Solution Brief

FPGA
Genomic Analytics

Acceleration of Sequence Alignment and Variant Calling for Genomic Analytics Using Intel® FPGAs

Authors

Natalia Poliakova
Technical Solution Sales Specialist
Intel Corporation

Ioannis Stamelos
Chief Solution Architect
InAccel

Elias Koromilas
Chief Technology Officer
InAccel

Aspasia Stavrianou
Senior FPGA Engineer
InAccel

Chris Kachris
CEO
InAccel

Calvin Hung
CEO
WASAI

Introduction

Genomic Analytics aligns a selected genome to a reference genome to detect genetic variants in that selected genome as compared to the reference genome. This technique is fundamental to the diagnosis and cure of rare, inherited diseases as well as for medical breakthroughs and personalized care. The medical industry is progressing towards personalized medicine, which will require the storage of many, many human genomes. There are 3 billion nucleotide base pairs in a human genome, which translates into immense data volumes needed to store everyone's genome.

As the Coronavirus has raced around the world, thousands of genome sequences of the virus have been shared on GISAID[1], an online global platform for genomic data. One Coronavirus genome sequence contains 26K to 32K bases in the RNA strand located inside the coronavirus. These shared sequence variants offer clues about how the virus, named SARS-CoV-2, is spreading and evolving. But because these shared sequences represent a tiny fraction of cases and show few tell-tale differences, they are easy to overinterpret.[2]

Virologist Eeva Broberg of the Centre for Disease Prevention and Control[3] states that "there are more plausible scenarios for how the disease reached northern Italy than an undetected spread from Bavaria."[3] This statement underscores the importance of fast sequence alignment of the Coronavirus mutations.

"*The very first SARS-CoV-2 sequence, in early January, answered the most basic question about the disease: What pathogen is causing it? The genomes that followed were almost identical, suggesting the virus, which originated in an animal, had crossed into the human population just once. If it had jumped the species barrier multiple times, the first human cases would show more variety. Some diversity is now emerging. Over the length of its 30,000-base-pair genome, SARS-CoV-2 accumulates an average of about one to two mutations per month. Using these little changes, researchers draw up phylogenetic trees, much like family trees, make connections between cases, and gauge whether there might be undetected spread of the virus.*"[3] Fast analysis and tracking of the mutations is critical for better protection of the population as the virus evolves and migrates between the countries and geographies.

Scientists will be scouring the genomic diversity of these viral genome sequences for signs that the virus is getting more dangerous. Caution is warranted. An analysis of 103 genomes published by Lu Jian of Peking University and colleagues on 3 March 2020 in the National Science Review argued that they fell into one of two distinct types, named S and L, and are distinguished by two mutations. Because 70% of sequenced SARS-CoV-2 genomes belong to a newer type L genome, the authors concluded that the type L genome has evolved to become more aggressive and spreads faster."[3]

Genomic scientists and researchers within different groups around the world have been trying to uncover genetic determinants of susceptibility, severity, and outcomes of COVID-19 from the genomes of COVID-19 patients. COVID-19 is the pandemic disease caused by the SARS-CoV-2 coronavirus.

Using genomic analytics on the sequences of novel viruses to track mutations is a computationally intensive algorithm and requires powerful processing platforms for the efficient processing of huge amounts of data. Fast genome sequencing will require hardware with more processing power, storage capacity, and network bandwidth to keep up. With larger genomic datasets that can exceed 300 GB of data per patient, one can only imagine the computing and speed demands of genomic analysis needed to deal with the genomic variants within 6 billion base pairs of human whole genomes from more than a million samples that are stored in global biobanks. These analyses can take months.

As the industry advances, we need scalable software and hardware resources to adapt to the growing performance demands. Cost is also a concern, because traditionally, more computing power means more hardware components, translating to higher cost. To speed up the process and reduce the expense, we can accelerate the process with specialized FPGA accelerators.

Introducing Intel®-based Acceleration of Sequence Alignment and Variant Calling

The main goal of the Sequence Alignment and Variant Calling Acceleration project is to develop high-performance genome analytics platforms using hardware accelerators that can be harnessed by researchers for the diagnosis and treatment of deadly diseases like coronavirus and its mutations. In this project, the accelerators, the integrated framework for the large deployment, and the interface with the infrastructure will be developed and evaluated. The infrastructure could be based either on a cloud infrastructure with FPGA resources or an on-premise infrastructure. The accelerators will be integrated with an FPGA orchestrator that will allow scaling to multiple servers and FPGAs. It will also allow resource sharing by multiple users and researchers.

Researchers are looking for a tailor-made architecture for FPGAs to study deadly viruses and human genomes. One of the project partners, WASAI Technology, will be introducing high-performance hardware accelerators for Genome Analytics. WASAI has announced the completion of a genome analytics accelerated solution based on the Intel® Programmable Acceleration Card with Intel® Arria® 10 GX FPGA (Intel® PAC with Intel® Arria® 10 GX FPGA). According to WASAI, this solution improves performance by 5X to 9X when compared to the traditional software solution for germline whole exome sequencing or whole genome sequencing[4]. Introduction of an Intel PAC based on the Intel® Stratix® 10 SX FPGA shortens the sequencing time even more.[4]

This project aims to integrate proprietary and open-source accelerators for genome analytics using an FPGA manager for scalable deployment of cloud clusters and HPC data centers that incorporate FPGA-based accelerators. Inaccel's Coral orchestrator serves as a framework that allows the distributed acceleration of large datasets across clusters of FPGA resources using simple programming models. Inaccel will integrate WASAI's FPGA-accelerated GATK pipeline in its Coral Framework and will manage features at the application layer, delivering highly available service on top of a cluster of Intel PACs for genomics research. This will enable scalable deployment of accelerated genome sequence alignment and variant calling within data centers and on the cloud, making this technology accessible to several research institutes, universities, and hospitals that need faster processing of sequence alignment.

Early evaluation shows that the accelerated sequence alignment algorithms (i.e. Smith-Waterman) can achieve up to 63X speedup using Intel FPGAs compared to single-thread CPU execution.[5]

Intel and the partners in this project are interested in discussing access to this cloud-based accelerator. Customers who are interested in early access to the solution should contact their Intel sales representative in the region to gain access to the application in the Intel Developer Cloud.

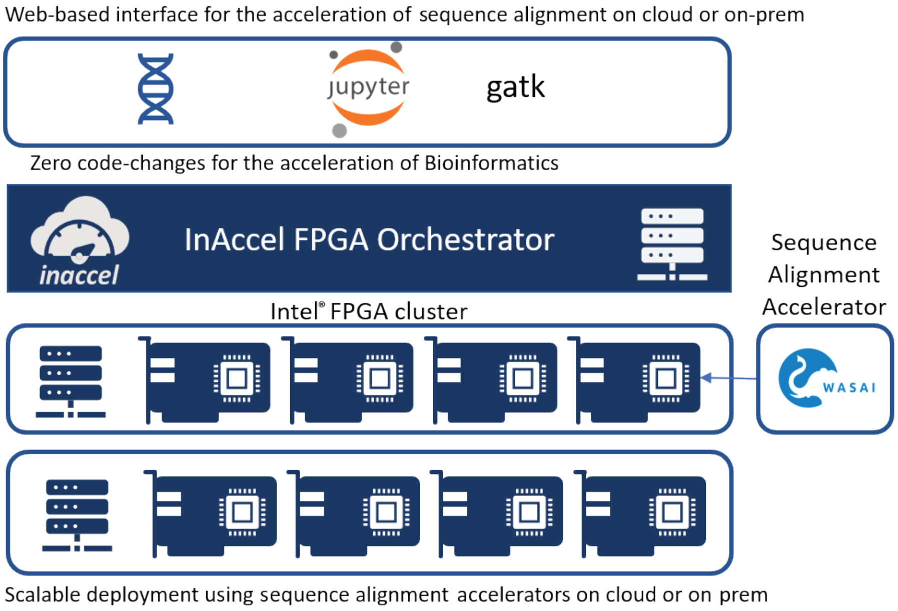
Web-based interface for the acceleration of sequence alignment on cloud or on-prem
jupyter
gatk
Zero code-changes for the acceleration of Bioinformatics
inaccel
InAccel FPGA Orchestrator
Intel® FPGA cluster
Sequence Alignment Accelerator
WASAI
Scalable deployment using sequence alignment accelerators on cloud or on prem

Figure 1. High-level overview of scalable deployment of sequence alignment using Intel hardware accelerators (FPGAs). The hardware accelerator seamlessly interfaces high-level bioinformatics frameworks used for genome analytics like GATK. (GATK is a trademark of the Broad Institute.)

See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Resources

1 https://www.gisaid.org/

2 https://science.sciencemag.org/content/367/6483/1176

3 https://www.sciencemag.org/news/2020/03/mutations-can-reveal-how-coronavirus-moves-they-re-easy-overinterpret

4 Performance improvement and sequencing time are based on estimates from internal WASAI data.
- https://www.wasaitech.com/post/announcing-wasai-lightning-plus-wgs-in-3-hours-with-intel-pac-d5005
- https://www.wasaitech.com/genomics

5 Testing performed by Inaccel and WASAI in August 2020. Configuration for the performance estimations (results below were simulated using the Inaccel CORAL framework):
- Reference CPU server: Dell PowerEdge T640 server, Intel® Xeon® Silver processor 4208 2.1 GHz, 8C/16T, 9.6GT/s, 11M cache, RAM: 64 GB DDR, 480 GB SSD SATA
- FPGA server: Dell PowerEdge T640 server, Intel Xeon Silver processor 4208 2.1 GHz, 8C/16T, 9.6GT/s, 11M cache, RAM: 64 GB DDR, 480 GB SSD SATA with 1x PAC Intel® Arria® 10 FPGA

Evaluated Dataset:

TARGET SEQ	QUERY SEQ	FPGA-ACCELERATED	CPU-ONLY	SPEEDUP
AF133821.1	AY352275.1	0m0.032s	0m0.845s	26X
NC_000898.1	NC_007605.1	0m1.525s	1m36.057s	63X
BA000035.2	BX927147.1	9m2.546s	7h17m1.528s	48X

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results may not reflect all publicly available security updates.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

All product plans and roadmaps are subject to change without notice.

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.