SOLUTION BRIEF

Intel® Scalable System Framework
High Performance Computing
June 2018 v2

Intel® Select Solutions for Genomics Analytics

Access performance, scale, and ease of deployment for genomics insight and discovery.

Intel Select Solutions for Genomics Analytics are based on the BIGstack 2.0* reference architecture.

Advancements in genomics are opening new doors for understanding human diseases, and they are increasingly informing innovative precision treatment plans. Discoveries are dependent on processing, storing, and analyzing a growing amount of genomic sequencing data. In 2015, worldwide sequencing storage capacity approached a petabyte per year, and it continues to double every seven months.[1,2] At this rate, genomics sequencing will generate hundreds of petabytes per year in the next five years, and could require nearly a zettabyte of storage per year by 2025.[1,2]

The Broad Institute of MIT and Harvard (broadinstitute.org) is one of the world's largest producers of human genomic data, creating about 24 TB of new data per day. Currently, Broad Institute manages more than 50 PB of data.

Researchers require tools to analyze these enormous volumes of data in a timely manner to gain insights into disease and possible treatments. They need tools like the Genome Analysis Toolkit* (GATK*), a set of leading software methods created by the Broad Institute and trusted by the majority of genomics centers worldwide.

Broad Institute will release GATK 4.0 as its next major version, under an open source license for all users, including for commercial purposes. An open source license will make GATK available to a wider audience of scientists and researchers and will help accelerate and advance genomics analytics worldwide.

Intel-Broad Center for Genomic Data Engineering

Intel and Broad Institute have collaborated on computing infrastructure and software optimization for years. In 2017, they launched a new effort—the Intel-Broad Center for Genomic Data Engineering is a five-year collaboration between the two organizations to simplify and accelerate genomics workflow execution using GATK, Burrow-Wheeler Aligner (BWA), Cromwell, Intel® Genomics Kernel Library (Intel® GKL), GenomicsDB*, and other tools and techniques. Together, experts from Broad Institute and Intel will build, optimize, and widely share tools and infrastructure to help scientists integrate and process genomic data. The result will be a growing set of optimized best practices in hardware and software for genomics analytics on Intel® architecture–based platforms that can be applied to research data sets stored in private data centers and that will extend to private, public, and hybrid clouds.

With the massive growth of genomics data, the collaboration makes use of technology to enable genomics analytics at scale. It has already resulted in Intel® Select Solutions for Genomics Analytics, a suite of optimized software, along with reference architectures for turnkey configuration, setup, and deployment to run genomics analysis that is qualified for GATK pipelines, Cromwell, and GenomicsDB.

Intel Select Solutions for Genomics Analytics

The Intel-Broad Center for Genomic Data Engineering works to optimize GATK on Intel architecture and technologies and to define a reference architecture for genomics analytics. The result is Intel Select Solutions for Genomics Analytics, developed by Intel and the Broad Institute, based on the BIGstack 2.0* reference architecture, and delivered by Intel solution providers. The solutions provide a five-times overall performance improvement running GATK 4.0 compared to previous versions of the genomics software, and they reduce setup time for deploying an infrastructure to accelerate genomics workflows.[3] Performance gains include a 75 percent speedup for the BWA using Intel® Solid-State Drives (SSDs) and a two-times speedup for HaplotypeCaller* using Intel® Field-Programmable Gate Arrays (Intel® FPGAs).[3] The validated performance and quality results have been certified by Broad Institute.

What Are Intel® Select Solutions?

Intel Select Solutions are verified hardware and software stacks that are optimized for specific software workloads across compute, storage, and network. The solutions are developed from deep Intel experience with industry solution providers, in addition to extensive collaboration with the world's leading data center and service providers.

To qualify as an Intel Select Solution, solution providers must:

1. Follow the software and hardware stack requirements outlined by Intel
2. Replicate or exceed Intel's reference benchmark-performance threshold
3. Publish a detailed implementation guide to facilitate customer deployment

Solution providers can develop their own optimizations to add further value to their solutions.

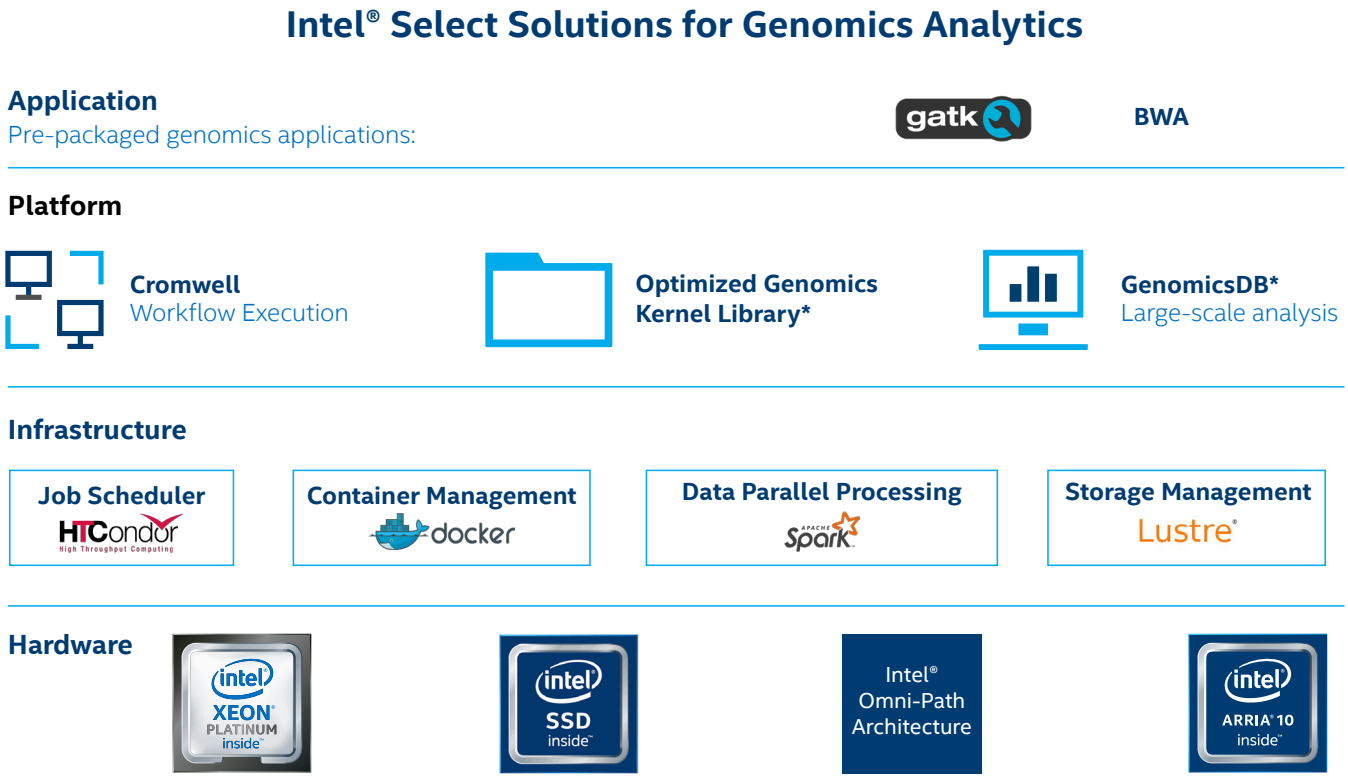


Figure 1. A high-level overview of the solution configuration

Photo credit: Len Rubenstein & Broad Institute

High-performance data analytics computing clusters and optimized workflows for genomics analytics are complicated hardware and software systems. Intel Select Solutions for Genomics Analytics are end-to-end optimized hardware and open source software configurations designed specifically to accelerate genomics analytics—both the deployment of systems and the software that runs on them—by providing verified stacks for setup and configuration of these complicated genomics pipelines.

Intel Select Solutions for Genomics Analytics are designed to scale from small to very large clustered supercomputers. The customized systems can quickly and dynamically be configured to meet specific needs. Organizations can scale as they grow their workloads. And Intel Select Solutions for Genomics Analytics include tools to discover, compose, and monitor resources with powerful, modern API-based software.

Table 1. The Base and Plus configurations for Intel® Select Solutions for Genomics Analytics

INGREDIENT	INTEL® SELECT SOLUTIONS FOR GENOMICS ANALYTICS **BASE CONFIGURATION**	INTEL SELECT SOLUTIONS FOR GENOMICS ANALYTICS **PLUS CONFIGURATION**
HEAD NODES		1 x head node
PLATFORM		Intel® Server Board S2600WFT
PROCESSOR		Intel® Xeon® Platinum 8180 processor (28 cores, 2.5 GHz)
MEMORY		128 GB DDR4-2666
HOST ADAPTERS		100HFA016LS Intel® Omni-Path Host Fabric Interface Adapter, Peripheral Component Interconnect Express* (PCIe*) x16
APPLICATION NODES	1 x application node (also gateway node)	1 x application node
PLATFORM	Intel Server Board S2600WFT	Intel Server Board S2600WFT
PROCESSOR	Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)	Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)
MEMORY	128 GB DDR4-2666	512 GB DDR4-2666
HOST ADAPTERS	100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16	100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16
COMPUTE FLEXIBLE NODES	4 x compute flexible nodes	8 x compute flexible nodes
PLATFORM	Intel Server Board S2600WFT	Intel Server Board S2600WFT
PROCESSOR	Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)	Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)
MEMORY	256 GB DDR4-2666	256 GB DDR4-2666

LOCAL STORAGE	7 x 4 TB Intel® SSD DC P4600 Series, PCIe HHHL	7 x 4 TB Intel SSD DC P4600 Series, PCIe HHHL
HOST ADAPTERS	100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16	100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16
ADD-IN CARDS	1 x Intel® Arria® 10 FPGA	1 x Intel Arria 10 FPGA
COMPUTE HIGH-DENSITY NODES	4 x compute high-density node	16 x compute high-density nodes
PLATFORM	Intel Server Board S2600BPB	Intel Server Board S2600BPB
PROCESSOR	Intel Xeon Platinum 8176 processor (28 cores, 2.1 GHz)	Intel Xeon Platinum 8176 processor (28 cores, 2.1 GHz)
MEMORY	256 GB DDR4-2666	256 GB DDR4-2666
LOCAL STORAGE	4 TB Intel SSD DC S4500 Series, Serial ATA (SATA) 2 TB Intel SSD DC P4501 Series, PCIe M.2	4 TB Intel SSD DC S4500 Series, SATA 2 TB Intel SSD DC P4501 Series, PCIe M.2
HOST ADAPTERS	100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16	100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16
ADD-IN CARDS	1 x Intel Arria 10 FPGA	1 x Intel Arria 10 FPGA
NETWORK INFRASTRUCTURE		
DATA NETWORK	Intel® Omni-Path Edge Switch 100 Series, 24-port	Intel Omni-Path Edge Switch 100 Series, 48-port
MANAGEMENT NETWORK	1 gigabit per second (Gbps) 48x port switch	1 Gbps 48x port switch
NFS STORAGE INFRASTRUCTURE		
PLATFORM	Intel Server Board S2600WFT	
PROCESSOR	Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)	
MEMORY	256 GB DDR4-2666	
DISKS	2 x 480 GB Intel SSD DC S3520 Series (mirrored OS)	
HOST ADAPTERS	12 Gbps Intel® RAID Controller RS3SC008, JBOD mode 100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16	
JBOD	Intel® Storage System JBOD2312S3SP x12 drive JBOD	
DRIVES	12–24 x Seagate* enterprise, capacity: 6–10 TB	
RAID	ZFS-on-RAID	
LUSTRE* STORAGE INFRASTRUCTURE		
LUSTRE MDS*		2 x metadata server
PLATFORM		Intel Server Board S2600WFT
PROCESSOR		Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)
MEMORY		256 GB DDR4-2666
DISKS		2 x 480 GB Intel SSD DC S3520 Series (mirrored OS)
HOST ADAPTERS		12 Gbps Intel RAID Controller RS3SC008, JBOD mode; 100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16
LUSTRE OSS		2 x object storage server
PLATFORM		Intel Server Board S2600WFT
PROCESSOR		Intel Xeon Platinum 8180 processor (28 cores, 2.5 GHz)
MEMORY		256 GB DDR4-2666
DISKS		2 x 480 GB Intel SSD DC S3520 Series (mirrored OS)
HOST ADAPTERS		12 Gbps Intel RAID Controller RS3SC008, JBOD mode; 100HFA016LS Intel Omni-Path Host Fabric Interface Adapter, PCIe x16
LUSTRE OST		Object storage target
JBOD		Colfax CX4270-JBOD* 44x drive with dual expanders
DRIVES		44 x Seagate enterprise, capacity: 6–10 TB
RAID		ZFS-on-RAID
LUSTRE MDT		Meta data target
JBOD		Intel Storage System JBOD2224S2DP, 2U, dual-port
SAS SSDS		4 x HGST* SAS 400 GB

SOFTWARE	GATK*, BWA, and GATK workflows optimized for Intel® technologies Optimized Cromwell workflow Intel® GKL with optimized routines for accelerating developer codes GenomicsDB*, specializing in large-scale variant analysis HTCondor* job scheduler for running clustered analytics jobs Docker* for running multiple jobs in isolated containers across a cluster Apache Spark* for big data analytics processing Lustre, the open source parallel file system, for high-performance storage	GATK, BWA, and GATK workflows optimized for Intel technologies Optimized Cromwell workflow Intel GKL with optimized routines for accelerating developer codes GenomicsDB, specializing in large-scale variant analysis HTCondor job scheduler for running clustered analytics jobs Docker for running multiple jobs in isolated containers across a cluster Apache Spark for big data analytics processing Lustre, the open source parallel file system, for high-performance storage
FIRMWARE AND SOFTWARE OPTIMIZATIONS	Intel® Advanced Vector Extensions 512 (Intel® AVX-512)	Intel AVX-512

Intel Select Solutions for Genomics Analytics: Software, Firmware, and Technology Configuration

Intel Select Solutions for Genomics Analytics take advantage of the high-performance capabilities of Intel architecture, including Intel® Xeon® Scalable processors, Intel SSD Data Center Family drives, Intel® Omni-Path Architecture high-performance fabric, and Intel FPGAs.

Table 1 shows hardware and software for both the "Base" and "Plus" configurations of Intel Select Solutions for Genomics Analytics. To refer to a solution as an Intel Select Solution, a server vendor or data center solution provider must use these or better configurations.

These solutions can be tailored with 2, 4, 16, 24, 36, or 48 of the specified compute devices and, when applicable, local and shared storage devices in order to meet the needs of individual environments.

Simplified Code Development with Intel® Genomics Kernel Library (Intel® GKL)

Intel GKL provides code used in genomics that is optimized for Intel architecture. Developers can call these routines to help accelerate their code performance. The library enables developers to focus on the function and operation of their code (instead of specific optimizations), while letting Intel GKL make use of the capabilities of Intel architecture.

Scalability Improvements with GenomicsDB*

GenomicsDB is a unique variant store capable of supporting up to hundreds of thousands of genome variant data. It was first developed by Intel Labs and customized for Broad Institute's use cases.

GenomicsDB will be packaged with GATK 4.0, which will help significantly accelerate workflows. It offers a large scaling benefit to the HaplotypeCaller Genomic VCF (GVCF) workflow. For example, without using GenomicsDB, Broad Institute took six weeks to generate a database from 2,300 whole genomes. With GenomicsDB, it was able to generate databases with five times more information in only two weeks.[4] That successfully enabled the Broad Institute–hosted Genomics Aggregation Database* (gnomAD*) project, which includes 15,000 whole genomes—one of the largest genomic data aggregations in the world.[4]

Continuing Development

There are large genomic databases around the world that can bring great benefits to worldwide research efforts. The ongoing work of the Intel-Broad Center for Genomic Data Engineering continues to develop Intel Select Solutions for Genomics Analytics to efficiently access those databases for analysis. In the future, incorporated technologies will provide the connectivity, performance, privacy, and security necessary for genomics in the cloud and shared environments.

Benefits of the Intel and Broad Institute Collaboration

The work of Intel and Broad Institute offers many benefits to the genomics community and to the technologists and business managers that support it, including:

Scientists Who Enjoy:

- Support for optimized and efficient pipelines
- Optimized, turnkey solutions
- Prepackaged workflow description languages (WDLs) scripts
- Peer application support
- Low-touch IT support
- Access to more in-house genomics data
- Increased statistical power
- Open source software
- Flexible application architecture

Photo credit: Len Rubenstein & Broad Institute

IT Departments Who Need:

- Ease of implementation
- Scalability
- Reduced setup time
- Open source software with no licensing costs
- Known reference architecture
- Vendor and solution support
- Optimal use of hardware versus workload (for example, prepackaged WDLs)

Business Owners Who Enjoy:

- The ability to scale the solution to fit a budget
- Low price/watt
- Preconfigured solutions to reduce setup time and support costs
- Maximized value for in-house genomic data
- No license fees
- Open source application software
- Extendability to other applications

OEM Partners—Simplifying Genomics Analytics Cluster Deployment

The introduction of Intel Select Solutions for Genomics Analytics makes it easier to run genomics workloads. It also enables accelerated deployment of predictable clusters designed for genomics analytics. Thus, many integrators of high-performance systems have partnered with Intel and are offering design and deployment of solutions that will meet the needs of their customers in the genomics community.

"Our goal is to reduce the challenges that researchers face to generate ever-more-meaningful insights from ever-larger sets of genomics data. For us, running GATK 4 on version 1.0 of the Intel Select Solutions for Genomics Analytics delivered a 5x performance gain right away. We're working with Intel to make the GATK Best Practices pipelines run even faster, at even greater scale, and with easier deployment for genomic research worldwide."

— Geraldine Van der Auwera, Associate Director of Outreach and Communications, Data Sciences Platform Group, Broad Institute

Access Performance, Scale, and Ease of Deployment for Genomics Analytics

The work of genomics science is critical to the understanding of disease and the creation of diagnostic tools and safe and effective therapies. Genomics data and analytics are quickly advancing as researchers use technology to build massive genomics data repositories and come to understand the power of that data. Broad Institute is one of the largest contributors of genomics data in the world, and its GATK software is the world's leading genome analysis tool for analytics and variant call research. The Intel-Broad Center for Genomic Data Engineering brings together science and technology to optimize genomics analytics codes and workflows and to define an optimized infrastructure—Intel Select Solutions for Genomics Analytics—to run those workloads. The results enable faster analysis and quicker times to deploy hardware solutions that are customized for genetics analysis. Several system integrators already offer services to install such systems that will continue to enable further discoveries through genetics.

Intel® Xeon® Scalable Processors

Intel Xeon Scalable processors:

- Offer high scalability for enterprise data centers
- Deliver performance gains for virtualized infrastructure compared to previous-generation processors
- Achieve exceptional resource utilization and agility
- Enable improved data and workload integrity and regulatory compliance for data center solutions

The family includes Intel Xeon Bronze processors, Intel Xeon Silver processors, Intel Xeon Gold processors, and Intel Xeon Platinum processors.

Learn More

The Intel-Broad Center for Genomic Data Engineering: **intel.com/broadinstitute**

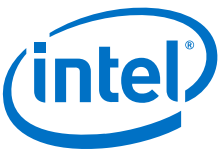
"Big Data Genomics and Optimized Genomics Code": **intel.com/content/www/us/en/healthcare-it/solutions/genomicscode.html**

Intel and Broad Institute white paper, "Infrastructure for Deploying GATK Best Practices Pipeline": **intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf**

Intel Select Solutions: **intel.com/selectsolutions**

Intel Xeon Scalable processors: **intel.com/xeonscalable**

Intel Select Solutions are supported by Intel® Builders: **http://builders.intel.com**. Follow us on Twitter: **#IntelBuilders**



[1] Stephens, Zachary D., et al. "Big Data: Astronomical or Genomical?" PLOS Biology. July 2015. https://doi.org/10.1371/journal.pbio.1002195.

[2] Robison, Reid J. "How Big Is the Human Genome?" Precision Medicine. January 2014. https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0.

[3] Intel. "Infrastructure for Deploying GATK Best Practices Pipeline." November 2016. intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf.

[4] Geraldine Van der Auwera, Ph.D. Broad Institute. Bio-IT World. May 2017.

Printed in USA 0618/JS/PRW/PDF Please Recycle 336731-003US