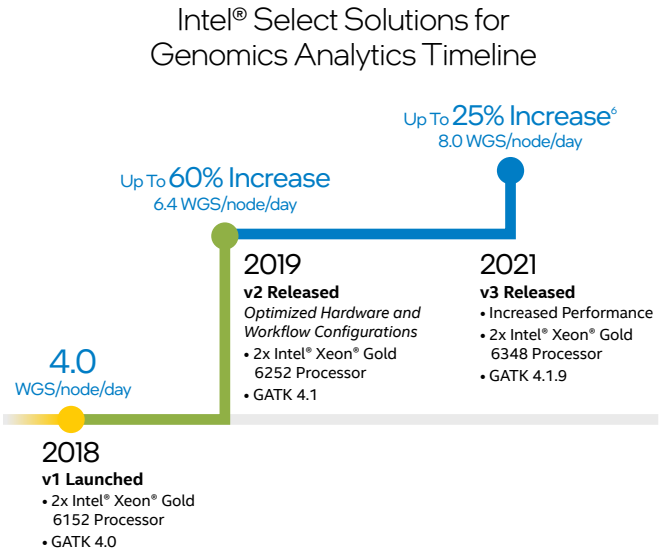

Figure 2. The latest release of Intel® Select Solutions for Genomics Analytics vastly improves the speed at which analysis can be performed, unlocking more genetic information that can be used to change lives.

Learn More

You may also find the following resources useful:

- Intel® Select Solutions for Genomics Analytics
- 3rd Generation Intel® Xeon® Scalable processors
- Broad Institute of MIT and Harvard
- Genome Analytics Toolkit

[1] MarketsandMarkets, "Genomics Market by Product & Service, Technology, Application, End User – Global Forecast to 2025," marketsandmarkets.com/Market-Reports/genomics-market-613.html

[2] **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 M50CYP2SB-003. **1x front-end configuration:** 2x Intel® Xeon® Gold 6348 processor (28 cores, 2.60 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; Microcode: 0x0b000280; CentOS Linux installation ISO (minimal or full) 8 build 2011; 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.
4x compute nodes 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; Microcode: 0x0b000280; CentOS Linux installation ISO (minimal or full) 8 build 2011; 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 front-end node and four compute nodes, all using Intel® Server Board S2600WFT. **Front-end node configuration:** 2x Intel® Xeon® Gold 6252 processor (24 cores, 2.10 GHz); total memory 64 GB (4 slots/16 GB/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® oneAPI 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); total memory 384 GB (12 slots/32 GB/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); Network devices: 1x Intel® C620 Series Chipset Ethernet Connection; Intel® Ethernet Adapter X722 onboard 10 GbE; Microcode: 0x500002c, BIOS: SE5C620.86B.02.01.0009.092820190230; CentOS Linux Installation ISO (minimal or full) 7.7 build 1910; 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.

1st Generation Intel® Xeon® Scalable Processor Configuration: Test by Intel as of October 15, 2018. **Single-node (compute and front-end node combined) configuration:** 2x Intel® Xeon® Gold 6152 processor (22 cores, 2.10 GHz); Intel® Server Board S2600WFT; total memory 192 GB (12 slots/16 GB/2666 MHz); boot storage: 2x 480 GB Intel® SSD DC S3520 Series; cache storage: 4x 4 TB Intel® SSD DC P4600 Series PCIe HHHL; capacity storage: 16 TB of 4x 4 TB Intel® SSD DC P4510 Series; Intel Hyper-Threading Technology = ON, Intel Turbo Boost Technology = ON; Microcode: 0x043; CentOS Linux installation 7.6.

[3] Strand NGS, "Guide to Storage and Computation Requirements," https://www.strand-ngs.com/support/ngs-data-storage-requirements#:~:text=Allowing%20for%20some%20extra%20analysis,sample%20to%20about%208%20GB.

[4] National Center for Biotechnology Information, "Big Data: Astronomical or Genomical?" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494865/

[5] Intel AVX-512 provides higher throughput to certain processor operations. Due to varying processor power characteristics, using Intel AVX-512 instructions may cause a) some parts to operate at less than the rated frequency, and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration. Learn more at intel.com/go/turbo.

[6] See endnote 2.

Performance varies by use, configuration and other factors. Learn more at intel.com/PerformanceIndex. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. No product or component can be absolutely secure. Your results may vary. Intel technologies may require enabled hardware, software or service activation. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. 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. © Intel Corporation 1121/GMCK/KC/PDF