White Paper
Intel® Xeon® processor E7 v2 Family

Accelerating Genome Analysis Using a Dense, Single-Server Solution

Rapid advances in genome sequencing are placing heavier demands on the high performance computing (HPC) clusters that are typically used for analysis, yet upgrading a cluster can be a complex and costly undertaking. Today's powerful, single-server solutions based on the Intel® Xeon® processor E7 v2 family provide an alternative approach to genome analysis that can help to improve performance, simplify the computing environment, and reduce total costs.

Based on tests performed by the Scripps Translational Science Institute and Intel, a single server based on the Intel Xeon processor E7 v2 family can provide up to 34 percent faster time to results for exome analysis than a typical, 35-node high-performance research cluster.

This paper provides detailed information about the tests, the results, and the potential benefits this solution offers for research organizations looking to grow their computing capability in the most cost-effective manner.

"Our performance tests demonstrate dramatic improvements in time to results for genome sequencing using Intel technology, with associated improvements in infrastructure utilization and efficiency."

– Ali Torkamani,
Director of Genome Informatics and Drug Discovery, Scripps Translational Science Institute

Opportunities and Challenges in Molecular Diagnostics

High-throughput sequencing technologies are revolutionizing the field of molecular diagnostics. Instead of targeting just one or a handful of the 20,000 to 25,000 genes that make up the human genome, it is now possible to sequence the majority of those genes in a single test. Clinical research teams can explore genetic profiles in more detail to better understand underlying disease mechanisms and to develop and test targeted therapies. These advanced capabilities will ultimately impact every area of clinical diagnostics. Today, they are particularly relevant in the areas of cancer, heart disease, neurological disease, pediatric disease, non-invasive prenatal testing, and carrier screening for family planning.

With state-of-the-art sequencing instruments, a human genome can now be sequenced in roughly 24 hours. The time required to collect and process samples and to analyze the resulting data typically extends the total turnaround time to at least a few days and sometimes weeks. Yet time-to-results can be a critical factor in clinical research settings. Speed is especially important when sequencing tumors, as studies have shown that early intervention can significantly impact patient survival, especially for aggressive and late stage tumors.[1]

New Computing Strategies Address Growing Sequencing Demands

Increasing the scope of genome analysis can substantially increase time to results. To help organizations make informed decisions as they move toward more complex genome testing, the Scripps Translational Science Institute worked with Intel to measure time to results for tumor genome sequencing on two different computing architectures. One system was a high-performance computing (HPC) cluster, a 35-node system comparable with the HPC clusters used today by many academic research organizations. The other was a single, four-processor server based on the Intel® Xeon® processor E7-4800 v2 family and configured with Intel® Solid-State Drives (Intel® SSDs).

As described in this paper, the single server improved time-to-results by approximately 34 percent for exome analysis and by almost 90 percent for the demanding, read mapping portion of a full genome analysis. The single-server computing platform also provides a simpler, more manageable, and more affordable computing environment, eliminating the need to deploy and maintain a large, clustered server infrastructure. Based on these results, organizations may want to consider this alternative strategy as they move toward next-generation genome sequencing.

34 Percent Faster Performance for Tumor Exome Analysis

In the initial series of tests, the team compared time to results for analyzing the raw sequencing data of a tumor exome, the portion of the tumor genome that contains genes. The same analysis was performed on the HPC cluster and on the single, four-socket server based on the Intel Xeon processor E7 v2 family.

The input data was analyzed using a standard analytical pipeline based on established best practices, including BWA for mapping,[2] the GATK Unified Genotyper for variant calling (under the most sensitive detection conditions),[3] and MuTecT[4] for somatic variant detection.

The total amount of raw sequencing data for the exome study was 14 GB, which provided approximately 150X exome coverage (Table 1). The raw sequencing data for the normal exome was 10 GB, corresponding to exome coverage of approximately 100X. Coverage refers to the amount of redundancy in the sequencing data, and the values used in this study are typical for sequencing of this nature. The large amount of redundancy is important to ensure that the majority of genes are accurately captured and that mutations in the tumor, which may be rare, are observable.

Although the raw genome data input size is an important measure, it does not define the amount of storage required in the computing system. Large intermediate files are created during analysis. Depending on how those files are managed, storage requirements can be several times larger than is required for the raw data.

Based on the results of the performance tests (Table 2), the single-server provided substantially faster analytical performance than the HPC cluster. It reduced the runtimes for all stages of the analysis by more than 30 percent, and reduced the total runtime for the analysis, in real linear time, by approximately 34 percent. The single server also reduced the load on data center resources compared with the HPC cluster, providing a 94 percent reduction in physical space requirements.

Table 1.

Tumor Exome – Data Characteristics
Raw Tumor Genome Data Size: 14 GB
Raw Normal Genome Data Size: 10 GB
Tumor Exome Coverage: ~150X
Normal Exome Coverage: ~100X

Significant Potential Cost Savings

Given the increasing use of cloud computing by research organizations, it can be useful to consider the performance results in terms of theoretical cloud costs. Multiplying the number of cores per system by the total runtime for the analysis provides a measure of core-hours consumed during the analysis. Based on a typical cloud cost of USD 0.05 per cpu-hour, the cost for analyzing a pair of genomes (tumor and normal) from raw sequence data to variant calls would be roughly USD 31.10 per patient on a comparable HPC cluster in the cloud and USD 7.70 per patient on a comparable single server hosted in the cloud. Note that this theoretical cost analysis is based on accessing resources from a non-HIPAA compliant cloud. Using a compliant cloud could lead to significant increases in cost.

Maintaining Control of Systems and Data

Many research organizations are exploring the use of cloud solutions for certain HPC workloads, but decision-makers often have concerns about data privacy and regulatory compliance in the cloud. An on-premise, single-server solution provides a way to modernize and streamline the computing environment

Table 2.

Tumor Exome – System and Performance Comparison			
	HPC Cluster	**Single Server**	**Single-Server Benefits**
Cores Utilized	560 (35 nodes)	60 (1 node)	90% fewer cores
Core Hours	622 hours	154 hours	75% fewer core hours
Physical Space Requirement	~50U	3U	94% less space
Mapping Runtime	9 min	6 minutes	33% faster
Realignment and Recalibration	198 min	142 minutes	31% faster
Variant Calling Runtime	27 min	6 minutes	78% faster
Total Runtime	**234 minutes**	**154 minutes**	**34% faster**

Table 3.

System Configurations		
	Single Server	**35-Node HPC Cluster[6]**
CPU	4 x Intel® Xeon® E7-4890 V2 (2.8 GHz) 60 Physical Cores, 120 threads	35 nodes of a much larger production cluster were allocated to the exome and genome analyses. The system is not designed to provide details of the hardware used to run specific processes. See footnote 2 for details of the large cluster configuration.
Memory	256 GB – 1600 MHz DDR3 Memory	
Storage	2 x 300 GB SAS Drives – 15K RPM	
Disk (SSD)	2 x 400 GB Intel® SSD DC S3700 series	
Cluster Network	N/A	

while maintaining control of systems and data. Solutions based on the Intel Xeon processor E7 v2 family add to the security of on-premise solutions, with integrated support for hardware-assisted security, including high-speed, low-overhead data encryption.

Although on-premise cost models will be different than cloud models, a single, on-premise server offers substantial efficiencies in comparison with a traditional cluster, due to reduced hardware requirements, a reduced data center footprint, and simpler infrastructure management. With cost-effective, security-enhanced, in-house resources, organizations have greater flexibility for managing cost versus risk.

The Importance of High-Speed Storage

Slow storage performance is a common and well-known bottleneck for research teams analyzing genomic data. Anytime large data files are read from storage disks into system memory or written back to disk, the I/O performance of the disks can potentially slow throughput, consume computing resources, and delay data analysis.

One way to ameliorate this bottleneck is to "stream data," by retaining intermediate files within system memory. This technique is particularly valuable for improving performance during the recalibration and realignment steps, in which the quality of the sequencing results are reassessed following the initial mapping. These steps result in multiple large data files, and streaming can significantly improve throughput by reducing the need for high-volume reads and writes.

Although streaming can reduce the impact of I/O performance limitations, certain phases of the analytical pipeline require complex processing of whole data sets that are too large to retain in system memory. For example, when transitioning from the mapping to the realignment and recalibration steps, large data sets must be converted into data chunks that can be streamed in parallel. At this point in the analysis, large volumes of data must be read from disk, and disk I/O performance becomes a potential bottleneck.

Additional tests were run to determine whether the use of high-speed Intel SSDs would help to reduce or eliminate storage bottlenecks and improve throughput. The single server based on the Intel Xeon processor E7 v2 family was configured with the Intel® Solid-State Drive DC S3700 series for these tests, as it was in the earlier performance tests. The same server was then reconfigured with standard Serial Attached SCSI (SAS) drives, and the tests were repeated.

As expected, the performance of the alignment step was most dramatically impacted, since it requires writing and merging multiple large data files so they can be split into data chunks based on chromosome region for parallel processing. The results showed that the use of Intel SSDs had a significant impact on performance, reducing the mapping runtime by 33 percent (from 9 minutes to 6 minutes). There was relatively little influence on other analytical steps.

Conclusion

Performance tests conducted by the Scripps Translational Science Institute in collaboration with Intel showed that a single server based on the Intel Xeon processor E7 v2 family and configured with sufficient memory and high-performance Intel SSDs could provide up to 34 percent faster time to results for exome analysis than a typical academic HPC cluster. The ability to achieve high levels of performance on a simpler and more cost-effective computing platform offers important advantages for organizations of all sizes as they increase the scope of their genome analyses.

Configurations: High performance computing cluster (HPC) with 35-node system and a single, four-processor server based on the Intel® Xeon® processor E7 v2 processor family and configured with Intel® Solid State Drives (Intel® SSDs). The Scripps Translational Science Institute worked with Intel to measure time to results for tumor genome sequencing on these two different computing architectures. For more information go to http://www.intel.com/performance

[1] Colleoni, M., et al., Early start of adjuvant chemotherapy may improve treatment outcome for premenopausal breast cancer patients with tumors not expressing estrogen receptors. The International Breast Cancer Study Group. J Clin Oncol, 2000. 18(3): p. 584-90.

[2] Li, H. and R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 2010. 26(5): p. 589-95.

[3] DePristo, M.A., et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 2011. 43(5): p. 491-8.

[4] Cibulskis, K., et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol, 2013. 31(3): p. 213-9.

[5] The cost of USD 0.05 per cpu-hour is used as a theoretical basis for comparison only. Actual cloud costs may differ significantly, and may also vary based on the vendor and the particular infrastructure and services that are accessed by the customer.

[6] The 35-node cluster was a portion of a much larger HPC cluster that is currently in production use by the Scripps Translational Science Institute. The system is not designed to provide details on the specific subset of the hardware that is utilized by individual processes. Although 35 nodes were allocated to the job, it is not possible to provide a detailed account of the specific server nodes used for the genotype analyses. Configuration of the larger cluster is as follows: 64 Dell PowerEdge* M610 blades with 2 x Intel® Xeon® processor E5530 (2.40 GHz) and 48GB ECC DDR3 memory, 96 Dell PowerEdge M610 blades with 2 x Intel® Xeon® processor E5520 (2.27 GHz) with 48GB ECC DDR3 memory, 96 Dell PowerEdge M600 blades with 2 x Intel® Xeon® processor E5430 (2.66 GHz) and 32GB ECC DDR2 memory, 200 Dell PowerEdge 1955 servers with 2 x Intel® Xeon® processor 5140 (2.33 GHz) and 8GB DDR2 memory. Total theoretical peak performance: 23.5 TFlops. Total memory: 12 TB. Cluster interconnect: 8 x Gigabit Ethernet Summit X450a* switches linked together by two 10 Gigabit Ethernet Summit X650* switches; Infiniband DDR interconnects are available on a large subset of the nodes for tightly-coupled parallel applications that require low latency and high bandwidth connections. Centralized storage infrastructure: A high performance SFA10K unit from Data Direct Networks* (DDN) provides 250 TB of distributed and persistent storage via the IBM* GPFS file system. Per-node storage: 72 GB or 146 GB local scratch space to store temporary data. Data are regularly backed-up and archived to a 4 PB central data facility, two large SGI* LINUX machines, a 32-bit LINUX cluster, and a 64-bit LINUX cluster. The SGI LINUX machine is an SGI* 3700 server with 128 x Intel® Itanium® 2 processors, 128 GB memory and one TB of local disk space.

 0215/RS/HBD/PDF 331604-001US