Solution Design Brief

Data Centers | Genomics Analytics
Intel Accelerated Solution

Accelerating Genomics Analytics with 5th Gen Intel® Xeon® Scalable Processors While Keeping Compute Costs Low

Business Challenge: How fast can your infrastructure provide answers to some of the world's most important questions about disease and treatment? Do you feel bogged down by sluggish hardware but are also concerned about rising analytics costs?

Up to
61%
Improvement in Throughput
On 5th Generation Intel® Xeon® Scalable processors vs 4th Generation

Solution Overview and Summary

Solution: Genomics analytics is at the heart of modern science. Researchers use genomics analytics to drive new drug discovery. Doctors can leverage it as part of the treatment regimen for patients with cancer and rare diseases. Genomics analytics is also being used worldwide for population sequencing, crop improvement and studying the human microbiome. The faster researchers can find answers, the faster they can change lives for the better.

This validated reference design for genomics analytics is an end-to-end hardware and software package developed in collaboration with the Broad Institute of MIT and Harvard. It's designed for optimized performance, scale and ease of deployment to allow customers to spend less time, effort and expense evaluating hardware and software options. All components are tuned to take advantage of the latest innovations in Intel® architecture.

Performance isn't the only reason to deploy this genomics analytics solution. It also delivers excellent total cost of ownership (TCO), costing only about USD 2.34 per genome (as calculated by the compute cost for the secondary analysis stage).[1] With this low per-genome cost, researchers can afford to run more analyses and uncover new insights.

Results: Genomics Analytics Toolkit (GATK) software tested on 5th Generation Intel® Xeon® Scalable processors demonstrates up to a 61% gen-over-gen improvement in throughput compared to 4th Gen Intel Xeon Scalable processors.[2] For details, refer to the full discussion of results on page 2.

Test Methodology

The test cluster setup includes one head node and four compute servers installed on an OpenHPC platform. (Detailed software information is available in the "Configuration Details" section.) Cromwell was used as the workflow management system, which is configured to use Slurm (backend scheduler) to submit throughput jobs (70 WGS samples). The publicly available NA12878 30x coverage whole genome sequence (WGS) was used as the input dataset. A single dataset size is about 85 GB and each WGS analysis consumes up to 480 GB when completed. Figure 1 shows the experimental setup.

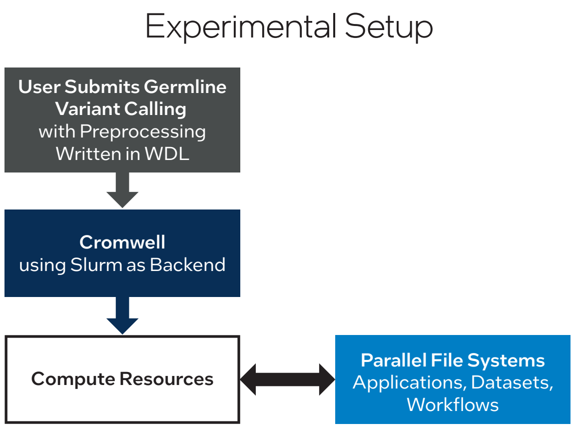

Figure 1. Experimental workflow setup.

Results

Customers who haven't recently refreshed their genomics analytics hardware and software can experience substantial improvement in throughput by upgrading to the latest release of Intel's Accelerated Solution for Genomics Analytics. In this validated reference design, Intel has demonstrated a throughput of up to 14.81 WGS samples per node per day using 5th Gen Intel Xeon Scalable processors. This is a 61% performance increase compared to the previous-generation solution (see Figure 2).[2] The speedup is a result of the following characteristics of the new generation of processor, compared to the prior generation:

- 1.5x more cores
- Higher frequency (2.3 GHz versus 2.0 GHz)
- Higher memory bandwidth (5600 MT/s versus 4800 MT/s), providing a 1.16x benefit
- Higher-capacity L3 cache

Additional performance benefits accrue as a result of software improvements from the Genomics Kernel Library (GKL)—developed by Intel—integrated with Broad's GATK to better take advantage of Intel® Advanced Vector Extensions 512 (Intel® AVX-512).

Compared to even earlier generations of Intel Xeon Scalable processors, upgrading to 5th Gen Intel Xeon Scalable processors enables you to process more than triple the number of WGS samples per node per day (see Figure 3).[2]

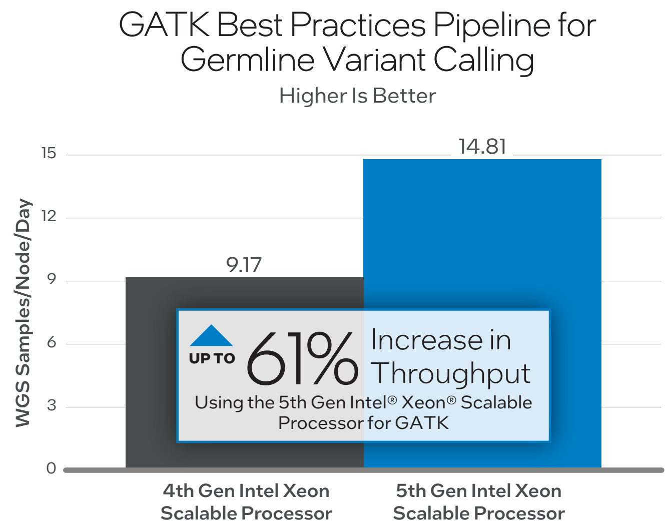

Figure 2. Boost your genomics analytics throughput by up to 61% by upgrading your genomics analytics clusters from 4th Gen Intel® Xeon® Scalable processors to the latest generation.[2]

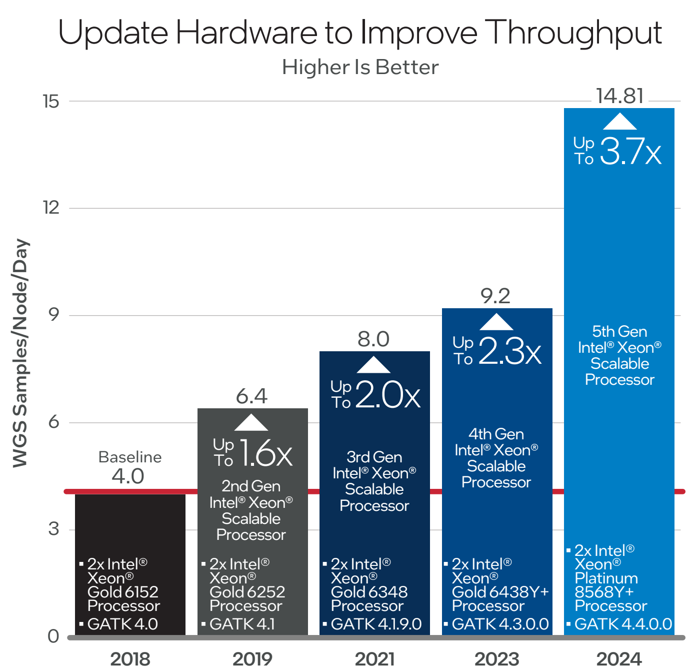

Figure 3. Compared to older hardware, the latest release of Intel's Accelerated Solution for Genomics Analytics more than triples the speed at which analysis can be performed, unlocking more genetic information that can be used to change lives.[2]

Configuration Details

The following tables provide information about components and settings of the infrastructure used for performance analysis and characterization testing with 4th Gen and 5th Gen Intel Xeon Scalable processors. See endnote 2 for configuration details for head nodes and previous-generation processors.

Hardware Configurations (per compute node)

	4th Generation Intel® Xeon® Scalable Processor	5th Generation Intel® Xeon® Scalable Processor
Processor	2x Intel Xeon Gold 6438Y+ processor (32 cores, 2.0 GHz, 60 MB L3 cache)	2x Intel Xeon Platinum 8568Y+ processor (48 cores, 2.3 GHz, 300 MB L3 cache)
Memory and Storage	512 GB (16x 32 GB DDR5 @ 4800 MT/s)	512 GB (16x 32 GB DDR5 @ 5600 MT/s)
Storage	1x Solidigm D3-S4510 Series	1x Solidigm D3-S4510 Series
Network Card	2x Intel® Ethernet Controller X710 for 10GBASE-T	2x Intel Ethernet Controller X710 for 10GBASE-T

Software Versions

	4th Generation Intel® Xeon® Scalable Processor	5th Generation Intel® Xeon® Scalable Processor
Workload	GATK Best Practices for Germline Variant Calling WholeGenomeGermlineSingleSample_v3.1.6	GATK Best Practices for Germline Variant Calling WholeGenomeGermlineSingleSample_v3.1.19
Applications	▪ WARP 3.1.6 ▪ GATK 4.3.0.0 ▪ Picard 3.0.0 ▪ Samtools 1.17 ▪ Burroughs-Wheeler Aligner (BWA) 0.7.17 ▪ VerifyBamID 2.0.1 ▪ MariaDB 10.3.35 ▪ Slurm 20.11.9 ▪ Cromwell 84	▪ WARP 3.1.19 ▪ GATK 4.4.0.0 ▪ Picard 3.1.1 ▪ Samtools 1.18 ▪ Burroughs-Wheeler Aligner (BWA) 0.7.17 ▪ VerifyBamID 2.0.1 ▪ MariaDB 10.3.39 ▪ Slurm 20.11.9 ▪ Cromwell 84
Libraries	▪ java-1.8.0-openjdk ▪ java-11-openjdk	▪ java-17-openjdk ▪ GKL8.11, gcc 8.5.0-18.el8

Important System Settings

Number of Nodes	1 head node + 4 compute nodes
Intel® Hyper-Threading Technology	Enabled on compute nodes only
Intel® Turbo Boost Technology	Enabled on compute nodes and header node

Intel® Processors Deliver Sustainable Genomics

Genomics is at the forefront of medical progress—but sustainability is also a priority for researchers. Balancing data center power consumption and high-performance genomics computing can be a concern. The good news is that 5th Gen Intel® Xeon® Scalable processors provide both power savings and the performance that's necessary to accelerate genomics research. The 61% gen-over-gen throughput improvement demonstrated in recent testing was achieved with a power consumption of only 0.341 kWh/WGS sample, which translates to a mere 120.7 grams of CO2 emissions per WGS sample.[3]

Only
.0341 kWh/WGS
per sample

Just
120.7 grams
of CO2 emissions per WGS sample

Conclusion

Genomics analytics helped find answers during the COVID-19 pandemic. It has transformed cancer treatments by enabling clinicians to choose treatments that are most likely to work with an individual's genetic makeup. It's even helping scientists determine how best to protect the world's most important pollinator—bees—from disease and environmental changes. We've come a long way from the first genome sequencing, which took 13 years. Now it is possible to process a WGS sample in just a few hours. Intel is committed to working with the genetics analytics ecosystem to continue to accelerate sequencing and drive down costs. This validated genomics analytics solution can contribute to these goals by lowering compute cost per genome, speeding up WGS sample processing, and making deployment of the solution quick and easy.

Further Information

- 5th Gen Intel® Xeon® Scalable Processors
- Intel® AVX-512
- Broad Institute of MIT and Harvard
- Genome Analysis Toolkit
- OpenHPC

Cloud & Enterprise Solutions Group Authors

Urszula Golowicz, Cloud Solutions Engineer

Michael J. McManus, Director, Precision Medicine & Principal Engineer

Paolo Narvaez, Sr. Director, Enterprise Solutions Engineering & Sr. Principal Engineer

Priyanka Sebastian, Cloud Solutions Architect

Read More About Running Genomics Analytics on Intel® Hardware:

- Deploying genomics workflows on HPC platforms
- Evaluating Genomics Pipelines on Azure

Contact your Intel representative to learn more about this solution.

Solution Provided By:

intel.

[1] Based on internal Intel analysis as of November 2023, the $2.27 per genome is calculated for processing a 30x whole genome sample (NA12878). Furthermore, this cost was derived as follows: Our benchmark of 14.81 WGS per compute node per day is converted to an annual number of 5,406 genomes (14.81 WGS/node/day * 365 days). Using the average cost of $38,000 for a 2-socket compute node from leading hardware companies, we further assume this cost is depreciated over a 3-year period. Our cost per genome is computed by taking the annual genome of number of 5,406 times 3 years for a 3-year total of 16,218 genomes. The 3-year genome total is divided into the server cost of $38,000 to yield per-genome cost of ~$2.34.

[2] **5th Generation Intel® Xeon® Scalable Processor Configuration:** Test by Intel as of February 8, 2024. 1x head node and 4x compute servers, all using Quanta Cloud Technology Inc. server board S6Q-MB-MPS.

1x Head Node Configuration: 2x Intel Xeon Gold 6448Y processor (32 cores, 2.1 GHz), Intel® Hyper-Threading Technology = OFF, Intel® Turbo Boost Technology = ON, total memory 512 GB (16x 32GB DDR5 4800 MT/s [4800 MT/s]), BIOS = 3A16.TEL0P1, microcode = 0x2b0004b1, Red Hat Enterprise Linux 8.8 (Ootpa), Kernel 4.18.0-477.10.1.el8_8.x86_64, storage: Solidigm D3-S4510 Series, network: 2x Ethernet Controller X710 for 10GBASE-T.

4x Compute Nodes Configuration: 2x Intel Xeon Platinum 8568Y+ processor (48 cores, 2.3 GHz), Intel Hyper-Threading Technology = ON, Intel Turbo Boost Technology = ON, total memory 512 GB (16x 32 GB DDR5 5600 MT/s [5600 MT/s]), BIOS = 3B05.TEL4P2, microcode = 0x21000161, Red Hat Enterprise Linux 8.8 (Ootpa), Kernel 4.18.0-477.10.1.el8_8.x86_64, storage: Solidigm D3-S4510 Series, network: 2x Ethernet Controller X710 for 10GBASE-T, workload: WholeGenomeGermlineSingleSample 3.1.19, GATK 4.4.0.0, java-17-openjdk, GKL8.11, gcc 8.5.0-18.el8, Cromwell 84, slurm 20.11.9, Samtools 1.18, VerifyBamID 2.0.1, BWA 0.7.17, Picard 3.1.1, MariaDB 10.3.39.

4th Generation Intel Xeon Scalable Processor Configuration: Test by Intel as of May 10, 2023. 1x application server and 4x compute servers, all using Quanta Cloud Technology Inc. server board S6Q-MB-MPS.

1x Front-End Node Configuration: same as compute node configuration.

4x Compute Nodes Configuration: 2x Intel Xeon Gold 6438Y+ processor (32 cores, 2.0 GHz), Intel Hyper-Threading Technology = ON, Intel Turbo Boost Technology = ON, total memory 512 GB (16x 32 GB @ 4800 MT/s), BIOS = WLYDCRB1.SYS.0021.P25.2107280557, microcode = 0xd000363, Rocky Linux 8.6, Kernel 4.18.0-372.32.1.el8_6.x86_64, storage: Solidigm D3-S4510 Series, network: 2x Intel® Ethernet Controller X710 for 10GBASE-T, workload: WholeGenomeGermlineSingleSample 3.1.6, GATK 4.3.0.0, java-1.8.0-openjdk & java-11-openjdk, Cromwell 84, Picard 3.0.0, Samtools 1.17, BWA 0.7.17, VerifyBamID 2.0.1, MariaDB 10.3.35.

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 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 256 GB (16x 16 GB @ 3200 MHz), BIOS = 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 Solidigm DC P4610 (formerly Intel) 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 (16x 32 GB @ 3200 MHz), BIOS = 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 Solidigm DC P4610 (formerly Intel) 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.

1x Front-End configuration: 2x Intel Xeon Gold 6252 processor (24 cores, 2.10 GHz), total memory 64 GB (4x 16 GB @ 2933 MHz), 1x Solidigm D3-S4510 Series (formerly Intel) 960 GB (2.5 in SATA 6 Gb/s, 3D2, TLC), 1x Solidigm DC P4610 (formerly Intel) (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 (12x 32 GB @ 2933 MHz), 1x Solidigm D3-S4510 Series (formerly Intel) 960 GB (2.5 in SATA 6 Gb/s, 3D2, TLC), 1x Solidigm DC P4610 (formerly Intel) 1.6 TB (2.5 in PCIe 3.1 x4, 3D2, TLC), network devices: 1x Intel® C620 Series Chipset Ethernet Connection and 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 (12x 16 GB @ 2666 MHz), boot storage: 2x Intel® SSD DC S3520 Series 480 GB, cache storage: 4x Intel® SSD DC P4600 Series 4 TB (PCIe HHHL), capacity storage: 4x Solidigm DC-P4510 Series (formerly Intel) 4 TB, Intel Hyper-Threading Technology = ON, Intel Turbo Boost Technology = ON, microcode = 0x043, CentOS Linux installation 7.6.

[3] **Calculating the kWh per WGS:** Number of 30x whole genomes = 70; total runtime (hours) = 28.37; average watts across all 4 nodes = 841.89; total power (runtime x average watts) = 23,88 watt-hours or 23.88 kWh; average power per whole genome = 23.88 kWh/70 WGS = 0.341 kWh/WGS.

Calculating the CO2 per WGS: We used the US average of 354 gCO2eq/kWh as of May 7 at 4:00 pm ET from this source: https://app.electricitymaps.com/map?lang=en

Total CO2 per 4 nodes = 23.88 kWh x 354 gCO2eq per kWh = 8,454 grams CO2 or 8.454 kg CO2.

Total CO2 per WGS = 8,454 grams/70 WGS = 120.7 grams CO2 per WGS.