SOLUTION BRIEF

Life Sciences and Healthcare
Data Driven

Genomics Analytics Speeds Time to Insight

Accelerate genomics analytics and reduce costs with Intel's reference architecture for genomics clusters designed using Intel® Scalable System Framework

This solution brief describes how to solve business challenges through investment in innovative technologies.

If you are responsible for...

- **Business strategy:** You will better understand how a genomics analytics solution will enable you to successfully meet your business outcomes.
- **Technology decisions:** You will learn how a genomics analytics solution works to deliver IT and business value.

Executive Summary

There are no silver-bullet treatments that work for all patients. However, caregivers have historically treated patients with similar conditions in the same way. The shift from population-based to personalized care and precision medicine acknowledges that optimal and efficient care depends on developing personalized treatment solutions.

Genomics is the key to unlocking the power of precision medicine. One study forecasts that the next-generation sequencing manufacturer market will reach $4.5 billion by 2019, driven primarily by increased genomics analytics in the clinical environment.[1] Genome sequencing can cost-effectively determine the genetic makeup of a patient, identify key genetic variations, and help to develop customized treatments for those genetic variations.

This brief describes how to accelerate the genome sequencing process and help reduce the cost of genomics alignment analysis and variant calling. These results are powered by software optimized for a multi-core, multi-processor infrastructure, high-performance computing (HPC) methodology, and Intel® Scalable System Framework.

Healthcare organizations that base their genomics clusters on Intel's reference architecture can position themselves to thrive in a healthcare industry where precision medicine is becoming the standard of care.

Intel Solution Architects

Michael McManus
Intel Solution Architect
Industry Sales Group

Chris Gough
Intel Solution Architect
Industry Sales Group

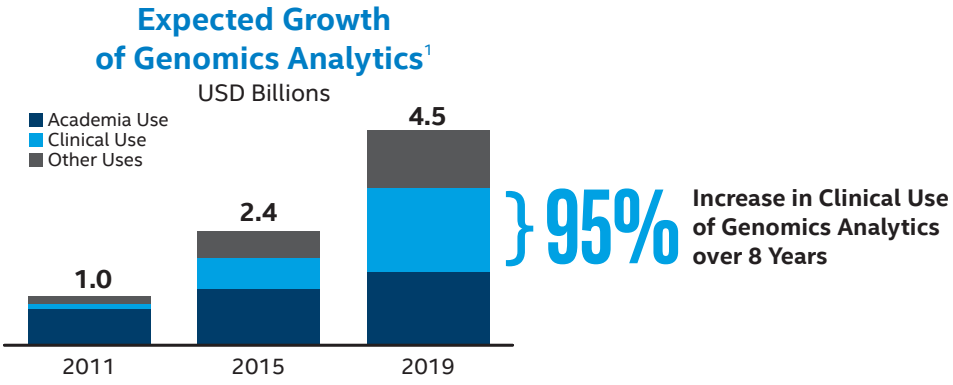

Figure 1. Clinical use of next-generation sequencing is growing exponentially, partly due to improved and more affordable analytics platforms.

Business Challenge: Accelerating Analysis to Support Precision Medicine

As the demand for genome sequencing grows, so does the amount of data that must be processed, stored, and managed. A whole genome sequence with 50X coverage, including FASTQ, BAM, and gVCF files, results in nearly 1 TB of data. Patients cannot afford to wait weeks or months to receive a treatment plan based on analysis of that data—it is critical to accelerate the genomics analytics process to produce results in less than a day. The faster the analytics can be performed, the faster a result can be delivered to determine a treatment plan.

As shown in Figure 2, once the sequence data are received from the sequencer, high-performance computing (HPC) clusters are used to quickly perform the genomics analytics. Results from the analysis are used to interpret any clinically significant genetic variants. After that the information is used to create a treatment plan, including any specific prescriptions for the patient.

In addition to requiring faster analytics, it is crucial that research institutions and clinics be able to reliably predict a genomics cluster's throughput. In this way, they can purchase the most efficient system possible and forecast the cluster's ability to accommodate increased processing requirements over several years, to maximize their investment.

Scalable Genomics Analytics Expedites both Research and Patient Treatment

Genomics analytics are used in three broad categories of use cases. The first two, research and development and clinical treatment, are described here.

- **Research and development setting.** The use of genomics for research and development has fueled a deeper understanding of how variations in human genes affect key biological functions. For example, exploring genetic variation has unlocked the relationships between human genomes, population-level genetic diversity, and the occurrence of disease in distinct sub-populations. As researchers discover new genetic variations, they are able to focus on subtle changes in protein structures and their impact on disease. The pharmaceutical industry can use this information to target the drug discovery process toward therapeutics tailored for these populations. For example, recent discovery of groups who do not benefit from, or have adverse reactions to, certain medications has helped the pharmaceutical industry to create diagnostic tests to identify who can safely receive a medication.
- **Clinical environment.** Currently, genomics is used primarily to treat cancer and inherited diseases. These patient types tend to experience a "diagnostic odyssey"—a long and often meandering journey from the onset of a problem to the discovery of an explanation or treatment. Performing genome sequencing earlier can help speed analytics, thus shortening the journey and eliminating much of the guesswork and unnecessary or ineffective procedures, because roughly 80 percent of all rare disorders have genetic origins.[2]

The third type of use case, using genomics to selectively choose patients to participate in clinical trials and to develop genetically-based diagnostics, is more complex and is outside the scope of this solution brief.

An End to One Diagnostic Odyssey

Children suffering from an unknown malady—and their parents—can attest to the power of genomics analytics. In one case, a child was suffering from seizures, developmental disabilities, and balance and movement issues. 37 years and 400 doctors and specialists later, there was still no explanation. Finally, a non-profit genetic research organization used genomics analytics to discover two mutations located on one of the patient's genes—the patient is believed to be the only known person to inherit two mutations (one from each of the parents) on a single gene.

Genomics Advanced Analytics Process

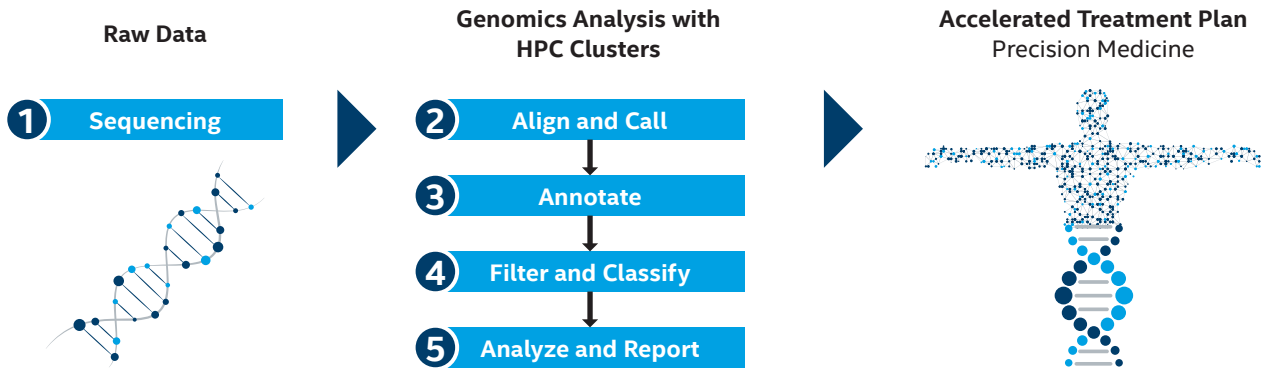

Figure 2. Accelerated genomics analytics using high-performance clusters can reduce the time it takes to get from raw sequencing data to a treatment plan.

Solution Value: Optimizing Genomics Alignment and Variant Calling

Intel and industry players are working closely to make it easier for researchers and clinics to accurately determine the size of genomics clusters they need and to reduce total cost of ownership for genomics analytics solutions.

Intel's scalable reference architecture for genomics clusters is designed to help build a more efficient hardware cluster so that fewer nodes are needed to process greater volumes of genomes. In one example, a solution based on Intel's reference architecture for genomics clusters helped QIAGEN use up to 62 percent fewer nodes than recommended by the sequencer manufacturer, thereby reducing the total cost of ownership of the genomics analytics solution by 47 percent.[3]

This reference architecture includes integrated compute, fabric, and storage components that are highly effective for genomics analytics workloads. Data are stored on a storage array that uses Intel® Enterprise Edition for Lustre*. Systems built with this reference architecture can substantially reduce the time required to process sequencing data, helping to reduce the "time to insight" in research and the "time to answer" for medical institutions.[4]

Equally important, the throughput of the cluster is predictable—a researcher or clinician can reliably estimate how long it will take to run a genome sequence. Using solutions based on Intel's reference architecture for genomics clusters and industry-standard genomics pipelines that have been optimized for Intel® multi-core CPUs, it is possible to run between 1.33 and 1.5 genomes per node.[5] Customers currently using this reference architecture are regularly processing between 48 and 54 genome sequences per day. Predictable throughput enables cluster scaling without guesswork. Performance is already validated so analysts can simply order more clusters if they need more capacity.

In addition to supplying a comprehensive reference architecture for genomics clusters, Intel® Parallel Computing Centers work with industry leaders, experts, and commercial and open-source authors of key genomics codes.[6] This work helps optimize leading industry genomics codes so that genome sequencing runs as efficiently as possible on Intel® architecture-based clusters. The result is a significant improvement in the speed and throughput of key genomics pipelines.

1.5 GENOMES PER NODE[5]

Customers using Intel's reference architecture for genomics clusters process between 48 to 54 genome sequences per day.[7]

Solution Architecture: Accelerated, Affordable Genome Sequencing

Intel's reference architecture for genomics clusters includes all the necessary components: compute, storage, memory, and network fabric (see Figure 3). The researcher or clinician can layer genome sequencing software on top of this hardware foundation to optimize results.

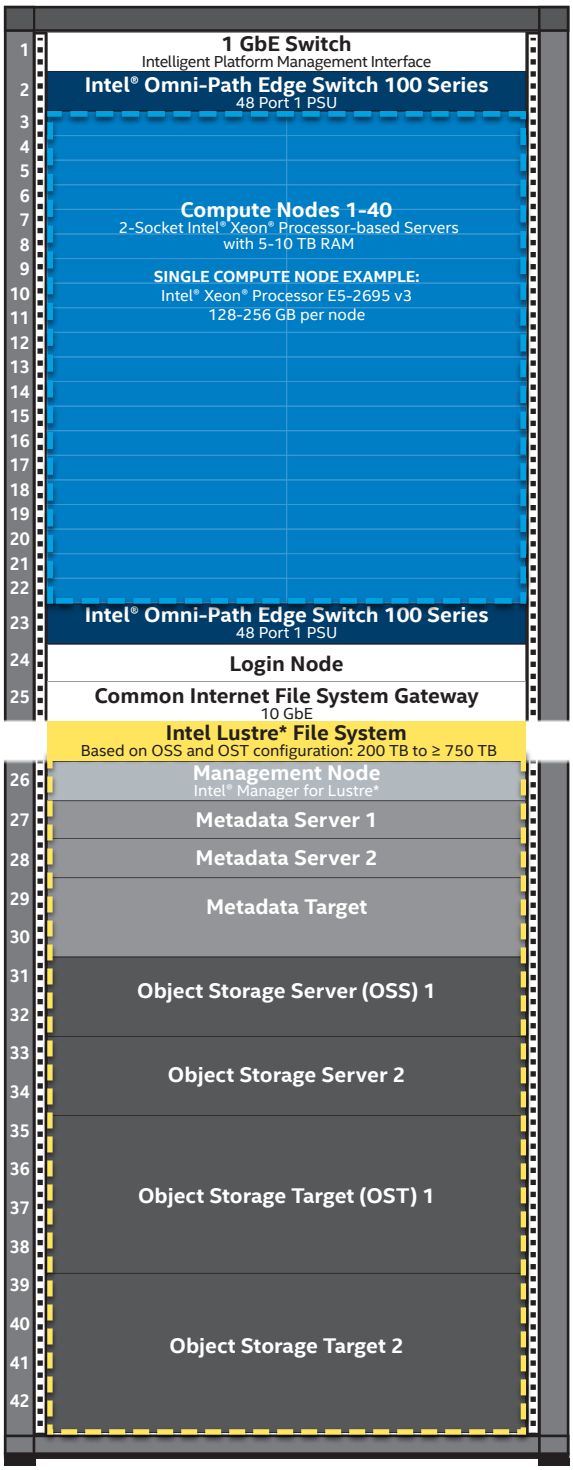

Figure 3. As shown in this rack diagram example, clinicians can layer genome sequencing software on this architecture and regularly process between 48 to 54 genome sequences per day.[7]

System components such as memory, fabric, and storage have evolved at varying rates over the past two decades, resulting in increasingly imbalanced systems. Bottlenecks between memory, fabric, and storage can impede analysis results. Intel® Scalable System Framework helps eliminate these barriers with tightly integrated components that deliver outstanding performance, scalability, efficiency, and reliability.

In particular, Intel Enterprise Edition for Lustre fast storage software plays a significant role in enabling both the predictability of the timing of results delivery and the storage and management of genetic data. The high-speed fabric (Intel® Omni-Path Architecture) prevents bottlenecks by providing enough bandwidth to maintain a balance between the compute nodes and Intel® solutions for Lustre software.

Conclusion: More Results in Less Time

Healthcare is transitioning from a "one size fits all" approach to a model based on personalized, precision treatment of diseases. Genomics is expanding out of the research lab and into the clinical environment; care providers are increasingly using genomics testing and analytics to treat individuals. But genomics testing generates vast quantities of data that must be processed, stored, and analyzed.

Intel SSF forms the basis for Intel's reference architecture for genomics clusters and includes powerful compute, storage, and fabric components capable of handling genomics workloads. This reference architecture, combined with genome sequencers optimized to run on Intel architecture, can dramatically reduce genome sequencing time as well as significantly reduce total cost of ownership for genomics clusters. Today, "do more with less" resonates with every enterprise. Intel's reference architecture for genomics clusters can cost-effectively accelerate genomics analytics and provide an affordable, holistic foundation to deliver breakthrough performance, balance, scalability, and resilience for genomics workloads, today and into the future.

Find the solution that is right for your organization. Contact your Intel representative or visit **intel.com/healthcare**.

Solutions Proven By Your Peers

Intel Solution Architects are technology experts who work with the world's largest and most successful companies to design business solutions that solve pressing business challenges. These solutions are based on real-world experience gathered from customers who have successfully tested, piloted, and/or deployed these solutions in specific business use cases. Solution architects and technology experts for this solution brief are listed on the front cover.

Learn More

You may find the following resources useful:

- Intel Life Sciences: intel.com/content/www/us/en/healthcare-it/life-sciences.html
- Optimized genomics code: intel.com/content/www/us/en/healthcare-it/solutions/genomicscode.html
- Intel® Scalable System Framework: intel.com/ssf
- "A Holistic Solution for Your HPC Needs: Intel® Scalable System Framework" solution brief
- "Accelerate Time to Insight for Whole Genome Analysis" white paper
- "Analyzing Whole Human Genomes for as Little as $22" white paper

[1] "Next-Generation Sequencing (NGS) Market Size, Growth and Trends (2011-2019)." November 3rd, 2015, 3rd Edition.

[2] Source: globalgenes.org/rare-diseases-facts-statistics

[3] Source: intel.com/content/www/us/en/healthcare-it/solutions/genomicscode-qiagen.html

[4] The actual processing time will vary based on several factors including the choice of CPU, the amount of installed memory, the specific genomics analytics pipeline and the coverage of the genomes being processed.

[5] The number of genomes per node will vary based on several factors including the choice of CPU (e.g., Intel® Xeon® processor E5-2600 v3 series), the amount of installed memory, the specific genomics analytics pipeline, and the coverage of the genomes being processed.

[6] "Big Data Genomics and Optimized Genomics Code," accessed on April 29, 2016, intel.com/content/www/us/en/healthcare-it/solutions/genomicscode.html

[7] Based on using 50X whole genomes from an Illumina HiSeq* instrument and also assumes the use of Intel® Xeon® processor E5-2600 v3 CPUs.

0417/JBLA/KC/PDF ♻ Please Recycle 334457-001US